Unsupervised Data Fusion with Deeper Perspective: A Novel Multi-Sensor Deep Clustering Algorithm

Kasra Rafiezadeh Shahi, *Student Member, IEEE,* Pedram Ghamisi, *Senior Member, IEEE,* Behnood Rasti, *Senior Member, IEEE,* Paul Scheunders, *Senior Member, IEEE,* and Richard Gloaguen

***Abstract*—The ever-growing developments in technology to capture different types of image data (e.g., hyperspectral imaging and Light Detection and Ranging (LiDAR)-derived digital surface model (DSM)), along with new processing techniques, have led to a rising interest in imaging applications for Earth observation. However, analyzing such datasets in parallel, remains a challenging task. In this paper, we propose a multi-sensor deep clustering (MDC) algorithm for the joint processing of multi-source imaging data. The architecture of MDC is inspired by autoencoder (AE)-based networks. The MDC paradigm includes three parallel networks, a spectral network using an autoencoder structure, a spatial network using a convolutional autoencoder structure, and lastly, a fusion network that reconstructs the concatenated image information from the concatenated latent features from the spatial and spectral network. The proposed algorithm combines the reconstruction losses obtained by the aforementioned networks to optimize the parameters (i.e., weights and bias) of all three networks simultaneously. To validate the performance of the proposed algorithm, we use two multi-sensor datasets from different applications (i.e., geological and rural sites) as benchmarks. The experimental results confirm the superiority of our proposed deep clustering algorithm compared to a number of state-of-the-art clustering algorithms. The code will be available at: https://github.com/Kasra2020/MDC.**

***Index Terms*—Multi-sensor Data Fusion; Deep Learning; Autoencoder; Convolutional Autoencoder; Remote Sensing**

I. Introduction

In recent years, we witnessed revolutionary advancements in imaging technologies (e.g., multi-spectral and hyperspectral imaging) [1]. Also, the number of platforms that can carry different sensors (e.g., unmanned aerial vehicles (UAVs) and satellites) grew fast [2]. These advancements allow users to acquire high-quality information of various aspects (i.e., spectral, spatial, and elevation) of on-ground materials and objects at various spatial scales (from close-range to space) [3]. Among the advanced imaging techniques, hyperspectral imaging is considered as the main source of high spectral resolution information. A hyperspectral image (HSI) contains hundreds of narrow spectral bands (channels), covering the visible and near-infrared (VNR, $0.4 - 1\ \mu m$) and shortwave infrared (SWIR, $1 - 2.5\ \mu m$) electromagnetic spectrum [4]. In this way, by employing an HSI, users can distinguish, identify, and track different materials and organisms. As a result, in the last decades, many studies in Earth science were devoted to the use of HSIs, e.g., in plant science [5], [6], urban-planning [7], [8], and geology [9], [10].

Despite the valuable information that an HSI provides on surface materials and objects, processing such data can be challenging [4]. In particular, HSIs suffer from (1) a high intrinsic dimensionality, which implies the existence of redundant features in an HSI, (2) the curse of dimensionality (also known as Hughes phenomenon), due to the imbalance between the number of dimensions and available training samples [11], and (3) highly mixed pixels [12]. In order to tackle the aforementioned challenges, several machine learning algorithms were extensively designed and proposed [4], [13], [14], in general such algorithms split up in two general categories: 1) conventional/shallow learning (CSL) algorithms and 2) deep learning (DL) algorithms [15].

In supervised CSL algorithms, hand-crafted features are initially extracted in an unsupervised manner, and subsequently fed into a supervised model to perform a specific task (i.e., classification, regression) [7], [8]. DL algorithms on the other hand, offer an end-to-end framework to process datasets [14], usually initialized via unsupervised learning and followed by fine-tuning in a supervised manner [16]. There has been an immense number of contributions on supervised DL algorithms in the recent years [13], [14], [17].

Both supervised CSL and DL techniques, despite their great performance, require a considerable number of training sample labels in the learning process, which is hard to acquire in most fields, specifically environmental applications [13]. This shortcoming led researchers to develop unsupervised learning algorithms [18], [19]. The most widely used unsupervised CSL algorithms (also known as clustering algorithm) are K-means [20] and Fuzzy C-means [21] that employ a distance measure (e.g., euclidean distance) to assign each data point to its closest cluster centroid [20]. These algorithms are iterative and rely on a random initialization of the centroids [22]. During the years, various enhanced clustering approaches have been proposed. Interested readers can find an extensive state of the art on CSL clustering algorithms for HSI analysis in [18]. These methods can be subdivided in four categories: 1) probability-based approaches [23] assume that data points from the same cluster follow a similar probability distribution; 2) density-based approaches [24] group data points into different clusters according to their local density and distance to each cluster centroid; 3) graph-based approaches [25] represent the data as a similarity graph (which represents the relations between pairs of points), on which spectral clustering is applied to generate the final clustering map; 4) subspace-

K. Rafiezadeh Shahi, P. Ghamisi, B. Rasti, and R. Gloaguen are with the Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, Freiberg 09599, Germany. Corresponding e-mail: (K.rafiezadeh-shahi@hzdr.de;)

Paul Scheunders is with the Imec-Visionlab, department of Physics, University of Antwerp, 2000 Antwerp, Belgium.

based approaches [26] assume that data points are drawn from several low-dimensional subspaces.

Since the latter assumption is realistic in real world datasets, subspace-based approaches received great attention [27]. One well-known subspace-based approach is the sparse subspace-based clustering (SSC) algorithm, which utilizes the self-expressiveness property that implies each data point can be written as a linear combination of other data points from the same subspace. The superiority of SSC in terms of accuracy, is counterbalanced with being computationally expensive and time-consuming compared to traditional approaches. Therefore, different studies have been devoted to address SSC's shortcomings [28], [29]. In [28], authors proposed a scalable exemplar-based subspace clustering (ESC) algorithm, in which a subset of representative samples (also known as exemplars) is used to construct the sparse representation, resulting in a drastic decline of the computational expenses. In [29], Rafiezadeh Shahi et al. recently proposed a hierarchical sparse subspace-based clustering algorithm (HESSC), which uses the sparse representation of an HSI to extract the lower dimensional subspaces information and to cluster the HSI into meaningful groups. In [30], authors proposed graph-based convolutional subspace clustering, in which a graph representation is combined with subspace clustering on (linear or non-linear) subspaces.

In addition, co-clustering approaches have been proposed to improve the performance of graph-based clustering approaches. Co-clustering approaches hence aim to cluster pixels and spectral features/bands simultaneously. For instance in [31], authors proposed a novel co-clustering approach based on bipartite graph partitioning with joint sparsity to analyze HSIs. Similarly, authors in [32] proposed a graph convolutional sparse subspace co-clustering that utilizes non-negative matrix factorization to reduce the computational power, and thus to allow analyzing large-scale HSIs.

All the above-mentioned CSL clustering approaches were applied on single sensor-based datasets. Most of them are pixel-wise, which implies that they do not consider spatial information from adjacent pixels. In [33], Rafiezadeh Shahi et al. proposed a multi-sensor sparse-based clustering (Multi-SSC) algorithm that exploits the spatial information derived from a complementary source of information, e.g., a high spatial resolution, multi-spectral image.

The main disadvantage of unsupervised CSL algorithms is that hand-crafted features need to be extracted first. Unsupervised DL algorithms on the other hand, offer an end-to-end framework to process datasets. In the last decade, there has been a remarkable number of contributions in computer vision regarding DL clustering architectures [34]–[36]. Autoencoders (AE) are regarded as the most prominent unsupervised DL architectures [19]. AE-based networks are capable to learn informative features without any need for supervision, which makes them highly suitable for clustering. A well-known AE-based clustering algorithms is the deep clustering network (DCN) [35]. DCN minimizes a loss function, which consists of the network reconstruction loss and a clustering loss. DL clustering algorithms have been proposed for the specific task of hyperspectral image clustering. In [37], a Laplacian regularized deep subspace clustering was proposed, that contains a Laplacian regularization to incorporate geometric information within the subspace clustering concept and a self-expressiveness layer in the architecture of a 3D deep convolutional autoecoder. Similarly, in [38], a 3D convolutional autoencoder architecture was presented in which the network is optimized according to two separate loss functions (i.e., network loss and clustering loss). In [39], authors proposed a deep spectral-spatial subspace-based clustering algorithm, in which various patches of an HSI are processed by parallel convolutional autoencoders (CAE), and in which the network parameters are simultaneously optimized. In [40], a deep clustering algorithm, utilizing an intraclass distance constraint within its network objective function was proposed. The authors in [41] proposed an automatic clustering approach using a two-branch convolutional neural network, one branch extracting spatial information, and the other branch extracting spectral information.

In all aforementioned studies using remote sensing datasets, DL clustering algorithms have been employed to analyze single sensor data (e.g., HSIs). Recently, in [42], authors proposed a multi-sensor CAE-based network to cluster urban areas. In their proposed framework, handcrafted features along with the products of normalized digital model, normalized difference vegetation index, and excess green are extracted, and fed to a boosted CAE network to produce a set of latent features that are passed through a mini-batch K-means algorithm.

