Spatial-Spectral Attention Bilateral Network for Hyperspectral Unmixing

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

***Abstract*— Autoencoders (AEs) are widely utilized in hyperspectral unmixing (HU) as an unsupervised learning model. In particular, convolutional AE networks are popular for processing multidimensional hyperspectral features. Nonetheless, the traditional convolutional AE network's receptive field is constrained in the unmixing task, and establishing the connection between the local spatial neighborhood and the local spectrum fails to improve unmixing performance significantly. To address these limitations, a bilateral global attention network based on both spatial and spectral information is proposed. It enables the network to obtain respective feature dependencies in the two dimensions and achieve optimal fusion of both features. The network comprises two information extraction branches. The spatial information extraction branch uses the Swin Transformer block to acquire the global spatial attention of the overall image, while the spectral information extraction branch designates a simplified spectral channel attention mechanism to gain spectral attention weight maps. The network's efficacy is demonstrated through a comparative study using a synthetic dataset and two real datasets. The code of this work is available at https://github.com/UPCGIT/SSABN.**

***Index Terms*— Attention, autoencoder (AE), bilateral network, hyperspectral unmixing (HU), spatial-spectral information.**

I. Introduction

THE hyperspectral image (HSI) is a 3-D cube with two spatial dimensions and one spectral dimension. There is a one-to-one correspondence between the color spots and the spectrum of HSI. The distribution of target objects can be obtained by analyzing the spatial dimension information and spectral structure characteristics of HSIs. At present, HSIs have been widely used in many fields, such as mineral detection, agricultural detection, target detection, medical treatment, national defense, and other fields [1]. However, due to the low spatial resolution of the sensor, the spectral features obtained in the same pixel are the mixture and superposition of multiple substances, which forms the mixed pixel phenomenon. Decomposing pixels to get sub-pixel information is an important research direction in the field of hyperspectral remote sensing [2]. The main aim of unmixing is to identify the spectral signal and proportions of each substance in collected mixed pixels, known as endmember (EM) extraction and abundance estimation. EM extraction is usually based on geometric assumptions [3], [4], while previous approaches use extracted EMs or prior spectral libraries for abundance estimation [5], [6], [7], making them neither unsupervised nor automated.

Blind hyperspectral unmixing (HU), also known as unsupervised HU, can simultaneously provide EMs and abundances in HSIs. Data-driven deep learning (DL) methods, particularly the autoencoder (AE) model, have demonstrated impressive potential in HU. AE model, a widely used DL model, comprises three major components: encoder, hidden layers, and decoder. The encoder learns critical features of the data via the hidden layer, thereby reducing data dimensionality, and the decoder reconstructs the high-dimensional representation of the data via the hidden layer. AEs have undergone a series of developments, resulting in improved models with various deformable structures [8], [9], [10], [11].

The majority of existing AE networks are trained based on spectral bands, but it is not guaranteed that the extracted features include the spatial information of the original data. Consequently, this may result in a loss of part of the original information during the unmixing process. To address this issue, it is effective to design the encoder to operate on training image patches rather than individual pixels, as this can leverage the spatial information in HSIs. Convolutional AE-based networks can be trained using patches of images, thus taking into account the spatial information of the patches [11], [12]. However, local convolutional filters fail to capture the relationships between distant pixels and spectra, resulting in the loss of essential non-local feature information necessary for unmixing. Fortunately, the Transformer architecture [13], [14] has been proposed to address the issue of limited convolution receptive field. It incorporates a multihead self-attention mechanism, which allows for the capture of global information. The attention mechanism utilizes specific feature maps as object-specific conditional weights to modify the behavior of preceding feature maps. This compels the network to prioritize the representation of the object of interest within the feature extraction block [15].

Therefore, a two-stream attention network is proposed, which considers both spatial and spectral information of HSI. The specific contributions of the proposed network are as follows: 1) the network combines the convolution filter, and the shifted windows Transformer block (STB), which can

Manuscript received 13 April 2023; revised 9 July 2023; accepted 11 July 2023. Date of publication 14 July 2023; date of current version 9 August 2023. 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: s21160036@s.upc.edu.cn; xumingming@upc.edu.cn; shanweiliu@163.com; sheng@upc.edu.cn; zhenghongxia@upc.edu.cn).

Digital Object Identifier 10.1109/LGRS.2023.3295437

simultaneously acquire the local and global spatial information of HSI; and 2) a simplified channel attention mechanism is designed to mine interdependencies between spectral feature channels.

