MS^{2}A-Net: Multiscale Spectral–Spatial Association Network for Hyperspectral Image Clustering

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

***Abstract*—Remote sensing hyperspectral cameras acquire high spectral-resolution data that reveal valuable composition information on the targets (e.g., for Earth observation and environmental applications). The intrinsic high dimensionality and the lack of sufficient numbers of labeled/training samples prevent efficient processing of hyperspectral images (HSIs). HSI clustering can alleviate these limitations. In this study, we propose a multiscale spectral–spatial association network (MS^{2}A-Net) to cluster HSIs. The backbone of MS^{2}A-Net is an autoencoder architecture that allows the network to capture the nonlinear relation between data points in an unsupervised manner. The network applies a multistream approach. One stream extracts spectral information by deploying a spectral association unit. The other stream derives multiscale contextual and spatial information by employing dilated (atrous) convolutional kernels. The obtained feature representation generated by MS^{2}A-Net is fed into a standard k-means clustering algorithm to produce the final clustering result. Extensive experiments on four HSIs for different types of applications (i.e., geological-, rural-, and urban-mapping) demonstrate the superior performance of MS^{2}A-Net over the state-of-the-art shallow/deep learning-based clustering approaches in terms of clustering accuracy.**

***Index Terms*—Convolutional autoencoder, deep learning, dilated (atrous) convolutions, hyperspectral imaging, multiscale information, remote sensing, spectral association.**

I. Introduction

REMOTE sensing (RS) has emerged as a prominent data source for Earth observation with various applications (e.g., agriculture [1], [2], urban mapping [3], [4] and geology [5]). Hyperspectral imaging is one of the most popular acquisition techniques, and creates hyperspectral images (HSIs) that contains hundreds of narrow spectral bands. An HSI usually covers the visible and near-infrared (VNIR) to the shortwave infrared (SWIR) range of the spectrum (0.38-2.50 μm) to enable users to observe and monitor materials and organisms of interest [6]. Visual interpretation and traditional approaches for processing HSIs require large amounts of man-power, time, and expenses [7]. The ever-growing demand for utilizing HSIs, encourages researchers to design and develop fast, yet robust analytical approaches. Recently, there has been a tremendous progress in the development of supervised and unsupervised machine/deep learning approaches to analyze HSIs. Such approaches have been successfully deployed to accomplish various HSI analysis tasks (e.g., feature extraction, classification, clustering). Despite the satisfactory performances obtained by supervised approaches, they require a considerable amount of labeled/training samples, which is not always easy to obtain. Thus, in recent years, unsupervised approaches have been receiving more attention [8].

One of the main tasks of unsupervised learning is clustering data points with similar characteristics into separate groups. For HSIs, the main objective is to group pixels that share similar spectral/spatial characteristics into distinct clusters. Overall, HSI clustering approaches can be categorized into two general groups (i.e., conventional shallow learning and deep learning). Conventional shallow learning (CSL) clustering approaches constitute the largest group [8], [9], [10], while deep learning (DL) clustering techniques have been developed more recently [11], [12]. The most widely used CSL-based clustering approach is the k-means clustering algorithm that iteratively clusters the data points by alternately assigning data points to the nearest cluster centroids and updating the cluster centroids, until convergence [13]. Density-based clustering techniques identify clusters by calculating the local densities of the feature space, assuming that each dense area represents a cluster [14]. In the last decade, sparse subspace clustering (SSC)-based approaches received significant attention [9], [10], [15], [16], [17]. These methods cluster data points, based on the self-expressiveness property, which indicates that each data point can be written as a linear combination of other data points from the same subspace [9]. SSC-based approaches initially generate a sparse representation of the data and then compute a similarity matrix on which spectral clustering is applied [18]. Despite the great success of CSL-based clustering approaches, their performance tend to deteriorate when it comes to processing complex datasets (e.g., HSIs), as they merely assume a linear relation between data points [8], [19].

Manuscript received 1 June 2022; revised 16 July 2022; accepted 28 July 2022. Date of publication 11 August 2022; date of current version 18 August 2022. This work is supported by the Imec-Vision Lab, University of Antwerp, Belgium. *(Corresponding author: Kasra Rafiezadeh Shahi.)*

Kasra Rafiezadeh Shahi and Paul Scheunders are with the Imec-Visionlab, Department of Physics, University of Antwerp, 2000 Antwerpen, Belgium (e-mail: rafizadehshahie.kasra@gmail.com; paul.scheunders@uantwerpen.be).

Pedram Ghamisi is with the Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, 09599 Freiberg, Germany, and also with the Institute of Advanced Research in Artificial Intelligence, 1030 Vienna, Austria (e-mail: p.ghamisi@gmail.com).

Behnood Rasti and Richard Gloaguen are with the Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, 09599 Freiberg, Germany (e-mail: b.rasti@hzdr.de; r.gloaguen@hzdr.de).

The code is available at: https://github.com/Kasra2020/MS2A-Net.

Digital Object Identifier 10.1109/JSTARS.2022.3198137

Due to the recent advances in computational technologies (e.g., GPUs) and inheriting concepts from the human neural network, DL-based clustering approaches have been developed, that attain superior performances compared to CSL-based approaches [8], [20]. Autoencoder (AE)-based networks are the pioneers of DL-based clustering approaches. A simple AE architecture consists of an encoder which extracts latent features from the original dataset, and a decoder that reconstructs the original dataset from the extracted latent features. Clustering is then performed on the latent features. The deep clustering network (DCN) is a representative DL-based clustering approach, which aims to learn k-means friendly features by minimizing both clustering and reconstruction losses simultaneously [11]. To further improve the performance of AE-based networks, convolutional AE (CAE)-based networks have been proposed [21], which reconstruct the original dataset by exploiting the spatial information. Clustering deep neural networks (CDNNs) merely use the clustering loss to train their network parameters, making it hard to extract informative and abstract features, and making them sensitive to network initialization [20]. Finally, variational AE (VAE)-based clustering approaches are generative models that enforce the extracted features to follow a predefined distribution [8], [12].

These developments in DL-based clustering found their way into the geoscience and remote sensing community. In [22], a deep clustering was proposed, which utilizes an intraclass distance constraint in its clustering loss, and employs a reconstruction loss as well, leading to cluster friendly latent features. In [23], a graph regularized residual subspace clustering network (GR-RSCNet) was proposed, which captures subspace information by learning the nonlinear relation between data points in an HSI. In [24], a spectral–spatial subspace clustering (DS^{3}C-Net) approach was proposed to analyze HSIs. DS^{3}C-Net is a multiscale approach, feeding patch blocks with different sizes with the same center pixel into parallel autoencoder networks. For the optimization, next to the reconstruction loss and a self-expressiveness loss at each individual stream, a collaborative self-expressiveness loss was employed to capture the subspace structure between various scales. Authors in [25] proposed to deploy 3-D convolutional kernels to capture the spatial information of HSIs and produce the latent features. Similar to DCN, the network parameters are optimized in accordance with both clustering and reconstruction losses. In [26], an HSI clustering network was designed to learn features by computing the set-to-set and sample-to-sample distances (LSSDs), from which the latent features are derived using different extraction approaches (i.e., pairs extraction, joint spectral–spatial feature extraction). Finally, density-based spectral clustering is applied on the learned features to produce the final clustering result.

Most of the aforementioned DL-based clustering approaches merely use the spectral information (e.g., AE, VAE). When spatial information is incorporated (e.g, CAE, DS^{3}C-Net), a single fixed convolutional operation has been employed [27] and less attention is paid to the spectral information [28]. To alleviate these shortcomings, and effectively exploit spectral as well as spatial information in the clustering process, the following approaches are suggested:

- Inspired by recent studies, we propose to employ dilated (atrous) convolutional operations. These operations effectively extract spatial information with a wider field of view, and require fewer number of learnable parameters [29].
- To capture subtle spectral information in the training process, we propose to include a spectral association unit into the current backbone of the networks. The spectral association unit is inspired by self-attention mechanisms and squeeze-and-excitation networks that allow the user to capture spectral information effectively [27], [28], [30].

Based on these two propositions, we present a multiscale spectral–spatial association network (MS^{2}A-Net) that combines a designed spectral-association stream to extract informative spectral features, and a multiscale spatial stream to capture the spatial information between data points.

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

1) To cluster HSIs in a more effective and accurate manner, spatial information is extracted with a wider field of view and fewer learnable parameters by using dilated convolutional operations.
2) Spectral information is optimally preserved in the reconstruction process, by deploying a spectral-association stream.
3) The network is optimized by fusing the spectral and spatial features and employing a loss function that contains a reconstruction loss term and a spectral mean constraint on the latent features.