In this work, we propose a novel multi-sensor deep clustering (MDC) algorithm for multi-source datasets. MDC is a multi-stream autoencoder-based framework for the clustering of multi-sensor data. More specifically, MDC uses AE and CAE networks to extract spectral information from the HSI and spatial information from an auxiliary image, e.g., a high spatial resolution image or LiDAR data, which contains a LiDAR-derived digital surface model (DSM), and thus consists of elevation information, respectively. Then, the computed latent features are concatenated and fed to a fusion network that reconstructs the concatenated images. A cost function is designed which optimizes the network parameters of all three networks simultaneously.

The main contributions of this study can be summarized as follows:

- We propose a novel multi-sensor deep clustering workflow to integrate multi-sensor remote sensing datasets (e.g., HSI, LiDAR-derived DSM) in a robust and effective manner. Furthermore, MDC can be regarded as a pioneer mechanism which offers an end-to-end framework to cluster multi-sensor remote sensing datasets.
- MDC benefits from (1) an AE-based network to process the spectral information of HSIs, (2) a CAE-based network to process the spatial information of auxiliary images, and (3) a fusion network to integrate different data modalities.
- We design a total loss deployed in MDC, where the spectral, spatial, and fusion network parameters (i.e., weights and bias) are optimized simultaneously, and in accordance with the proposed total loss. The total loss

consists of three reconstruction losses that are computed using the aforementioned networks. In addition, MDC has a control on the impact of the spectral and spatial networks, in which a contributing weight is assigned to their corresponding losses.

The rest of the paper is structured as follows: in section II, we explain the proposed methodology. In section III, the datasets and the experimental setup are described. Section IV gives a quantitative and qualitative assessments of the obtained experimental results, and a discussion follows in section V. Conclusions are provided in section VI.

II. Methodology

In this section, we initially describe the notations utilized throughout this paper. Subsequently, to comprehend the MDC's architecture, prior to introduce its structure, we provide an elaboration on the deployed AE and CAE networks in MDC's paradigm. Following, we present the proposed loss function which is utilized to optimize the parameters (i.e., weights and bias) in MDC.

A. Notation

Throughout the paper $\mathbf{X} \in \mathbb{R}^{N \times \mathcal{D}}$ expresses an input image (e.g., an HSI) where N and $\mathcal{D}$ represent the number of pixels and the spectral dimension of the image. A column vector in $\mathbf{X}$ is presented as $\mathbf{x}_i, i = \{1, 2, \cdots, N\}$. The concatenation of images acquired from different sources is further presented as $Fused = Sensor_1 + Sensor_2$ (e.g., $Fused = HSI + LiDAR$). Let $\mathbf{H} \in \mathbb{R}^{N \times \mathcal{M}}$ denote the generated latent features, with $\mathcal{M}$ the number of latent features. The reconstructed image of $\mathbf{X}$ is represented as $\mathbf{R} \in \mathbb{R}^{N \times \mathcal{D}}$. Furthermore, we use the following notations throughout the manuscript: $\mathbf{X_1}$ as a high spectral resolution image, $\mathbf{X_2}$ as a high spatial resolution image, and $\mathbf{X_3}$ as the concatenation of both. We extend this notation for the reconstructed images.

B. Autoencoder (AE)-based network

An AE consists of three main sections, an encoder, a bottleneck and a decoder (Fig. 1). The encoder is a multilayer perceptron with the original image as the input and latent features as the output. These features are stored in the bottleneck section. In this study, we use an AE-based network as the spectral network with an encoder section which consists of three fully-connected layers and a rectified linear unit (ReLU) as its activation function. The decoder has the mirror architecture of the encoder and reconstructs the original image from the latent features. The reconstruction loss can be computed as the mean squared difference between the reconstructed and original image.

Formally, let us formulate an AE-based network as follows. The encoder generates $\mathbf{H}$ from $\mathbf{X}$ using a nonlinear mapping process. The encoder function can be formulated as:

$$\mathbf{H} = f_\theta(\mathbf{X}), \tag{1}$$

where $f_\theta(.)$ expresses an encoder nonlinear mapping function and θ represents the set of parameters (i.e., weights and biases), to be optimized during the encoding process. The decoder uses the latent features to reconstruct the input $\mathbf{X}$ by a reverse mapping:

$$\mathbf{R} = f_\phi(\mathbf{H}), \tag{2}$$

where $f_\phi(.)$ denotes a decoder nonlinear mapping function with ϕ the set of parameters to be optimized during the decoding procedure. The reconstruction loss $\mathcal{L}_{rec}$ is defined as the mean squared error (MSE) between $\mathbf{X}$ and $\mathbf{R}$. The network is constrained to minimize the reconstruction loss:

$$\arg\min_{\theta,\ \phi} \mathcal{L}_{rec} = \arg\min_{\theta,\ \phi} \frac{1}{N} \sum_{i\ =\ 1}^{N} ||\mathbf{x}_i\ -\ f_\phi(f_\theta(\mathbf{x}_i))||_2^2, \tag{3}$$

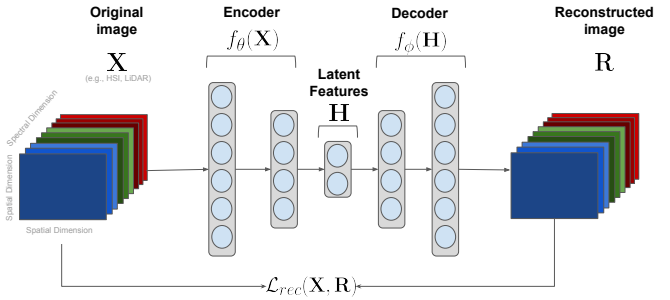


Fig. 1. A fully connected autoencoder network for HSI analysis.

C. Convolutional autoencoder (CAE)-based network

Convolutional autoencoder (CAE)-based networks (see Fig. 2) inherit the general architecture (i.e., encoder, bottleneck, and decoder) of an AE-based network. In the encoder and decoder parts of a CAE network, each fully connected layer is replaced with a (de)convolutional layer. Each (de)convolutional layer contains convolutional filters, batch normalization steps, and an activation function (Fig. 3). In a CAE-based network, the main objective is the minimization of the reconstruction loss, similar as in an AE. The main difference is that a CAE can exploit spatial information from neighbouring pixels. In this regard, a CAE-based network has become the desired architecture to inject spatial and contextual information in the processing workflow, and the (de)convolutional layers play an important role to extract distinct features which conserve the spatial continuity between neighboring pixels [34].

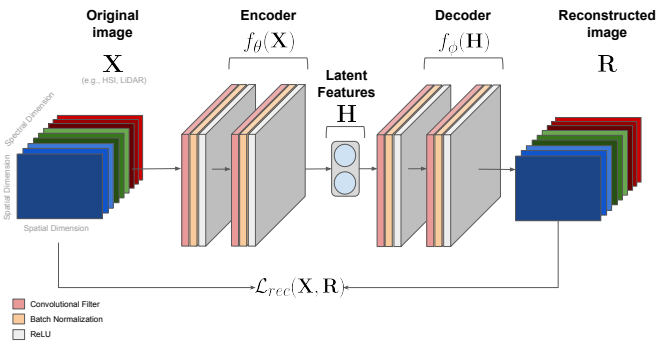


Fig. 2. A convolutional autoencoder network for HSI analysis.

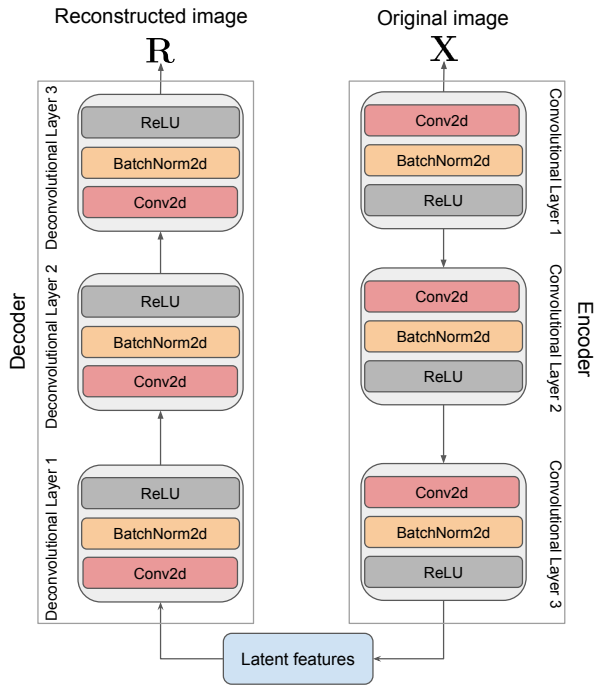


Fig. 3. The architecture of a CAE. Each (de)convolutional layer includes Conv2d, BatchNorm2d, and ReLU, which represent 2-dimensional convolution operations, 2-dimensional batch normalization, and a rectified linear unit as the activation function, respectively.

D. Multi-sensor deep clustering (MDC)

The proposed multi-sensor deep clustering (MDC) algorithm offers a workflow to simultaneously extract spectral and spatial features by fusing information derived from high spectral and spatial resolution data sources. Such a workflow mitigates the absence of spatial information in spectral-based deep clustering algorithms, and at the same time, allows the network to maintain the balance between spectral and spatial features.