The structure of this letter is as follows. In Section II, the typical AE methods and our proposed method are elaborated. In Section III, we present the performance of the algorithm and its implementation on synthetic and real data. Finally, a conclusion is made in Section IV.

II. Related Works

In this section, the linear mixed model (LMM) and classical LMM-based AE network are introduced. Furthermore, we present the details of the proposed spatial-spectral attention bilateral network (SSABN).

A. LMM Autoencoder

The LMM can be expressed as follows:

$$\boldsymbol{X} = \boldsymbol{E}\boldsymbol{A} + \boldsymbol{\xi} \tag{1}$$

where $\boldsymbol{X}$ represents an HSI, $\boldsymbol{X} \in \boldsymbol{R}^{L\times n}$. L represents the number of bands and n represents the number of pixels, $n = h \times w$. The h and w are the length and width of the HSI, respectively. $\boldsymbol{E}$ represents the EM matrix, $\boldsymbol{E} \in \boldsymbol{R}^{L\times p}$. $\boldsymbol{A}$ represents the abundance matrix, $\boldsymbol{A} \in R^{p\times n}$. $\boldsymbol{\xi}$ is the noise matrix. Normally, the influence of noise will be ignored, that is, $\boldsymbol{X} \approx \boldsymbol{E}\boldsymbol{A}$.

In actual scenarios, the reflectivity of EMs in each band is non-negative, so the EMs need to satisfy the non-negative constraint (ENC). In addition, abundance needs to meet two conditions: 1) abundance nonnegativity constraint (ANC), each abundance value is non-negative; and 2) abundance sum-to-one constraint (ASC), that is, the sum of the abundance values of each EM is equal to $\mathbf{1}$.

The above unmixing problem can be achieved using an AE. An AE network consists of two parts: an encoder and a decoder.

The encoder $f(\boldsymbol{X})$ encodes the input data $\boldsymbol{X}$ into a low-dimensional representation, and the output of the hidden layer of the encoder is this low-dimensional representation, i.e., abundance $\boldsymbol{A}$

$$\boldsymbol{A} = f(\boldsymbol{X}) = \boldsymbol{\sigma}(\boldsymbol{W}_1\boldsymbol{X}) \tag{2}$$

where $\boldsymbol{\sigma}(x)$ is the activation function of the hidden layer, and $\boldsymbol{W_1}$ is the weight of the hidden layer connecting the input layer. For a stacked AE network, i.e., an encoder structure containing n hidden layers, the encoder is represented as follows:

$$\boldsymbol{A} = f(\boldsymbol{X}) = \sigma_n(\boldsymbol{W}_n \cdots \sigma_2(\boldsymbol{W}_2\sigma_1(\boldsymbol{W}_1\boldsymbol{X}))). \tag{3}$$

The decoder $g(\boldsymbol{A})$ decompresses the encoded vector to reconstruct the original data $\hat{\boldsymbol{X}}$, i.e.,

$$\hat{\boldsymbol{X}} = g(\boldsymbol{A}) = \boldsymbol{E}\boldsymbol{A} \tag{4}$$

where $\boldsymbol{E}$ is the decoder weight matrix representing the weights connecting the hidden layer and output layer and $\hat{\boldsymbol{X}}$ denotes the reconstructed data.

B. Proposed Network

The proposed network, which is called SSABN, comprises two branches: spatial information extraction branch and spectral information extraction branch. The design details of the network are as follows.

1) Spatial Self-Attention Mechanism Module: In the branch of spatial information extraction, we incorporate the core component of the Swin Transformer [14], known as the STB. The STB is a standard multihead self-attention mechanism modified with a shifted window. This modification leads to a significant reduction in computational complexity for the multihead self-attention mechanism. The STB structure has been presented in Fig. 1, composed of window multihead self-attention (W-MSA) and shifted window MSA (SW-MSA) combination. First, the feature $\mathbf{z}^{l-1}$ is normalized via layer of normalization (LN), and then it is learned using W-MSA. The output $\hat{\mathbf{z}}^l$ is obtained from a residual operation, followed by another LN, a multilayer perceptron (MLP), and a residual operation to produce the output feature $\mathbf{z}^l$ of that layer. The SW-MSA layer exhibits a similar structure to that of the W-MSA layer, with the computation of the feature segment assigned to SW-MSA and W-MSA, respectively. W-MSA measures the multihead self-attention within the window, i.e., uses the window as a discrete global region for computing the two-by-two attention of each token in the window. Whereas, SW-MSA shifts the window span to measure the attention between windows. The self-attention is calculated as follows:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{SoftMax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}} + \mathbf{B}\right)\mathbf{V} \tag{5}$$

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \boldsymbol{R}^{v\times d}$ denote the query, key, and value matrices. v and d are the numbers of patches in a window and the dimension of the query or key, respectively. And the values in $\mathbf{B}$ come from the bias matrix.

