Multiscale Convolutional Mask Network for Hyperspectral Unmixing

Mingming Xu, *Member, IEEE*, Jin Xu, Shanwei Liu, Hui Sheng, and Zhiru Yang

***Abstract*—Deep learning has gained popularity in hyperspectral unmixing (HU) applications recently due to its powerful learning and data-fitting capabilities. As an unmixing baseline network, the autoencoder (AE) framework performs well in HU by automatically learning low-dimensional embeddings and reconstructing data. Nevertheless, there are spectral variability and nonlinear mixing problems in the highly mixed region of hyperspectral images, which can cause interference to structures using only AE. Therefore, inspired by the effectiveness of mask modeling, we propose a multiscale convolutional mask network (MsCM-Net) for HU with two new strategies. First, we propose a mixed region mask strategy suitable for the HU task, and a multiscale convolutional AE is adopted as the unmixing baseline network to apply the mask strategy, making the method more robust in solving ill-posed unmixing problems. In addition, a new initialization strategy is used in which vertex component analysis (VCA) is combined with density-based spatial clustering of applications with noise (DBSCAN) to mitigate the impact of outliers and noise on initialization. The proposed MsCM-Net performs more accurately than state-of-the-art methods by comparison experiments on one synthetic and three real hyperspectral data sets. The effectiveness of the mixed region mask strategy and DBSCAN-VCA initialization is also demonstrated by ablation experiments.**

***Index Terms*—Autoencoder (AE), hyperspectral unmixing (HU), initialization, mixed region mask, multiscale.**

I. Introduction

THE hyperspectral image (HSI) with its high spectral resolution allows for the identification and differentiation of various materials by leveraging abundant spectral information, resulting in numerous applications, such as target detection [1], [2], [3], image classification [4], [5], [6], and feature fusion [7], [8], [9]. However, the low spatial resolution of HSI and the presence of mixing effects during image acquisition generally result in mixed pixels, which impedes the development and application of hyperspectral technology. Hyperspectral unmixing (HU) is an essential technique for addressing the issues earlier by decomposing the HSI into pure material spectra (i.e., endmember) and a set of abundance fractions representing each endmember's proportions within the mixed pixels [10].

The development of HU algorithms is based on different mixing models, i.e., linear mixing model (LMM) [11] and non-LMM (NLMM) [12]. LMM assumes that the mixed spectrum is a linear combination of the endmember spectra, and the NLMM considers the multiple reflections of light between objects. Although NLMM is more realistic and has better interpretability, it faces challenges such as the lack of appropriate measures for nonlinear degrees and prior knowledge. Therefore, LMM is still currently the most widely used model due to its simple modeling and low computational complexity.

Deep learning (DL) has developed rapidly in recent years with the emergence of the big data era and has been used in HU tasks [13]. Compared with traditional unmixing approaches, DL can implement various regularization constraints with arbitrary deep or nonlinear structures [14], [15], [16], [17], [18], [19], [20]. The generative network provides a new perspective for solving endmember variation [21], [22], [23], and the convolutional network can use spatial filtering in unmixing [24], [25], [26], [27], [28]. Autoencoder (AE) serves as the fundamental framework network for HU tasks, consisting of two main components: an encoder that gets the latent representation (i.e., abundance) of the input HSI and a decoder that reconstructs the HSI using appropriate weights (i.e., endmember) [29]. Nowadays, numerous DL unmixing methods based on AE have been proposed. For instance, Palsson et al. [17] proposed a deep AE with different activation functions for unmixing. Wang et al. [18] proposed a postnonlinear AE structure to focus on the nonlinear factors present in HSI. Shi et al. [23] developed a variational AE (VAE)-based generative unmixing structure to fit any endmember distribution with the help of VAE's properties of identifying endmember variability. However, these methods are not considered spatial information. HSI's inherent abundant spatial information in unmixing is important because the pixels in HSI are strongly correlated with their neighbors. Ghosh et al. [25] combined convolutional AE with Transformers to take advantage of the transformer's capacity to capture global feature dependencies, enhancing the quality of endmember spectra and abundance maps. Yu et al. [27] used a multistage convolutional AE in HU tasks and demonstrated its effectiveness in acquiring extensive contextual information without sacrificing details.

However, DL-based unmixing methods face the following challenges: (1) Unmixing is a nonconvex problem [30], and DL

Manuscript received 8 November 2023; revised 28 December 2023; accepted 6 January 2024. Date of publication 10 January 2024; date of current version 31 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62071492, and in part by the Shandong Natural Science Foundation under Grant ZR2023MD115. *(Corresponding author: Mingming Xu.)*

The authors are with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266580, China (e-mail: xumingming@upc.edu.cn; s23160009@s.upc.edu.cn; shanweiliu@163.com; sheng@upc.edu.cn; s21160036@s.upc.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3352080

is inherently a black box mechanism, which leads to uncertainty in the starting point and direction of gradient descent [31]. A good initialization will improve this [19], [32], [33]. In HU tasks, this primarily refers to the initialization of decoder weights, which represent endmembers. Random initialization was used in [17] and [24]. However, this will make the results of the network uncertain. In this regard, existing unmixing methods commonly used vertex component analysis (VCA) [34] to extract endmembers for initialization [25], [26], [27], [33], [35]. For example, Xiong et al. [33] used VCA for initialization to introduce prior knowledge and then unmixing by deep alternating network. However, ground materials often contain some outliers. As a pure pixel method, VCA has limitations in handling outliers and tends to suffer from significant endmember estimation biases, which can heavily interfere with the initialization results. Therefore, Dou et al. [19] proposed an outlier detection method based on the heat kernel similarity to improve initialization by excluding outliers. But it was considered from the global perspective of the image and could not exclude as many outliers as possible. Hong et al. [32] used K-means [36] for the VCA results to obtain spectral bundles for the subsequent endmember guidance network. However, K-means is based on the Euclidian distance [37] and has a weak ability to distinguish spectra. (2) The global training goal of the AE is to recover the original image, but due to the presence of noise, the network may eventually converge to a local minimum [38], [39]. To solve this, the authors in [18] and [23] considered the nonlinear and spectral variation to deal with noise, respectively. However, the measurement of the degree of both was not taken into account, which in fact introduced additional noise. Guo et al. [15] used denoising AE (DAE) to deal with noise. However, by adding noise to the whole image and training it to restore to the original image for denoising, this method also introduced additional interference to the area with low noise level. Moreover, these methods do not consider the spatial information, which leads to an unsatisfactory abundance map. How to use the spatial information of HSI and at the same time deal with the areas that will cause interference to the unmixing in a targeted manner is very important. Based on the abovementioned, we propose a multiscale convolutional mask network (MsCM-Net) for HU using two new strategies to solve these problems.

Initialization is very important in DL training, which determines the convergence speed and performance of the network [40]. Proper initialization can make the unmixing network start from a more accurate starting point and find the optimal value of the HU task more easily [41]. To address this issue, we propose a novel initialization method for outlier removal that combines VCA with density-based spatial clustering of applications with noise (DBSCAN) [42], [43]. The basic motivation is that typical materials have continuous spatial distributions, and neighboring pixels of pure endmembers will likely belong to the same class [44], [45]. The same classes should have similar spectra and lower cosine distance. DBSCAN is a density-based clustering algorithm, unlike classic methods like K-means, which can detect outliers and use cosine distance to better identify spectral differences. We first divide the HSI into multiple blocks and the DBSCAN clustering is applied to exclude outliers. The outliers can be better removed by dividing the image into small pieces for local consideration. VCA is then run on the remaining pixels to obtain endmembers.

The region containing strong noise which could affect unmixing result is defined as the highly mixed region. These regions usually contain anomalies generated by factors such as nonlinearity and spectral variation, and LMM cannot reconstruct them well [32], [46]. If these regions are trained directly, these redundant and interfering information will also affect the gradient descent direction, making it easier for the unmixing results to shift in the direction of the noise. Mask modeling has made significant progress in the field of computer vision (CV) [47], [48], [49], [50] in recent years. The mask mechanism is essentially a noise type of DAE, which can extract discriminant representations from masked images [51], [52]. But it has no relevant application in the HU field at present. We propose a mixed region mask (MRM) strategy that is suitable for HU tasks, aiming to reduce noise effects. By hiding the information of some pixels in the highly mixed region obtained by OSTU [53] automatic thresholding, the network can focus more on learning the mapping relationship between endmember and the abundance of relatively pure region, so as to obtain more accurate unmixing results. Inspired by the multiscale convolutional autoencoder (MuCAE) [27], the baseline network was built to apply MRM. By considering multiscale information, invariant features at different scales can be mined to help obtain more useful features [54]. The application of MRM in MuCAE realizes the use of spatial information while resisting the interference of noise, and further improves the accuracy of the results.

In summary, the main contribution of this article has threefold.

1) We propose a multiscale convolutional mask network called MsCM-Net. A new mask strategy for mixed region, MRM, which is suitable for HU tasks, is proposed and applied to MuCAE. MRM is a mask mechanism that is executed at the pixel level. It helps AE to have the ability to resist the interference of noise in the highly mixed region, resulting in greater accuracy and stability than previous methods.
2) A new initialization method called DBSCAN-VCA to initialize decoder weights is presented. Compared with only using VCA, DBSCAN-VCA initialization can exclude outliers from local areas, thereby providing better initialization results and promoting better convergence of unmixing networks.
3) Ablation experiment and comparison experiment prove the effectiveness and necessity of the proposed MRM and DBSCAN-VCA initialization, and also prove the effectiveness of multiscale network applied to MRM.

The rest of this article is organized as follows. Section II describes the proposed MsCM-Net method. Section III reports the experimental results in synthetic and real data sets. Section IV provides relevant discussion to verify the effect of the proposed strategy. Finally, Section V, concludes this article.

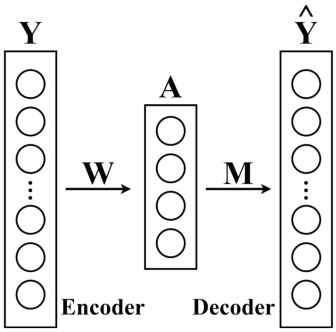

Fig. 1. Framework of AE-based unmixing network. **A** represents abundance maps and **M** represents the endmembers.