To the best of our knowledge, this study is the first attempt to utilize the concepts of self-attention mechanisms and dilated convolutional operations for the purpose of HSI clustering. Experimental results on four HSIs for different types of applications (i.e., geological-, rural-, and urban- mapping) demonstrate the superior performance of MS2A-Net over several state-of-the-art shallow/deep learning-based clustering approaches in terms of clustering accuracy.

The remaining of this article is organized as follows. In Section II, we describe the proposed approach in detail. Section III is devoted to the description of the data. The presentation and discussion of the experimental results are elaborated in Section IV. Conclusions are provided in Section VI.

II. Multiscale Spectral–Spatial Association Network (MS^{2}A-Net)

Motivated by the architecture of autoencoder-based networks and dilated convolutions [27], [29], we propose a multiscale spectral–spatial association network (MS^{2}A-Net) for HSI clustering. MS^{2}A-Net has a simple yet effective architecture, as shown in Fig. 1. MS^{2}A-Net aims to extract spatial and contextual information from an HSI at various scales, while preserving the spectral information. In the following section, we describe the two main streams deployed in MS^{2}A-Net.

A. Notation

Throughout the article, bold upper case characters ($\mathbf{X} \in \mathbb{R}^{a\times b\times c}$) denote tensors of rank 3, and $\mathbf{X}_i \in \mathbb{R}^{a\times b}$ is a matrix, denoting layer i ($i = 1, \ldots, c$) of that tensor. $\mathbf{X} \in \mathbb{R}^{h\times w\times \mathcal{D}}$

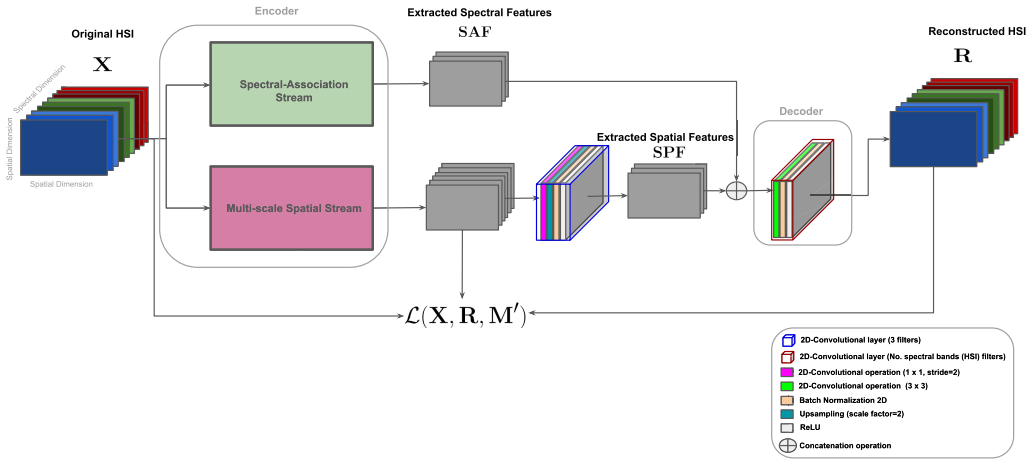


Fig. 1. Illustration of the proposed multiscale spectral–spatial association network.

and $\mathbf{R} \in \mathbb{R}^{h \times w \times \mathcal{D}}$ express an HSI and its reconstructed image, respectively, with spatial dimensions h (height) and w (width), and the number of spectral bands $\mathcal{D}$. $\mathbf{X}_s \in \mathbb{R}^{h \times w}$ represents the sth spectral band of $\mathbf{X}$. A vector is denoted with a lower case character $\mathbf{x}$, and its components as x_i.

B. Multiscale Spatial Stream

A normal 2D-convolutional layer on an HSI can be formulated as:

$$\mathbf{M}_i = \sigma\left(\mathbf{BN}\left(\sum_{s=1}^{\mathcal{D}} \mathbf{X}_s * \mathbf{W}_i + b_i\right)\right), \quad (1)$$

$(i = 1, \ldots, d_1)$. Here, for each value of i, all spectral bands of $\mathbf{X}$ are convolved with the same filter $\mathbf{W}_i \in \mathbb{R}^{r \times r}$, with a predefined kernel size r, after which all bands are summed up (sum and convolution can be swapped) and a bias b_i is added. Then, a batch normalization ($\mathbf{BN}$) is applied to guarantee a fast and stable training process. Finally, a rectified linear (ReLU) function is applied as the nonlinear mapping function σ to obtain the ith extracted feature map $\mathbf{M}_i \in \mathbb{R}^{h \times w}$, with $i = 1, 2, \ldots, d_1$. The number of filters (d_1), is predefined by the user (in this study, $d_1 = 12$).

The receptive field of the kernels is bound to close-range neighbors. In order to capture spatial information with a wider receptive field, the kernel size of the convolutional layers can be increased, and the extracted feature maps from multiple kernel sizes can be stacked together. However, this strategy requires high computational power and results in a significant increase in the number of learnable parameters.

As a remedy to this issue, we will deploy dilated convolutional layers. Dilated convolutional layers with different dilation rates enlarge the receptive field and capture multiscale spatial information, while keeping the number of learnable parameters under control [27], [29]. The idea is to apply several streams of 2-D convolutional layers, with different dilation rates l:

$$\mathbf{M}_i'^{l} = \sigma\left(\mathbf{BN}\left(\sum_{s=1}^{\mathcal{D}} \mathbf{X}_s * \mathbf{W}_i'^{l} + b_i^l\right)\right) \quad (2)$$

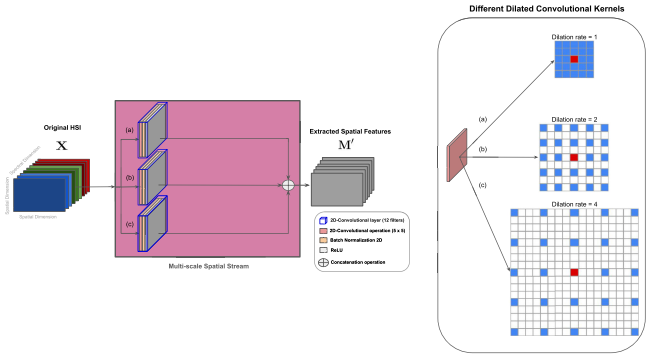


Fig. 2. Illustration of the multiscale spatial stream, where the streams (a), (b), and (c) employ filters with dilation rates of 1, 2, 4, respectively.

$(i = 1, \ldots, d_1)$, with $\mathbf{W}_i'^{l} \in \mathbb{R}^{(lr-l+1)\times(lr-l+1)}$ represents the weights corresponding to the ith convolutional filter with dilation rate l ($l = 1$ leads to Eq. (1)). As shown in Fig. 2, we employ three streams (a), (b), and (c), with dilation rates $l = 1$, 2, and 4, respectively. In this study, we fix r to 5. Since we aim to use the multiscale extracted features both in the reconstruction and clustering processes, we concatenate $\mathbf{M}_i'^{l}$, with $i = 1, \ldots, 12$ and $l = 1, 2, 4$ to shape $\mathbf{M}' \in \mathbb{R}^{h \times w \times 36}$.

$\mathbf{M}'$ is subject to a final 2-D convolutional layer, consisting of filters with $r = 1$ and stride$= 2$ (see Fig. 1), generating d_2 output features. To restore the original spatial dimensions of the original HSI, upsampling with a scale factor of 2 is performed before the batch normalization and the employment of the nonlinear activation function (σ). The final extracted multiscale spatial feature map is denoted by $\mathbf{SPF} \in \mathbb{R}^{h \times w \times d_2}$.

C. Spectral-Association Stream

Apart from exploiting the spatial information using the multiscale spatial stream, it is important to preserve and effectively incorporate the spectral information from the original HSI during the reconstruction process [29]. For this, MS^{2}A-Net employs a spectral association unit (henceforth we will call it spectral-association stream), which contains two phases

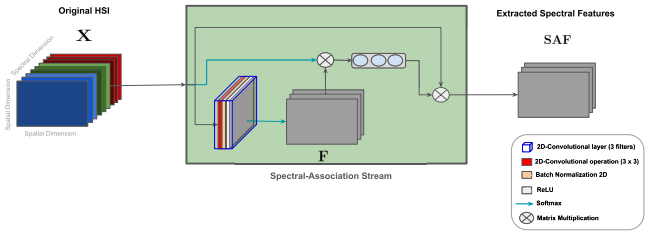


Fig. 3. Spectral-association stream schema, where we extract the most informative spectral features ($\mathbf{SAF}$) from the original HSI ($\mathbf{X}$).

