Multimodal Hyperspectral Unmixing: Insights From Attention Networks

Zhu Han, *Student Member, IEEE*, Danfeng Hong, *Senior Member, IEEE*, Lianru Gao, *Senior Member, IEEE*, Jing Yao, Bing Zhang, *Fellow, IEEE*, and Jocelyn Chanussot, *Fellow, IEEE*

***Abstract*— Deep learning (DL) has aroused wide attention in hyperspectral unmixing (HU) owing to its powerful feature representation ability. As a representative of unsupervised DL approaches, autoencoder (AE) has been proven to be effective to better capture nonlinear components of hyperspectral images than the traditional model-driven linearized methods. However, only using hyperspectral images for unmixing fails to distinguish objects in complex scene, especially for different endmembers with similar materials. To overcome this limitation, we propose a novel multimodal unmixing network for hyperspectral images, called MUNet, by considering the height differences of light detection and ranging (LiDAR) data in a squeeze-and-excitation (SE)-driven attention fashion to guide the unmixing process, yielding performance improvement. MUNet is capable of fusing multimodal information and using the attention map derived by LiDAR to aid network that focuses on more discriminative and meaningful spatial information regarding scenes. Moreover, attribute profile (AP) is adopted to extract the geometrical structures of different objects to better model the spatial information of LiDAR. Experimental results on synthetic and real datasets demonstrate the effectiveness and superiority of the proposed method compared with several state-of-the-art unmixing algorithms. The codes will be available at https://github.com/hanzhu97702/IEEE_TGRS_MUNet, contributing to the remote sensing community.**

***Index Terms*— Attention, autoencoder (AE), deep learning (DL), hyperspectral unmixing (HU), light detection and ranging (LiDAR), multimodality.**

I. Introduction

HYPERSPECTRAL imagery (HSI) provides hundreds of contiguous narrow spectral bands, which enables various objects to be identified and discriminated in remote sensing (RS) applications [1]. However, owing to the relatively low spatial resolution of sensors, many pixels in HSI usually contain reflections from multiple types of materials, inevitably degrading the performance of high-level data processing. Hyperspectral unmixing (HU) aims at separating the mixed pixels into a set of endmember signatures and their corresponding abundances [2].

Depending on the photon interaction mechanism in the scene, two mixing assumptions are applied in HU: linear mixing model (LMM) and nonlinear mixing model (NLMM) [3]. LMM assumes that the observed pixel is a linear combination of different endmembers weighted by their fractional abundances, but it does not consider intimate reflections and multiple scattering interactions, especially in desert and urban areas [4]. To deal with these nonlinear interactions, numerous NLMMs have been applied to complex real scenarios by modeling different order scattering effects and produce more accurate unmixing results [5]–[10]. Nevertheless, they usually require some prior knowledge about nonlinear characteristics to establish the NLMM. In recent years, the deep learning (DL) methods solve this bottleneck and can automatically extract robust and high-level features from a data-driven perspective [11]. As a representative of unsupervised DL approaches, autoencoder (AE) has become a hotspot for HU, because it can simultaneously learn low-dimensional representation (e.g., the abundance) of data and the corresponding weight base (e.g., the endmember) by minimizing the reconstruction error [12]. To further improve the unmixing performance, a handful of improvements have been applied to the AE framework, such as denoising [13], sparsity [14], spatiality [15]–[18], generative adversarial module [19], and self-supervised deep prior [20]–[22]. Despite being able to efficiently unmix, these aforementioned AE-based approaches only focus on single modality and fail to accurately discriminate different objects produced by the same material, e.g., concrete road and concrete roof [23]. Therefore, it is crucial to develop and incorporate multi-source data to assist HSI for better unmixing results.

Up until now, aerospace and aerial RS data can be acquired through different sensors with the rapid development of imaging techniques [24], e.g., light detection and ranging (LiDAR) [25], synthetic aperture radar (SAR) [26], and passive devices, e.g., multispectral imagery (MSI) [27] and HSI [28], [29], which provides diverse characteristics about various

Manuscript received December 15, 2021; revised January 28, 2022; accepted February 16, 2022. Date of publication March 2, 2022; date of current version April 5, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62161160336 and Grant 42030111, in part by MIAI@Grenoble Alpes under Grant ANR-19-P3IA-0003, and in part by the AXA Research Fund. *(Corresponding author: Lianru Gao.)*

Zhu Han and Bing Zhang are with the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: hanzhu19@mails.ucas.ac.cn; zb@radi.ac.cn).

Danfeng Hong, Lianru Gao, and Jing Yao are with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: hongdf@aircas.ac.cn; gaolr@aircas.ac.cn; yaojing@aircas.ac.cn).

Jocelyn Chanussot is with the University Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, 38000 Grenoble, France, and also with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: jocelyn@hi.is).

Digital Object Identifier 10.1109/TGRS.2022.3155794

objects or materials in the scene. Moreover, there is an increased interest in exploiting more effective spectral and spatial techniques by means of multimodal RS data. For example, Hang *et al.* [30] proposed a coupled convolutional neural network (CNN) for collaborative classification of HSI and LiDAR. To enhance joint representations of different modalities, Mohla *et al.* [31] used attention mask produced by one modality to highlight features in HSI and this attention mechanism can effectively learn the associated feature representations. Hu *et al.* [32] proposed a semi-supervised manifold alignment method by fusing optimal and SAR data, which showed superior performance in land use classification and local climate zone classification. Hong *et al.* [33] further explored the cross-modality learning DL framework in RS image classification applications and achieved more compact modality blending. Inspired by the success of multimodal data processing technology, Uezato *et al.* [34] first incorporated external LiDAR to adjust standard spatial regularization in the unmixing process. To handle the spatial similarity among the neighborhood pixels, hypergraph regularization was further introduced to improve abundance estimation [35]. However, these above-mentioned LiDAR spatial regularization unmixing methods only considered the elevation information of neighboring pixels and lacked sufficient exploration of high-dimensional features in LiDAR data, which leads to unmixing results susceptible to endmember variability [36]. In addition, in the process of solving abundances, these approaches ignored the exploration of endmember extraction from the perspective of multimodal data, and the setting of endmembers still needed to be manually given in advance.

How to acquire complete and meaningful information of multimodal data still faces great challenges. With a growing demand for intrinsic properties of multimodal data, it is difficult to meet the requirements by relying on manually designed feature extraction techniques. Recently, the attention-based methods have broadly replaced handcraft approaches in many domains, such as object detection [37]–[39] and image classification [40]–[42], which are able to effectively extract the most valuable and informative features in a given scene. Attention is originally derived by the study of human vision, and its goal is to use limited visual resources to select core parts in the image. By assigning different weight coefficients to these parts, the image is guided to adaptively focus on the detailed information of the specific target while suppressing other irrelevant information. Considering the spectral dimension data in HSI, the attention mechanism has been proven to be effective in capturing spectral correlations between adjacent spectra so far. For instance, Zeng *et al.* [43] designed an attention-based residual network for HU with limited training samples, and the attention architecture can help the unmixing network pay attention to important features in HSI. Qing *et al.* [44] used an efficient channel attention classification method based on multi-scale residual CNN and solved the problem of gradient dispersion and sample information redundancy. Sun *et al.* [45] proposed a spectral–spatial attention method by embedding attention modules into CNN to extract discriminative spectral–spatial features for HSI classification. Xue *et al.* [46] adopted a spectral–spatial self-attention module to adaptively calibrate weight coefficients of different scale features in multimodal data, thereby improving the overall accuracy of HSI classification. Although these attention-based approaches can play a role in capturing features, current researches on the multimodal attention mechanism in HSI are still scarce. In fact, the multimodal RS datasets within the same scene contain rich land-cover information, and there is still room for improvement in effectively integrating the multimodality features with attention mechanism.

To this end, we propose a multimodal unmixing network, called MUNet, in which the squeeze-and-excitation (SE) attention is incorporated into the AE unmixing network, to effectively fuse HSI and LiDAR features in an unsupervised fashion. Compared with the existing multimodal unmixing methods that only consider low-dimensional elevation information of neighboring pixels, MUNet is capable of focusing on the most important and useful feature information extracted by LiDAR and guiding the encoder of AE to obtain more accurate abundance results. More specifically, the major contributions can be summarized as follows.

1) We propose an end-to-end multimodal unmixing network for the HU task, MUNet for short, by integrating the height differences of LiDAR data into HSI to enhance unmixing performance. Considering the performance bottleneck only using the single modality (e.g., hyperspectral data) for HU, the proposed MUNet can make full use of the height information obtained from LiDAR data as prior knowledge to better guide the unmixing process more accurately. To the best of our knowledge, this is the first time to investigate the multimodal unmixing task using DL.
2) We propose to embed the height information obtained from LiDAR data into the AE-based unmixing architecture in an attention fashion. More specifically, an SE-driven attention mechanism is designed to represent the height knowledge by the way of weighted multiplications in the process of unmixing HSIs, yielding a significant performance improvement.
3) The attribute profile (AP) is introduced to better model the spatial information of LiDAR data and assist the subsequent attention mechanism to converge quickly. Compared with inputting single LiDAR data, the performance of this attribute strategy is effectively verified on both the synthetic and real multimodal datasets.

The remaining of this article is organized as follows. Section II briefly introduces the related multimodal unmixing method and the AE framework. Section III details the design of the proposed MUNet method. Section IV validates the proposed method with experiments in one synthetic and two real multimodal datasets. Section V concludes this article with some remarks and presents the perspective of the future work.