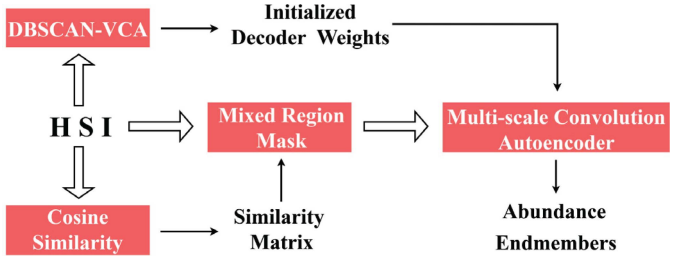

Fig. 2. Flowchart of the proposed method.

II. Proposed Method

A. Related Work About AE-Based Method for Unmixing

According to the LMM, the unmixing problem can be formulated as follows:

$$\mathbf{Y} = \mathbf{MA} + \mathbf{N} \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{L\times N}$ is the observed HSI with L bands and N pixels. $\mathbf{N} \in \mathbb{R}^{L\times N}$ is the noise matrix. $\mathbf{M} \in \mathbb{R}^{L\times p}$ represents the endmember matrix with p endmember categories, and $\mathbf{A} \in \mathbb{R}^{p\times N}$ denotes the corresponding abundance matrix. Furthermore, the abundance vectors a_j should satisfy the abundance nonnegative constraint (ANC) and the abundance sum-to-one constraint (ASC) by the following equations:

$$\mathrm{a}_j \geq 0 \tag{2}$$

$$\sum_{i=1}^{p} \mathrm{a}_{ij} = 1. \tag{3}$$

Since the structure of AE is suitable for solving the unmixing problem, it is widely used in HU tasks. An AE generally consists of an encoder and a decoder. As shown in Fig. 1, the encoder converts the input HSI into a latent low-dimensional representation. The decoder reconstructs the HSI from the learned latent representation. The difference between the reconstructed image $\hat{\mathbf{Y}}$ and the original observed data $\mathbf{Y}$ is used to train the AE. Meanwhile, the low-dimensional representation is regarded as the obtained abundance. The weight matrix of the decoder is regarded as the endmember results. ANC and ASC can be implemented using a suitable activation function like Softmax.

B. Multiscale Convolutional Mask Unmixing Network

Our method mainly includes the following process: First, the initialized decoder weights are obtained through DBSCAN-VCA. This will be elaborated in the following Section II-C. The next part is the MsCM-Net, which is introduced in this section. The flowchart of the method is shown in Fig. 2.

By introducing mask modeling and the multiscale mechanism, the proposed MsCM-Net can obtain contextual information from different scales and eliminate noise interference in highly mixed regions. Fig. 3 illustrates the proposed MsCM-Net architecture. The proposed network consists of two structures: mixed region mask strategy and multiscale convolutional AE.

1) Mixed Region Mask: The nonhomogeneous regions are unbalanced parts of the data set, and complex DL models with numerous parameters may amplify the effects of these unbalanced parts. The existing mask strategy is usually to randomly select a certain percentage of patches in the entire image for masking, but for HSI, this may lose lots of details and introduce extra noise. So we propose MRM, a mask strategy performed at the pixel level in the mixed region.

Some regions containing factors, such as nonlinearity and endmember variation become noise or anomalies in the image due to the complex mixing mechanism. These regions are defined as highly mixed regions, which often interfere with the optimization direction of the network. They have significant differences compared with the surrounding relatively pure pixels. Inspired by this, a similarity matrix can be learned by measuring the similarity of the neighboring pixels over the entire image, and the similar (i.e., relatively pure) region and the nonsimilar (i.e., highly mixed) region are divided by threshold segmentation. Then we randomly mask the pixels in nonsimilar areas to reduce noise effects, and take the resulting masked image as the input image. Random mask means to select a mask ratio and mask pixels of this ratio in nonsimilar areas. This is to avoid completely ignoring the details that these areas contain. The target image is the corresponding original HSI. Specifically, for the *i*th pixel, its value in the similarity matrix d_i could be estimated by measuring the similarity between spatially neighboring pixels as follows:

$$\mathrm{d}_i = \sum_{j\in\mathcal{N}_i} \mathrm{s}_{ij} \tag{4}$$

where $\mathcal{N}_i$ is the neighborhood of the *i*th pixel that includes surrounding four pixels. s_{ij} is the similarity between the *i*th pixel y_i and its neighboring pixel y_j. The cosine similarity is used to calculate similarity, which is a classical distance measure in HU study [55], [56]. The value range is $[-1, 1]$, where positive and negative values indicate that the vectors are in the same direction or opposite direction. A value closer to 1 indicates that the two vectors are more similar. The calculation is as follows:

$$\mathrm{s}_{ij} = \frac{\mathrm{y}_i \mathrm{y}_j}{\| \mathrm{y}_i \|_2 \, \| \mathrm{y}_j \|_2}. \tag{5}$$

After obtaining the similarity matrix, the OSTU algorithm is used to get the threshold to distinguish similar and nonsimilar regions automatically. OSTU uses the greatest between-class

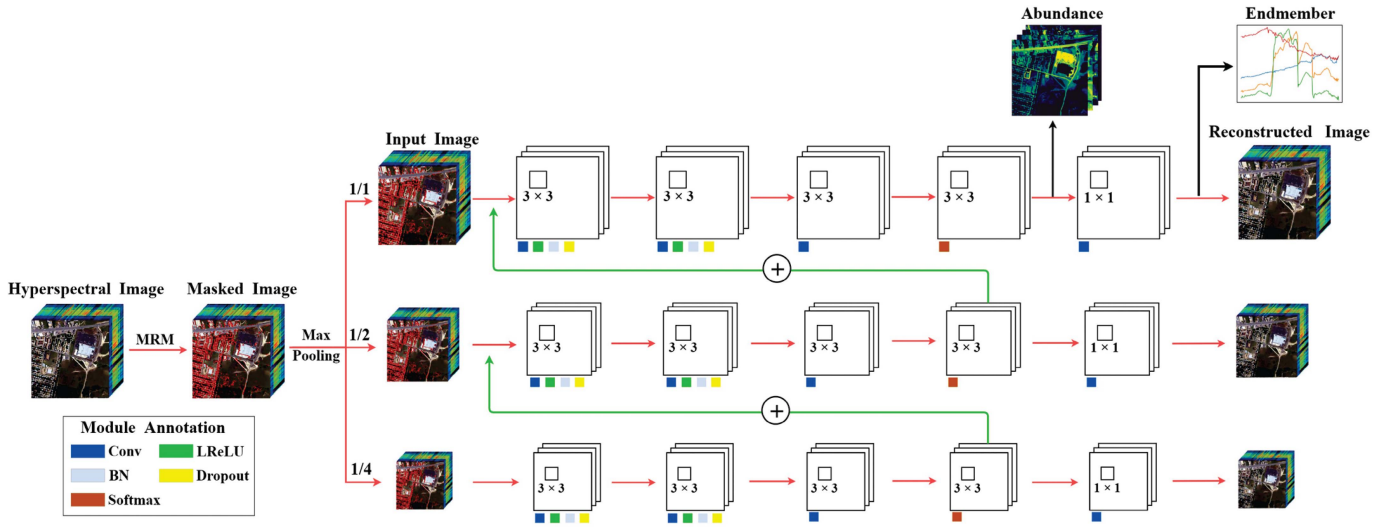

Fig. 3. Overview of the MsCM-Net. It is mainly composed of mixed region mask and multiscale convolutional AE. The red area in masked image indicates the mask area.

Fig. 4. Original image (left) and the masked image obtained (right). The mask is implemented at the pixel level. The masked pixels are in red. (a) Samson. (b) Urban.

variance as the criterion for automatically selecting the threshold [57]. The goal of OSTU is as follows:

$$\max g(t) = w_0 (u_0 - u)^2 + w_1 (u_1 - u)^2 \tag{6}$$

where t is the segmentation threshold. w_0 and u_0 are the proportions and mean values of similar regions in the matrix, respectively. w_1 and u_1 are the proportions and mean values of nonsimilar regions in the matrix, respectively. u is the mean value of all regions. The similarity values greater than t are similar regions, while those less than t are nonsimilar regions. Next, a mask label matrix of the same size as the HSI is constructed at each epoch. All values are set to 1 at first. Then the position index of a certain number of pixels in mixed areas is randomly selected and change the mask label matrix value at that location to 0. The number selected is the total number of pixels in mixed areas multiplied by the mask ratio. Therefore, the mask ratio is only applied to the highly mixed region obtained by OSTU. The mask ratio range is [0, 1]. Finally, we multiply the obtained mask label and whole HSI. The similar regions remain unchanged, and all band values of a certain ratio of pixels in the mixed region become 0 to obtain the masked image. In this way, since the network does not know the original content of the pixels that become **0** (zero vector), it focuses on learning the features of the relatively pure region. Fig. 4 shows the masked image obtained and the original image. The red area is the masked pixels, which can be seen that it usually appears in the transition area.

TABLE I
NETWORK CONFIGURATION FOR EACH SCALE IN THE PROPOSED MSCM-NET ARCHITECTURE

Pathway	Layer composition	Unit
Block 1	3 × 3 Convolution LeakyReLU Batch normalization Dropout	L + p (L for 1/4 scale)
Block 2	3 × 3 Convolution LeakyReLU Batch normalization Dropout	96
Block 3	3 × 3 Convolution	48
Block 4	Softmax	p
Block 5	1 × 1 Convolution	L

2) Multiscale Convolutional AE: We modify and use MuCAE to validate and further improve the effect of MRM. What we input to the network at both training and testing time are masked image. The network configuration for each scale in the MuCAE is shown in Table I. Blocks 1–4 represent the encoder, and block 5 is the decoder. It is worth noting that when the scale is 1/4, the unit in block 1 is the band number L. Because this layer does not perform feature fusion. The encoder uses 3 × 3 convolution with LeakyReLU as the activation function. Batch normalization is to prevent the activation value after LeakyReLU from being too large and causing the gradient to explode. Dropout is to prevent overfitting. The last layer of the encoder uses the softmax activation function to meet the ANC and ASC at the same time. The abundance map and endmember are obtained from the output of the encoder in the 1/1 scale (i.e., the original scale).