(see Fig. 3) [31]. The initial step in the spectral-association stream is to extract the spatial information at the local level. To extract spatial information correlated with each pixel, we deploy a normal close-range 2-D convolutional layer with d_3 filters with $r = 3$ on $\mathbf{X}$:

$$\mathbf{F}_j = \text{soft}\left(\sigma\left(\mathbf{BN}\left(\sum_{s=1}^{\mathcal{D}} \mathbf{X}_s * \mathbf{W}_j + b_j\right)\right)\right) \quad (3)$$

$(j = 1, \ldots, d_3)$, with $\mathbf{F} \in \mathbb{R}^{h\times w\times d_3}$. To make the process faster and produce a spectral-association matrix, we deploy a softmax function (soft) to rescale each output feature of σ between 0 and 1. Subsequently, a spectral-association matrix $\mathbf{SA} = \text{reshape}(\text{soft}(\mathbf{X}^T)) \times \text{reshape}(\mathbf{F}) \in \mathbb{R}^{\mathcal{D}\times d_3}$ is produced, where reshape (.) unfolds a tensor of rank 3 into a matrix with $h.w$ rows, and $\times$ denotes the matrix multiplication operation. $\mathbf{SA}$ describes the contribution of each spectral band to the extracted spatial features in $\mathbf{F}$, derived in Eq. (3).

Originally, the spectral-association matrix was proposed in [31] to reconstruct the original HSI. However, in this study, we propose to use $\mathbf{SA}$ to extract spectral features as follows:

$$\mathbf{SAF} = \text{reshape}^{-1}\left(\text{reshape}\left(\mathbf{X}\right) \times \mathbf{SA}\right) \quad (4)$$

where $\mathbf{SAF} \in \mathbb{R}^{h\times w\times d_3}$, and reshape^{-1} (.) folds a matrix into a tensor of rank 3.

D. Reconstruction Process

To train the network in an unsupervised manner, the original HSI needs to be reconstructed. For this, the extracted spatial ($\mathbf{SPF}$) and spectral ($\mathbf{SAF}$) feature maps are concatenated to $\mathbf{EF} \in \mathbb{R}^{h\times w\times d}$, where $d = d_2 + d_3$. To equally contribute spectral and spatial information in the reconstruction process, we set $d_2 = d_3 = 3$ in this study. However, d_2 and d_3 can be varied, depending on the application at hand. Thereafter, we feed $\mathbf{EF}$ into the decoder section, which consists of a normal close-range 2-D convolutional layer, with $\mathcal{D}$ filters of 3×3 kernel size, 2D batch normalization, and nonlinear mapping function σ, and that finally generates the reconstructed HSI $\mathbf{R}$ from $\mathbf{EF}$.

E. Optimization of MS^{2}A-Net

In order to train MS^{2}A-Net in an unsupervised manner and stabilize its performance, we designed the following loss function:

$$\arg\min_{\mathbf{W},\mathbf{b}} \mathcal{L} = ||\mathbf{X} - \mathbf{R}||_F^2 + \lambda\, ||\overline{\mathbf{X}} - \overline{\mathbf{M}}'||_F^2 \quad (5)$$

where $\mathcal{L}$ represents the loss function, minimized with respect to all weights and biases of the entire network. $\overline{\mathbf{X}} \in \mathbb{R}^{h\times w}$ and $\overline{\mathbf{M}}' \in \mathbb{R}^{h\times w}$ denote computed averages over the spectral dimension from $\mathbf{X}$ and $\mathbf{M}'$, respectively. In addition, $||.||_F$ represents the $Frobenius$-norm. In Eq. (5), the first term denotes the mean squared error (MSE) between the original ($\mathbf{X}$) and reconstructed ($\mathbf{R}$) images. The second term defines the spectral mean constraint, which denotes the MSE between the averaged feature values from $\mathbf{M}'$ and the spectrally averaged image $\mathbf{X}$. λ is a tradeoff parameter to control the impact of the spectral mean constraint. Since MS^{2}A-Net ultimately aims to cluster the multiscale spatial features ($\mathbf{M}'$), the spectral mean constraint on the generated $\mathbf{M}'$ is included in the designed loss function, to assure that the generated latent features have a direct impact on the training process, in such a way that they are enforced to preserve the mean spectral information of the original image.

As the final step of MS^{2}A-Net, the clustering is applied. To be more specific, we apply k-means clustering on the generated multiscale spatial features ($\mathbf{M}'$) to cluster $\mathbf{X}$. The optimization of the MS^{2}A-Net assures a more effective exploitation of spatial as well as spectral information for a better description of the relations between pixels in an HSI, leading to an improved clustering map.

III. Hyperspectral Data Description

We evaluate the performance of our proposed algorithm on four real HSIs, covering three different application domains (i.e, rural, urban, and geological sites).

A. Trento Dataset

This dataset is acquired by the AISA Eagle sensor over a rural area in the south of the city of Trento, Italy. The HSI is composed of 166×600 pixels with a spatial resolution of 1 m, and 63 spectral bands ranging between 0.40 and 0.98 μm. The acquired HSI along with its corresponding ground truth dataset are presented in Fig. 4. The Trento dataset contains six classes: 1) Apple trees, 2) Buildings, 3) Ground, 4) Wood, 5) Vineyard, and 6) Roads.

B. Houston 2013 Dataset

This dataset is acquired over the University of Houston campus and the neighboring urban area by the Compact Airborne Spectrographic Imager (CASI) on June 23, 2012. In this work, we utilized a subset of this scene, composed of 300×300 pixels (indices on spatial dimensions range within [40:340,500:800]), with a spatial resolution of 2.5 m and 144 spectral bands ranging between 0.38–1.05 μm. The HSI along with its corresponding ground truth dataset are presented in Fig. 5. More details on the Houston 2013 dataset can be found in [32]. The Houston 2013 subset contains six classes: 1) healthy grass, 2) soil, 3) residential area, 4) road, 5) parking lot, and 6) tennis court.

C. Geological Finland Dataset

The geological Finland dataset was captured over an outcrop of the Archean Siilinjärvi glimmerite-carbonatite complex in

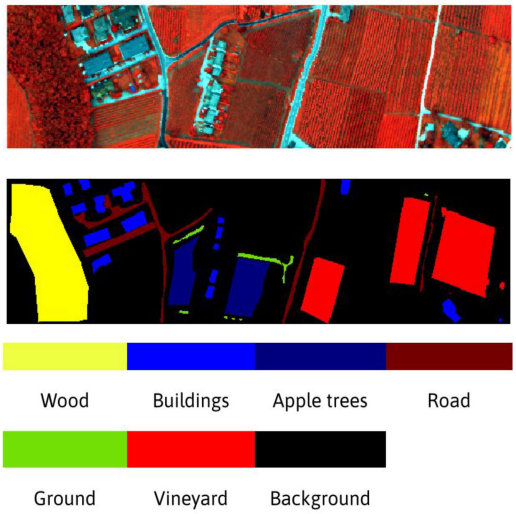


Fig. 4. Trento dataset. From top to bottom: false color-composite image of the HSI using bands R:40, G:20, B:10; ground truth along with the class legends.

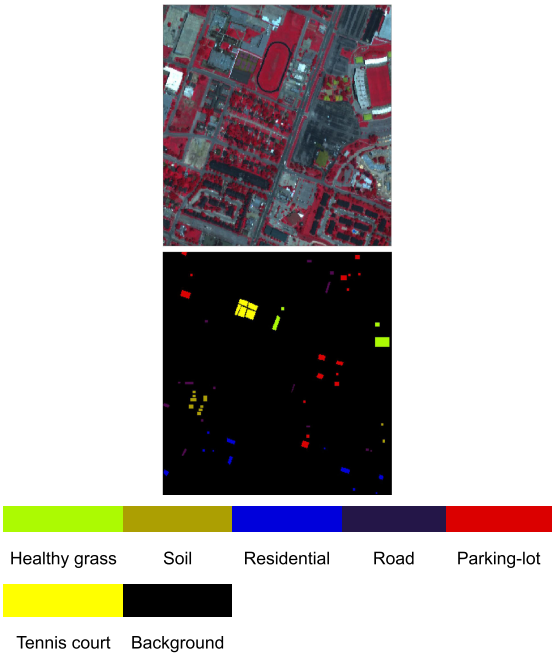


Fig. 5. Houston 2013 subset. Top: the false color-composite image using spectral bands: R:105, G:61, B:40; middle: corresponding ground truth maps, and bottom: class labels.

Finland [33], by a hyperspectral frame-based camera (0.6 Mp Rikola Hyperspectral Imager), mounted on a hexacopter (Aibotix Aibot X6v2). The HSI is composed of 300×900 pixels and contains 50 spectral bands covering the range between 0.50 and 0.90 μm. The geological Finland dataset contains five classes: 1) Clay, 2) Glimmerite, 3) Dark-rocks (which is a mixture of soil and Glimmerite), 4) Dust, and 5) Water. The RGB image of the scene and its corresponding reference map are shown in Fig. 6. More elaborated and detailed information on the geological Finland dataset can be found in [5].