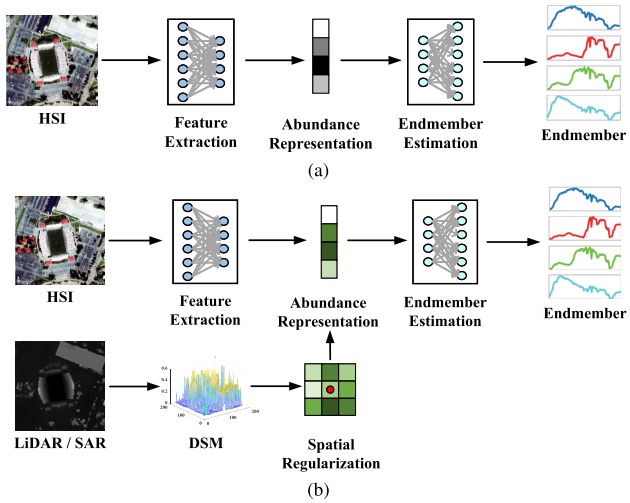


Fig. 1. Illustration to clarify the similarities and differences between single-modality unmixing method and multimodal unmixing method using DL. (a) Workflow for the single-modality unmixing method. (b) Workflow for the multimodal unmixing method.

II. Related Work

In this section, we first outline the existing multimodal unmixing framework for LMM problem and then provide a detailed description of the AE unmixing approaches.

A. Existing Multimodal Unmixing Approaches

The general description of the LMM can be formulated as follows:

$$\begin{aligned} &\mathbf{Y} = \mathbf{MA} + \mathbf{N} \\ &\text{s.t. } \mathbf{1}_S^T \mathbf{A} = \mathbf{1}_N^T, \quad \mathbf{A} \geq \mathbf{0}, \ \mathbf{M} \geq \mathbf{0} \end{aligned} \tag{1}$$

where $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N] \in \mathbb{R}^{L\times N}$ represents the input HSI with L spectral bands and N pixels. $\mathbf{y}_i$ stands for the ith observed spectrum. $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_P] \in \mathbb{R}^{L\times P}$ is the endmember matrix with P endmember categories and $\mathbf{m}_i$ denotes the ith endmember vector. $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N] \in \mathbb{R}^{P\times N}$ is the abundance matrix and $\mathbf{a}_i$ denotes the corresponding fractional abundance of the ith observed pixel. Each abundance vector $\mathbf{a}_i$ should satisfy the abundance sum-to-one constraint (ASC) and the abundance non-negativity constraint (ANC). $\mathbf{M}$ is also required to satisfy the endmember non-negative constraint (ENC). $\mathbf{N} \in \mathbb{R}^{L\times N}$ denotes the additive noise matrix.

Fig. 1 briefly illustrates the similarities and differences between the single-modality unmixing method and the multimodal unmixing method. It is noted that both these types of unmixing approaches obtain abundance representation through feature extraction, and then use the extracted abundances to estimate the endmember results. However, in the multimodal unmixing framework, different modal datasets are introduced into feature extraction, such as LiDAR and SAR, and abundance representation is obtained by applying spatial regularization derived from the multimodal data to abundance extraction, thereby acquiring the corresponding endmember results. Compared with the single-modality unmixing method, the advantage of the multimodal unmixing method is that it can make full use of the latent features from different modalities to improve the unmixing accuracy and avoid the shortcoming of single-modality data missing significant object information in complex scenes.

In this article, LiDAR is considered as the external multimodal data to enhance unmixing performance in HSI, because LiDAR can provide essential height information to distinguish spectrally similar materials. By combining the guidance map derived from the LiDAR data, the HU problem can fuse more spatial information to obtain ideal unmixing results. Since the existing multimodal methods only focus on applying the extracted guidance map to abundance regularization, a set of endmembers $\mathbf{M}$ should be extracted by the traditional geometric endmember extraction methods in advance. Based on the given endmembers, the estimation of abundances is solved by the following optimization problem:

$$\min_{\mathbf{A}} \frac{1}{2}\|\mathbf{Y} - \mathbf{MA}\|_F^2 + \lambda\phi(\mathbf{A}) \tag{2}$$

where $\phi(.)$ is the spatial regularization function based on LiDAR data, and λ is the tradeoff parameter to balance the reconstruction term and spatial regularization.

The height information of neighboring pixels is considered in (2) by defining the total variation (TV) spatial regularization [34] as

$$\phi(\mathbf{A}) = \sum_{i=1}^{N} \sum_{j\in\mathcal{N}(i)} w_{ij}\|\mathbf{a}_i - \mathbf{a}_j\|_1 \tag{3}$$

where $\mathcal{N}(i)$ denotes the set of neighboring pixels for the ith pixel, and w_{ij} represents the weight coefficient of height similarity, which is given by

$$w_{ij} = e^{-\frac{1}{\sigma^2}\frac{(h_i - h_j)^2}{(h_i + h_j)^2}} \tag{4}$$

where σ^2 is the controller parameter to balance the weight range. h_i and h_j denote the heights corresponding to the ith and jth pixels provided by LiDAR, respectively. Note that the weight should satisfy $\sum_{j\in\mathcal{N}(i)} w_{ij} = 1$ for the ith pixel.

In addition, the guidance map derived from HSI can also be introduced in (4) to realize the joint abundance solution, such as TV and spatial hypergraph (SH) regularization [35], [47]. Although current multimodal unmixing approaches have proven that using the multimodal prior knowledge can help improve unmixing performance, the design of spatial regularization only focuses on shallow features and lacks sufficient exploration of high-dimensional representations in multimodal data.

B. AE-Based Unmixing Networks

Owing to its powerful representation and reconstruction capabilities, AE has become a typical representative of unsupervised DL models in the field of HU. In general, the AE consists of two parts, namely, an encoder and a decoder. The encoder part learns the input pixel $\mathbf{y}_i \in \mathbb{R}^N$ into a hidden low-dimensional representation $\mathbf{v}_i \in \mathbb{R}^P$, which can be expressed as

$$\mathbf{v}_i = f_E(\mathbf{y}_i) = f(\mathbf{W}^{(e)T}\mathbf{y}_i + \mathbf{b}^{(e)}) \tag{5}$$

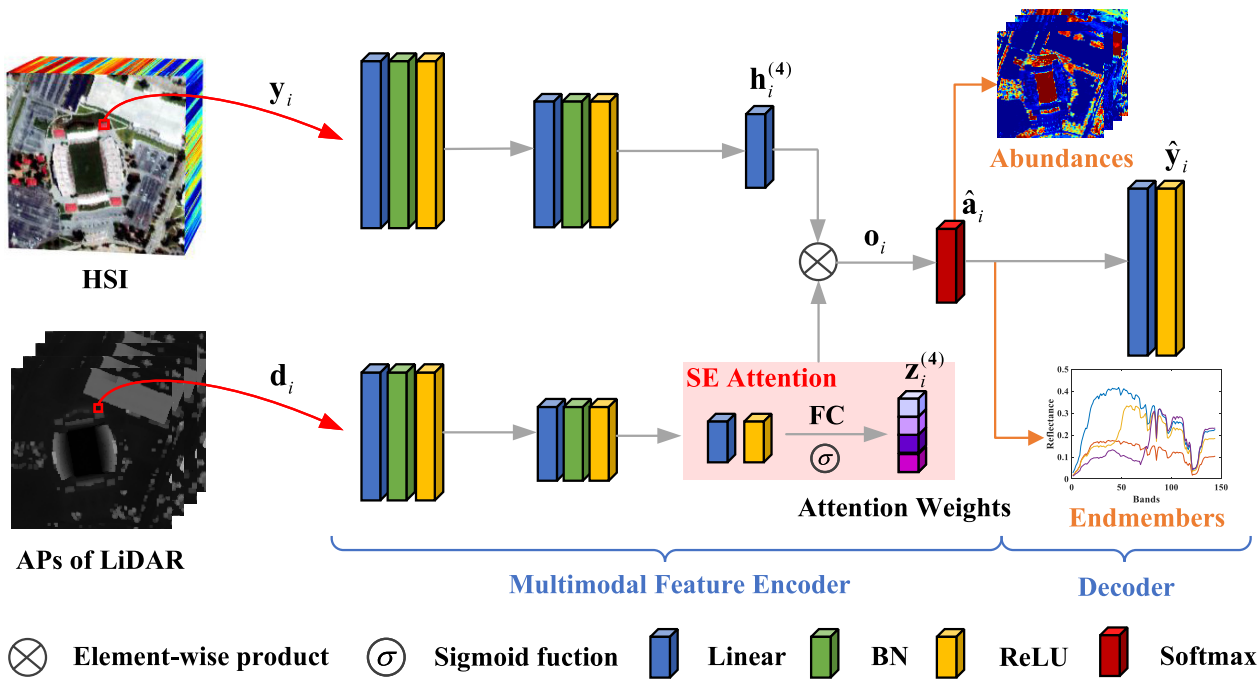


Fig. 2. Architecture of the proposed framework (MUNet), which consists of two-stream multimodal feature encoders and one decoder. The AP strategy and SE attention are used to learn morphological and high-dimensional features of LiDAR to further enhance unmixing performance. BN and FC stand for batch normalization and fully connected network, respectively.

where $f(.)$ is the nonlinear activation function, such as the rectified linear unit (ReLU) and the sigmoid function. $\mathbf{W}^{(e)}$ and $\mathbf{b}^{(e)}$ denote the weight and the bias in the eth encoder part, respectively.