2) Spectral Channel Attention Mechanism: In the spectral information extraction branch, a simplified version of the channel self-attention mechanism is designed, which inputs $\boldsymbol{Y} \in \boldsymbol{R}^{h\times w\times c}$ to the spectral channel attention mechanism (SCAM) and directly calculates the spectral self-attention weight distribution Attn $\in \boldsymbol{R}^{c\times c}$ from the feature map. First, reshape $\boldsymbol{Y}$ to $R^{n\times c}$, then do matrix multiplication between $\boldsymbol{Y}$ and the transpose of $\boldsymbol{Y}$, and finally, use the normalization layer to obtain the spectral self-attention weight distribution. The specific formula is as follows:

$$\text{Attn}_{ji} = \frac{\exp\left(\boldsymbol{Y}_i \cdot \boldsymbol{Y}_j\right)}{\sqrt{c}\sum_{i=1}^{C}\exp\left(\boldsymbol{Y}_i \cdot \boldsymbol{Y}_j\right)} \tag{6}$$

where Attn_{ji} calculates the influence of the ith spectrum on the jth spectrum. Then do a matrix multiplication between Attn and the transpose of $\boldsymbol{Y}$, and reshape the result to $\boldsymbol{R}^{h\times w\times c}$. The SCAM models the long-term dependencies between feature maps, which can effectively capture the correlation features between spectra.

3) Encoder and Decoder: The encoder's specific structure is depicted in Fig. 1. The spatial information feature extraction branch utilizes a 3×3 convolution to extract the local spatial information while reducing the dataset's dimensionality.

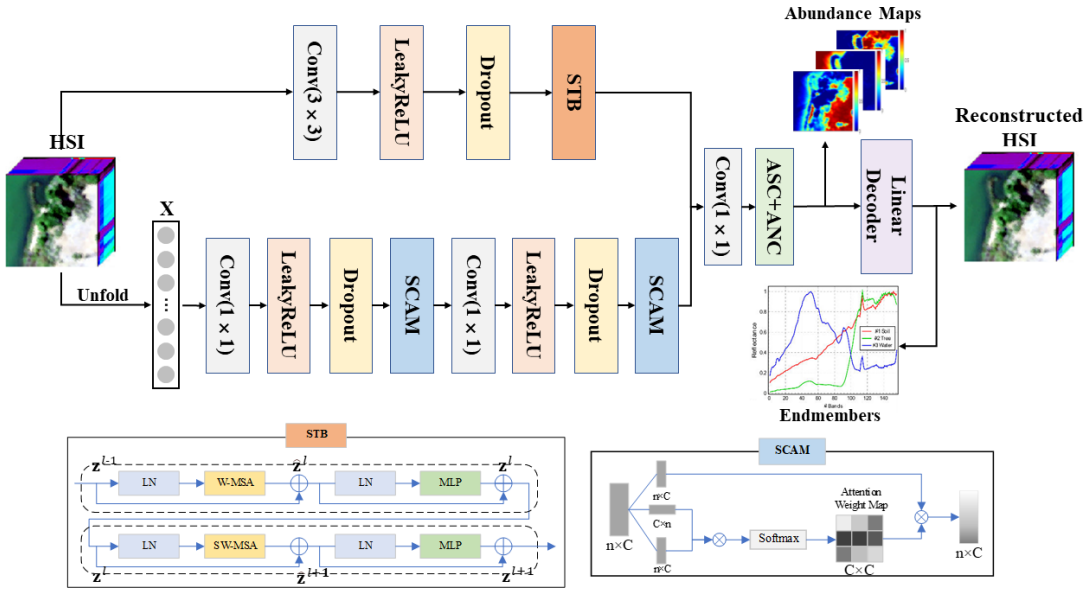

Fig. 1. SSABN, with the STB and SCAM.