To be more precise, an AE-based network is employed to extract spectral information from the high spectral resolution image (e.g., HSIs), while a CAE-based network is implemented to extract spatial information from high spatial resolution data (e.g., LiDAR-derived DSM). In the original AE-based and CAE-based networks, the corresponding reconstruction losses are optimized to find a set of optimal parameters (i.e., weights and bias), after which the produced latent features are passed through K-means clustering to generate a final clustering map. We employ K-means on the lower dimensional but informative latent features due to its fast process and to preserve the geometric correlations between data points from the same cluster.

In MDC, we propose to use AE-based and CAE-based networks and their corresponding loss functions. A straightforward approach that might be effective for data fusion is to train each network individually, and generate the clustering map by concatenating the extracted latent features and feed them into the K-means clustering algorithm. We propose a more sophisticated scheme for the fusion process.

A third network can be regarded as a decoder phase which aims to minimize the reconstruction error between reconstructed and original concatenation image inputs to AE and CAE (see Fig. 4). Moreover, all three networks are trained simultaneously, by minimizing a loss function that is the weighted sum of the loss functions of the three networks. This allows to control the contribution of the spectral and spatial networks on the fusion process.

Formally, the loss function of the spectral AE is given by:

$$\mathcal{L}_{Spectral} = \frac{1}{N}\sum_{i\,=\,1}^{N} ||\mathbf{x_1}_i \;-\; f_{\phi_1}(f_{\theta_1}(\mathbf{x_1}_i))||_2^2, \qquad (4)$$

where $\mathbf{x_1}_i$ represents the i-th column vector of the HSI $\mathbf{X_1}$; θ_1 and ϕ_1 denote the set of network parameters for the encoding and decoding parts of the spectral network, respectively.

The loss function of the spatial CAE is given by:

$$\mathcal{L}_{Spatial} = \frac{1}{N}\sum_{i\,=\,1}^{N} ||\mathbf{x_2}_i \;-\; f_{\phi_2}(f_{\theta_2}(\mathbf{x_2}_i))||_2^2, \qquad (5)$$

where $\mathbf{x_2}_i$ denotes the i-th column vector of the high spatial resolution image ($\mathbf{X_2}$) and θ_2 and ϕ_2 express the set of network parameters in the spatial network.

For the fusion purpose, the latent features extracted from the spectral and spatial networks are fused (concatenated) and presented as $\mathbf{H}_3$ at the input to a decoder of a AE-based network, which will be referred to as the fusion network. That network is trained to reconstruct $\mathbf{X_3} \in \mathbb{R}^{N\times(\mathcal{D}+\mathbf{B})}$, which is the concatenation of the images $\mathbf{X_1}$ and $\mathbf{X_2}$. The loss function of that network is given by:

$$\mathcal{L}_{Fusion} = \frac{1}{N}\sum_{i\,=\,1}^{N} ||\mathbf{x_3}_i \;-\; f_{\phi_3}(\mathbf{h_3}_i))||_2^2, \qquad (6)$$

where $\mathbf{x_3}_i$ and $\mathbf{h_3}_i$ are the i-th column vectors of $\mathbf{X_3}$ and $\mathbf{H}_3$, respectively. ϕ_3 represents the network parameters of the decoding phase of the fusion network.

Rather than training each network individually, we propose to train them simultaneously, by minimizing the following loss function:

$$\underset{\theta_{\{1,2,3\}},\ \phi_{\{1,2,3\}}}{\arg\min} \quad \{\mathcal{L}_{Total} = \lambda_1\mathcal{L}_{Spectral} + \lambda_2\mathcal{L}_{Spatial} + \mathcal{L}_{Fusion}\} \qquad (7)$$

where λ_1 and λ_2 are the weights of the spectral and spatial networks, ranging between 0 and 1, respectively.

When the networks are trained, the obtained latent features $\mathbf{H}_3$ are fed to a K-means clustering algorithm to provide the final clustering map. Furthermore, for the optimization, we employ the adaptive moment estimation (Adam) in the back-propagation step as a stochastic optimization approach to optimize all parameters [43].

III. Data description and Experimental setup

We evaluate the performance of our proposed algorithm in two different application domains (i.e, geological and rural sites), for which already co-registered multi-source datasets are available (i.e., HSI, RGB and LiDAR-derived DSM).

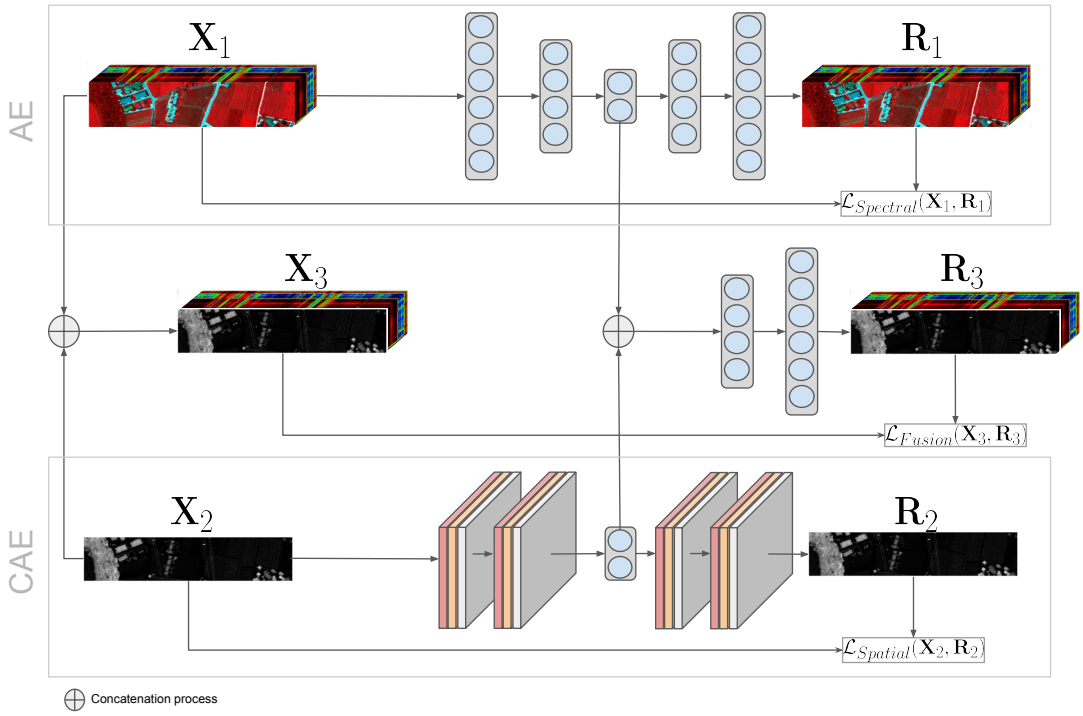


Fig. 4. The scheme of our proposed multi-sensor deep clustering algorithm. In the figure, $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$, $\mathbf{R}_1$, $\mathbf{R}_2$, and $\mathbf{R}_3$ represent the original images of HSI, LiDAR-derived DSM, and the concatenation of HSI and LiDAR-derived DSM data and their reconstructed ones, respectively. The connection operation $\oplus$ denotes the concatenation process of extracted features from the AE and CAE networks.

A. The Trento dataset

The dataset is acquired over a rural area in the south of the city of Trento, Italy. The dataset includes HSI and LiDAR-derived DSM data, composed of 166×600 pixels with a spatial resolution of 1 m. The AISA Eagle sensor was employed to capture the HSI with 63 spectral bands ranging between 0.40 and 0.98 μm. The Optech ALTM 3100EA sensor was utilized to capture the LiDAR-derived DSM data. The acquired HSI and LiDAR-derived DSM data are presented in Fig. 5. In the Trento dataset, there are 6 classes: (1) Apple trees, (2) Buildings, (3) Ground, (4) Wood, (5) Vineyard, and (6) Roads.

B. The geological Finland dataset

The geological Finland dataset was captured over an outcrop of the Archean Siilinjärvi glimmerite-carbonatite complex in Finland [44]. A hyperspectral frame-based camera (0.6 Mp Rikola Hyperspectral Imager), which was mounted on a hexacopter unmanned aerial vehicle (UAV; Aibotix Aibot X6v2) is utilized to capture the HSI, containing 50 spectral bands in the range between 0.5 and 0.9 μm. A senseFly S.O.D.A. RGB camera, mounted on a fixed-wing UAV is engaged to acquire the RGB image. In the geological Finland dataset, the high spatial resolution RGB image with a spatial resolution of 1.5 cm, was downsampled to the HSI with a spatial resolution of 3.3 cm. After co-registration and resampling, the HSI and RGB images are composed of 300×900 pixels. The geological Finland dataset contains 5 classes: (1) Clay, (2) Glimmerite, (3) Dark-rocks (which is a mixture of soil and Glimmerite), (4) Dust, and (5) Water. The captured RGB image and its corresponding reference map are shown in Fig. 6. More elaborated and detailed information on the geological Finland dataset can be found in [45].