The decoder aims to transform the extracted hidden representation into the original input pixel based on LMM, and the reconstructed pixel $\hat{\mathbf{y}}_i \in \mathbb{R}^N$ is denoted by

$$\hat{\mathbf{y}}_i = f_D(\mathbf{v}_i) = \mathbf{W}^{(d)T}\mathbf{v}_i \tag{6}$$

where $\mathbf{W}^{(d)}$ represents the weight matrix in the decoder part. Since the solution of the decoder part is consistent with LMM in (1), the results of the extracted endmember matrix $\hat{\mathbf{M}}$ and the estimated abundance vector $\hat{\mathbf{a}}_i$ correspond to $\mathbf{W}^{(d)}$ and $\mathbf{v}_i$, respectively.

The objective function of the AE unmixing network is realized by minimizing the reconstruction error between $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ in different measurement forms, such as mean square error (MSE) and spectral angle distance (SAD), given by

$$J_{\mathrm{MSE}}(\hat{\mathbf{y}}_i, \mathbf{y}_i) = \frac{1}{N}\sum_{i=1}^{N}\|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2 \tag{7}$$

$$J_{\mathrm{SAD}}(\hat{\mathbf{y}}_i, \mathbf{y}_i) = \arccos\left(\frac{\hat{\mathbf{y}}_i^T\mathbf{y}_i}{\|\hat{\mathbf{y}}_i\|_2\|\mathbf{y}_i\|_2}\right). \tag{8}$$

III. Problem Formulation and Method

The architecture of the proposed MUNet framework is shown in Fig. 2, containing two-stream multimodal feature encoder parts and one decoder part. The former aims at learning hierarchical representations of the hyperspectral and LiDAR data by integrating the AP technique and the SE attention mechanism. The latter is a common decoder architecture, which uses the extracted abundances to reconstruct HSI. In the following parts, we specifically detail the proposed MUNet framework.

A. Multimodal Feature Encoder

To enhance multimodal unmixing performance, we propose two-stream multimodal feature encoders to learn discriminative representations of hyperspectral and LiDAR data. First, the LiDAR image D is extended to multi-band profiles by AP [48], which can model the spatial information of D by applying S attribute thinning (γ^T) and S attribute thickening (ρ^T) operations, given by

$$D_{\mathrm{AP}} = \left\{\gamma_S^T(D), \ldots, \gamma_1^T(D), D, \rho_1^T(D), \ldots, \rho_S^T(D)\right\} \tag{9}$$

where $D_{\mathrm{AP}} \in \mathbb{R}^{H_D \times W_D \times (2S+1)}$ is the AP result for the LiDAR data. H_D, W_D, and $2S+1$ represent the height, the width, and the dimensional number of D_{AP}, respectively.

The overall network configuration in the proposed MUNet is shown in Table I. The architecture of MUNet is divided into six blocks. Among them, blocks 1–5 represent the two-stream multimodal feature encoder, and block 6 is the decoder. Given the input hyperspectral and LiDAR pixel, denoted as $\{\mathbf{y}_i\}_{i=1}^N \in \mathbb{R}^L$ and $\{\mathbf{d}_i\}_{i=1}^N \in \mathbb{R}^{2S+1}$, blocks 1–4 perform feature extraction by the following transformation:

$$\mathbf{h}_i^{(e)} = \begin{cases} f\left(BN_{\gamma,\beta}\left(\mathbf{W}_h^{(e)T}\mathbf{y}_i + \mathbf{b}_h^{(e)}\right)\right), & e = 1 \\ f\left(BN_{\gamma,\beta}\left(\mathbf{W}_h^{(e)T}\mathbf{h}_i^{(e-1)} + \mathbf{b}_h^{(e)}\right)\right), & e = 2, 3 \\ \mathbf{W}_h^{(e)T}\mathbf{h}_i^{(e-1)} + \mathbf{b}_h^{(e)}, & e = 4 \end{cases} \tag{10}$$

$$\mathbf{z}_i^{(e)} = \begin{cases} f\left(BN_{\gamma,\beta}\left(\mathbf{W}_d^{(e)T}\mathbf{d}_i + \mathbf{b}_d^{(e)}\right)\right), & e = 1 \\ f\left(BN_{\gamma,\beta}\left(\mathbf{W}_d^{(e)T}\mathbf{z}_i^{(e-1)} + \mathbf{b}_d^{(e)}\right)\right), & e = 2, 3 \\ g_{\mathrm{SE}}\left(\mathbf{z}_i^{(e-1)}, \mathbf{W}_d^{(e)}, \mathbf{b}_d^{(e)}\right), & e = 4 \end{cases} \tag{11}$$

TABLE I
NETWORK CONFIGURATION OF THE PROPOSED MUNet

Architecture	Pathway	Layer composition		Unit	
		HSI	LiDAR	HSI	LiDAR
Multimodal Feature Encoder	Block 1	Linear BN ReLU	Linear BN ReLU	L	2S + 1
	Block 2	Linear BN ReLU	Linear BN ReLU	L / 2	S
	Block 3	Linear BN ReLU	Linear BN ReLU	L / 4	P
	Block 4	Linear - - -	Linear ReLU Linear Sigmoid	P - - -	P / 2 P
	Block 5	Softmax		P	
Decoder	Block 6	Linear ReLU		L	

where $\mathbf{h}_i^{(e)}$ and $\mathbf{z}_i^{(e)}$ denote the extracted hierarchical representations of the hyperspectral and LiDAR data in the eth encoder block, respectively. $\{\mathbf{W}_h^{(e)}, \mathbf{b}_h^{(e)}\}$ and $\{\mathbf{W}_d^{(e)}, \mathbf{b}_d^{(e)}\}$ are the set of weights and biases in the encoder part of two modal datasets, respectively. $f(.)$ is the ReLU nonlinear activation function. $BN_{\gamma,\beta}(\mathbf{x}_i) = \gamma\hat{\mathbf{x}}_i + \beta$ represents the batch normalization (BN) layer to speed up the parameter learning and avoid the problem of vanishing gradients in the training phase [49]. Block 4 in the LiDAR stream is a SE attention layer g_{SE}, which aims at using the learned channel relationship to emphasize high-dimensional features of LiDAR data. Different from the original research in [50], the proposed SE attention mechanism does not consider the global average pooling (GAP) part, because the GAP operation will not only reduce the convergence speed of the unmixing network but also lose some characteristic information, such as edges and outliers. Therefore, the improved SE attention operation is designed as

$$\mathbf{z}_i^{(4)} = g_{\mathrm{SE}}(\mathbf{z}_i^{(3)}, \mathbf{W}_d^{(4)}, \mathbf{b}_d^{(4)}) = \sigma(\mathbf{W}_2 f(\mathbf{W}_1\mathbf{z}_i^{(3)} + \mathbf{b}_1) + \mathbf{b}_2) \tag{12}$$

where $\mathbf{W}_1 \in \mathbb{R}^{P/2\times P}$ and $\mathbf{W}_2 \in \mathbb{R}^{P\times P/2}$ are the weight matrices of two successive linear layers in block 4. $\sigma(.)$ denotes the sigmoid activation function, aiming to generate the attention coefficients between 0 and 1.

With the results of (10) and (12), the recalibrated output of block 4 can be formulated as

$$\mathbf{o}_i = \mathbf{h}_i^{(4)} \odot \mathbf{z}_i^{(4)} \tag{13}$$

where $\odot$ is the element-wise product between different feature vectors.

Moreover, to guarantee ANC and ASC constraint, the softmax function is applied in (13) to obtain the estimated abundance result, which is given by

$$\hat{\mathbf{a}}_i = \frac{e^{\mathbf{o}_i}}{\sum_{j=1}^{P} e^{\mathbf{o}_j}} \tag{14}$$

where $\hat{\mathbf{a}}_i$ represents the ith estimated abundance vector.

B. Decoder

The decoder is designed to reconstruct the input pixels by integrating the estimated abundances and the corresponding endmembers, which is written as

$$\hat{\mathbf{y}}_i = f(\mathbf{W}^{(d)T}\hat{\mathbf{a}}_i) = \hat{\mathbf{M}}\hat{\mathbf{a}}_i \tag{15}$$

where $\hat{\mathbf{M}} \in \mathbb{R}^{L\times P}$ and $\hat{\mathbf{y}}_i \in \mathbb{R}^L$ denote the estimated endmember matrix and the reconstructed pixel, respectively. Note that to promote the training of the decoder part, the vertex component analysis (VCA) algorithm is adopted to initialize the weights of the decoder $\mathbf{W}^{(d)}$ in this article [51].

C. Objective Function

As stated before, the objective function of the proposed MUNet is realized by minimizing the reconstruction error between $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$. Here, the SAD measure is adopted in the objective function of MUNet, which is given by

$$L_R = \frac{1}{N}\sum_{i=1}^{N} J_{\mathrm{SAD}}(\hat{\mathbf{y}}_i, \mathbf{y}_i) \tag{16}$$

where $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ denote the ith pixel in the input HSI $\mathbf{Y}$ and the reconstructed HSI $\hat{\mathbf{Y}}$, respectively.

Since the softmax activation function cannot produce sparse abundance results, $L_{1/2}$ sparsity regularization [52] is introduced in (16), denoted as follows:

$$L_{\mathrm{sp}} = \|\hat{\mathbf{A}}\|_{1/2} = \sum_{i=1}^{N}\sum_{j=1}^{P}\left|\hat{a}_{ji}\right|^{1/2} \tag{17}$$

where $\hat{a}_{ji}$ is the abundance element in the jth row and the ith column of the abundance matrix $\hat{\mathbf{A}}$.