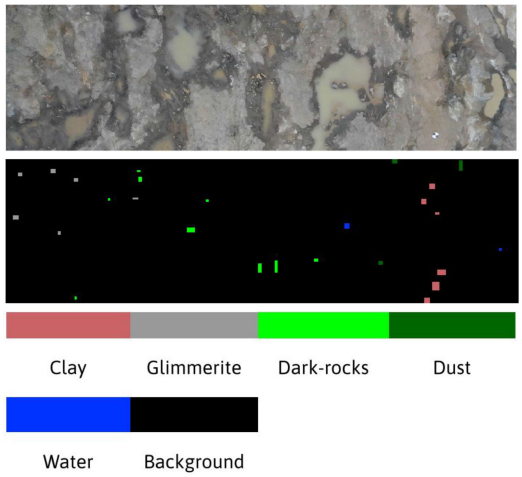


Fig. 6. Geological Finland dataset, captured over Siilinjärvi in Finland. Top: RGB image; bottom: ground truth along with the class legends.

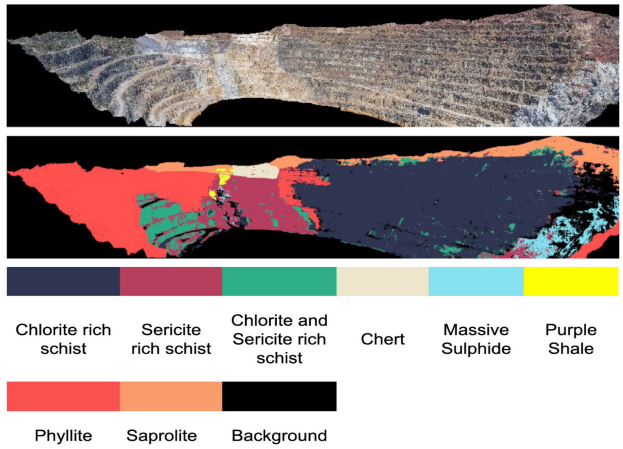


Fig. 7. Geological Spain dataset, captured over the Rio Tinto area in Spain. Top: RGB image; bottom: ground truth along with the class legends.

D. Geological Spain Dataset

The Rio Tinto area is located 70 km north of Huelva in the Iberian Pyrite Belt, in Spain. The area has a rich mining history dating back to the Bronze Age, while currently, significant resources remain and mining operations still take place. Panoramic outcrop scans were acquired by an AISA-FENIX camera, mounted on a tripod [34]. The captured HSI is composed of 300×1416 pixels and 190 spectral bands, covering the range between 0.38 and 2.50 μm. The geological Spain dataset consists of eight classes: 1) Chlorite rich schist, 2) Sericite rich schist, 3) Chlorite + Sericite rich schist, 4) Chert, 5) Massive Sulphide, 6) Purple Shale, 7) Phyllite, and 8) Saprolite. The RGB image along with the corresponding ground truth of the scene is displayed in Fig. 7.

E. Evaluation Metrics

We evaluate the clustering performance of the studied approaches using three widely used classification metrics: overall

accuracy (OA), average accuracy (AA), and Kappa. Remark that the proposed clustering approach is entirely unsupervised, i.e., only unlabeled data are used for optimizing the network and the clustering. However, for a quantitative evaluation of the clustering results, labeled data are applied for validation. $\mathbf{Y} = [y_1, y_2, \ldots, y_N]$ represents the true class labels. $\mathbf{C} = [c_1, c_2, \ldots, c_N]$ denotes the obtained cluster labels, where $c_i = \{1, \ldots, k\}$, with k the number of clusters. To evaluate the clustering performance, a matching function $c_i' = bestMap(y_i, c_i)$ between the predicted cluster labels and true class labels is constructed by the Hungarian algorithm [35]. Subsequently, OA is computed 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.

In addition, we report two commonly applied unsupervised evaluation metrics, namely, the normalized mutual information (NMI) and the adjusted rand index (ARI). NMI is based on the common/mutual information between two clusters and is defined as

$$\frac{\sum_{ij} n_{ij} \log \frac{n_i n_{ij}}{n_{i+} n_{+j}}}{\sqrt{\left(\sum_i n_{i+} \log \frac{n_{i+}}{n}\right)\left(\sum_j n_{+j} \log \frac{n_{+j}}{n}\right)}} \tag{6}$$

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 [36].

ARI computes the similarity (or dissimilarity) between two clusters and is a adopted from the original rand index [37]. It is defined as

$$\frac{\sum_{ij} \binom{n_{ij}}{2} - \sum_i \binom{n_{i+}}{2} \sum_j \binom{n_{+j}}{2} / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i+}}{2} + \sum_j \binom{n_{+j}}{2}\right] - \sum_i \binom{n_{i+}}{2} \sum_j \binom{n_{+j}}{2} / \binom{n}{2}}. \tag{7}$$

The value of ARI is smaller than 1 and can be negative, in which case two clusters are less similar than what can be expected from a random result.

IV. Experimental Results

A. Implementation Details

We implemented MS^{2}A-Net in Python, version 3.8 using the Pytorch library on a workstation with an i9-7900X CPU, 128 GB RAM, NVIDIA GeForce RTX 2080 Ti 11 GB GPU. We adopted the Adam optimizer with default parameters for both streams (i.e., spectral-association and spatial multiscale). The parameters of the Adam optimizer are set as: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and weight decay $= 0$. The implementation of MS^2 A-Net is available online at: https://github.com/Kasra2020/MS2A-Net.

B. Comparison to State-of-the-art Clustering Approaches

In this section, we provide a quantitative and qualitative assessment of the obtained clustering results from the studied clustering approaches on different datasets. All the experiments are reported and analyzed based on five runs, with different random initialization of the learnable parameters. Please note that for each dataset, the entire ground truth dataset was utilized for validation during the performance evaluation.

We compare clustering performance of our proposed MS^{2}A-Net to ten other representative CSL/DL-based clustering approaches:

1) Traditional CLS-based clustering approaches: k-means [13] and spectral clustering on the sparse coefficients (SC-SC) [18].
2) Advanced sparse subspace-based clustering approaches that have proven to be effective for analyzing complex datasets: hierarchical sparse subspace clustering (HESSC) [38], scalable exemplar-based subspace clustering (ESC) [10], and elastic net subspace clustering (EnSC) [39].
3) DL-based clustering approaches: AE [40], CAE [40], VAE [12], DCN [11] and deep multiresolution clustering network (DMC-Net).

In AE, CAE, VAE, DCN, DMC-Net, the clustering results are generated by employing k-means on the latent features. To have a fair comparison, the number of latent features for all aforementioned DL-based approaches is set to 36.

1) Quantitative Results on Trento Dataset: Table I reports the quantitative assessment of the studied clustering approaches applied on the Trento dataset. Among the CSL-based approaches, k-means and HESSC obtained the highest OAs (57.95 % and 56.76 %, respectively), while among the DL-based clustering approaches, the approaches that incorporate spatial and contextual information (i.e., CAE, DMC-Net, and MS^{2}A-Net) attained higher OAs. Among the approaches that merely use spectral information (i.e., AE, VAE, and DCN), DCN is superior and obtained comparable results as CAE. The inferior performance of VAE (OA $=$ 50.12 %) indicates that better hyperparameter tuning for such an approach is required. Overall, DMC-Net and MS^{2}A-Net yielded the highest OAs (82.35% and 84.16%, respectively). This demonstrates a substantial performance improvement by deploying a multiscale spatial stream. However, integrating spectral information by employing a spectral-association stream improved the clustering performance even further. With respect to individual class accuracies, most approaches failed to capture the "Ground" class accurately, except for EnSC (97.29 %). One can argue that the cause of this problem is insufficient test samples for the "Ground" class.