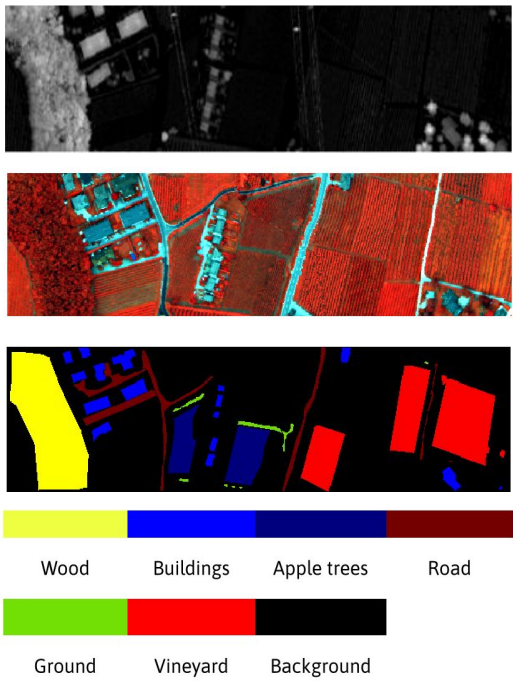


Fig. 5. Trento dataset. From top to bottom: LiDAR-derived DSM rasterized dataset; false color-composite image of the HSI using bands R:40, G:20, B:10; ground truth along with the class legends.

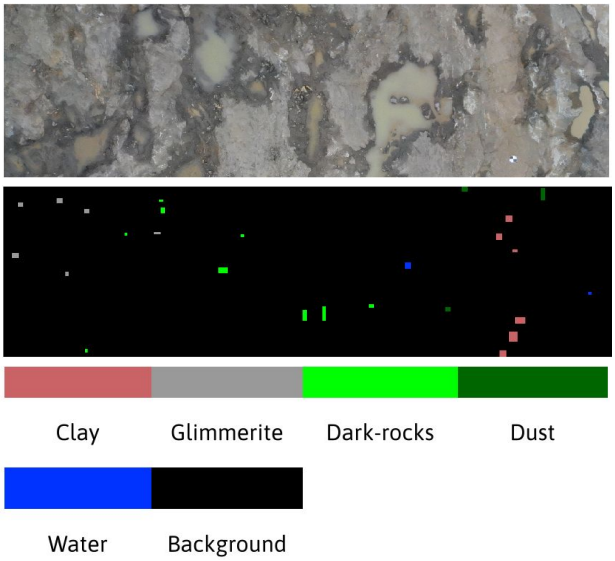


Fig. 6. Geological Finland dataset, captured over Siilinjärvi in Finland. Top: RGB image; bottom: ground truth along with the class legends.

C. Experimental setup

To validate the generality of our proposed approach, we tested the MDC's performance on the aforementioned datasets. We additionally investigated the effect of different hyperparameters (e.g., λ_1, λ_2) in the performance of MDC.

Adam optimizer with default parameters is used for optimizing each of the networks (i.e., spectral, spatial, fusion). The parameters of Adam are set as follows: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and weight decay is equal to 0. We choose $\{64, 128, 40\}$ as the number of nodes for each of the layers of both spectral and spatial networks. The proposed algorithm is implemented in Python 3.8 using PyTorch library. The implementation of MDC will be available online at: https://github.com/Kasra2020/MDC. For all experiments, we run the algorithm 5 times, and the average results are presented.

D. Evaluation metrics

For the validation, three commonly used evaluation metrics are employed: overall accuracy (OA), average accuracy (AA), and Kappa. The reference map is denoted as $\mathbf{Y} = [y_1, y_2, ..., y_N]$ and the clustering map as $\mathbf{C} = [c_1, c_2, ..., c_N]$, where $c_i = \{1, ..., k\}$, with k the number of clusters. To validate the performance of a clustering approach, a matching function $c'_i = bestMap(y_i, c_i)$, is required to match the cluster labels c_i and reference labels y_i. The employed matching function is based on the Hungarian algorithm, and c'_i is the clustering map for which the best match between y_i and c_i is produced by $bestMap(.)$ [46]. OA is then calculated as $\Sigma_{i=1}^{N}\Gamma(c'_i, y_i)/N$, where $\Gamma(c'_i, y_i)$ is 1 if $y_i = c'_i$ and 0 otherwise.

We additionally report two commonly applied unsupervised evaluation metrics, namely, the adjusted rand index (ARI) and the normalized mutual information (NMI). NMI is based on the common/mutual information between two clusters and is defined by:

$$\frac{\sum_{ij} n_{ij} \log \frac{n_i n_{ij}}{n_{i+} n_{+j}}}{\sqrt{(\sum_i n_{i+} \log \frac{n_{i+}}{n})(\sum_j n_{+j} \log \frac{n_{+j}}{n})}} \tag{8}$$

where $n_{ij} = |c'_i \cap y_j|$, n_{i+} and n_{+j} are defined as $\sum_{j=1}^{N} n_{ij}$ and $\sum_{i=1}^{N} n_{ij}$, respectively. In order to compare different approaches, the mutual information is normalized between 0 and 1 [47].

ARI computes the similarity (or dissimilarity) between two clusters and is a adopted from the original rand index [48]. It is defined a:

$$\frac{\sum_{ij}\binom{n_{ij}}{2} - \sum_i \binom{n_{i+}}{2}\sum_j \binom{n_{+j}}{2}/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{n_{i+}}{2} + \sum_j \binom{n_{+j}}{2}] - \sum_i \binom{n_{i+}}{2}\sum_j \binom{n_{+j}}{2}/\binom{n}{2}} \tag{9}$$

The value of ARI is smaller than 1 and can be negative, which implies that 2 clusters have even less similarity than what can be expected from a random result.

E. Comparison with the state-of-the-art approaches

We will compare the performance of the proposed approach with a number of state of the art clustering approaches. Since these approaches are all single-sensor approaches, we will apply each of these approaches twice; once on the HSI alone, and once on the concatenation of the multi-sensor images.

The following clustering approaches have been applied:

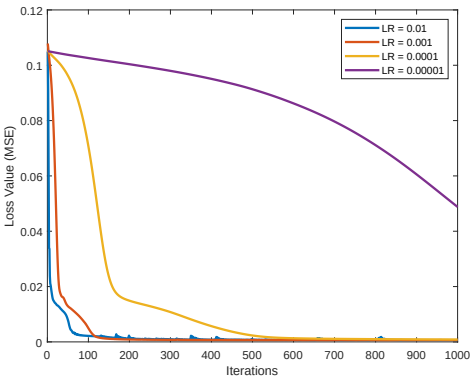
- K-means clustering algorithm [20].
- AE [15], applied on the HSI and the concatenated images respectively. The same architecture is applied as in the spectral network of the proposed method.
- CAE [15], applied on the HSI and the concatenated images respectively. The same architecture is applied as in the spatial network of the proposed method.
- Variational AE (VAE) [49] can be regarded as a variant of AE and is a deep generative learning approach, aiming to force the latent features to follow a predefined distribution.
- DCN [35], combining an AE reconstruction loss and a clustering loss.
- the multi-sensor sparse-based clustering (Multi-SSC) algorithm [33], another multi-sensor clustering approach, which takes the same multi-sensor image data as input as the proposed approach.

We use K-means to generate the final clustering maps in AE, CAE, and VAE. It also should be brought to attention that all the samples in the ground-truth dataset are used in the test phase, and none is used to train the unsupervised networks.

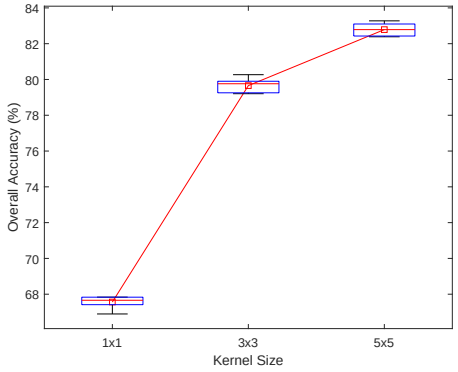
IV. Experiments

A. Hyperparameters evaluation of MDC

Here, we investigate the effect of the hyperparameters deployed in MDC and the evaluation is carried out on the Trento dataset, because of its availability of rich ground truth data. The hyperparameter values that will be selected as the optimal values are applied in all consecutive experiments, on both datasets.



(a)



(b)

Fig. 7. Sensitivity of MDC to, (a) various learning rates (LR) and (b) different kernel sizes.

1) Sensitivity of MDC to learning rate (LR): We conducted an experiment to identify the optimal value for the LR with respect to the model's loss value. Fig. 7(a) illustrates the effect of different LR values on MDC's performance. It can be observed that in general, a higher LR value leads to a lower loss value and a lower number of required iterations before convergence. Remark that the plotted loss values in Fig. 7(a) are smooth, due to the fact that the entire scene is fed to the network. If the employed LR is too high, it might cause trapping of the model in local minima. On the other hand, deploying too low values for the LR can cause a slow convergence. In this respect and based on empirical results, we employ 0.001 as the optimal LR value. With $LR = 0.001$, the algorithm converges after a few hundreds of iterations. To accelerate the procedure, we fix the number of iterations to 500.

2) The influence of the kernel sizes on the performance of MDC: We evaluated the effect of different kernel sizes of the convolutional filters applied in the CAE of the MDC approach by performing MDC with 1×1, 3×3 and 5×5 kernels respectively. Fig. 7(b) displays the obtained OAs as a function of the different kernel sizes. The 5×5 kernel achieved the highest OA, and will be applied in all consecutive experiments.