In addition, minimum volume constraint (MVC) can effectively deal with the endmember extraction problem [53] and help find the compact simplex enclosed by endmembers. Therefore, based on the measurement of endmember distance [54], MVC regularization is used in the decoder part to obtain robust endmember results, which can be expressed as

$$L_{\mathrm{MVC}} = \frac{1}{\mathrm{LP}}\sum_{j=1}^{P}\left\|\hat{\mathbf{m}}_i - \frac{1}{P}\sum_{i=1}^{P}\hat{\mathbf{m}}_i\right\|_2^2. \tag{18}$$

Finally, the overall loss of MUNet can be formulated as

$$L = L_R + \lambda L_{\mathrm{sp}} + \delta L_{\mathrm{MVC}} \tag{19}$$

where λ and δ are the hyperparameters to balance these three types of objective functions.

IV. EXPERIMENT

In this section, we conduct the experiments to assess the performance of the proposed method in the synthetic and real multimodal datasets. In addition, six classic and state-of-the-art unmixing approaches related to the blind HU task are selected for comparison, mainly including three categories:

TABLE II
NETWORK CONFIGURATION OF THE PROPOSED MUNet

Dataset	λ	δ	l_{en}	l_{de}	epoch
Synthetic	0	0	$1e-4$	$1e-4$	120
Muffle	$3e-2$	1	$3e-4$	$1e-4$	50
Houston	$8e-2$	0.5	$1e-4$	$5e-4$	40

1) Non-AE-Based Unmixing Method: Multiscale sparse unmixing algorithm with simple linear iterative clustering (MUA-SLIC) [55] and spatial group sparsity-regularized non-negative matrix factorization (SGSNMF) [56]. MUA-SLIC is a spatial-regularized sparse unmixing method and introduces two multiscale domain transformations to capture more spectral–spatial contextual information in HSI. SGSNMF is a traditional NMF-based unmixing method and aims to use the prior knowledge of the group structure and abundance sparsity to enhance unmixing performance.

2) AE-Based Unmixing Method: Deep AE unmixing (DAEU) [12], untied denoising AE with sparsity (uDAS) [13], and cycle consistency unmixing network (CyCU-Net) [21]. DAEU is a basic AE unmixing method and the SAD objective function is used to solve the unmixing results. uDAS is a tied-weighted AE unmixing method and the denoising module is incorporated to reduce noise interference. CyCU-Net is a self-supervised AE method and cycle consistency regularization is adopted to preserve high-level semantic information of HSI.

3) Multimodal Unmixing Method: Weighted spatial regularization from DSM (w-DSM) [34]. w-DSM is a LiDAR-aided unmixing method by integrating the guidance map of multimodal data to improve the unmixing results.

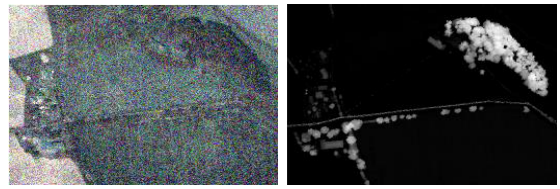
Note that the parameter settings of these comparison approaches refer to the original literature in our experiments, and the initial endmembers are extracted by VCA for fair comparison.

A. Experimental Setup

1) Hyperparameter Settings: In our case, the proposed MUNet is implemented on the PyTorch platform with i7-6850K CPU and a 1080Ti 11GB GPU. The number of endmembers is estimated using hyperspectral signal identification by minimum error (HySime) [57], and the endmembers are initialized by VCA in the training phase. The Adam optimizer is adopted to update the network parameters with a mini-batch size of 256 in the Houston dataset and 128 in other datasets. The learning rates of the encoder part l_{en} and the decoder part l_{de} are empirically set in different datasets and decay by multiplying a factor of 0.8 after each 20 epochs. The specific hyperparameter settings of each multimodal dataset are displayed in Table II.

2) Evaluation Metrics: For both the datasets, quantitative unmixing results are evaluated by two evaluation metrics, including the abundance root MSE (aRMSE)

$$\mathrm{aRMSE}(\hat{\mathbf{a}}_j, \mathbf{a}_j) = \sqrt{\frac{1}{\mathrm{pn}} \sum_{j=1}^{N} \left\| \hat{\mathbf{a}}_j - \mathbf{a}_j \right\|_2^2} \tag{20}$$



(a) (b)

Fig. 3. RGB image of synthetic multimodal data. (a) Hyperspectral data. (b) LiDAR data.

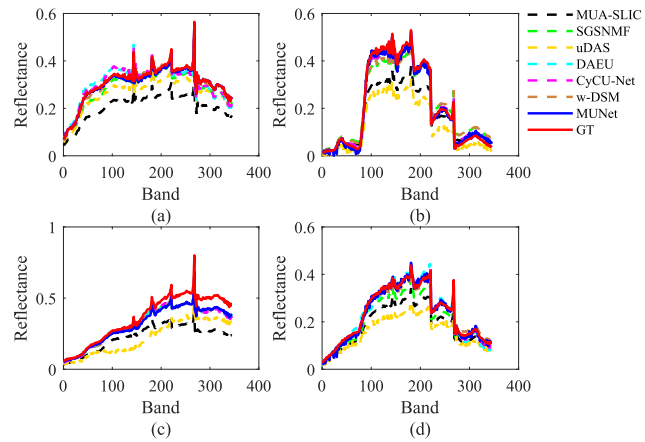


Fig. 4. Comparison of different endmember extraction algorithms on the synthetic multimodal data. (a) Material 1. (b) Material 2. (c) Material 3. (d) Material 4.

and the endmember SAD (eSAD)

$$\mathrm{eSAD}(\hat{\mathbf{m}}_i, \mathbf{m}_i) = \arccos\left(\frac{\hat{\mathbf{m}}_i^T \mathbf{m}_i}{\|\hat{\mathbf{m}}_i\|_2 \|\mathbf{m}_i\|_2} \right) \tag{21}$$

where $\hat{\mathbf{a}}_i$ and $\mathbf{a}_i$ represent the estimated abundance and the corresponding abundance ground truth (GT), respectively. $\hat{\mathbf{m}}_j$ and $\mathbf{m}_j$ denote the extracted endmember and the reference endmember, respectively. Here, the acquisition of the reference GT in different multimodal datasets follows our previous work in [21].

B. Experiment With Synthetic Multimodal Data

1) Data Description: The synthetic dataset has been applied in [34] to quantitatively evaluate unmixing performance, namely, SIM2, and the corresponding RGB image is shown in Fig. 3. In this studied scene, four main endmember references are manually extracted from a real hyperspectral image, acquired by the HySpex hyperspectral camera over Saint-André, France, and the LiDAR data are simultaneously acquired by real LiDAR measurements. The hyperspectral data in SIM2 are generated by following LMM with the extracted endmembers and the estimated abundance maps. In addition, the addictive Gaussian noise with SNR = 20 dB is introduced to model more realistic hyperspectral scene. The synthetic multimodal data contain 260 × 180 pixels and 345 bands ranging from 0.414 to 2.398 μm. Refer to [34] for more details regarding these data.

2) Results and Discussion: Table III lists the quantitative results of different algorithms in terms of aRMSE, eSAD for

TABLE III

QUANTITATIVE RESULTS FOR THE SYNTHETIC MULTIMODAL DATASET, WHERE eSAD FOR EACH MATERIAL, MEAN eSAD, AND aRMSE ARE REPORTED. THE BEST RESULTS ARE SHOWN IN BOLD

Methods		MUA-SLIC	SGSNMF	uDAS	DAEU	CyCU-Net	w-DSM	MUNet
eSAD	Material 1	0.0354	0.0369	0.0135	0.0736	0.0631	0.0277	**0.0126**
	Material 2	0.1007	0.1117	0.0429	0.0430	0.0471	0.1062	**0.0415**
	Material 3	0.1321	0.0636	0.1280	0.0576	0.0753	0.0588	**0.0362**
	Material 4	**0.0185**	0.0293	0.0615	0.0706	0.0334	0.0393	0.0370
Mean eSAD		0.0717	0.0604	0.0615	0.0612	0.0547	0.0580	**0.0318**
aRMSE		0.1232	0.0892	0.0654	0.0963	0.0816	0.0506	**0.0482**

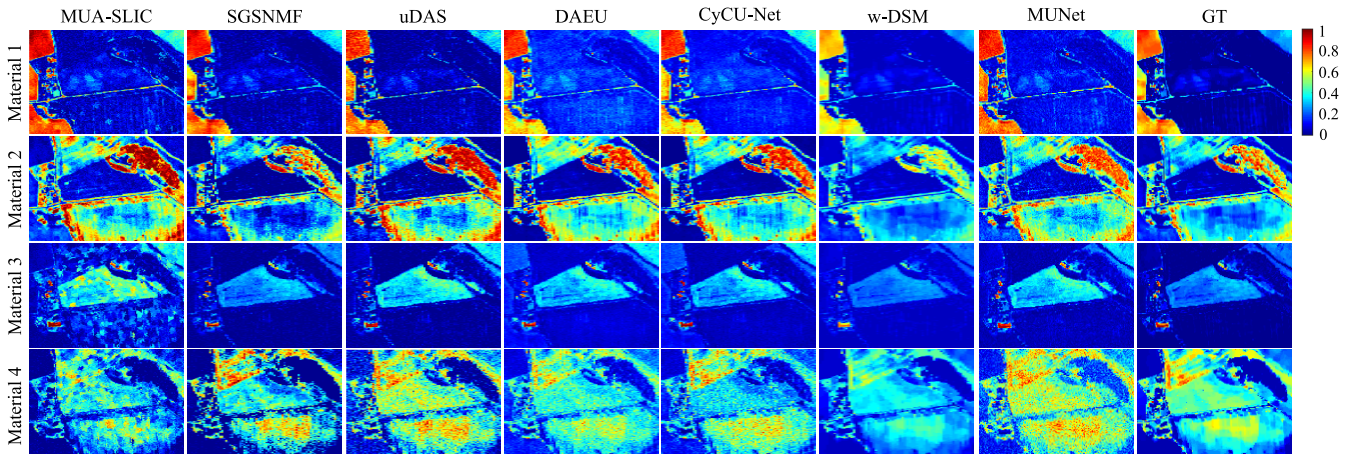


Fig. 5. Abundance maps of four materials from the synthetic multimodal data obtained by different algorithms.