2) Quantitative Results on Houston 2013 Dataset: We quantitatively assessed the clustering performance of studied approaches on Houston 2013 dataset (see Table II). Among the CSL-based approaches, ESC obtained the highest OA (67.18%), revealing that the selected exemplars are sufficiently representative to describe the entire dataset. SC-SC obtained the poorest performance (OA $=$ 43.38%), meaning that the representative dictionary utilized in SC-SC, is not well-built, and further tuning of its parameters is required. MS^{2}A-Net attained the highest OA (73.06%) among all studied clustering approaches. In the Houston 2013 dataset, all DL-based clustering approaches distinguish the "Healthy grass" class perfectly (100%). Except for VAE (class accuracy $=$ 87.09%), the performance of the DL-based approaches is poor for the "Road" class. In addition, the poor performance of some DL-based approaches could be due to the

TABLE I
QUANTITATIVE ASSESSMENT OF ALL CONSIDERED CLUSTERING APPROACHES ON THE TRENTO DATASET

Clusters	# Test samples	Clustering approaches										
		Shallow					Deep					
		k-means	HESSC	ESC	SC-SC	EnSC	AE	CAE	VAE	DCN	DMC-Net	MS^{2}A-Net
Apple Trees	4034	27.49	0.00	64.43	10.88	78.88	16.63	0.00	32.56	33.81	61.43	**87.57**
Buildings	2903	58.63	**80.26**	0.00	37.04	41.30	54.85	70.54	52.05	57.60	68.74	55.80
Ground	479	0.00	21.71	96.99	0.00	**97.29**	43.33	14.87	0.00	0.00	4.01	0.00
Wood	9123	61.14	91.69	46.61	23.44	40.28	55.05	41.45	73.20	68.29	**98.83**	97.21
Vineyard	10,501	**98.21**	43.70	34.03	60.74	44.51	97.81	66.78	81.04	97.57	93.81	94.67
Roads	3174	72.88	55.45	91.05	45.09	75.11	83.63	70.85	**99.21**	69.68	47.93	46.12
OA (%)		57.95	56.76	45.60	37.93	51.57	50.87	59.54	50.12	60.89	82.35	**84.16**
AA (%)		53.06	48.80	55.52	29.54	62.90	58.55	44.08	56.34	54.49	62.46	**63.58**
Kappa		0.46	0.41	0.31	0.16	0.37	0.38	0.43	0.39	0.50	0.76	**0.79**
NMI		0.43	0.49	0.43	0.26	0.44	0.45	0.55	0.47	0.47	0.73	**0.76**
ARI		0.28	0.37	0.26	0.11	0.25	0.28	0.39	0.28	0.34	0.77	**0.81**
Computing time (seconds)		**3.90**	478.94	120.60	1233.23	4738.00	15.16	87.71	1054.82	3053.10	114.16	36.95

TABLE II
QUANTITATIVE ASSESSMENT OF ALL CONSIDERED CLUSTERING APPROACHES ON THE HOUSTON 2013 DATASET

Clusters	# Test samples	Clustering approaches										
		Shallow					Deep					
		k-means	HESSC	ESC	SC-SC	EnSC	AE	CAE	VAE	DCN	DMC-Net	MS^{2}A-Net
Healthy grass	364	**100.00**	**100.00**	95.44	62.86	77.20	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Soil	189	26.88	28.57	**80.00**	0.00	0.00	4.66	18.31	47.72	16.83	71.01	11.64
Residential	248	**100.00**	0.00	20.00	8.39	94.76	**100.00**	98.47	0.00	**100.00**	11.77	**100.00**
Road	234	0.00	8.97	35.64	0.00	0.00	0.00	0.00	**87.09**	36.07	10.43	0.00
Parking lot	496	21.45	79.40	52.46	80.00	39.11	35.73	48.99	13.63	24.68	**88.87**	75.60
Tennis court	428	85.84	83.69	99.11	47.48	**100.00**	90.75	73.79	88.13	20.56	85.56	98.64
OA (%)		58.02	60.80	67.18	43.38	58.09	60.56	61.34	56.29	47.1	69.36	**73.06**
AA (%)		55.70	50.11	63.78	33.13	51.84	55.19	56.59	56.10	49.69	61.27	**64.31**
Kappa		0.48	0.50	0.59	0.26	0.48	0.51	0.51	0.47	0.37	0.61	**0.66**
NMI		0.58	0.65	0.71	0.41	0.61	0.60	0.62	0.56	0.49	0.64	**0.72**
ARI		0.49	0.48	0.57	0.18	0.45	0.44	0.44	0.39	0.30	0.55	**0.60**
Computing time (seconds)		**8.42**	469.64	63.76	649.63	1799.80	18.30	22.22	1120.61	2874.25	163.39	53.99

TABLE III
QUANTITATIVE ASSESSMENT OF ALL CONSIDERED CLUSTERING APPROACHES ON THE GEOLOGICAL FINLAND DATASET

Clusters	# Test samples	Clustering approaches										
		Shallow					Deep					
		k-means	HESSC	ESC	SC-SC	EnSC	AE	CAE	VAE	DCN	DMC-Net	MS^{2}A-Net
Clay	767	89.05	72.36	**100.00**	89.28	98.44	85.10	84.40	89.35	89.98	87.64	99.27
Glimmerite	381	72.43	**100.00**	67.11	16.00	73.33	91.37	69.00	48.26	66.50	94.67	**100.00**
Dark-rocks	659	79.09	49.08	0.00	5.98	2.62	73.12	**80.53**	61.38	78.80	41.89	47.40
Dust	282	39.44	91.05	0.00	10.32	47.80	49.20	42.62	38.19	36.19	**99.79**	99.39
Water	135	45.84	28.72	0.00	18.01	0.00	67.63	**68.32**	4.51	27.03	34.82	20.50
OA (%)		64.52	70.05	38.56	38.13	53.01	70.76	68.37	56.43	60.80	77.13	**80.48**
AA (%)		65.17	68.24	33.42	27.92	44.44	73.28	68.89	48.34	59.07	71.76	**73.31**
Kappa		0.55	0.60	0.08	0.09	0.32	0.61	0.59	0.42	0.49	0.69	**0.73**
NMI		0.51	0.57	0.36	0.11	0.36	0.58	0.58	0.47	0.51	0.69	**0.75**
ARI		0.44	0.52	0.08	0.05	0.25	0.53	0.49	0.38	0.42	0.66	**0.72**
Computing time (seconds)		**8.34**	2321.30	1953.20	12133.64	31232.00	25.36	112.85	2880.62	8216.10	163.20	57.84

fact that default hyperparameter values were used, or because the ground truth dataset was too small.

3) Quantitative Results on Geological Finland Dataset: The quantitative assessment of the different clustering approaches applied on the geological Finland dataset is reported in Table III. HESSC is capable of clustering the geological Finland scene (OA $=$ 70.05 %) more accurately than other CSL-based approaches. Despite the good performance of CAE in other datasets, in the geological Finland dataset, it performed slightly weaker (OA $=$ 70.76%) than AE (OA $=$ 68.37%). Similarly as in the Houston 2013 dataset, this may be caused by the low number of available test samples. MS^{2}A-Net attained the highest OA (80.48%) among all studied clustering approaches. DL-based approaches distinguished the "Dark-rocks" class better than the majority of the CSL-based approaches.

4) Quantitative Results on Geological Spain Dataset: The performance of the studied clustering approaches applied on the geological Spain dataset is reported in Table IV. Interestingly, the availability of a high number of test samples for various classes reveals valuable information. We can observe that some CSL-based approaches (i.e., ESC, SC-SC) are not applicable on this dataset. In the case of SC-SC, deriving sparse representation for a large-scale dataset is too computationally expensive. In the case of ESC, one can reduce the number of exemplars, but

TABLE IV
QUANTITATIVE ASSESSMENT OF ALL CONSIDERED CLUSTERING APPROACHES ON THE GEOLOGICAL SPAIN DATASET. "**OM**" DENOTES OUT OF MEMORY

Clusters	# Test samples	Clustering approaches										
		Shallow					Deep					
		k-means	HESSC	ESC	SC-SC	EnSC	AE	CAE	VAE	DCN	DMC-Net	MS^{2}A-Net
Chlorite rich schist	126526	44.14	35.46	**OM**	**OM**	51.16	45.66	59.91	55.46	62.81	79.67	**87.54**
Sericite rich schist	31007	44.11	32.52	**OM**	**OM**	1.07	43.57	44.54	32.86	18.10	46.84	**47.26**
Chlorite + Sericite rich schist	17215	12.41	20.00	**OM**	**OM**	0.82	0.47	0.22	**26.49**	13.23	23.51	15.45
Chert	3054	0.92	**60.51**	**OM**	**OM**	10.54	59.10	25.58	1.57	26.44	19.48	6.97
Massive Sulphide	9314	22.44	17.96	**OM**	**OM**	0.00	20.65	24.92	0.00	19.91	65.29	**97.71**
Purple Shale	1020	0.98	0.98	**OM**	**OM**	1.41	0.29	0.00	**21.86**	0.59	0.00	1.80
Phyllite	60523	33.94	27.75	**OM**	**OM**	39.15	33.86	**46.59**	33.05	17.88	32.46	40.50
Saprolite	20124	51.22	**65.38**	**OM**	**OM**	54.59	51.25	18.68	55.79	21.60	51.08	56.31
OA (%)		38.93	34.18	**OM**	**OM**	37.29	39.40	46.40	43.32	39.14	58.03	**64.25**
AA (%)		26.27	32.57	**OM**	**OM**	19.84	31.86	27.55	28.40	22.57	39.79	**43.44**
Kappa		0.21	0.22	**OM**	**OM**	0.16	0.22	0.27	0.26	0.10	0.42	**0.52**
NMI		0.22	0.21	**OM**	**OM**	0.22	0.23	0.23	0.22	0.18	0.34	**0.47**
ARI		0.20	0.14	**OM**	**OM**	0.14	0.21	0.22	0.22	0.08	0.41	**0.61**
Computing time (seconds)		**86.93**	1890.80	**OM**	**OM**	36498.00	101.70	289.74	7215.57	19470.43	892.93	299.67