3) The contribution of the spectral and spatial networks: To investigate the contribution of the individual losses $\mathcal{L}_{Sepctral}$ and $\mathcal{L}_{Spatial}$ in MDC, we varied the values of λ_1 and λ_2 to be $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. Fig. 8 shows the obtained

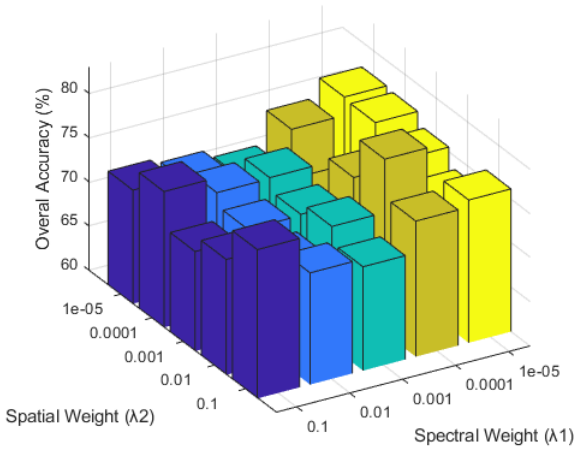


Fig. 8. The contribution of each network in MDC.

OAs. The best overall result was obtained with $\lambda_1 = 0.0001$ and $\lambda_2 = 0.0001$. Therefore, these values were applied as the default values for the consecutive experiments, for both datasets. Remark that the relative contribution of the spectral and spatial information may vary depending on the application at hand, resulting in different optimal values of λ_1 and λ_2.

4) The influence of the number of latent features on the MDC performance: We have evaluated the performance of MDC as a function of the number of latent features (i.e., the number of nodes in the bottleneck of the networks), varying to be $\{10, 20, 30, 40, 50, 60, 70, 80, 90\}$, of which half are extracted from the spectral network, and the other half from the spatial network. Fig. 9 shows the results in terms of OA. When the number of latent features is too low, information is lost during the encoding, while a high number of latent features generates redundancy. Based on the obtained quantitative result, 40 is selected as the optimal number of latent features for MDC. In addition, the number of parameters in each network (i.e., spectral, spatial, and fusion) are reported in Table I. For both datasets, the number of required parameters in the spatial network is higher than in the spectral and fusion networks. This implies that the spatial network requires more time to be optimized compared to the other employed networks.

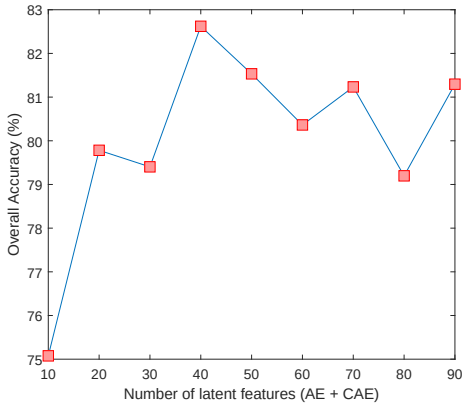


Fig. 9. The impact of different number of latent features on the performance of MDC.

TABLE I
NUMBER OF NETWORK PARAMETERS REQUIRING OPTIMIZATION IN MDC.

Dataset	Different streams designed in MDC		
	Spectral Network	**Spatial Network**	**Fusion Network**
Trento	30035	542015	17664
Finland	28358	548421	16949

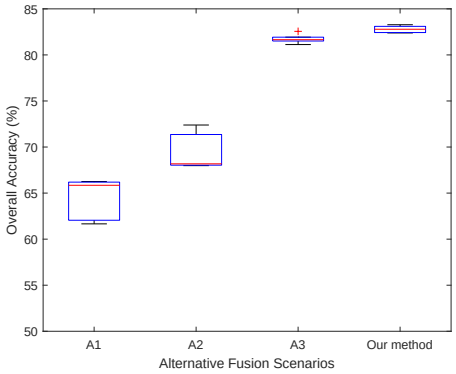


Fig. 10. Comparing the performance of MDC using different fusion scenarios.

5) The performance of MDC by employing different fusion scenarios: The performance of the proposed fusion strategy is validated by comparing it to three alternative fusion strategies:

- Alternative 1 (A1): the AE and CAE networks are trained individually, and their corresponding latent features are concatenated and fed into the K-means clustering algorithm.
- Alternative 2 (A2): the AE and CAE networks are trained simultaneously, and the loss function is given by $\mathcal{L}_{Total} = \mathcal{L}_{Spectral} + \mathcal{L}_{Spatial}$. No fusion network is applied.
- Alternative 3 (A3): only the fusion loss ($\mathcal{L}_{Total} = \mathcal{L}_{Fusion}$) is employed to optimize all three networks of MDC.

Fig. 10 compares the performance of the MDC fusion approach with these alternative fusion scenarios. The proposed fusion scenario surpasses the alternative scenarios, which confirms the effectiveness of the proposed fusion technique. Both alternatives A1 and A2 show a large variance in their results. A2 is superior to A1, showing that the inclusion of spatial information from the CAE in the training phase of the AE boosts the final performance, and vice versa. The performance of A3 is close to the performance of the proposed fusion strategy, showing that the fusion loss ($\mathcal{L}_{Fusion}$) plays an important role in the performance of MDC.

B. Comparison to state of the art

1) Experimental results on Trento dataset: The performance of the different clustering approaches applied on the Trento dataset are quantitatively compared in Table II. Overall, the inclusion of the LiDAR-derived DSM data along with the HSI data in the single-source clustering approaches, led to improving the results. For instance, in terms of OA, when CAE was applied on HSI+LiDAR, a 10% increase can be observed in comparison to when CAE applied on the HSI alone. Such observations confirm the importance of amalgamating the information derived from different sensors as well as incorporating the information of adjacent pixels. In VAE however, the fusion of HSI and LiDAR-derived DSM data did not improve the result; in addition, VAE poorly performs in the clustering task compared to all studied approaches. MDC outperformed the single-source approaches and the Multi-SSC. In particular, the Apple Trees and Wood classes have been much better clustered by MDC.

For a visual comparison, the obtained clustering maps are shown in Fig. 11. Noisy clustering maps are generated, except for the CAE-based clustering approaches, including MDC, and Multi-SSC. These clustering approaches (Fig. 11 (e), (f), (k), (l)) employ both spatial and spectral information. Only in MDC (Fig. II (l)), the Apple Trees class is clearly visible. The smooth clustering result generated by MDC is due to employing convolutional operators. To even more smooth out the clustering results in MDC, the size of the kernel size needs to be increased; however, such a strategy might result in losing local pixels' details.

The total required processing time of all studied clustering approaches are reported in Table II. K-means is the fastest algorithm, as it merely requires the computation of the Euclidean distances between the centroids and the remaining pixels in the dataset. MDC (79.76 seconds) has a multi-stream structure, but is able to compete with the single-sensor approaches. Despite its good performance, the most expensive approach is DCN, as it needs to optimize both reconstruction and clustering losses.

2) Experimental results on geological Finland dataset: The quantitative assessment of the geological Finland dataset is reported in Table III. In terms of OA, MDC attained the highest performance. This observation supports the generalization capability of MDC to different types of datasets. Similar to the Trento dataset, VAE attained the lowest performance. Remarkably, AE and CAE performed reasonably well on HSI, but worse on HSI+RGB, indicating that the mere concatenation of features of multi-sensor datasets is not the best strategy. On the contrary, DCN on HSI+RGB outperforms DCN on HSI.

For a visual comparison, the obtained clustering maps of the geological Finland dataset are presented in Fig. 12. Similar as in the Trento dataset, the CAE-based networks, Multi-

TABLE II
QUANTITATIVE ASSESSMENT OF ALL CONSIDERED CLUSTERING APPROACHES ON THE TRENTO DATASET. IN THE TABLE, FUSED INDICATES THE CONCATENATION OF HSI AND LIDAR-DERIVED DSM DATA.