each endmember, and mean eSAD on the synthetic multimodal data. Figs. 4 and 5 present the extracted abundance maps and the corresponding endmember results of different algorithms in the synthetic multimodal dataset. Overall, MUA-SLIC yields poor unmixing performance for both endmember extraction and abundance estimation, because the application of large spectral libraries causes unmixing problem to be sensitive to noise. Unlike MUA-SLIC, SGSNMF considers abundance sparsity in the form of group structures, bringing certain performance improvement in terms of aRMSE and mean eSAD. Compared with the traditional methods, some DL-based unmixing approaches can generally show better endmember and abundance results, such as uDAS and CyCU-Net, due to the introduction of denoising and self-supervised techniques. w-DSM can obtain relatively smaller eSAD and aRMSE results than the traditional and DL-based methods, which validates the effectiveness of the spatial regularization derived by the multimodal dataset. It can be seen that the proposed MUNet achieves the best performance in terms of eSAD, mean eSAD, and RMSE, demonstrating the superiority of the combination of attention mechanism and DL network in the multimodal unmixing task.

C. Experiment With Real Multimodal Data

1) Data Description: In this section, two types of real multimodal datasets are adopted to validate the unmixing results of different algorithms. The first one is the Muffle dataset,[1] collected over the campus of Southern Mississippi-Gulfpark [58]. The original image has 325 × 220 pixels and 64 bands in the spectral range from 0.375 to 1.050 μm. We select a popular region of interest (ROI) with a size of 130 × 90 pixels, as shown in Fig. 6. With reference to the marked scene label in [59], five dominated materials in this scene are investigated: *#1 Roof*, *#2 Grass*, *#3 Tree*, *#4 Shadow*, and *#5 Asphalt*.

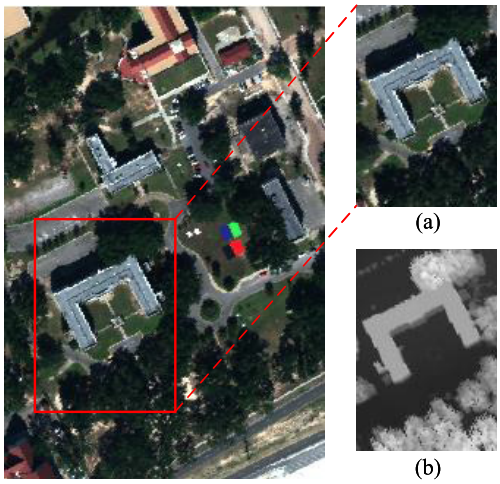


Fig. 6. RGB image of the Muffle multimodal data. (a) Hyperspectral data. (b) LiDAR data.

The second data are the Houston data, acquired by the ITRES CASI-1500 sensor over the University of Houston campus, TX, USA, in June 2012. This dataset was originally released by the 2013 IEEE GRSS data fusion contest,[2] and it

[1] https://github.com/GatorSense/MUUFLGulfport

[2] http://hyperspectral.ee.uh.edu

TABLE IV

QUANTITATIVE RESULTS FOR THE MUFFLE DATASET, WHERE eSAD FOR EACH MATERIAL, MEAN eSAD, AND aRMSE ARE REPORTED. THE BEST RESULTS ARE SHOWN IN BOLD

Methods		MUA-SLIC	SGSNMF	uDAS	DAEU	CyCU-Net	w-DSM	MUNet
eSAD	Roof	0.1322	0.0720	0.2857	0.0549	0.0386	0.1812	**0.0117**
	Grass	0.1479	0.1716	0.0735	0.1917	**0.0123**	0.0676	0.1073
	Tree	0.0740	0.1001	0.1725	**0.0416**	0.0806	0.1434	0.1215
	Shadow	0.2764	0.1391	0.1330	0.1370	0.2186	0.1220	**0.0526**
	Asphalt	0.1283	0.1538	0.1680	0.2653	0.0731	0.1251	**0.0380**
Mean eSAD		0.1518	0.1273	0.1666	0.1381	0.0846	0.1279	**0.0662**
aRMSE		0.2270	0.2344	0.3048	0.1931	0.1913	0.2213	**0.1765**

(a) (b)

Fig. 7. RGB image of the Houston multimodal data. (a) Hyperspectral data. (b) LiDAR data.

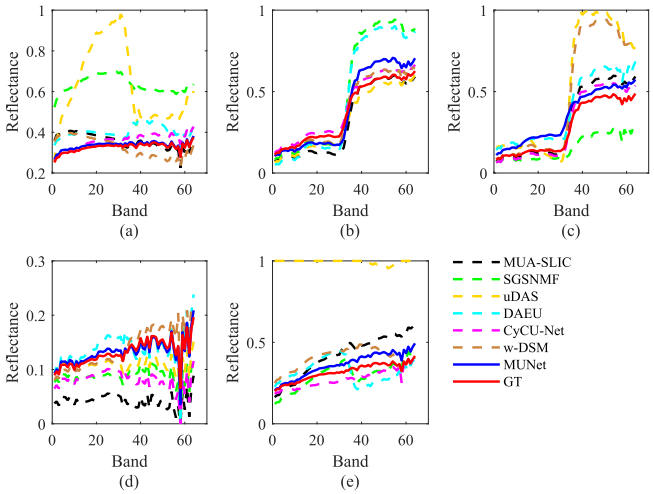


Fig. 8. Comparison of different endmember extraction algorithms on the Muffle multimodal data. (a) Roof. (b) Grass. (c) Tree. (d) Shadow. (e) Asphalt.

has been widely applied for evaluating the performance of land cover classification. The original image is 349 × 1905 pixels recorded in 144 bands ranging from 0.364 to 1.046 μm. We investigate a 170 × 170 pixel subimage cropped from the original image, visualized in Fig. 7. The four endmembers in this scene are *#1 Parking lot1*, *#2 Parking lot2*, *#3 Running track*, and *#4 Grass healthy*.

2) Results and Discussion: The quantitative results on the Muffle and the Houston multimodal datasets are reported in Tables IV and V, where the best results are illustrated in bold. For illustrative purposes, the extracted endmember signatures and the corresponding abundance maps of different algorithms on these two real datasets are depicted in Figs. 8–11. It can be clearly observed that uDAS cannot perform well on these two datasets, because the real scene usually contains complex noise distribution, and it is difficult to model the measured noise based on the assumption of linear transformation in designed

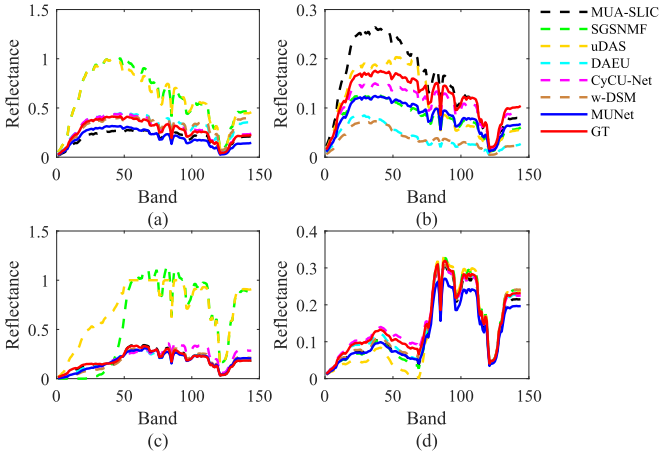


Fig. 9. Comparison of different endmember extraction algorithms on the Houston multimodal data. (a) Parking lot1. (b) Parking lot2. (c) Running track. (d) Grass healthy.

denoising module. As for MUA-SLIC, the unmixing results on the Muffle dataset are not good as those of SGSNMF, DAEU, and w-DSM, but it outperforms these three comparison approaches on the Houston dataset in terms of mean eSAD. The reason may be that the Muffle data contain certain endmember variability, such as grass and tree signatures, which make it hard for MUA-SLIC to construct an accurate spectral library. On the contrary, the spectral differences of various materials in the Houston data are quite large, which can help the sparse-based unmixing methods obtain better endmember results. Compared with SGSNMF and w-DSM, DAEU and CyCU-Net have superior unmixing performance in real multimodal datasets, further proving the effectiveness of the DL-based methods. Although MUNet does not obtain optimal eSAD results for each endmember, the mean eSAD by considering all the endmembers is the best and all the extracted endmember results of MUNet are close to the optimal ones on the Houston data, respectively, illustrating the stability and effectiveness of the proposed MUNet. By synthesizing the evaluation performance of aRMSE and mean eSAD in multiple datasets, the proposed MUNet can yield more accurate endmember and abundance results compared with other approaches as a whole, indicating its superiority for the multimodal unmixing task in real scenarios.