Since the pixel value of the highly mixed region in the masked image is replaced with 0, the average pooling may not

be representative. Therefore, the masked image of the original scale is downsampled to half and quarter of the original size, respectively, by using max pooling to obtain 1/4 scale and 1/2 scale. For 1/4 scale to 1/2 scale, each scale layer inputs the output of block 4 to the latter large-scale layer according to the following equation to integrate multiscale information:

$$\mathbf{O}^* = \mathbf{O} \oplus \text{TransConv}(\mathbf{F}) \tag{7}$$

where TransConv denotes the 2×2 transposed convolution and $\oplus$ represents the concatenate operation. $\mathbf{F} \in \mathbb{R}^{p \times N}$ represents feature maps of encoder after softmax operation at previous small-scale layer, using transposed convolutional layer for increasing scale. Increasing the scale by transposed convolution can lead to better feature maps through network learning. $\mathbf{O} \in \mathbb{R}^{L \times N}$ represents the original mask input for the current layer. $\mathbf{O}^*$ represents the fused information.

Our objective function has two loss terms, including spectral angle distance loss (L_S) in (8) and $\mathrm{L}_{1/2}$ sparse loss (L_P) in (9)

$$\mathrm{L}_S = \arccos\left(\frac{\langle \mathbf{Y}, \widehat{\mathbf{Y}} \rangle}{\| \mathbf{Y} \|_2 \| \widehat{\mathbf{Y}} \|_2}\right) \tag{8}$$

$$\mathrm{L}_P = \sum_{i,j=1}^{N,p} \mathrm{a}_{ij}^{1/2} \tag{9}$$

where $\mathbf{Y}$ represents the original HSI, and $\widehat{\mathbf{Y}}$ represents the reconstructed HSI. It should be noted that $\mathbf{Y}$ is the unmasked image, and the image input into the network for training and testing is the whole image after the random mask. $\mathrm{L}_{1/2}$ sparsity loss gains more zero elements in the abundance by penalizing nonzero elements, since usually most pixels in HSI are mixed by only a few endmembers [58]. It plays a sparse constraint on the abundance estimation, which is helpful to obtain more accurate unmixing results. Finally, the overall loss of MsCM-Net can be formulated as follows:

$$\mathrm{L} = \mathrm{L}_S + \alpha \mathrm{L}_P \tag{10}$$

where α is the tradeoff parameter. Note that the loss function is applied to all scales.

C. DBSCAN-VCA Initialization

Because the unmixing problem is nonconvex, the optimization process could benefit from a good initialization of decoder weight. However, outliers in the HSI would lead to poor initialization and strongly interfere with the unmixing solutions. Thus, we propose a new initialization method called DBSCAN-VCA initialization.

DBSCAN is a clustering algorithm based on density space. Its clustering principle is simply that the density of each cluster is higher than the density around the cluster, and the noise density is smaller than that of any cluster. The schematic diagram is shown in Fig. 5. The parameters of DBSCAN include the distance threshold D and the quantity threshold Q. If a sample contains more than Q points in D-domain, the sample is a core point. Suppose there is a data set C, f is a sample in it, then the

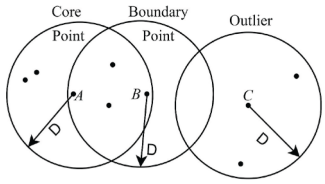

Fig. 5. DBSCAN diagram. A is the core point, B is the boundary point, and C is the outlier. D is the distance threshold and the quantity threshold Q is 5.

expression for the D-domain of sample f is

$$\mathrm{D}_{domain}(f) = \{\mathrm{g} \in \mathrm{C} \mid \mathrm{dist}(f, \mathrm{g}) \leq \mathrm{D}\} \tag{11}$$

where g is the set of points in the field of f. If the number of points of a sample in D-domain is less than Q but it falls in the domain of the core point, the sample is a boundary point; if a sample is neither core nor boundary point, the sample is an outlier. The distance measurement method of DBSCAN is chosen as cosine similarity in (5).

Since the distribution of ground objects is continuous, pixels in locally smaller neighborhoods should be more similar. Therefore, DBSCAN can be used to exclude pixels with large spectral differences in smaller regions. DBSCAN-VCA first divides the HSI into several patches in a small size of 4×4, then uses DBSCAN clustering within each patch. A small distance threshold and a large quantity threshold are selected, and pixels with large spectral differences between the surrounding pixels in a 4×4 region are excluded as HSI outliers. Next, VCA is used on the remaining pixels to extract the initial endmembers.

III. Experiments

In this section, we evaluate the performance of the proposed method on several data sets. One synthetic and three real data sets are used for testing. First, we analyze the noise robustness and hyperparameter sensitivity of the method on synthetic data. The effect of mask ratio on unmixing is also tested. And then, to evaluate the accuracy and consistency of the method, the experiments are repeated fifteen times and report the average and standard deviation of the results. The average value represents the accuracy of the method, that is, whether it can produce correct unmixing results. The standard deviation represents the consistency of the method, i.e., whether it can produce the same result in each run.

Two classical and five state-of-the-art DL unmixing methods are chosen for comparison: VCA [32], $\mathrm{L}_{1/2}$ sparsity-constrained nonnegative matrix factorization ($\mathrm{L}_{1/2}$-NMF) [56], deep AE unmixing network (DAEU) [17], nonlinear AE unmixing network (NLAEU) [18], probabilistic generative model for hyperspectral unmixing (PGMSU) [23], transformer AE unmixing network (TAEU) [25], and multistage convolution AE unmixing network (MuCAEU) [27].

A. Data Description

1) Synthetic Data Set: The synthetic data set is simulated using five endmember references with 200 bands selected from the advanced spaceborne thermal emission and reflection radiometer

Fig. 6. RGB image of four datasets. (a) Synthetic data. (b) Samson. (c) Urban. (d) Washington DC.

TABLE II
PARAMETER SETTINGS IN OUR EXPERIMENTS

	α	Learning rate	S	G	OSTU threshold	TN
Synthetic data	0.01	0.001	45	80	0.897	2469
Samson	0.001	0.03	25	60	0.989	340
Urban	0.001	0.004	50	40	0.976	14717
Washington DC	0.001	0.004	50	40	0.984	16953

(ASTER) spectral library [59], and the corresponding abundance maps satisfy the ANC and ASC with a size of 60 × 60 pixels. It should be noted that the abundance maps follow a Dirichlet distribution.

2) Samson: The SAMSON sensor [60] acquired this widely used real data set. It is an image cropped from a larger image, having 95 × 95 pixels with 156 channels, covering the 0.4–0.9 μm wavelength range. There are three material types: Water (#1), Soil (#2), and Tree (#3).

3) Urban: This popular real data has 307 × 307 pixels, with 210 bands covering the 0.4–2.5 μm wavelength range. After removing the destroyed and noisy bands, 162 bands remain. Four materials are observed in the scene: Asphalt (#1), Grass (#2), Tree (#3), and Roof (#4).

4) Washington DC: The real data are captured from the hyperspectral digital image collection experiment (HYDICE) sensor [61]. It has 191 bands covering the wavelength ranging from 0.4 to 2.4 μm. Due to the large size of the raw HSI, a subimage is cropped with a size of 290 × 290. Five materials are observed in the scene: Grass (#1), Water (#2), Roof (#3), Road (#4), and Tree (#5).

The real data and the corresponding ground truth (GT) are publicly available data sets, that can be obtained from the website https://rslab.ut.ac.ir/data. The RGB image of four data sets is shown in Fig. 6.

B. Experimental Setup

1) Implementation Details: We use an Adam optimizer with a weight decay of 1e-4 for Urban data set and 1e-3 for others. The number of epochs is 700. The learning rate decreases by G% after every S epoch is trained. Set the slope of the LeakyReLU to 0.2. Drop out 0.25 is used to prevent the fitting. For DBSCAN, we set D to 1e-3 for the real data sets, D to 0.2 for the synthetic data set, and Q equal to 13. Other more specific details are illustrated in Table II. The threshold value obtained automatically by OSTU in the experiment and the total number of pixels divided into the highly mixed region (represented by TN) are also listed in the table. The mask ratio is set to 0.9 for all experiments, which means that 90% of the TN pixels in the highly mixed region are masked. HU is an unsupervised task that does not require labeled samples for training, and the network evaluates the quality of the unmixing results based on the difference between the input and output. So the test can be done directly on the training set. In addition, selecting subsets for training may result in missing certain endmembers (i.e., some ground objects are not present in the subimage) [33]. So each unlabeled entire data set after MRM processing is simultaneously used as training set and test set to train the network and test the effect.

2) Evaluation Metrics: In the experiment, two commonly used evaluation metrics are introduced to assess the performance of algorithms: spectral angle distance (SAD) and root mean square error (RMSE), which are defined as follows:

$$\mathrm{SAD}\left(\widehat{\mathrm{m}}_i, \mathrm{m}_i\right) = \arccos\left(\frac{\widehat{\mathrm{m}}_i^{\mathrm{T}} \mathrm{m}_i}{\| \widehat{\mathrm{m}}_i \|_2 \| \mathrm{m}_i \|_2}\right) \tag{12}$$

$$\mathrm{RMSE}\left(\hat{\mathrm{a}}_j, \mathrm{a}_j\right) = \sqrt{\frac{1}{\mathrm{N}} \sum_{j=1}^{N} \| \hat{\mathrm{a}}_j - \mathrm{a}_j \|_2^2} \tag{13}$$

where $\widehat{\mathrm{m}}_i$ and m_i denote the extracted endmember and the reference endmember, respectively. $\hat{\mathrm{a}}_j$ and a_j are the estimated abundance and the reference abundance, respectively. The small values of SAD and RMSE imply better unmixing.