Clusters	Test	Different clustering approaches											
		K-means		AE		CAE		VAE		DCN		Multi-SSC	MDC
		HSI	Fused	HSI	Fused	HSI	Fused	HSI	Fused	HSI	Fused		
Apple Trees	4034	27.49	54.45	16.63	54.95	0.00	0.00	32.56	33.24	33.81	44.01	0.00	**76.14**
Buildings	2903	58.63	50.43	54.85	56.07	70.54	**92.21**	52.05	52.79	57.60	69.91	57.68	90.88
Ground	479	0.00	0.16	**43.33**	0.17	14.87	0.19	0.00	0.00	0.00	0.21	9.22	1.94
Wood	9123	61.14	62.88	55.05	64.64	41.45	47.41	73.20	72.84	68.29	62.09	**90.95**	79.77
Vineyard	10,501	**98.21**	88.91	97.81	87.45	66.78	85.80	81.04	80.50	97.57	86.80	68.89	94.23
Roads	3174	72.88	98.36	83.63	98.17	70.85	75.19	99.21	**99.82**	69.68	97.95	75.37	89.83
OA (%)		57.95	61.62	50.87	64.23	59.54	68.25	50.12	50.56	60.89	63.79	71.90	**82.79**
AA (%)		53.06	59.20	58.55	60.24	44.08	50.13	56.34	56.53	54.49	60.16	50.35	**72.13**
Kappa		0.46	0.50	0.38	0.53	0.43	0.56	0.39	0.39	0.50	0.53	0.61	**0.77**
NMI		0.43	0.48	0.45	0.49	0.55	0.61	0.47	0.47	0.47	0.48	0.60	**0.69**
ARI		0.28	0.37	0.28	0.39	0.39	0.48	0.28	0.28	0.34	0.37	0.60	**0.63**
A full iteration (t $seconds$)		**3.90**	6.65	15.16	15.19	87.71	87.13	1054.82	1146.71	3053.10	3093.29	518.63	79.76

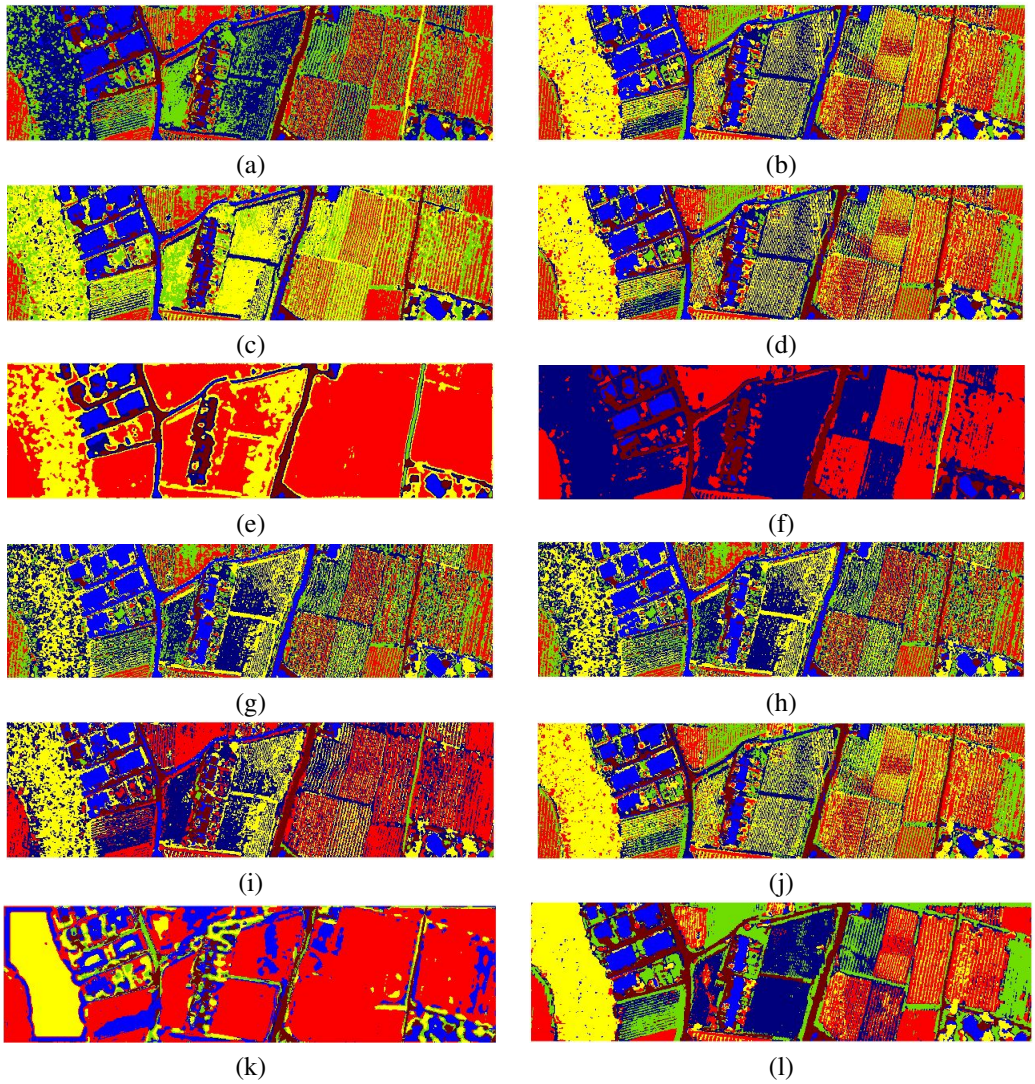


Fig. 11. Clustering maps of Trento dataset obtained by: (a) K-means on HSI, (b) K-means on HSI+LiDAR; (c) AE on HSI, (d) AE on HSI+LiDAR; (e) CAE on HSI, (f) CAE on HSI+LiDAR; (g) VAE on HSI, (h) VAE on HSI+LiDAR; (i) DCN on HSI, (j) DCN on HSI+LiDAR; (k) Multi-SSC; (l) MDC.

SSC, and MDC (Fig. 12(e), (f), (k), and (i)) yield smoother clustering maps compared to the others. VAE produced noisy clustering maps and was not able to separate relevant clusters. While MDC was able to distinguish Water and Clay, Multi-SSC could not. Despite the noisy clustering maps produced by DCN in both scenarios, it has the capability of distinguishing different clusters well.

The total required processing time of all studied clustering approaches on the geological dataset are reported in Table III. Similar to the Trento dataset, among all applied approaches, K-

TABLE III
QUANTITATIVE ASSESSMENT OF ALL CONSIDERED CLUSTERING APPROACHES ON THE GEOLOGICAL FINLAND DATASET. IN THE TABLE, FUSED INDICATES THE CONCATENATION OF HSI AND RGB DATA.

Clusters	Test	Different clustering approaches											
		K-means		AE		CAE		VAE		DCN		Multi-SSC	MDC
		HSI	Fused	HSI	Fused	HSI	Fused	HSI	Fused	HSI	Fused		
Clay	767	89.05	89.34	85.10	81.09	84.40	92.13	89.35	79.64	89.98	90.92	78.18	**92.95**
Glimmerite	381	72.43	76.36	91.37	**100.00**	69.00	78.90	48.26	52.92	66.50	67.34	99.74	93.97
Dark-rocks	659	79.09	79.33	73.12	58.30	80.53	84.45	61.38	82.68	78.80	**86.93**	67.27	71.67
Dust	282	39.44	37.37	49.20	**59.30**	42.62	48.87	38.19	39.82	36.19	43.75	31.95	40.81
Water	135	45.84	32.30	67.63	**100.00**	68.32	17.77	4.51	8.73	27.03	21.91	0.00	**100.00**
OA (%)		64.52	61.67	70.76	68.89	68.37	60.83	56.43	59.94	60.80	68.12	62.14	**79.69**
AA (%)		65.17	62.94	73.28	79.74	68.89	64.42	48.34	52.76	59.07	62.17	55.43	**79.88**
Kappa		0.55	0.52	0.61	0.58	0.59	0.51	0.42	0.47	0.49	0.59	0.50	**0.72**
NMI		0.51	0.49	0.58	0.52	0.58	0.53	0.47	0.52	0.51	0.59	0.44	**0.68**
ARI		0.44	0.39	0.53	0.42	0.49	0.41	0.38	0.43	0.42	0.56	0.40	**0.67**
A full iteration (t $seconds$)		**8.34**	9.08	25.36	26.62	112.85	134.42	2880.62	2911.11	8216.10	8592.59	2112.90	136.13