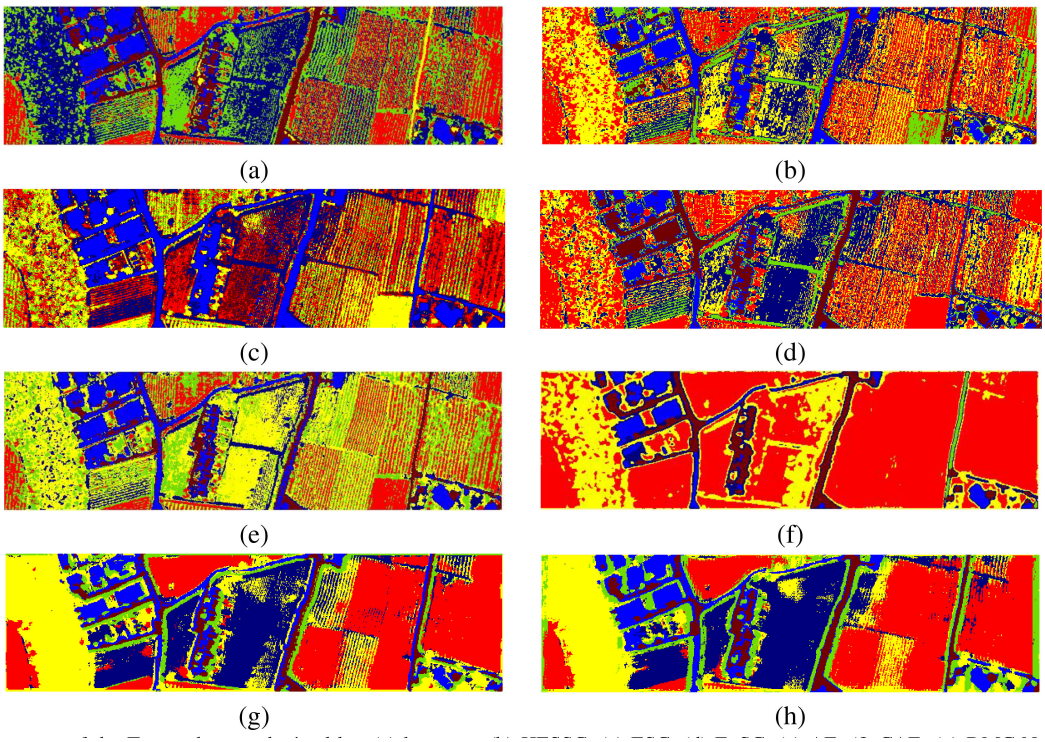


Fig. 8. Clustering maps of the Trento dataset obtained by (a) k-means, (b) HESSC, (c) ESC, (d) EnSC, (e) AE, (f) CAE, (g) DMC-Net, (h) MS^{2}A-Net.

this results in poor performance when the number of selected exemplars is not sufficient to represent the entire dataset. In addition, overall, CSL-based approaches have inferior performance compared to DL-based approaches. Furthermore, incorporating spatial information (i.e., CAE, DMC-Net, MS^{2}A-Net) ameliorates the clustering performance of DL-based approaches, compared to when merely spectral information is deployed.

5) Processing Time: All tables report required processing times on the four datasets. k-means is the fastest clustering approaches on all studied datasets, since it merely needs to compute the Euclidean distance between data points and centroids. AE is the fastest DL-based approach. However, comparing all other CSL- and DL-based approaches, MS^{2}A-Net is faster than any other approach. Among all approaches, EnSC and DCN are the most time consuming CSL- and DL-based clustering approaches, respectively.

6) Qualitative Assessment on Trento, Houston 2013, Geological Finland, and Geological Spain Datasets: Figs. 8–11 display clustering maps generated by the most representative CSL- and DL-based clustering approaches, which can handle complex and large scale datasets, on Trento, Houston 2013, geological Finland and geological Spain, respectively. There is a general trend, in which the spectral-based approaches (e.g., k-means, AE) tend to produce "noisy" maps in comparison

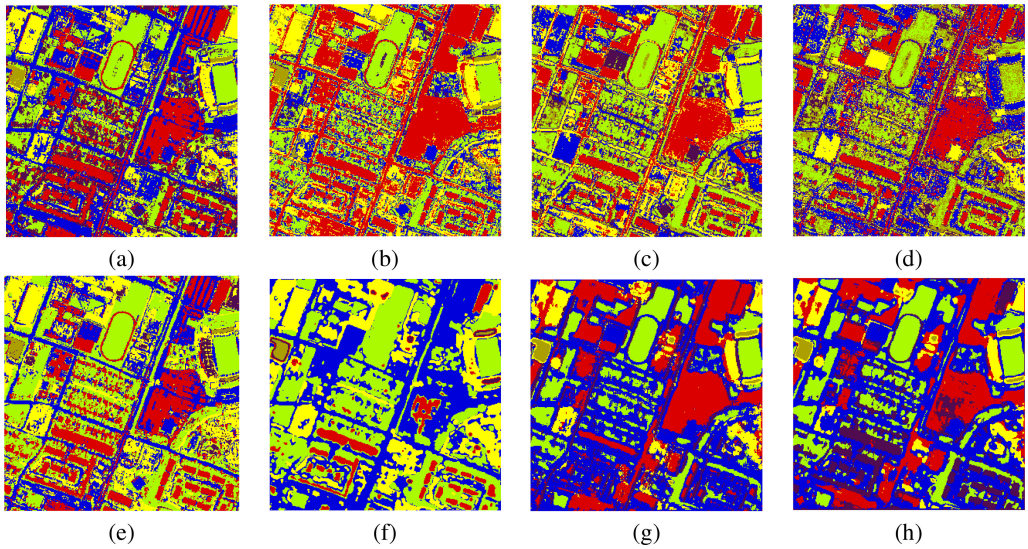


Fig. 9. Clustering maps of the Houston 2013 dataset obtained by (a) k-means, (b) HESSC, (c) ESC, (d) EnSC, (e) AE, (f) CAE, (g) DMC-Net, (h) MS^{2}A-Net.

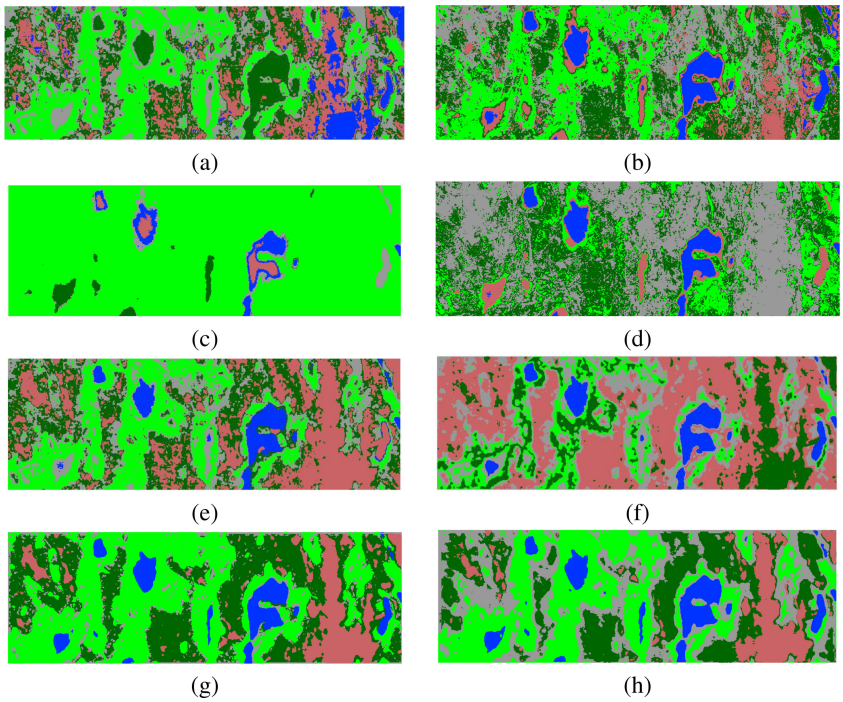


Fig. 10. Clustering maps of the geological Finland dataset obtained by (a) k-means, (b) HESSC, (c) ESC, (d) EnSC, (e) AE, (f) CAE, (g) DMC-Net, (h) MS^{2}A-Net.

to the ones (e.g., CAE, MS^{2}A-Net), which incorporate spatial information. Although CAE, DMC-Net, and MS^{2}A-Net generate smooth clustering maps, it can be observed that DMC-Net and MS^{2}A-Net provide more detailed clustering maps compared to CAE. This observation reveals the essence of using both spectral and spatial information in the clustering process. Further investigation of Fig. 8(g) and (h) demonstrates that MS^{2}A-Net more efficiently utilizes both spectral and spatial information compared to DMC-Net. For instance, several "Apple trees" pixels have been misclustered as "Wood" by DMC-Net, while this effect is reduced when MS^{2}A-Net is employed. The same trend is observed in Fig. 11(f) and (g), where the "Chlorite rich schist" class is not well clustered by DMC-Net, whereas MS^{2}A-Net can distinguish this class.