C. Experiment With Synthetic Data Set

1) Noise Robustness Analysis: To investigate the robustness of the proposed method, different signal-to-noise ratio (SNR) values from 5 to 30 dB are added in the synthetic experiment. Table III illustrates the results of the synthetic data set with different SNR values in terms of the SAD and RMSE. At high noise intensity (SNR = 5, 10 dB), as expected, the proposed MsCM-Net method achieves the best average SAD and average RMSE compared with those of other state-of-the-art methods. Also, the standard deviation of MsCM-Net is the lowest, outperforming the other state-of-the-art methods by a large margin. This is because MsCM-Net masks the pixels that interfere with the unmixing results. At relatively low noise intensity (SNR = 20, 30 dB), it still has the best average and standard deviation in terms of SAD. For the abundance comparisons, the proposed method produces the best standard deviation and achieves the comparable average RMSE. The main reason for the relative decline in abundance estimation ability is that MsCM-Net will

TABLE III
SAD AND RMSE FROM THE SYNTHETIC DATA SET WITH DIFFERENT SNR

Method		VCA	$L_{1/2}$-NMF	DAEU	NLAEU	PGMSU	TAEU	MuCAEU	MsCM-Net
SAD	SNR = 5 dB	0.128±0.007	0.145±0.006	0.100±0.015	0.132±0.011	0.118±0.033	0.045±0.014	0.062±0.004	**0.031±0.001**
	SNR = 10 dB	0.066±0.008	0.079±0.005	0.070±0.023	0.071±0.005	0.103±0.034	0.032±0.006	0.034±0.003	**0.021±0.001**
	SNR = 20 dB	0.036±0.002	0.040±0.003	0.048±0.012	0.040±0.005	0.101±0.059	0.027±0.007	0.022±0.002	**0.015±0.001**
	SNR = 30 dB	0.036±0.003	0.033±0.002	0.045±0.010	0.038±0.004	0.090±0.042	0.024±0.007	0.020±0.002	**0.016±0.001**
RMSE	SNR = 5 dB	0.132±0.013	0.151±0.013	0.190±0.018	0.118±0.008	0.129±0.003	0.098±0.011	0.116±0.006	**0.077±0.003**
	SNR = 10 dB	0.085±0.013	0.101±0.013	0.131±0.035	0.078±0.014	0.090±0.005	0.077±0.015	0.076±0.004	**0.061±0.002**
	SNR = 20 dB	**0.042**±0.004	0.066±0.004	0.123±0.037	0.056±0.013	0.062±0.005	0.070±0.013	0.055±0.003	0.058±**0.001**
	SNR = 30 dB	**0.040**±0.005	0.066±0.003	0.110±0.043	0.056±0.013	0.059±0.004	0.067±0.010	0.053±0.003	0.050±**0.002**

Mean and standard deviation are reported. Best results are reported in bold.

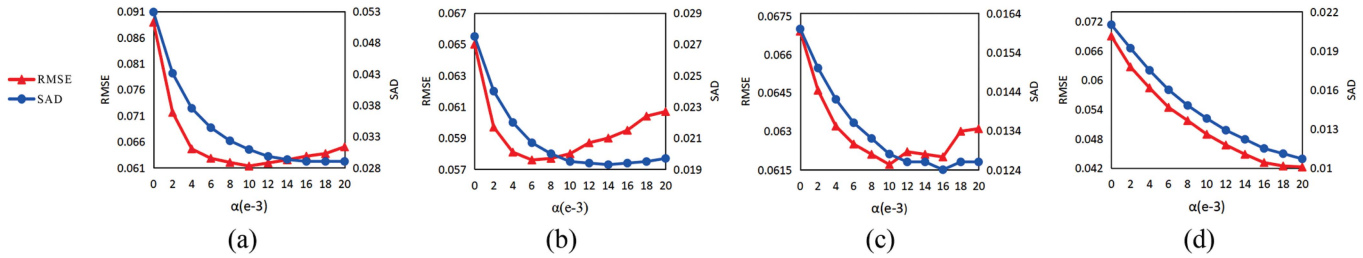

Fig. 7. Parameter sensitivity analysis of the proposed MsCM-Net method in the synthetic data set. (a) SNR = 5 dB. (b) SNR = 10 dB. (c) SNR = 20 dB. (d) SNR = 30 dB.

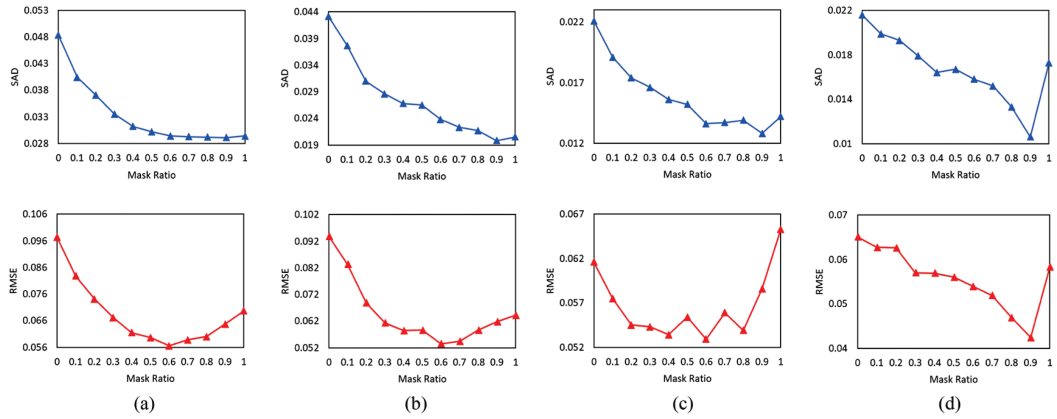

Fig. 8. Mask ratio analysis of the proposed MsCM-Net method in the synthetic data set. (a) SNR = 5 dB. (b) SNR = 10 dB. (c) SNR = 20 dB. (d) SNR = 30 dB.

mask some pixels, which will lose some information under low noise. Generally, methods that use convolutional networks to consider spatial information, including TAEU, MuCAEU, and MsCM-Net, commonly perform better. It shows that it is essential and correct to make full use of spatial information. Overall, the proposed method exhibits competitive robustness in abundance and endmember estimation under different noise levels, especially in high noise level scenarios.

2) Parameter Analysis: The performance of the proposed MsCM-Net method is somewhat sensitive to the setting of regularization parameter α. For this reason, the corresponding experiments are conducted to investigate the effects of parameter setting, as shown in Fig. 7. It can be observed that the general trend is that the best RMSE result comes from α between 6 and 10. And the SAD results change very little after $\alpha = 12$. However, SAD and RMSE results consistently improved when SNR = 30 dB. This may be because the synthetic data set is more sparse in higher SNR circumstance. There is a noticeable improvement in unmixing performance when setting the α from zero to nonzero. The improvement validates the necessity of adding the sparse regularization term.

3) Mask Ratio Analysis: We conduct corresponding experiments to verify whether the mask modeling is valid and the effect of the mask ratio on the unmixing solutions. The result is shown in Fig. 8. It can be obviously observed that when the mask is added, that is, when the mask ratio is from zero to nonzero, the results of SAD and RMSE are significantly improved at different noise levels. For endmember extraction, the general trend is that the larger the mask ratio, the better. This is because the remaining pixels can be better unmixed by a linear decoder. In terms of abundance estimation, RMSE shows a trend of decreasing first and then increasing with the increase of mask ratio. This may be

TABLE IV
SAD AND RMSE FROM THE SAMSON DATA SET

Method		VCA	$L_{1/2}$-NMF	DAEU	NLAEU	PGMSU	TAEU	MuCAEU	MsCM-Net
SAD	Soil	0.028±0.011	0.031±**0.004**	0.057±0.058	0.060±0.093	0.062±0.047	0.016±0.014	0.043±0.069	**0.012**±0.006
	Water	0.042±0.002	0.052±**0.001**	0.061±0.029	0.112±0.235	0.041±0.009	0.039±0.005	0.059±0.082	**0.030**±0.003
	Tree	0.378±0.371	0.096±0.006	0.056±0.030	0.133±**0.002**	0.188±0.062	0.077±0.021	0.045±0.004	**0.035**±0.012
	Average	0.149±0.127	0.060±**0.001**	0.058±0.032	0.102±0.109	0.097±0.026	0.045±0.011	0.049±0.051	**0.026**±0.005
RMSE	Soil	0.252±0.027	0.172±0.004	0.179±0.085	0.219±0.058	0.218±0.018	0.163±0.032	0.088±0.084	**0.055±0.003**
	Water	0.264±0.023	0.179±0.008	0.192±0.075	0.236±0.071	0.231±0.007	0.065±0.011	0.078±0.091	**0.047±0.004**
	Tree	0.424±0.012	0.509±0.070	0.162±0.112	0.339±0.032	0.359±0.012	0.177±0.027	0.037±0.023	**0.030±0.006**
	Average	0.313±0.008	0.287±0.022	0.178±0.061	0.264±0.037	0.270±0.010	0.135±0.022	0.068±0.066	**0.044±0.003**

Mean and standard deviation are reported. Best results are reported in bold.

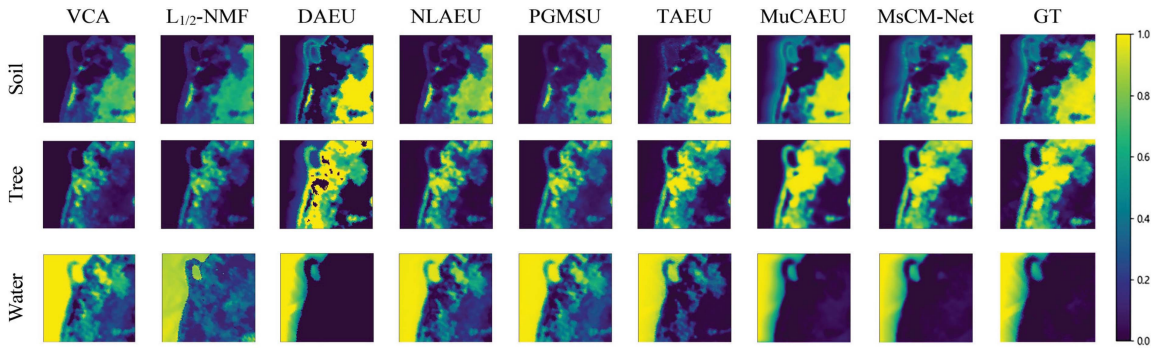

Fig. 9. Abundance maps from the Samson data set obtained by the different methods.