D. Model Analysis

1) Ablation Study on Network Modules: To validate the essentiality of the proposed MUNet network, as shown in Table VI, the ablation study on different network modules is

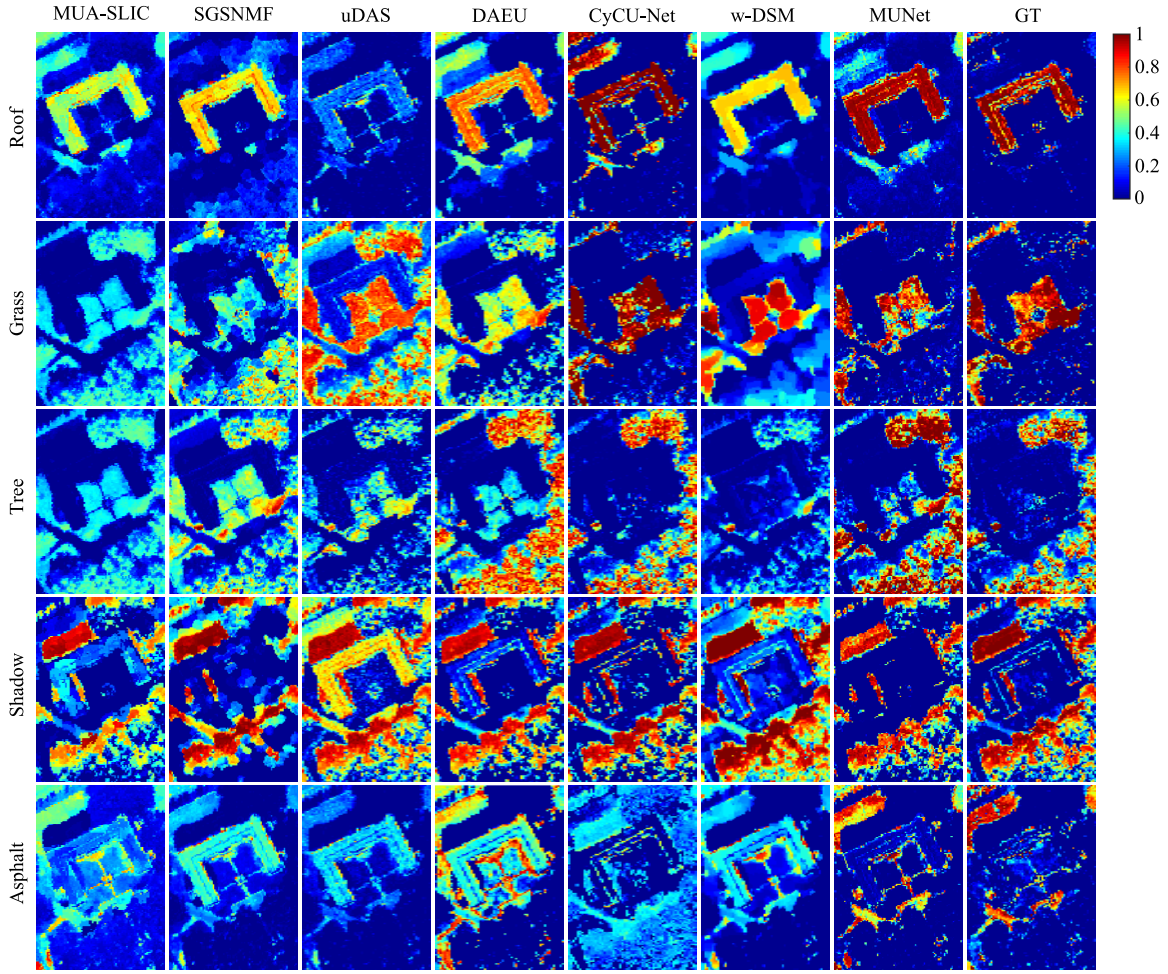


Fig. 10. Abundance maps of five materials from the Muffle multimodal data obtained by different algorithms.

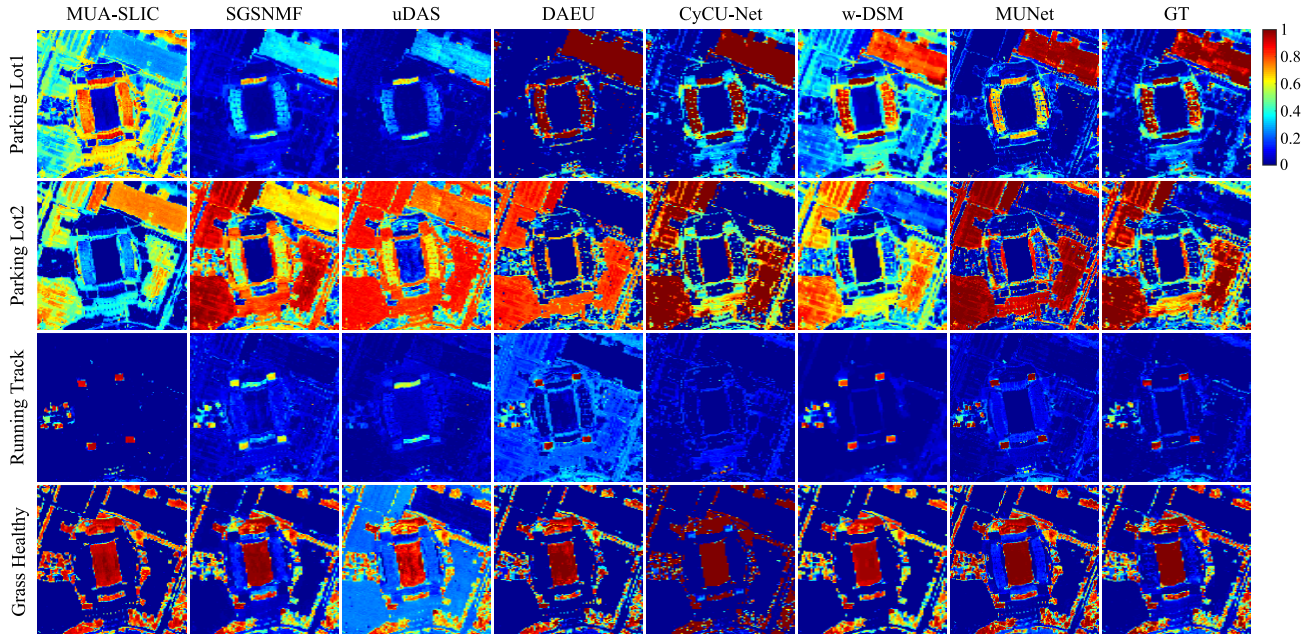


Fig. 11. Abundance maps of four materials from the Houston multimodal data obtained by different algorithms.

investigated in this section, including AP and SE attention modules. For fair comparison, the hyperparameter settings under different network configurations are consistent and the optimal unmixing performance is selected for comparative analysis. It can be seen from Table VI that MUNet after removing AP and SE attention modules yields the worst

TABLE V

QUANTITATIVE RESULTS FOR THE HOUSTON DATASET, WHERE eSAD FOR EACH MATERIAL, MEAN eSAD, AND aRMSE ARE REPORTED. THE BEST RESULTS ARE SHOWN IN BOLD

Methods		MUA-SLIC	SGSNMF	uDAS	DAEU	CyCU-Net	w-DSM	MUNet
eSAD	Parking lot1	0.0838	0.0535	0.0572	0.1073	**0.0052**	0.1944	0.0547
	Parking lot2	0.0983	0.1021	0.2699	0.3263	**0.0214**	0.2389	0.0384
	Running track	**0.0949**	0.3604	0.1517	0.1007	0.2539	0.1613	0.1181
	Grass healthy	0.0836	0.1305	0.2279	0.0575	**0.0433**	0.0740	0.0581
Mean eSAD		0.0901	0.1616	0.1767	0.1479	0.0809	0.1672	**0.0673**
aRMSE		0.2607	0.2017	0.2366	0.1159	0.1154	0.1254	**0.1039**

TABLE VI

ABLATION ANALYSIS OF THE PROPOSED MUNet WITH A COMBINATION OF DIFFERENT NETWORK MODULES ON THE HOUSTON DATASET

Module		eSAD				mean eSAD	aRMSE
AP	SE attention	Parking lot1	Parking lot2	Running track	Grass healthy		
✗	✗	0.0859	0.5251	0.1031	0.0846	0.1997	0.2579
✓	✗	0.0560	0.0643	0.4816	**0.0468**	0.1622	0.2464
✗	✓	0.0868	0.1964	**0.0674**	0.0922	0.1107	0.1938
✓	✓	**0.0547**	**0.0384**	0.1181	0.0581	**0.0673**	**0.1039**

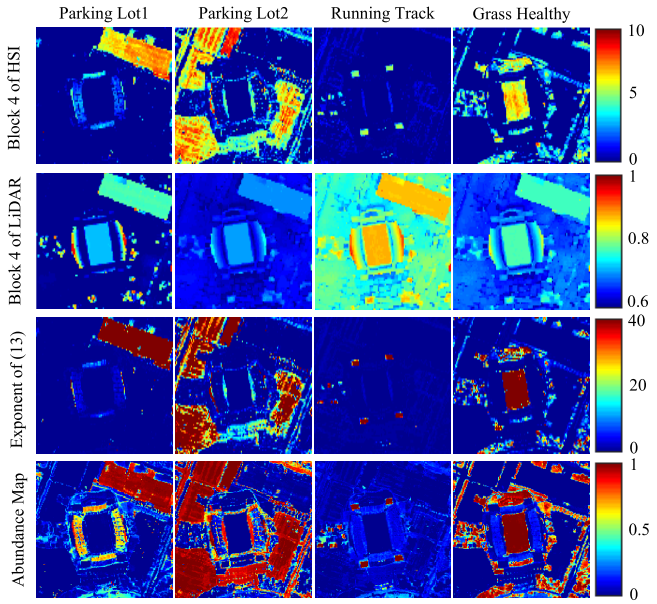


Fig. 12. Visualization of the extracted features obtained by different sources using the proposed MUNet. From top to bottom: the encoder output of HSI, the encoder output of LiDAR, the exponent result of (13), and the extracted abundance map.

unmixing performance, which to some extent indicates that the single two-stream AE network may not be suitable for HU. By introducing either AP or SE attention techniques into the two-stream AE model, the integrated MUNet has a certain improvement in the estimation of endmembers and abundances. Note that since the AP technique pays more attention to characterizing the spatial information of materials with more height differences and large areas, it can effectively improve the unmixing results of Parking lot1, Parking lot2, and Grass healthy. The introduction of SE attention can reasonably embed more detailed information, which can further bring a dramatic enhancement of different materials in terms of aRMSE and eSAD. This might be well-explained that the joint exploration of AP and SE attention in MUNet is capable of learning more high-dimensional multimodal features and aiding the unmixing results toward a more accurate direction. Therefore, the design of AP and SE attention techniques plays an important role in the field of multimodal unmixing.