V. Discussion: Ablation Study and Hyperparameter Evaluation

In this section, we evaluate the impact of different hyperparameters on the performance of MS^{2}A-Net. In order to tune the MS^{2}A-Net and identify its corresponding optimal hyperparameters, we utilize the Trento dataset that has a richer and more

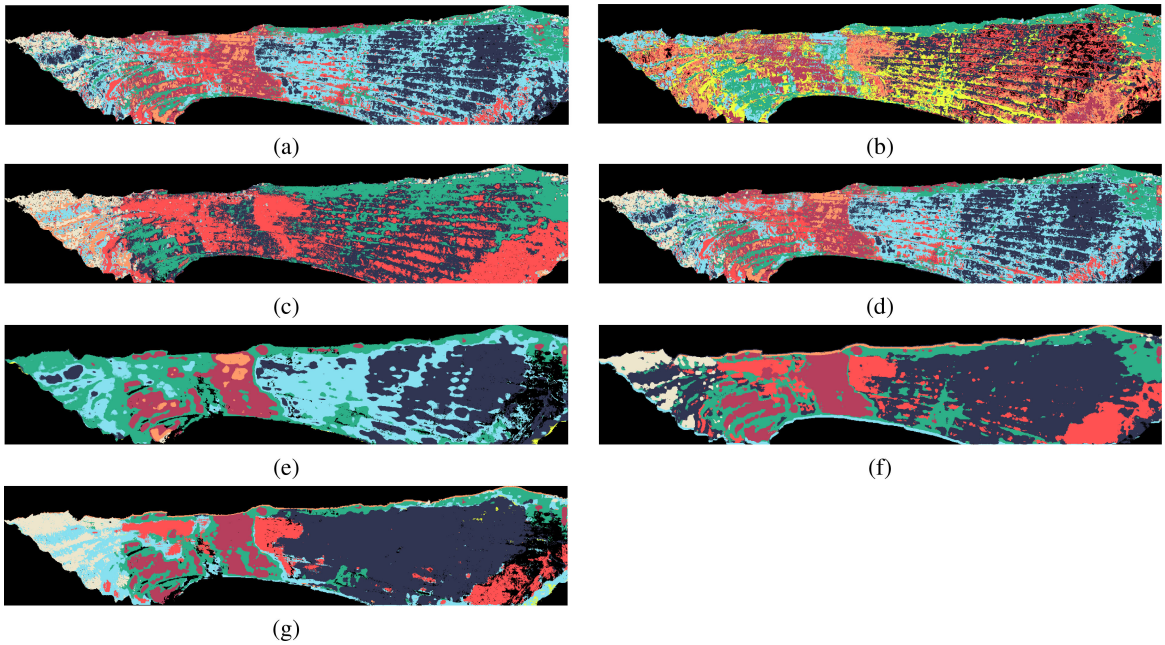


Fig. 11. Clustering maps of the geological Spain dataset obtained by (a) k-means, (b) HESSC, (c) EnSC, (d) AE, (e) CAE, (f) DMC-Net, (g) MS^{2}A-Net.

TABLE V
NUMBER OF LEARNABLE PARAMETERS IN MS^{2}A-NET

Different convolutional architectures	Datasets			
	Trento	Houston 2013	Geological Finland	Geological Spain
Dilated convolution	**62226**	**141930**	**49434**	**187194**
Normal convolution	304146	694890	241434	916794

balanced ground truth dataset, compared to the other studied datasets.

A. *Influence of the Multiscale Spatial Stream on the Number of Parameters*

In MS^{2}A-Net, dilated convolutions are utilized to cover a larger receptive field while requiring fewer learnable parameters. As reported in Table V, we computed the number of learnable parameters for two scenarios. In the first scenario, dilated convolutions are deployed as described in Section II. In the second scenario, we increase the kernel size of normal convolutions, to cover the same receptive field as their corresponding dilated convolutions at different scales. The results reveal that the deployment of dilated convolutions reduces the required number of learnable parameters (i.e., weights and biases) by approximately a factor of 5. Consequently, the computational effort of MS^{2}A-Net is reduced proportionally compared to its version with normal convolutions. Furthermore, in the studied datasets (i.e., rural, geological, and urban applications), pixels which lie within a close-range neighborhood from each other tend to be drawn from the same class. Therefore, to primarily capture the local neighboring information, we limited the receptive fields by selecting the dilated rates ($\{1, 2, 4\}$). However, depending on the applications at hand, other choices for these hyperparameters can be made.

B. *Impact of the Spectral Mean Constraint*

In the proposed approach, the ultimate goal is to cluster the extracted multiscale features ($\mathbf{M}'$). To include $\mathbf{M}'$ implicitly in the training process, we defined the spectral mean constraint on the generated $\mathbf{M}'$, to assure that the generated latent features have a direct impact on the training process, in such a way that they are enforced to preserve the mean spectral information of the original image. We evaluated the impact of the spectral mean constraint in Eq. (5) for the following values of λ: $\{0, 0.0001, 0.001, 0.01, 0.1\}$. In Fig. 12, the results are displayed in terms of OA(%). From the obtained results, one can conclude that $\lambda = 0.1$ leads to the highest OA and the lowest variation.

C. *Influence of the Number of Latent Features on the MS^{2}A-Net Performance*

In Fig. 13, the performance of MS^{2}A-Net is validated in terms of OA(%) by utilizing different numbers of extracted latent

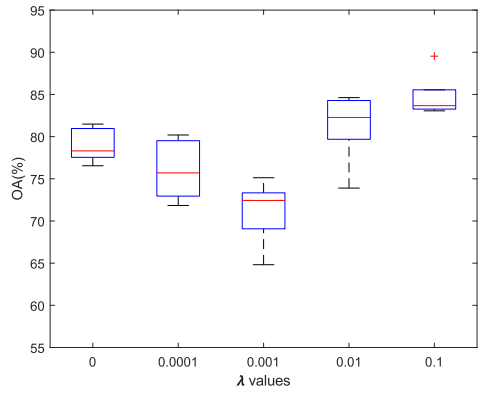


Fig. 12. Impact of the spectral mean constraint for different values of λ.

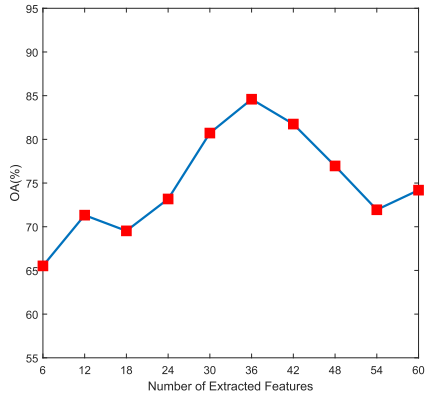


Fig. 13. Impact of the number of extracted latent features on the performance of MS^{2}A-Net.

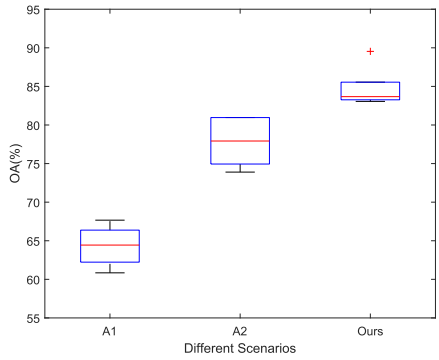


Fig. 14. Comparing the performance of MS^{2}A-Net by using different scenarios.

features ($d = 6, 12, 18, 24, 30, 36, 42, 48, 54, 60$). One can observe that the best performance is obtained using 36 features.

Thus, we propose to use 36 as the number of extracted features for all datasets.

D. Impact of the Multiscale Spatial and Spectral-Association Streams

We evaluated the effectiveness of the streams deployed in MS^{2}A-Net by using various alternative scenarios:

1) *Alternative* 1 (A1): In this alternative, only the spectral-association stream is deployed in the MS^{2}A-Net architecture, hereby completely ignoring the effect of the multiscale spatial stream in the training process.
2) *Alternative* 2 (A2): In this alternative, merely the multiscale spatial stream is deployed in the MS^{2}A-Net training process.
3) *Alternative* 3 (A3): This alternative is the proposed approach that deploys both spectral-association and multiscale spatial streams.