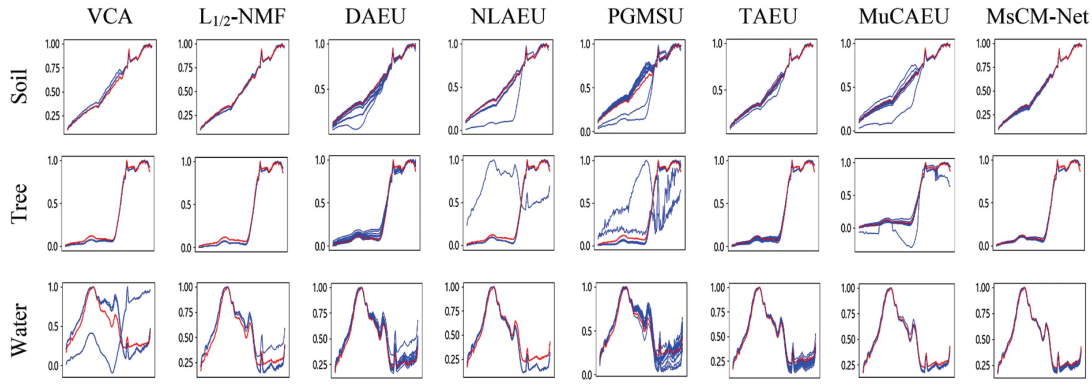

Fig. 10. Plots of all extracted endmembers by all methods (blue) and the reference endmembers (red) for the Samson data set.

because a larger mask ratio will lose more spatial information. In high SNR circumstance, the entire mask (mask ratio = 1) produces poor unmixing results because the highly mixed region is not involved in training at all, which also affects the unmixing performance. Therefore, a general trend for the mask ratio setting can be summarized empirically. It is recommended to set a higher ratio, which will extract better endmembers. Although this may not necessarily be the optimal abundance result, it will be within the acceptable range.

D. Experiment With Samson Data Set

Table IV and Fig. 9 present the unmixing results and the corresponding abundance maps, respectively, in the Samson data set. It can be seen that the proposed MsCM-Net method achieves the best performance in terms of the SAD and RMSE, which validates the effectiveness of the proposed method. In addition, the abundance maps obtained by MsCM-Net are visually similar to the ground truth in Fig. 9. Fora illustrative purposes, all endmembers extracted by all methods, with corresponding GTs, are shown in Fig. 10. From the figure, we can see that MsCM-Net has good accuracy and consistency compared with other methods. This means that, it can stably extract accurate endmembers in multiple runs.

E. Experiment With Urban Data Set

The Urban data set has a larger size than the Samson data set mentioned earlier. Table V illustrates the results of different algorithms in the Urban data set, and the corresponding abundance maps are depicted in Fig. 11. As seen from Table V, the proposed MsCM-Net method significantly improves performance and also achieved the best results in terms of endmember and abundance estimation, including standard deviation. Especially in terms

TABLE V
SAD AND RMSE FROM THE URBAN DATA SET

Method		VCA	$L_{1/2}$-NMF	DAEU	NLAEU	PGMSU	TAEU	MuCAEU	MsCM-Net
SAD	Asphalt	0.137±0.015	0.206±0.028	0.088±0.016	0.160±0.031	0.119±0.021	0.122±0.020	0.090±0.019	**0.065±0.003**
	Grass	0.413±0.030	0.406±0.034	0.080±0.029	0.424±0.043	0.211±0.068	0.165±0.030	0.197±0.130	**0.046±0.003**
	Tree	0.715±0.176	0.837±0.032	0.051±0.015	0.592±0.198	0.055±0.009	0.086±0.016	0.104±0.194	**0.029±0.002**
	Roof	0.759±0.158	0.785±0.547	0.264±0.102	0.488±0.318	0.188±0.090	0.154±0.030	0.114±0.171	**0.025±0.001**
	Average	0.506±0.083	0.559±0.143	0.121±0.025	0.416±0.132	0.143±0.033	0.132±0.020	0.126±0.116	**0.041±0.001**
RMSE	Asphalt	0.341±0.013	0.336±0.014	0.182±0.026	0.300±**0.004**	0.289±0.012	0.176±0.006	**0.142**±0.032	0.143±0.011
	Grass	0.299±0.014	0.314±0.012	0.174±0.016	0.306±0.009	0.261±0.041	0.240±0.029	0.185±0.055	**0.139±0.006**
	Tree	0.444±**0.001**	0.435±0.009	0.118±0.046	0.412±0.015	0.252±0.042	0.200±0.018	0.132±0.022	**0.107**±0.005
	Roof	0.445±0.071	0.231±0.014	0.100±0.020	0.445±0.032	0.388±0.036	**0.083**±0.007	0.101±0.021	0.098±**0.003**
	Average	0.382±0.018	0.329±0.008	0.143±0.017	0.366±0.014	0.297±0.018	0.175±0.012	0.140±0.016	**0.122±0.005**

Mean and standard deviation are reported. Best results are reported in bold.

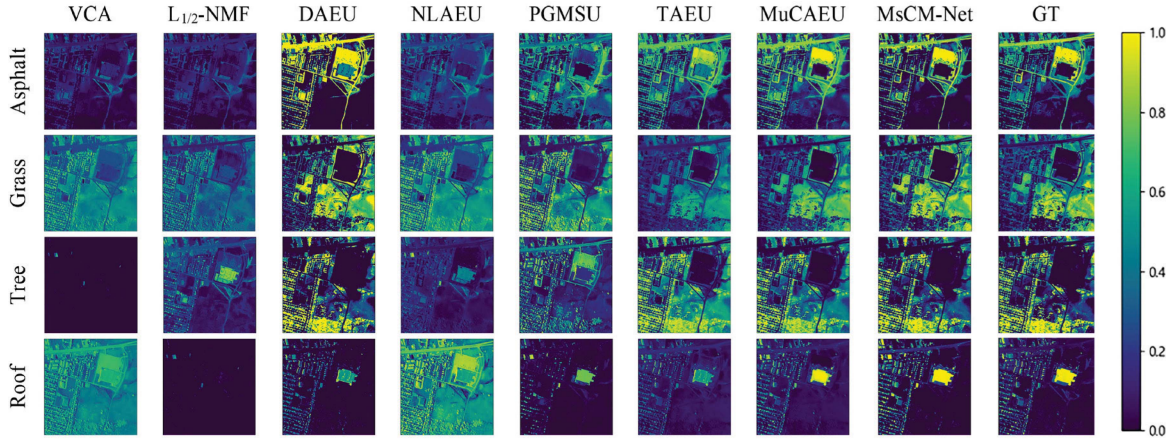

Fig. 11. Abundance maps from the Urban data set obtained by the different methods.

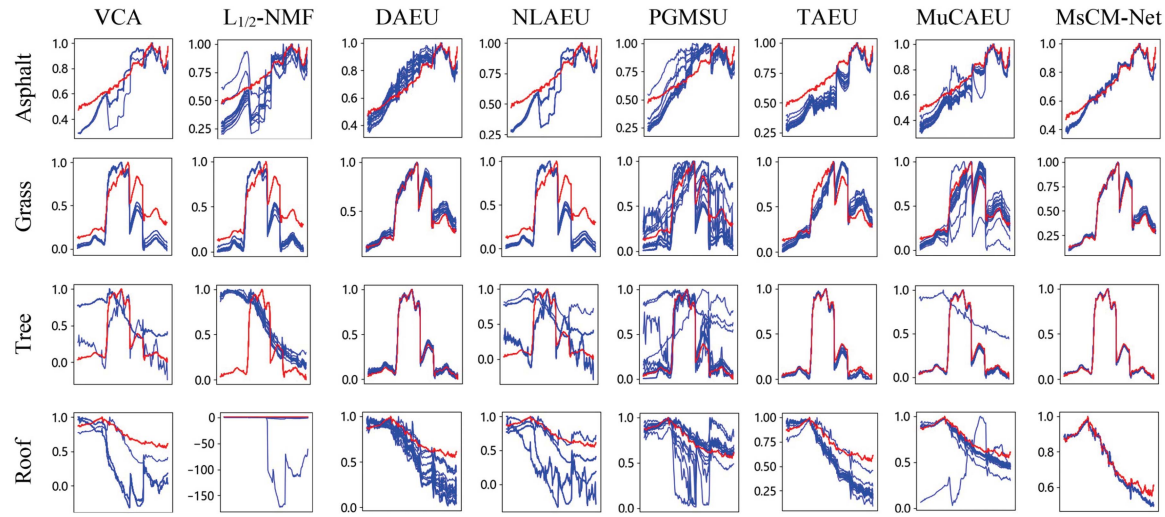

Fig. 12. Plots of all extracted endmembers by all methods (blue) and the reference endmembers (red) for the Urban data set.

of endmember extraction, the best SAD values and standard deviations are achieved on all materials. Compared with the second-best value, the average SAD error is reduced by 66.1% and the standard deviation by 95%. For illustrative purposes, all endmembers extracted by all methods, with corresponding GTs, are shown in Fig. 12. It is obvious that the MsCM-Net can extract more accurate endmembers in each run, which further validates its effectiveness for HU in real scenarios.