TABLE VII

COMPUTATIONAL COST OF ALL COMPARISON METHODS ON DIFFERENT MULTIMODAL DATASETS IN TERMS OF SECONDS (s)

Method	Synthetic	Muffle	Houston
MUA-SLIC	108.41	**3.41**	46.09
SGSNMF	84.09	42.74	56.18
uDAS	350.89	31.61	180.28
DAEU	122.61	10.63	33.53
CyCU-Net	**41.83**	16.28	**21.57**
w-DSM	134.78	10.91	30.22
MUNet	323.90	52.86	34.87

2) Feature Visualization: Fig. 12 visualizes the extracted features obtained by different sources on the Houston dataset, including the encoder output of HSI and LiDAR in block 4, the exponent result of (13), and the extracted abundance map, where each column represents different endmember types. Note that according to the definition of the softmax function in (14), the final abundance maps are obtained by dividing the sum of the exponential results in (13). As shown in the first row of Fig. 12, only relying on HSI cannot obtain accurate abundance features, because different materials with a certain spectral similarity are hard to effectively distinguish, such as Parking lot1 and Parking lot2. After the aid of LiDAR in the third row of Fig. 12, the extracted abundance features are more separate and optimal, which demonstrates that the features extracted from HSI and LiDAR are complementary to each other in the process of fusion. For the redundant information derived from LiDAR, MUNet can adaptively select the most effective and meaningful features based on the encoder output of HSI, thereby realizing the enhancement of unmixing results. This also demonstrates the effectiveness and superiority of the proposed MUNet from the visual perspective.

3) Computational Cost: The comparisons of average computational cost for different multimodal datasets are illustrated in Table VII. Note that all these unmixing approaches are carried out in the same hardware environment. It can be seen from Table VII that the proposed MUNet mainly depends on the size of the input multimodal image. Due to the introduction of more modality data and the training way of the two-stream architecture, the computational cost of MUNet is higher than that of the traditional single-modality-based unmixing approaches in large datasets, e.g., synthetic and Muffle, but

it is also acceptable in practical applications. It should be noted, however, that the computational cost of the MUNet is comparable to other comparison methods in small datasets, e.g., Houston. Overall, the computation cost of MUNet is acceptable for all multimodal datasets.

V. Conclusion

In this article, we propose an end-to-end multimodal network for HU, called MUNet, by integrating the height differences of LiDAR data into HSI to enhance unmixing performance. Benefiting from the AP and SE attention techniques, the proposed MUNet method can learn essential high-dimensional and spatial information of LiDAR, aiding the unmixing network toward a more accurate extraction direction. Experiments with synthetic and real multimodal datasets validate the effectiveness and superiority of the proposed MUNet compared with the state-of-the-art unmixing methods. The combination of DL network and attention technology provides great possibilities for multimodal unmixing tasks in the future.

References

[1] D. Hong *et al.*, "Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 52–87, Jun. 2021.

[2] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 44–57, Jan. 2002.

[3] J. M. Bioucas-Dias *et al.*, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 354–379, Apr. 2012.

[4] R. Heylen, M. Parente, and P. Gader, "A review of nonlinear hyperspectral unmixing methods," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 1844–1868, Jun. 2014.

[5] W. Fan, B. Hu, J. Miller, and M. Li, "Comparative study between a new nonlinear model and common linear model for analysing laboratory simulated-forest hyperspectral data," *Int. J. Remote Sens.*, vol. 30, no. 11, pp. 2951–2962, Jan. 2009.

[6] A. Halimi, Y. Altmann, N. Dobigeon, and J.-Y. Tourneret, "Nonlinear unmixing of hyperspectral images using a generalized bilinear model," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4153–4162, Nov. 2011.

[7] Y. Altmann, A. Halimi, N. Dobigeon, and J.-Y. Tourneret, "Supervised nonlinear spectral unmixing using a postnonlinear mixing model for hyperspectral imagery," *IEEE Trans. Image Process.*, vol. 21, no. 6, pp. 3017–3025, Jun. 2012.

[8] R. Heylen and P. Scheunders, "A multilinear mixing model for nonlinear spectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 240–251, Jan. 2016.

[9] M. Tang, B. Zhang, A. Marinoni, L. Gao, and P. Gamba, "Multiharmonic postnonlinear mixing model for hyperspectral nonlinear unmixing," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1765–1769, Nov. 2018.

[10] B. Yang and B. Wang, "Band-wise nonlinear unmixing for hyperspectral imagery using an extended multilinear mixing model," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6747–6762, Nov. 2018.

[11] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[12] B. Palsson, J. Sigurdsson, J. R. Sveinsson, and M. O. Ulfarsson, "Hyperspectral unmixing using a neural network autoencoder," *IEEE Access*, vol. 6, pp. 25646–25656, 2018.

[13] Y. Qu and H. Qi, "uDAS: An untied denoising autoencoder with sparsity for spectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1698–1712, Mar. 2019.

[14] S. Ozkan, B. Kaya, and G. B. Akar, "EndNet: Sparse AutoEncoder network for endmember extraction and hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 482–496, Jan. 2019.

[15] M. M. Elkholy, M. Mostafa, H. M. Ebied, and M. F. Tolba, "Hyperspectral unmixing using deep convolutional autoencoder," *Int. J. Remote Sens.*, vol. 41, no. 12, pp. 4799–4819, Mar. 2020.

[16] F. Khajehrayeni and H. Ghassemian, "Hyperspectral unmixing using deep convolutional autoencoders in a supervised scenario," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 567–576, 2020.

[17] B. Palsson, M. O. Ulfarsson, and J. R. Sveinsson, "Convolutional autoencoder for spectral–spatial hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 535–549, Jan. 2021.

[18] Z. Hua, X. Li, Y. Feng, and L. Zhao, "Dual branch autoencoder network for spectral-spatial hyperspectral unmixing," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Jul. 2022, doi: 10.1109/LGRS.2021.3091858.

[19] Q. Jin, Y. Ma, F. Fan, J. Huang, X. Mei, and J. Ma, "Adversarial autoencoder network for hyperspectral unmixing," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 8, 2021, doi: 10.1109/TNNLS.2021.3114203.

[20] Z. Han, D. Hong, L. Gao, B. Zhang, and J. Chanussot, "Deep half-siamese networks for hyperspectral unmixing," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 11, pp. 1996–2000, Nov. 2021.

[21] L. Gao, Z. Han, D. Hong, B. Zhang, and J. Chanussot, "CyCU-Net: Cycle-consistency unmixing network by learning cascaded autoencoders," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, doi: 10.1109/TGRS.2021.3064958.

[22] D. Hong *et al.*, "Endmember-guided unmixing network (EGU-Net): A general deep learning framework for self-supervised hyperspectral unmixing," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 28, 2021, doi: 10.1109/TNNLS.2021.3082289.

[23] S. Kahraman and R. Bacher, "A comprehensive review of hyperspectral data fusion with LiDAR and SAR data," *Annu. Rev. Control*, vol. 51, pp. 236–253, Mar. 2021.

[24] M. D. Mura, S. Prasad, F. Pacifici, P. Gamba, J. Chanussot, and J. A. Benediktsson, "Challenges and opportunities of multimodality and data fusion in remote sensing," *Proc. IEEE*, vol. 103, no. 9, pp. 1585–1601, Sep. 2015.

[25] R. Huang, D. Hong, Y. Xu, W. Yao, and U. Stilla, "Multi-scale local context embedding for LiDAR point cloud classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 721–725, Apr. 2020.

[26] Z. Zeng, J. Sun, C. Xu, and H. Wang, "Unknown SAR target identification method based on feature extraction network and KLD–RPA joint discrimination," *Remote Sens.*, vol. 13, no. 15, p. 2901, Jul. 2021.

[27] B. Huang, B. Zhao, and Y. Song, "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery," *Remote Sens. Environ.*, vol. 214, pp. 73–86, Sep. 2018.

[28] X. Tao, T. Cui, A. Plaza, and P. Ren, "Simultaneously counting and extracting endmembers in a hyperspectral image based on divergent subsets," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8952–8966, Dec. 2020.

[29] D. Hong *et al.*, "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, doi: 10.1109/TGRS.2021.3130716.

[30] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and LiDAR data using coupled CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4939–4950, Jul. 2020.

[31] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "FusAtNet: Dual attention based SpectroSpatial multimodal fusion network for hyperspectral and LiDAR classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 92–93.

[32] J. Hu, D. Hong, and X. X. Zhu, "MIMA: MAPPER-induced manifold alignment for semi-supervised fusion of optical image and polarimetric SAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9025–9040, Nov. 2019.

[33] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, Aug. 2021.

[34] T. Uezato, M. Fauvel, and N. Dobigeon, "Hyperspectral image unmixing with LiDAR data-aided spatial regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 4098–4108, Jul. 2018.