As shown in Fig. 14, poor results are obtained by scenario A1, compared to the other scenarios. The spectral-association stream mainly uses spectral information, and the lack of sufficient spatial information strongly deteriorates the clustering performance. With scenario A2, the clustering performance improves by the beneficial influence of spatial information. However, one can observe that optimal clustering performance is obtained in scenario A3, when both multiscale spatial and spectral-association streams are deployed.

VI. Conclusion

HSI clustering is a challenging task, which can provide valuable insight into datasets. Unlike CSL-based clustering approaches, DL-based approaches can capture nonlinear intrinsic relationships between data points complex datasets. Furthermore, most CSL- and DL-based HSI clustering approaches merely utilize spectral information, while neighboring pixels likely share the same characteristics. Hence, in this article, we proposed a multiscale spectral–spatial association network, which effectively exploits spectral and spatial information for HSI clustering. MS^{2}A-Net contains two main streams (i.e., a spectral-association and a multiscale spatial stream). The spectral-association stream aims to efficiently extract spectral information, whereas the multiscale spatial stream deploys dilated convolutions to capture spatial information at various scales. We have demonstrated that MS^{2}A-Net outperforms the state-of-the-art CSL- and DL-based clustering approaches with competitive processing times on four real hyperspectral datasets.

In the future, we will work on different optimization strategies to further improve the clustering performance. In addition, we intend to design even lighter networks with a reduced number of learnable parameters, and consequently processing time.

Acknowledgment

The authors would like to thank L. Bruzzone of the University of Trento for providing the Trento dataset. In addition, the authors would like to acknowledge the National Center for Airborne Laser Mapping (NCALM) for providing the Houston dataset. The authors would also like to thank the Hyperspectral Image Analysis Group, University of Houston and the IEEE GRSS DFC 2013 for providing the CASI University of Houston dataset.

REFERENCES

[1] I. H. Khan et al., "Early detection of powdery mildew disease and accurate quantification of its severity using hyperspectral images in wheat," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3612.

[2] A. D. Rocha, T. A. Groen, and A. K. Skidmore, "Spatially-explicit modelling with support of hyperspectral data can improve prediction of plant traits," *Remote Sens. Environ.*, vol. 231, 2019, Art. no. 111200.

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

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

[5] R. Jackisch et al., "Integrated geological and geophysical mapping of a carbonatite-hosting outcrop in Siilinjärvi, Finland, using unmanned aerial systems," *Remote Sens.*, vol. 12, no. 18, 2020, Art. no. 2998. [Online]. Available: https://www.mdpi.com/2072-4292/12/18/2998

[6] P. Ghamisi et al., "The potential of machine learning for a more responsible sourcing of critical raw materials," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8971–8988, 2021.

[7] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, Mar. 2017.

[8] H. Zhai, H. Zhang, P. LI, and L. Zhang, "Hyperspectral image clustering: Current achievements and future lines," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 35–67, Dec. 2021.

[9] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[10] C. You, C. Li, D. P. Robinson, and R. Vidal, "Self-representation based unsupervised exemplar selection in a union of subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2698–2711, May 2022.

[11] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong, "Towards k-means-friendly spaces: Simultaneous deep learning and clustering," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3861–3870.

[12] A. Tasissa, D. Nguyen, and J. M. Murphy, "Deep diffusion processes for active learning of hyperspectral images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2021, pp. 3665–3668.

[13] D. Arthur and S. Vassilvitskii, "k-means plus plus: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2006, pp. 1027–1035.

[14] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[15] R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, Mar. 2011.

[16] H. Zhang, H. Zhai, L. Zhang, and P. Li, "Spectral-spatial sparse subspace clustering for hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3672–3684, Jun. 2016.

[17] S. Matsushima and M. Brbic, "Selective sampling-based scalable sparse subspace clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[18] J. Bruton and H. Wang, "Dictionary learning for clustering on hyperspectral images," *Signal, Image, Video Process.*, vol. 15, no. 2, pp. 255–261, 2021.

[19] K. R. Shahi, P. Ghamisi, B. Rasti, P. Scheunders, and R. Gloaguen, "Unsupervised data fusion with deeper perspective: A novel multisensor deep clustering algorithm," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 284–296, 2022.

[20] 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. 39501–39514, 2018.

[21] X. Guo, X. Liu, E. Zhu, and J. Yin, "Deep clustering with convolutional autoencoders," in *Proc. Int. Conf. Neural Inf. Process.*, 2017, pp. 373–382.

[22] J. Sun et al., "Deep clustering with intraclass distance constraint for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4135–4149, May 2021.

[23] Y. Cai, M. Zeng, Z. Cai, X. Liu, and Z. Zhang, "Graph regularized residual subspace clustering network for hyperspectral image clustering," *Inf. Sci.*, vol. 578, pp. 85–101, 2021.

[24] J. Lei, X. Li, B. Peng, L. Fang, N. Ling, and Q. Huang, "Deep spatial-spectral subspace clustering for hyperspectral image," *IEEE Trans. Circuits, Syst. Video Technol.*, vol. 31, no. 7, pp. 2686–2697, Jul. 2021.

[25] 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 Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1948–1952, Nov. 2020.

[26] Y. Qin, L. Bruzzone, and B. Li, "Learning discriminative embedding for hyperspectral image clustering based on set-to-set and sample-to-sample distances," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 473–485, Jan. 2019.

[27] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[28] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[29] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[31] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral-spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5514715.

[32] C. Debes et al., "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.

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

[34] M. Kirsch et al., "Hyperspectral outcrop models for palaeoseismic studies," *Photogrammetric Rec.*, vol. 34, no. 168, pp. 385–407, 2019.

[35] M. Rezaei and P. Fränti, "Set matching measures for external cluster validity," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 8, pp. 2173–2186, Aug. 2016.

[36] J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for k-means clustering," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*. New York, NY, USA:ACM, 2009, pp. 877–886. [Online]. Available: http://doi.acm.org/10.1145/1557019.1557115

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

[38] 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 Sens.*, vol. 12, no. 15, 2020, Art. no. 2421.

[39] C. You, C.-G. Li, D. P. Robinson, and R. Vidal, "Oracle based active set algorithm for scalable elastic net subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3928–3937.

[40] B. Rasti et al., "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, Dec. 2020.

Kasra Rafiezadeh Shahi (Student Member, IEEE) received the B.Eng. degree in computer engineering from the Urmia University of Technology, Urmia, Iran, in 2015 and the M.Sc. degree in geoinformation science and earth observation from the faculty of ITC, University of Twente, The Netherlands, in 2018. He is currently working toward the Ph.D. degree in developing unsupervised learning techniques for image and signal processing using remote sensing techniques with the Faculty of Physics, University of Antwerp, Belgium.

Furthermore, he works as a Researcher in the Machine Learning Group, Department of Exploration at Helmholtz–Zentrum Dresden–Rossendorf (HZDR), Germany. His research interests include clustering, and multisensor data fusion using unsupervised (shallow/deep) learning techniques, particularly for remote sensing applications. He serves as a Reviewer for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL), and Remote Sensing (Multidisciplinary Digital Publishing Institute).

Pedram Ghamisi (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Iceland, Reykjavik, Iceland, in 2015.

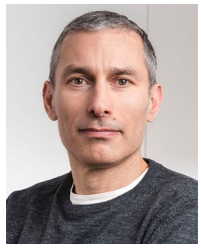
He works as the Head of the Machine Learning Group, Helmholtz–Zentrum Dresden–Rossendorf (HZDR), Germany, and a visiting Professor and a group leader of AI4RS with the Institute of Advanced Research in Artificial Intelligence (IARAI), Austria. He is a cofounder of VasoGnosis Inc. with two branches in San Jose and Milwaukee, USA. He was the cochair of IEEE Image Analysis and Data Fusion Committee (IEEE IADF) between 2019 and 2021. 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 JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSINGand IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. For detailed info, please see http://pedram-ghamisi.com/.

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 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.

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 Postdoctoral 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 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.



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, London, U.K., and Göttingen University, Göttingen, Germany, in 2000.

He was a Marie Curie Postdoctoral Research Associate with the Royal Holloway University of London from 2000 to 2003. He led the Remote Sensing Group, 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 noninvasive exploration. His research interests focus on multisource and multiscale remote sensing integration.

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.

He became a Research Associate with the Vision Lab, Department of Physics, University of Antwerp, in 1991, where he is a Full Professor. His research interest includes remote sensing and hyperspectral image processing. He has authored more than 200 papers in international journals and proceedings in the field of image processing, pattern recognition, and remote sensing. He is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSINGand has served as a program committee member for numerous international conferences.