F. Experiment With Washington DC Data Set

The Washington DC data set is similar to the Urban data set in that it has a large size and complex material distribution. Table VI lists the performance assessment for all algorithms, and the corresponding abundance maps are displayed in Fig. 13. The proposed method achieves the best results on average SAD, average RMSE, and their standard deviations. The average SAD value is the only one less than 0.1. The average RMSE of

TABLE VI
SAD AND RMSE FROM THE WASHINGTON DC DATA SET

Method		VCA	$L_{1/2}$-NMF	DAEU	NLAEU	PGMSU	TAEU	MuCAEU	MsCM-Net
SAD	Grass	0.177±**0.001**	0.126±0.044	**0.073**±0.017	0.182±0.044	0.127±0.034	0.166±0.038	0.118±0.022	0.140±0.008
	Tree	0.321±0.014	0.341±0.018	**0.180**±0.040	0.360±0.021	0.253±0.055	0.228±0.055	0.201±0.006	0.214±**0.005**
	Road	0.180±0.010	0.212±0.022	0.220±0.011	0.261±**0.006**	0.202±0.023	0.162±0.086	0.108±0.104	**0.035**±0.033
	Water	0.475±0.118	0.398±0.178	0.503±0.033	0.338±0.095	0.344±0.249	0.212±0.226	0.023±**0.002**	**0.022**±0.003
	Roof	0.160±0.026	0.286±0.090	0.401±0.186	0.216±0.037	0.316±0.221	0.166±0.163	0.115±0.086	**0.062±0.025**
	Average	0.263±0.029	0.273±0.051	0.275±0.036	0.271±0.025	0.248±0.080	0.187±0.068	0.113±0.023	**0.090±0.005**
RMSE	Grass	0.376±0.054	0.308±0.008	0.354±0.034	**0.244**±0.032	0.277±0.052	0.309±0.038	0.297±0.040	0.253±**0.005**
	Water	0.377±**0.001**	0.341±0.021	**0.236**±0.038	0.367±0.005	0.341±0.075	0.248±0.022	0.279±0.038	0.245±0.005
	Roof	0.227±**0.009**	0.207±0.022	0.221±0.031	0.235±0.014	0.219±0.041	0.181±0.038	0.206±0.019	**0.148**±0.041
	Road	0.408±0.014	0.402±0.018	0.190±0.106	0.346±**0.013**	0.292±0.017	0.172±0.034	0.190±0.016	**0.162**±0.019
	Tree	0.449±0.176	0.176±0.054	0.396±0.085	0.180±0.033	0.212±0.037	0.141±0.061	0.141±0.053	**0.106±0.016**
	Average	0.367±0.040	0.287±0.016	0.279±0.032	0.274±0.012	0.268±0.037	0.210±0.020	0.223±0.026	**0.179±0.010**

Mean and standard deviation are reported. Best results are reported in bold.

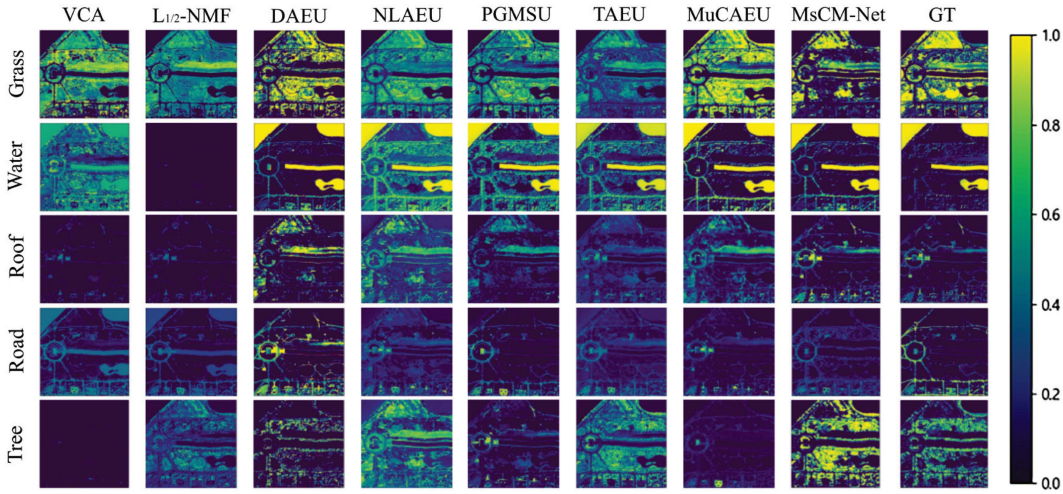

Fig. 13. Abundance maps from the Washington DC data set obtained by the different methods.

TABLE VII
EFFECTS OF COMPONENTS IN OUR DESIGN

	Synthetic (10 dB)		Samson		Urban		Washington DC	
	SAD	RMSE	SAD	RMSE	SAD	RMSE	SAD	RMSE
Single-scale+VCA	0.049±0.003	0.108±0.003	0.137±0.200	0.098±0.085	0.170±0.107	0.228±0.085	0.109±0.004	0.224±0.007
Single-sclae+MRM+VCA	0.027±0.001	0.091±0.003	0.045±0.030	0.067±0.050	0.075±0.038	0.174±0.013	0.097±0.008	0.184±0.006
Single-sclae+MRM+DBSCAN-VCA	0.024±0.001	0.090±0.004	0.032±0.006	0.053±0.003	0.045±0.002	0.131±0.004	0.093±0.006	0.189±**0.001**
Two-sclae+MRM+DBSCAN-VCA	0.022±0.001	0.072±0.004	0.028±0.004	0.048±0.005	0.042±0.001	0.123±0.004	0.091±0.005	0.183±0.010
Three-sclae+MRM+DBSCAN-VCA	**0.021±0.001**	**0.061±0.002**	**0.026**±0.005	0.044±0.003	**0.041±0.001**	0.122±0.005	**0.090**±0.005	0.179±0.010
Four-sclae+MRM+DBSCAN-VCA	0.024±0.001	0.067±0.005	0.027±**0.002**	**0.043±0.002**	0.043±0.002	**0.121±0.003**	0.092±**0.003**	**0.178**±0.005

Mean and standard deviation are reported. Best results are reported in bold.

abundance map is the only one less than 0.2. It is further proved that the proposed MsCM-Net method has outstanding advantages in dealing with complex situations. For illustrative purposes, all endmembers extracted by all methods, with corresponding GTs, are shown in Fig. 14.

IV. DISCUSSION

A. Ablation Experiment

We gradually add different components in our design to investigate the effectiveness of each component. Table VII lists the results. The single-scale model is the baseline method. It only considers the information at the original scale, and the model architecture is consistent with Table I. Two-scale, three-scale, and four-scale refer to continue to add 1/2 scale, 1/4 scale, and 1/8 scale information, respectively. It can be seen that after the addition of MRM, the unmixing accuracy of each data set is significantly improved. Especially in terms of the average SAD value, which is increased by 67% on Samson and 56% on Urban. After adding DBSCAN-VCA initialization, the accuracy and stability of the method are further improved. When the 1/2 scale, 1/4 scale, and 1/8 scale are gradually added, it can

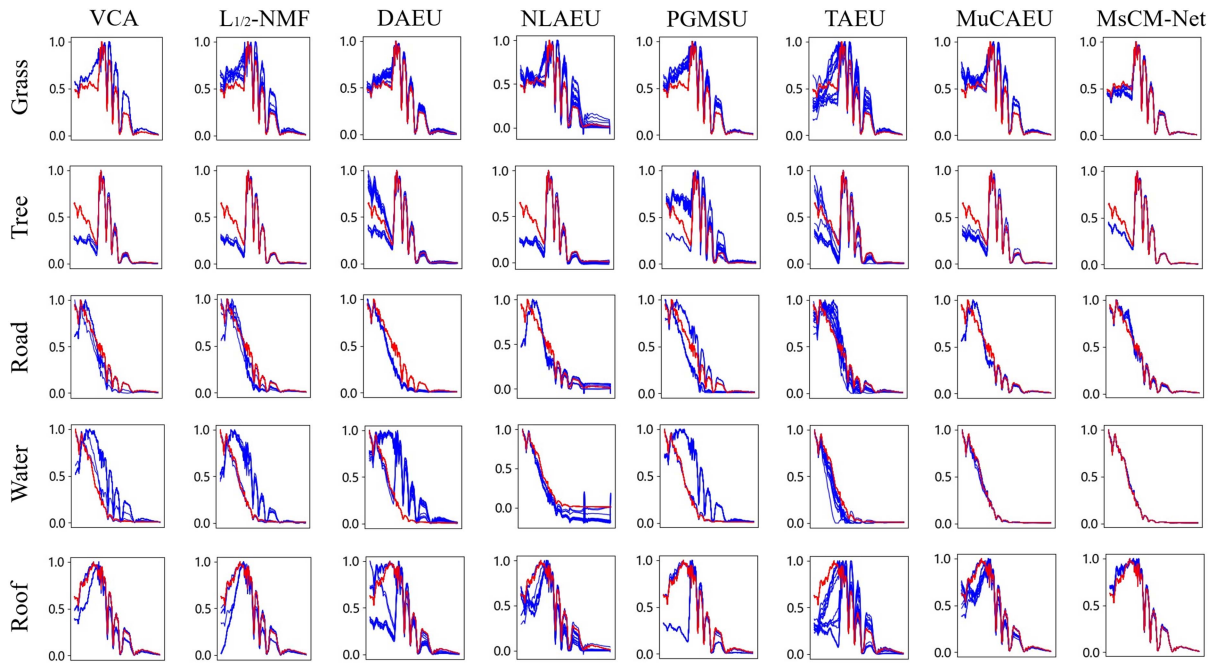

Fig. 14. Plots of all extracted endmembers by all methods (blue) and the reference endmembers (red) for the Washington DC data set.