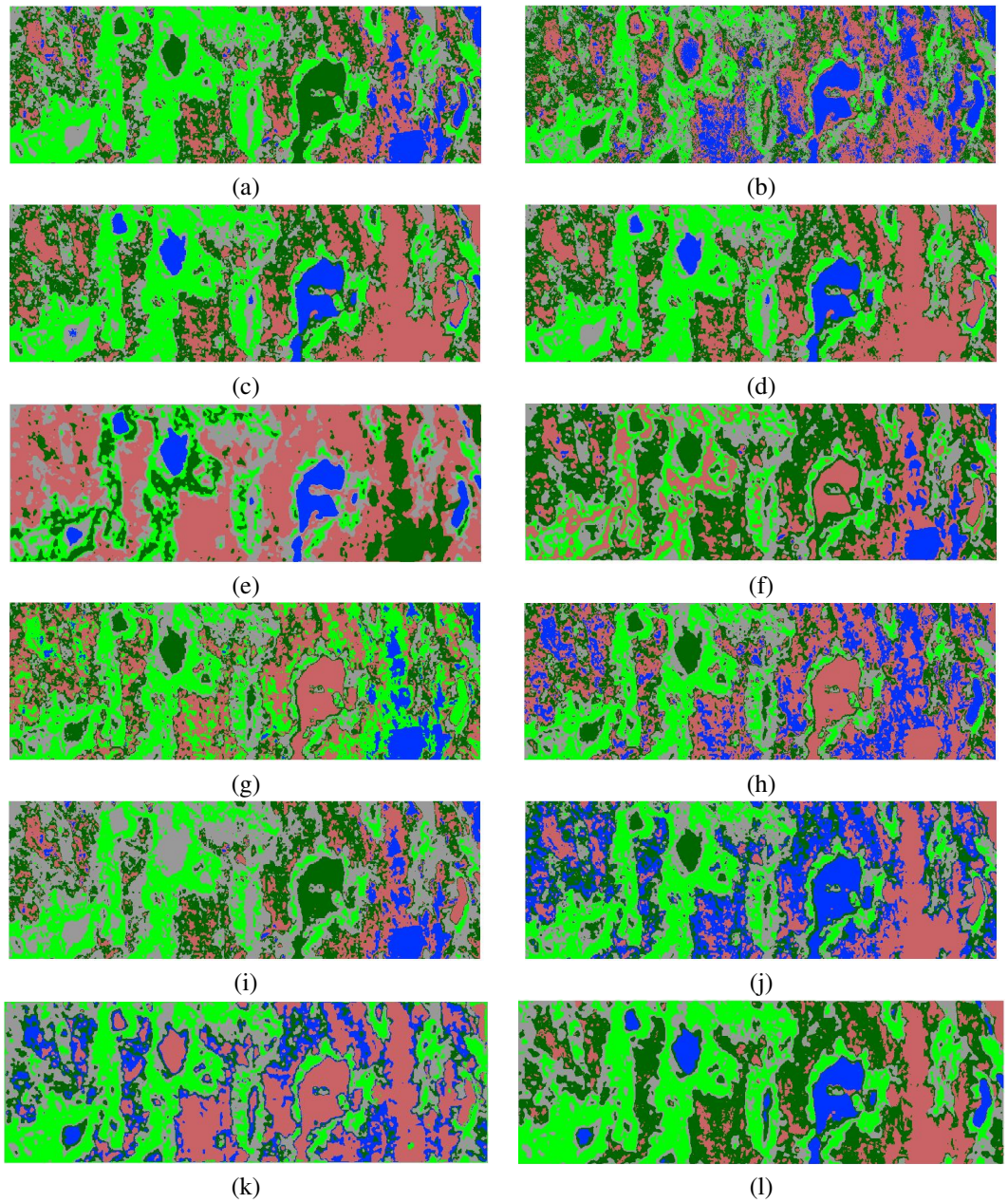


Fig. 12. Clustering maps of geological Finland dataset obtained by: (a) K-means on HSI, (b) K-means on HSI+RGB; (c) AE on HSI, (d) AE on HSI+RGB; (e) CAE on HSI, (f) CAE on HSI+RGB; (g) VAE on HSI, (h) VAE on HSI+RGB; (i) DCN on HSI, (j) DCN on HSI+RGB; (k) Multi-SSC; (l) MDC.

means is the fastest. MDC processed the multi-sensor dataset in 136.13 seconds, which is reasonably fast, considering the number of parameters that needs to be trained. Similar to the Trento data, DCN is the slowest approach, it analyzed the HSI and the HSI+LiDAR-derived DSM in 3053.10 and 3093.29 seconds, respectively.

V. Discussion

We evaluated and compared the performance of MDC with a conventional multi-sensor clustering approach (i.e., Multi-SSC) and some single source-based deep learning clustering approaches on two different types of datasets (i.e., geological and rural areas). Experimental results confirm the superiority of MDC over the Multi-SSC and the state-of-the-art deep learning-based clustering algorithms. From these observations, we can conclude that the proposed fusion strategy is more reliable and effective than a mere concatenation of the multi-sensor datasets, initial to the clustering procedure. In addition, it was shown that clustering approaches, which incorporate spatial information of neighboring pixels, produce less noisy clustering maps. Among the state-of-the-art deep learning-based approaches without including the information of adjacent pixels, AE and DCN performed relatively strong, in particular DCN, which indicates its effective architecture design to extract clustering friendly features as well as its great potential for clustering of remote sensing datasets. The poor performance of VAE can be explained by its pixel-wise framework that generates non-spatial and non-contextual latent features in the encoding phase.

We conducted experiments to select the optimal learning rate (LR), which highly influences the pace of the training phase. According to obtained results. LR = 0.001 is selected as the optimal value. Furthermore, in this study, we investigated the impact of the hyperparameters in MDC's architecture. Regarding the convolutional kernel sizes of MDC, we propose 5×5 as the optimal kernel size. The lower kernel sizes degraded the performance of MDC, because the spatial and spectral information is not efficiently exploited. A kernel size of 1×1 achieved weak results, since MDC performs as a pixel-wise approach.

The impact of the fusion strategy was investigated, by comparing the proposed strategy with a number of alternatives (explained in section IV-A5). From this comparison, it was clear that the simultaneous training of the spectral and spatial networks was advantageous over training them separately. The fusion network has the highest impact on the performance. The combination of all three networks, with a minor contribution of the spatial and spectral networks, provided the best results. Depending on the application at hand, the optimal relative contribution of the spectral and spatial networks may vary.

In this study, we proposed to use a baseline clustering approach (i.e., K-means) to produce the final clustering map. As future work, we will investigate the performance of MDC when combined with more sophisticated clustering approaches (e.g., spectral clustering).

With respect to the geological Finland dataset, we should note that the theory remains the same, however, a more effective approach is to upsample the dataset with a lower spatial resolution; nonetheless, due to the availability of the ground truth dataset at the lower spatial resolution, we downsampled the RGB image.

VI. Conclusions

In this paper, we proposed a multi-sensor deep clustering algorithm that exploits spectral and spatial information from multi-sensor datasets (i.e., HSI, LiDAR-derived DSM, RGB). MDC includes three architectures; an AE-based network which extracts the spectral information from a HSI, a CAE-based network that extracts spatial information from a high spatial resolution image, and a fusion network that takes the concatenated features from the former networks as input to reconstruct the concatenated image data. Subsequently, MDC computes three different losses (i.e., spectral, spatial, fusion losses) to find the optimal network parameters (i.e., weights, bias). The fusion loss was observed as the main contributor, but MDC additionally benefits from the spectral and spatial losses in the training phase. In future work, we will combine this strategy with more sophisticated clustering approaches. Among all applied DL-based clustering approaches, AE, CAE, and DCN performed well, and have high potential to be further explored for application in multi-sensor deep clustering frameworks. This work may lead to an enhanced effort to further explore DL-based unsupervised multi-sensor approaches for remote sensing applications.

Acknowledgment

The authors would like to thank L. Bruzzone of the University of Trento for providing the Trento dataset. The authors would like to acknowledge the Federal Ministry of Education and Research (BMBF) for funding this study, in the client II program, within the MoCa project under grant assignment No.033R189B.

References

[1] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, P. M. Atkinson, and J. A. Benediktsson, "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 1, pp. 6–39, 2019.

[2] C. Pohl and J. L. Van Genderen, "Review article multisensor image fusion in remote sensing: concepts, methods and applications," *International journal of remote sensing*, vol. 19, no. 5, pp. 823–854, 1998.

[3] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: A comparative review of the recent literature," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 2, pp. 29–56, 2017.

[4] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 1, pp. 8–32, 2017.

[5] M. A. Cho, A. Skidmore, F. Corsi, S. E. van Wieren, and I. Sobhan, "Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression," *International Journal of Applied Earth Observation and Geoinformation*, vol. 9, no. 4, pp. 414–424, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030324340700013X

[6] R. Darvishzadeh, A. Skidmore, M. Schlerf, C. Atzberger, F. Corsi, and M. Cho, "Lai and chlorophyll estimation for a heterogeneous grassland using hyperspectral measurements," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 63, no. 4, pp. 409–426, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271608000166

[7] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 480–491, 2005.

[8] M. D. Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *International Journal of Remote Sensing*, vol. 31, no. 22, pp. 5975–5991, 2010. [Online]. Available: https://doi.org/10.1080/01431161.2010.512425

[9] P. Duan, J. Lai, P. Ghamisi, X. Kang, R. Jackisch, J. Kang, and R. Gloaguen, "Component decomposition-based hyperspectral resolution enhancement for mineral mapping," *Remote Sensing*, vol. 12, no. 18, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/18/2903

[10] S. Lorenz, P. Ghamisi, M. Kirsch, R. Jackisch, B. Rasti, and R. Gloaguen, "Feature extraction for hyperspectral mineral domain mapping: A test of conventional and innovative methods," *Remote Sensing of Environment*, vol. 252, p. 112129, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425720305022

[11] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions Information Theory*, vol. IT, no. 14, pp. 55 – 63, 1968.

[12] M.-D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 6, pp. 2014–2039, 2011.

[13] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6690–6709, 2019.

[14] N. Audebert, B. Le Saux, and S. Lefèvre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, pp. 159–173, June 2019.

[15] B. Rasti, D. Hong, R. Hang, P. Ghamisi, X. Kang, J. Chanussot, and J. A. Benediktsson, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 4, pp. 60–88, 2020.

[16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[17] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.

[18] H. Zhai, H. Zhang, P. LI, and L. Zhang, "Hyperspectral image clustering: Current achievements and future lines," *IEEE Geoscience and Remote Sensing Magazine*, pp. 0–0, 2021.

[19] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, "A survey of clustering with deep learning: From the perspective of network architecture," *IEEE Access*, vol. 6, pp. 39 501–39 514, 2018.

[20] D. Arthur and S. Vassilvitskii, "k-means plus plus: the advantages of careful seeding," in *Proceedings of the Eighteenth Annual Acm-Siam Symposium on Discrete Algorithms*, 2006, pp. 1027–1035.