[35] S. Kahraman, Y. Xu, J. Chanussot, and A. Tangel, "LiDAR data-aided hypergraph regularized multi-modal unmixing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 696–699.

[36] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[37] N. Liu, J. Han, and M. H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Aug. 2018, pp. 3089–3098.

[38] S. Zhou *et al.*, "Hierarchical U-shape attention network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 8417–8428, 2020.

[39] X. Wu, W. Li, D. Hong, J. Tian, R. Tao, and Q. Du, "Vehicle detection of multi-source remote sensing data using active fine-tuning network," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 39–53, Sep. 2020.

[40] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Aug. 2015, pp. 842–850.

[41] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8065–8080, Oct. 2019.

[42] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Proc. ECCV*. Cham, Switzerland: Springer, 2020, pp. 208–224.

[43] Y. Zeng, C. Ritz, J. Zhao, and J. Lan, "Attention-based residual network with scattering transform features for hyperspectral unmixing with limited training samples," *Remote Sens.*, vol. 12, no. 3, p. 400, Jan. 2020.

[44] Y. Qing and W. Liu, "Hyperspectral image classification based on multi-scale residual network with attention mechanism," *Remote Sens.*, vol. 13, no. 3, p. 335, Jan. 2021.

[45] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, May 2020.

[46] Z. Xue, X. Yu, X. Tan, B. Liu, A. Yu, and X. Wei, "Multiscale deep learning network with self-calibrated convolution for hyperspectral and LiDAR data collaborative classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022, doi: 10.1109/TGRS.2021.3106025.

[47] H. Yuan and Y. Y. Tang, "Learning with hypergraph for hyperspectral image feature extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 8, pp. 1695–1699, Aug. 2015.

[48] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. X. Zhu, "Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3791–3808, Jun. 2020.

[49] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2015, pp. 448–456.

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

[51] J. M. P. Nascimento and J. M. Bioucas-Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 898–910, Apr. 2005.

[52] Y. Qian, S. Jia, J. Zhou, and A. Robles-Kelly, "Hyperspectral unmixing via $L_{1/2}$ sparsity-constrained nonnegative matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4282–4297, Nov. 2011.

[53] L. Miao and H. Qi, "Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 3, pp. 765–777, Mar. 2007.

[54] Y. Yu and W. Sun, "Minimum distance constrained non-negative matrix factorization for the endmember extraction of hyperspectral images," *Proc. SPIE*, vol. 6790, Nov. 2007, Art. no. 679015.

[55] R. A. Borsoi, T. Imbiriba, J. C. M. Bermudez, and C. Richard, "A fast multiscale spatial regularization for sparse hyperspectral unmixing," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 598–602, Apr. 2019.

[56] X. Wang, Y. Zhong, L. Zhang, and Y. Xu, "Spatial group sparsity regularized nonnegative matrix factorization for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6287–6304, Nov. 2017.

[57] J. M. Bioucas-Dias and J. M. P. Nascimento, "Hyperspectral subspace identification," *IEEE Trans. Geosci. Remote Sensing*, vol. 46, no. 8, pp. 2435–2445, Aug. 2008.

[58] P. Gader, A. Zare, R. Close, J. Aitken, and G. Tuell, "MUUFL gulfport hyperspectral and LiDAR airborne data set," Univ. Florida, Gainesville, FL, USA, Tech. Rep. REP-2013-570, Oct. 2013.

[59] X. Du and A. Zare, "Technical report: Scene label ground truth map for MUUFL gulfport data set," Univ. Florida, Gainesville, FL, USA, Tech. Rep. 20170417, Apr. 2017.

Zhu Han (Student Member, IEEE) received the B.S. degree in electrical engineering from the North China University of Technology, Beijing, China, in 2019. She is currently pursuing the Ph.D. degree in cartography and geographic information system with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing.

Her research interests include hyperspectral image processing, deep learning, and artificial intelligence.

Danfeng Hong (Senior Member, IEEE) received the M.Sc. degree *(summa cum laude)* in computer vision from the College of Information Engineering, Qingdao University, Qingdao, China, in 2015, and the Dr.-Ing. degree *(summa cum laude)* from Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany, in 2019.

He has been a Research Scientist and led a Spectral Vision Working Group with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany. He was also an Adjunct Scientist at GIPSA-Lab, Grenoble INP, CNRS, University Grenoble Alpes, Grenoble, France. He is currently a Professor with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences (CAS), Beijing, China. His research interests include signal/image processing and analysis, hyperspectral remote sensing, machine/deep learning, artificial intelligence, and their applications in earth vision.

Dr. Hong is an Editorial Board Member of *Remote Sensing* and a Topical Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS). He was a recipient of the Best Reviewer Award of the IEEE TGRS in 2021, the Jose Bioucas Dias Award for recognizing the Outstanding Paper at the Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS) in 2021, and the Remote Sensing Young Investigator Award in 2022. He is also a Leading Guest Editor of the *International Journal of Applied Earth Observation and Geoinformation*, the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS, and *Remote Sensing*.

Lianru Gao (Senior Member, IEEE) received the B.S. degree in civil engineering from Tsinghua University, Beijing, China, in 2002, and the Ph.D. degree in cartography and geographic information system from Institute of Remote Sensing Applications, Chinese Academy of Sciences (CAS), Beijing, in 2007.

He has been a Visiting Scholar with the University of Extremadura, Cáceres, Spain, in 2014, and Mississippi State University (MSU), Starkville, MS, USA, in 2016. He is currently a Professor with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, CAS. His research focuses on hyperspectral image processing and information extraction. In the past 10 years, he was the PI of ten scientific research projects at national and ministerial levels, including projects by the National Natural Science Foundation of China (2016–2019, 2018–2020, 2022–2025) and the Key Research Program of the CAS (2013–2015). He has authored or coauthored more than 180 peer-reviewed articles, and there are more than 100 journal papers included by Science Citation Index (SCI). He was a coauthor of three academic books including *Hyperspectral Image Information Extraction* (Science Press). He received 29 National Invention Patents in China.

Dr. Gao was awarded the Outstanding Science and Technology Achievement Prize of the CAS in 2016, was supported by the China National Science Fund for Excellent Young Scholars in 2017, and won the Second Prize of The State Scientific and Technological Progress Award in 2018. He received the recognition of the Best Reviewers of the IEEE JSTARS in 2015 and the Best Reviewers of the IEEE TGRS in 2017.

Jing Yao received the B.Sc. degree from Northwest University, Xi'an, China, in 2014, and the Ph.D. degree from the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, in 2021.

From 2019 to 2020, he was a Visiting Student with Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany, and with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany. He is currently an Assistant Professor with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include low-rank modeling, hyperspectral image analysis, and deep-learning-based image processing methods.

Bing Zhang (Fellow, IEEE) received the B.S. degree in geography from Peking University, Beijing, China, in 1991, and the M.S. and Ph.D. degrees in remote sensing from the Institute of Remote Sensing Applications, Chinese Academy of Sciences (CAS), Beijing, in 1994 and 2003, respectively.

He is currently a Full Professor and the Deputy Director of the Aerospace Information Research Institute, CAS, where he has been leading lots of key scientific projects in the area of hyperspectral remote sensing for more than 25 years. He has developed five software systems in image processing and applications. His creative achievements were awarded ten prizes from the Chinese government and special government allowances of the Chinese State Council. He has authored more than 300 publications, including more than 170 journal articles. He has edited six books/contributed book chapters on hyperspectral image processing and subsequent applications. His research interests include development of mathematical and physical models and image processing software for analysis of hyperspectral remote sensing data in many different areas.

Dr. Zhang was awarded the National Science Foundation for Distinguished Young Scholars of China in 2013 and was awarded the 2016 Outstanding Science and Technology Achievement Prize of the Chinese Academy of Sciences, the highest level of Awards for CAS scholars. He is currently serving as the Associate Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING. He has been serving as a Technical Committee Member of the IEEE Workshop on Hyperspectral Image and Signal Processing since 2011, as the President of hyperspectral remote sensing committee of China National Committee of International Society for Digital Earth since 2012, and as the Standing Director of Chinese Society of Space Research (CSSR) since 2016. He was a Student Paper Competition Committee member in IGARSS from 2015 to 2019.

Jocelyn Chanussot (Fellow, IEEE) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from the Université de Savoie, Annecy, France, in 1998.

Since 1999, he has been with Grenoble INP, where he is currently a Professor of signal and image processing. He has been a Visiting Scholar with Stanford University, Stanford, CA, USA, KTH, Sweden, and NUS, Singapore. Since 2013, he has been an Adjunct Professor with the University of Iceland, Iceland, U.K. From 2015 to 2017, he was a Visiting Professor with the University of California, Los Angeles (UCLA), Los Angeles, CA, USA. He holds the AXA Chair in remote sensing and is an Adjunct Professor with the Chinese Academy of Sciences, Aerospace Information Research Institute, Beijing. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence.

Dr. Chanussot is a member of the Institut Universitaire de France from 2012 to 2017 and is a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters). He was the Founding President of the IEEE Geoscience and Remote Sensing French chapter from 2007 to 2010, which received the 2010 IEEE GRSS-S Chapter Excellence Award. He has received multiple outstanding paper awards. He was the Vice-President of the IEEE Geoscience and Remote Sensing Society, in charge of meetings and symposia from 2017 to 2019. He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote sensing (WHISPERS). He was the Chair from 2009 to 2011 and a Co-Chair of the GRS Data Fusion Technical Committee from 2005 to 2008. He was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society from 2006 to 2008 and the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing in 2009. He is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the PROCEEDINGS OF THE IEEE. He was the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2011 to 2015. In 2014 he served as a Guest Editor of the *IEEE Signal Processing Magazine*.