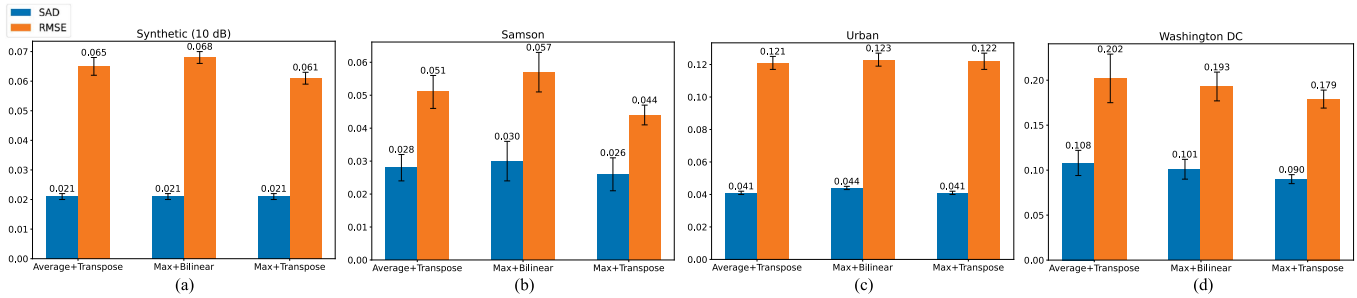

Fig. 15. Experimental results of different network configurations. The height of the column and the values on the column represent the average. The black line is the error bar and the length represents the size of the standard deviation. (a) Synthetic (10 dB). (b) Samson. (c) Urban. (d) Washington DC.

be found that when considering the information of the three scales, the unmixing accuracy is further improved compared with that of the single-scale network. However, after the addition of the fourth scale, there is no significant increase or even decrease. In summary, the ablation experiment demonstrates the effectiveness of the proposed MRM and DBSCAN-VCA strategies. The performance can be improved by using multiscale network after applying the new strategies. Next, the rationality of the network structure setting is also tested. We use three configurations for the experiment, namely average pooling and transposed convolution, max pooling and bilinear upsampling, and max pooling and transposed convolution (i.e., the setup for MsCM-Net). The experimental results are shown in Fig. 15. It can be seen that max pooling and transposed convolution can obtain better unmixing results, indicating that the choice of this configuration is correct. Then we will discuss the effect of MRM and DBSCAN-VCA further.

1) MRM: The patches mask (PM) is commonly used in CV field, which divides the image into several small patches proportionally and masks them. While MRM is masking at the pixel scale. In addition, the existing PM models usually use the masked image in network training and the original image in network testing. However, the main purpose of MRM is to mask the interference of noise in highly mixed regions, so it inputs masked images in both processes. In order to verify the applicability of the proposed MRM in HU, the single-scale network is used to test the effect of PM with a mask ratio of 0.15 and 0.05, respectively. The effect when MRM inputs the original image in network testing (called OMRM) is also tested. The results are shown in Table VIII. To explore only the effects of the mask mechanism, all experiments in the table are initialized using the original VCA. It can be seen that for the Samson data set, PM introduces a lot of additional noise, resulting in unstable results. For the other three data sets, decreasing the mask ratio leads to an improvement in abundance, but the endmember extraction accuracy does not improve. This may be because with a lower mask ratio, less information is ignored and the abundance naturally improves. But the masked area is

TABLE VIII
RESULTS OF EFFECTIVENESS EXPERIMENTS BETWEEN PM AND MRM

	Synthetic (10 dB)		Samson		Urban		Washington DC	
	SAD	RMSE	SAD	RMSE	SAD	RMSE	SAD	RMSE
Single-scale	0.049±0.003	0.108±0.003	0.137±0.200	0.098±0.085	0.170±0.107	0.228±0.085	0.109±**0.004**	0.224±0.007
Single-scale+0.15 PM	0.032±0.001	0.103±**0.002**	0.167±0.164	0.216±0.139	0.110±0.056	0.192±0.015	0.112±0.032	0.192±0.025
Single-scale+0.05 PM	0.036±0.002	0.093±0.003	0.110±0.124	0.191±0.122	0.111±**0.034**	**0.168±0.008**	0.114±0.041	**0.181**±0.015
Single-scale+OMRM	**0.024±0.001**	0.097±0.004	0.068±0.101	0.131±0.127	0.123±0.042	0.175±0.009	0.107±0.035	0.195±0.031
Single-scale+MRM	0.027±0.001	**0.091**±0.003	**0.045±0.030**	**0.067±0.050**	**0.075**±0.038	0.174±0.013	**0.097**±0.008	0.184±**0.006**

Mean and standard deviation are reported. Best results are reported in bold.

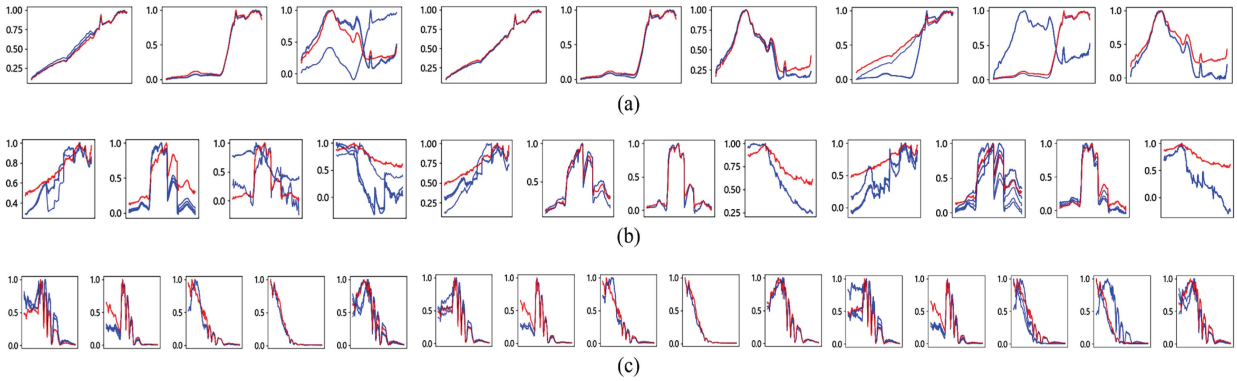

Fig. 16. Each row from left to right shows the endmember extraction results (blue curve) and reference endmembers (red) from VCA initialization, DBSCAN-VCA initialization, and HKS-VCA initialization of the data set. (a) Samson. (b) Urban. (c) Washington DC.

TABLE IX
RESULTS OF DBSCAN-VCA NECESSITY EXPERIMENTS

	Samson		Urban		Washington DC	
	SAD	RMSE	SAD	RMSE	SAD	RMSE
TAEU+VCA	0.045±0.011	0.135±0.022	0.132±0.020	0.175±0.012	0.187±0.068	0.210±0.020
TAEU+DBSCAN-VCA	0.026±0.004	0.105±0.014	0.120±0.018	0.135±0.009	0.153±0.040	0.210±0.011
NLAEU+VCA	0.102±0.109	0.264±0.037	0.416±0.132	0.366±0.014	0.271±0.025	0.274±0.012
NLAEU+DBSCAN-VCA	0.065±0.003	0.237±0.013	0.339±0.117	0.373±0.004	0.178±0.012	0.276±0.004

also reduced and is not necessarily the mixed area that would interfere with the unmixing. These areas are still input into the network during testing and still affect the unmixing. It is also better to input the masked image both for training and testing than to switch to the original image for testing. In summary, MRM that is specific to mixed regions is more suitable for HU tasks.

2) DBSCAN-VCA: We use three real data sets to verify the effectiveness of DBSCAN-VCA initialization. Fig. 16 shows the results of using VCA initialization, DBSCAN-VCA initialization, and the VCA initialization based on heat kernel similarity (HKS-VCA) proposed in [19]. The number of runs is 15. It is obvious that the initial results of adding DBSCAN are more similar to ground truths and more stable than the other two methods.

In order to further verify whether the good initialization generated by DBSCAN-VCA is beneficial to other methods, we choose two methods, TAEU and NLAEU, to conduct experiments on three real data sets in combination with DBSCAN-VCA. The results are shown in Table IX. It can be seen that good initialization plays an obvious role in improving both methods, which proves the necessity of this strategy.

B. Computational Cost

We compare the computational complexity of all test methods, and the result of the running time is shown in Table X. All the experiments are conducted on the same PC with an Intel 12th i7 CPU, 16-GB memory, and one NVIDIA GeForce RTX 3060 graphic card. It can be seen that the proposed MsCM-Net method has a competitive computational cost compared with other methods on relatively small data sets, such as the Samson data set and the Synthetic data set. However, it takes more time to run on large data sets, including the Urban data set and the Washington DC data set, because of the DBSCAN-VCA initialization. But the computational cost is still acceptable.

TABLE X
COMPUTATIONAL COST OF ALL METHODS ON DIFFERENT DATA SETS IN TERMS OF SECONDS (S)

Method	Synthetic (10 dB)	Samson	Urban	Washington DC
VCA	0.6	0.7	7.1	23
$L_{1/2}$-NMF	14	17.6	47.9	42
DAEU	35	40.5	97.8	105
NLAEU	10	14.1	61.7	72
PGMSU	20	22.1	89.2	114
TAEU	6	7.1	74.8	94
MuCAEU	9	12.7	51	53
MsCM-Net	15	21	91	94

V. CONCLUSION

In this article, we propose a new multiscale convolutional mask unmixing network named MsCM-Net. We discuss the challenges faced by DL-based unmixing methods and give the corresponding solutions. Through the mixed region mask strategy suitable for HU tasks, MsCM-Net can reduce the influence of noise present in highly mixed regions, making the unmixing process more robust. Based on the consideration of local similarity, a new initialization strategy combining DBSCAN and VCA is proposed. Benefiting from the good initialization of DBSCAN-VCA, more reasonable and superior unmixing results are further yielded. Extensive experiments on synthetic and real data sets demonstrate the effectiveness and robustness of the proposed method. Especially on low SNR or relatively complex data sets, the proposed method has outstanding advantages over other state-of-the-art unmixing approaches. The ablation experiment also demonstrates the advantages of MRM in HU tasks compared with existing patches mask. The benefits of good initialization are also proven. In the future, we will further study the application of mask modeling in HU.

REFERENCES

[1] L. Zhang, L. Zhang, D. Tao, and X. Huang, "Sparse transfer manifold embedding for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1030–1043, Feb. 2014.

[2] C.-I. Chang, "An effective evaluation tool for hyperspectral target detection: 3D receiver operating characteristic curve analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5131–5153, Jun. 2021.