[21] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.

[22] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, Nov 2013.

[23] C. A. Shah, M. K. Arora, and P. K. Varshney, "Unsupervised classification of hyperspectral data: an ica mixture model based approach," *International Journal of Remote Sensing*, vol. 25, no. 2, pp. 481–487, 2004.

[24] J. M. Murphy and M. Maggioni, "Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1829–1845, 2019.

[25] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[26] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, 2011.

[27] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, Nov 2013.

[28] C. You, C. Li, D. P. Robinson, and R. Vidal, "Self-representation based unsupervised exemplar selection in a union of subspaces," 2020.

[29] K. Rafiezadeh Shahi, M. Khodadadzadeh, L. Tusa, P. Ghamisi, R. Tolosana-Delgado, and R. Gloaguen, "Hierarchical sparse subspace clustering (hessc): An automatic approach for hyperspectral image analysis," *Remote Sensing*, vol. 12, no. 15, p. 2421, 2020.

[30] Y. Cai, Z. Zhang, Z. Cai, X. Liu, X. Jiang, and Q. Yan, "Graph convolutional subspace clustering: A robust subspace clustering framework for hyperspectral image," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[31] N. Huang, L. Xiao, and Y. Xu, "Bipartite graph partition based co-clustering with joint sparsity for hyperspectral images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 12, pp. 4698–4711, 2019.

[32] N. Huang, L. Xiao, J. Liu, and J. Chanussot, "Graph convolutional sparse subspace coclustering with nonnegative orthogonal factorization for large hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2021.

[33] K. Rafiezadeh Shahi, P. Ghamisi, B. Rasti, R. Jackisch, P. Scheunders, and R. Gloaguen, "Data fusion using a multi-sensor sparse-based clustering algorithm," *Remote Sensing*, vol. 12, no. 23, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/23/4007

[34] X. Guo, X. Liu, E. Zhu, and J. Yin, "Deep clustering with convolutional autoencoders," in *International conference on neural information processing*. Springer, 2017, pp. 373–382.

[35] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong, "Towards k-means-friendly spaces: Simultaneous deep learning and clustering," in *international conference on machine learning*. PMLR, 2017, pp. 3861–3870.

[36] L. Zhou and W. Wei, "Dic: Deep image clustering for unsupervised image segmentation," *IEEE Access*, vol. 8, pp. 34 481–34 491, 2020.

[37] M. Zeng, Y. Cai, X. Liu, Z. Cai, and X. Li, "Spectral-spatial clustering of hyperspectral image based on laplacian regularized deep subspace clustering," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 2694–2697.

[38] J. Nalepa, M. Myller, Y. Imai, K. I. Honda, T. Takeda, and M. Antoniak, "Unsupervised segmentation of hyperspectral images using 3-d convolutional autoencoders," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 11, pp. 1948–1952, 2020.

[39] J. Lei, X. Li, B. Peng, L. Fang, N. Ling, and Q. Huang, "Deep spatial-spectral subspace clustering for hyperspectral image," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2020.

[40] J. Sun, W. Wang, X. Wei, L. Fang, X. Tang, Y. Xu, H. Yu, and W. Yao, "Deep clustering with intraclass distance constraint for hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4135–4149, 2021.

[41] Y. Li, Q. Xu, W. Li, and J. Nie, "Automatic clustering-based two-branch cnn for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2020.

[42] M. Rahimzad, S. Homayouni, A. Alizadeh Naeini, and S. Nadi, "An efficient multi-sensor remote sensing image clustering in urban areas via boosted convolutional autoencoder (bcae)," *Remote Sensing*, vol. 13, no. 13, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/13/2501

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[44] H. O'Brien, E. Heilimo, and P. Heino, "Chapter 4.3 - the archean siilinjärvi carbonatite complex," in *Mineral Deposits of Finland*, W. D. Maier, R. Lahtinen, and H. O'Brien, Eds. Elsevier, 2015, pp. 327 – 343. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780124104389000133

[45] R. Jackisch, S. Lorenz, M. Kirsch, R. Zimmermann, L. Tusa, M. Pirttijaervi, A. Saartenoja, H. Ugalde, Y. Madriz, M. Savolainen, and R. Gloaguen, "Integrated geological and geophysical mapping of a carbonatite-hosting outcrop in siilinjärvi, finland, using unmanned aerial systems," *Remote Sensing*, vol. 12, no. 18, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/18/2998

[46] M. Rezaei and P. Fränti, "Set matching measures for external cluster validity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2173–2186, 2016.

[47] J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for k-means clustering," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09. New York, NY, USA: ACM, 2009, pp. 877–886. [Online]. Available: http://doi.acm.org/10.1145/1557019.1557115

[48] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, Dec 1985. [Online]. Available: https://doi.org/10.1007/BF01908075

[49] A. Tasissa, D. Nguyen, and J. Murphy, "Deep diffusion processes for active learning of hyperspectral images," *arXiv preprint arXiv:2101.03197*, 2021.

Kasra Rafiezadeh Shahi (Student Member, IEEE) received the B.Eng. degree in computer engineering at the Urmia University of Technology, Iran, in 2015. He received the M.Sc. degree in Geo-information Science and Earth Observation at the faculty of ITC, University of Twente, the Netherlands, in 2018. He is currently pursuing his Ph.D. in developing unsupervised learning techniques for image and signal processing using remote sensing datasets at the faculty of Physics, University of Antwerp, Belgium. Furthermore, he works as a researcher in the machine learning group of the Department of Exploration at Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Germany. His research interests include clustering, and multi-sensor data fusion using unsupervised (shallow/deep) learning techniques, particularly for remote sensing applications.

He serves as a Reviewer for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS and Remote Sensing (Multidisciplinary Digital Publishing Institute).

Pedram Ghamisi (S'12, M'15, SM'18) graduated with a Ph.D. in electrical and computer engineering at the University of Iceland in 2015. He works as (1) the head of the machine learning group at Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Germany and (2) visiting professor and group leader of AI4RS at the Institute of Advanced Research in Artificial Intelligence (IARAI), Austria. He is a co-founder of VasoGnosis Inc. with two branches in San Jose and Milwaukee, the USA. He was the co-chair of IEEE Image Analysis and Data Fusion Committee (IEEE IADF) between 2019 and 2021.

Dr. Ghamisi was a recipient of the IEEE Mikio Takagi Prize for winning the Student Paper Competition at IEEE International Geoscience and Remote Sensing Symposium (IGARSS) in 2013, the first prize of the data fusion contest organized by the IEEE IADF in 2017, the Best Reviewer Prize of IEEE Geoscience and Remote Sensing Letters in 2017, and the IEEE Geoscience and Remote Sensing Society 2020 Highest-Impact Paper Award. His research interests include interdisciplinary research on machine (deep) learning, image and signal processing, and multisensor data fusion. He is an associate editor of IEEE JSTARS and IEEE GRSL. For detailed info, please see http://pedramghamisi.com/.

Behnood Rasti (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees both in electronics-electrical engineering from the Electrical Engineering Department, University of Guilan, Rasht, Iran, in 2006 and 2009, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Iceland, Reykjavik, Iceland, in 2014. In 2015 and 2016, he worked as a Post-Doctoral Researcher and a Seasonal Lecturer with Electrical and Computer Engineering Department, University of Iceland. From 2016 to 2019, he has been a Lecturer with the Center of Engineering Technology and Applied Sciences, Department of Electrical and Computer Engineering, University of Iceland. His research interests include signal and image processing, machine/deep learning, remote sensing, and artificial intelligence.

Dr. Rasti won the prestigious "Alexander von Humboldt Research Fellowship Grant" in 2019 and started his work in 2020 as a Humboldt Research Fellow with Machine Learning Group, Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Freiberg, Germany. He was the Valedictorian as an M.Sc. Student in 2009 and he won the Doctoral Grant of The University of Iceland Research Fund and was awarded "The Eimskip University fund," in 2013. He serves as an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL).

Paul Scheunders (Senior Member, IEEE) received the M.S. and Ph.D. degrees in physics, with work in the field of statistical mechanics, from the University of Antwerp, Antwerp, Belgium, in 1986 and 1990, respectively. In 1991, he became a Research Associate with the Vision Lab, Department of Physics, University of Antwerp, where he is a Full Professor. His research interest includes remote sensing and hyperspectral image processing. He has authored over 200 papers in international journals and proceedings in the field of image processing, pattern recognition, and remote sensing.

Dr. Scheunders is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and has served as a program committee member for numerous international conferences.

Richard Gloaguen received the Ph.D. degree (Doctor Communitatis Europae) in marine geosciences from the University of Western Brittany, Brest, France, in collaboration with the Royal Holloway University of London, U.K., and Göttingen University, Germany, in 2000.

He was a Marie Curie Post-Doctoral Research Associate at the Royal Holloway University of London from 2000 to 2003. He led the Remote Sensing Group at University Bergakademie Freiberg, Freiberg, Germany, from 2003 to 2013. Since 2013, he has been leading the division "Exploration Technology" at the Helmholtz-Institute Freiberg for Resource Technology, Freiberg. He is currently involved in UAV-based multisource imaging, laser-induced fluorescence, and non-invasive exploration. His research interests focus on multisource and multiscale remote sensing integration.