[3] D. Zhu, B. Du, and L. Zhang, "Two-stream convolutional networks for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6907–6921, Aug. 2021.

[4] L. Zhang, L. Zhang, D. Tao, and X. Huang, "On combining multiple features for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 879–893, Mar. 2012.

[5] 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.

[6] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2021, Art. no. 5518615.

[7] S. Huang and D. W. Messinger, "An unsupervised Laplacian pyramid network for radiometrically accurate data fusion of hyperspectral and multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5527517.

[8] J. Li, K. Zheng, J. Yao, L. Gao, and D. Hong, "Deep unsupervised blind hyperspectral and multispectral data fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Feb. 2022, Art. no. 6007305.

[9] G. Iyer, J. Chanussot, and A. L. Bertozzi, "A graph-based approach for data fusion and segmentation of multimodal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4419–4429, May 2021.

[10] 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, Feb. 2013.

[11] J. Li, A. Agathos, D. Zaharie, J. M. Bioucas-Dias, A. Plaza, and X. Li, "Minimum volume simplex analysis: A fast algorithm for linear hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 9, pp. 5067–5082, Sep. 2016.

[12] 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.

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

[14] G. A. Licciardi and F. Del Frate, "Pixel unmixing in hyperspectral data by means of neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4163–4172, Nov. 2011.

[15] R. Guo, W. Wang, and H. Qi, "Hyperspectral image unmixing using autoencoder cascade," in *Proc. Workshop Hyperspectral Image Signal Process.: Evol. Remote Sens.*, 2015, pp. 1–4.

[16] F. Palsson, J. Sigurdsson, J. R. Sveinsson, and M. O. Ulfarsson, "Neural network hyperspectral unmixing with spectral information divergence objective," in *Proc. Dig. Int. Geosci. Remote Sens. Symp.*, 2017, pp. 755–758.

[17] 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.

[18] M. Wang, M. Zhao, J. Chen, and S. Rahardja, "Nonlinear unmixing of hyperspectral data via deep autoencoder networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1467–1471, Sep. 2019.

[19] Z. Dou, K. Gao, X. Zhang, H. Wang, and J. Wang, "Hyperspectral unmixing using orthogonal sparse prior-based autoencoder with hyper-Laplacian loss and data-driven outlier detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6550–6564, Sep. 2020.

[20] K. T. Shahid and I. D. Schizas, "Unsupervised hyperspectral unmixing via nonlinear autoencoders," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5506513.

[21] Y. Su, J. Li, A. Plaza, A. Marinoni, P. Gamba, and S. Chakravortty, "DAEN: Deep autoencoder networks for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4309–4321, Jul. 2019.

[22] 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.*, vol. 34, no. 8, pp. 4555–4569, Aug. 2023.

[23] S. Shi, M. Zhao, L. Zhang, Y. Altmann, and J. Chen, "Probabilistic generative model for hyperspectral unmixing accounting for endmember variability," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5516915.

[24] 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.

[25] P. Ghosh, S. K. Roy, B. Koirala, B. Rasti, and P. Scheunders, "Deep hyperspectral unmixing using transformer network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 5535116.

[26] 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, Mar. 2022, Art. no. 5503914.

[27] Y. Yu, Y. Ma, X. Mei, F. Fan, J. Huang, and H. Li, "Multi-stage convolutional autoencoder network for hyperspectral unmixing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 113, p. 102981, Sep. 2022.

[28] B. Rasti, B. Koirala, P. Scheunders, and J. Chanussot, "MiSiCNet: Minimum simplex convolutional network for deep hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5522815.

[29] J. S. Bhatt and M. V. Joshi, "Deep learning in hyperspectral unmixing: A review," in *Proc. Dig. Int. Geosci. Remote Sens. Symp.*, 2020, pp. 2189–2192.

[30] X.-R. Feng, H.-C. Li, R. Wang, Q. Du, X. Jia, and A. Plaza, "Hyperspectral unmixing based on nonnegative matrix factorization: A comprehensive review," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4414–4436, May 2022.

[31] M. Gaur, K. Faldu, and A. Sheth, "Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable?," *IEEE Internet Comput.*, vol. 25, no. 1, pp. 51–59, Jan./Feb. 2021.
[32] 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.*, vol. 33, no. 11, pp. 6518–6531, Nov. 2022.
[33] F. Xiong, J. Zhou, S. Tao, J. Lu, and Y. Qian, "SNMF-Net: Learning a deep alternating neural network for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2021, Art. no. 5510816.
[34] J. M. Nascimento and J. M. 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.
[35] B. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Blind hyperspectral unmixing using autoencoders: A critical comparison," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1340–1372, Jan. 2022.
[36] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, Aug. 2020, Art. no. 1295.
[37] L. Wang, Y. Zhang, and J. Feng, "On the Euclidean distance of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1334–1339, Jun. 2005.
[38] J. Zou, J. Lan, and Y. Shao, "A hierarchical sparsity unmixing method to address endmember variability in hyperspectral image," *Remote Sens.*, vol. 10, no. 5, May 2018, Art. no. 738.
[39] L. Zhou et al., "Subspace structure regularized nonnegative matrix factorization for hyperspectral unmixing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4257–4270, Jul. 2020.
[40] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1139–1147.
[41] B. Kazimipour, X. Li, and A. K. Qin, "A review of population initialization techniques for evolutionary algorithms," in *Proc. IEEE Congr. Evol. Computation*, 2014, pp. 2585–2592.
[42] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. Int. Conf. Knowl. Discov. Data Mining*, 1996, pp. 226–231.
[43] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "DBSCAN: Past, present and future," in *Proc. Int. Conf. Appl. Digit. Inf. Web Technol.*, 2014, pp. 232–238.
[44] J. Li, Y. Li, R. Song, S. Mei, and Q. Du, "Local spectral similarity preserving regularized robust sparse hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7756–7769, Oct. 2019.
[45] F. Zhu, Y. Wang, B. Fan, G. Meng, S. Xiang, and C. Pan, "Spectral unmixing via data-guided sparsity," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5412–5427, Dec. 2014.
[46] B. Rasti, B. Koirala, P. Scheunders, and P. Ghamisi, "UnDIP: Hyperspectral unmixing using deep image prior," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5504615.
[47] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15979–15988.
[48] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9653–9663.
[49] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-BERT: Pre-training 3D point cloud transformers with masked point modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19313–19322.
[50] M. Assran et al., "Masked siamese networks for label-efficient learning," in *Lect. Notes Comput. Sci.*, Tel Aviv, Israel, 2022, pp. 456–473.
[51] Y. Liu, S. Zhang, J. Chen, K. Chen, and D. Lin, "PixMIM: Rethinking pixel reconstruction in masked image modeling," 2023, *arXiv:2303.02416*.
[52] C. Zhang, C. Zhang, J. Song, J. S. K. Yi, K. Zhang, and I. S. Kweon, "A survey on masked autoencoder for self-supervised learning in vision and beyond," 2022, *arXiv:2208.00173*.
[53] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.
[54] Q. Sun, X. Liu, S. Bourennane, and B. Liu, "Multiscale denoising autoencoder for improvement of target detection," *Int. J. Remote Sens.*, vol. 42, no. 8, pp. 3002–3016, Jan. 2021.
[55] X. Liu, W. Xia, B. Wang, and L. Zhang, "An approach based on constrained nonnegative matrix factorization to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 757–772, Feb. 2010.
[56] 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, Jun. 2011.
[57] Q. Zhu, L. Jing, and R. Bi, "Exploration and improvement of OSTU threshold segmentation algorithm," in *Proc. 8th World Congr. Intell. Control Automat.*, 2010, pp. 6183–6188.
[58] F. Xiong, J. Zhou, J. Lu, and Y. Qian, "Nonconvex nonseparable sparse nonnegative matrix factorization for hyperspectral unmixing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 6088–6100, Oct. 2020.
[59] G. P. Asner and K. B. Heidebrecht, "Spectral unmixing of vegetation, soil and dry carbon cover in arid regions: Comparing multispectral and hyperspectral observations," *Int. J. Remote Sens.*, vol. 23, no. 19, pp. 3939–3958, Nov. 2002.
[60] C. O. Davis, M. Kavanaugh, R. Letelier, W. P. Bissett, and D. Kohler, "Spatial and spectral resolution considerations for imaging coastal waters," *Proc. SPIE*, vol. 6680, 2007, pp. 196–207.
[61] L. J. Rickard, R. W. Basedow, E. F. Zalewski, P. R. Silverglate, and M. Landers, "HYDICE: An airborne system for hyperspectral imaging," *Proc. SPIE*, vol. 1937, 1993, pp. 173–179.

Mingming Xu (Member, IEEE) received the B.S. degree in surveying and mapping engineering from the China University of Petroleum, Qingdao, China, in 2011, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Lab of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2016.

She is currently an Associate Professor with the College of Oceanography and Space Informatics, China University of Petroleum. Her research interests include hyperspectral image processing and wetland remote sensing.

Jin Xu received the bachelor's degree in geomatics engineering in 2023 from the China University of Petroleum (East China), Qingdao, China, where he is currently working toward the master's degree in geomatics science and technology with the College of Oceanography and Space Informatics.

His research focuses on hyperspectral unmixing.

Shanwei Liu received the Ph.D. degree in environmental science from the Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Beijing, China, in 2011.

He is currently a Professor with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao, China. His research interests include satellite altimetry, ocean remote sensing, hyperspectral image processing, and GIS application.

Hui Sheng received the Ph.D. degree in geological resources and geological engineering from the China University of Petroleum (East China), Qingdao, China, in 2010.

He is currently an Associate Professor with the College of Oceanography and Space Informatics, China University of Petroleum (East China). His research interests include ocean remote sensing, hyperspectral image processing, and photogrammetry.

Zhiru Yang received the bachelor's degree in geomatics engineering in 2021 from the China University of Petroleum (East China), Qingdao, China, where she is currently working toward the master's degree in geomatics science and technology with the College of Oceanography and Space Informatics.

Her research focuses on hyperspectral unmixing.