LeakyReLU is employed as the activation function. To prevent overfitting, Dropout layers are integrated into the network to eliminate certain neurons, which enhances the network's generalization ability for diverse datasets. In the spectral information extraction branch, a 1 × 1 convolution is implemented to reduce the spectral channel dimensions. Subsequently, the SCAM is utilized to acquire the attention features for the spectral dimension. To extract more detailed spectral features, dimensionality reduction and channel attention are employed twice. Finally, a 1 × 1 convolutional layer is incorporated to fuse spatial and spectral features. This substitution of a fully connected layer with a convolutional layer is intended to minimize the parameter count. To satisfy the ASC and ANC, a softmax layer operates on the features to produce the abundance result

$$\boldsymbol{A} = \text{Softmax}(\boldsymbol{h}). \tag{7}$$

The decoder is a simple linear layer, and we add a non-negativity constraint to satisfy the ENC.

4) Loss Function: Using the mean squared error (MSE) as a loss function to evaluate network training may result in features that lack practical significance. Therefore, the spectral angular distance (SAD) is used as the loss function for our network

$$\boldsymbol{Loss} = \text{Arccos}\left(\frac{\boldsymbol{X}^T\hat{\boldsymbol{X}}}{||\boldsymbol{X}||_2||\hat{\boldsymbol{X}}||_2}\right). \tag{8}$$

III. Experiments

In this section, we present the evaluation of our proposed network using both synthetic and real datasets. Six algorithms are selected as comparison methods, namely, vertex component analysis (VCA) [3], $L_{1/2}$ sparsity-constrained non-negative matrix factorization ($L_{1/2}$-NMF) [16], deep AE unmixing network (DAEU) [17], an untied denoising autoencoder (uDAS) [8], convolutional AE unmixing network (CNNAEU) [11], and convolutional AE combined with a transformer AE unmixing network (TAEU) [18]. Of the aforementioned methods, VCA is a conventional EM extraction algorithm, and the abundance estimation algorithm that we employ is fully constrained least squares (FCLSs) [5]. While $L_{1/2}$-NMF is a blind decomposition algorithm that relies on NMF, DAEU, and uDAS to represent AE methods based on pixel training. Last, CNN and TAEU denote AE methods founded on the learning of spatial information.

A. Experiment Settings

The parameters set in the network are as follows: the number of self-attention heads for the STB is 5, and the window size is 4. The spatial information extraction branch employs a convolutional layer with 40 filters, while the spectral information extraction branch uses two convolutional layers with 40 and 16 filters, respectively. We repeatedly stack each data 20 times into the network. That is, the patch size of each experimental dataset corresponds to the size of the original image data, and the training set comprises 20 patches. The batch size of the network is set to 2. The primary adjusting parameters for the network are the learning rate and the number of epochs. In addition, SAD and the root mean square error (RMSE) are used as evaluation indicators of EMs and abundances.

B. Synthetic Dataset Experiments

The production method of the simulated dataset follows [19]. The synthetic dataset is comprised of 60 × 60 pixels, and five distinct EMs sourced from the ASTER spectral library [20]. The dataset comprises 200 bands over the band range of 0.4–14 μm. One material is selected as the background, and the remaining four materials are assigned to each of the dataset's corners. The abundance values are generated following the Dirichlet distribution. Fig. 2(a) and (b) illustrate the synthesized dataset and its corresponding EMs. To test the

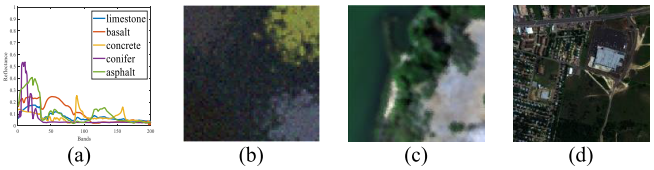

Fig. 2. (a) EMs of the synthetic dataset. (b) Synthetic dataset. (c) Samson dataset. (d) Urban dataset.

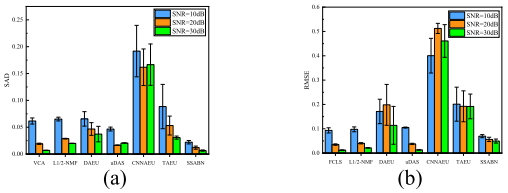

Fig. 3. Experimental results with different SNR. (a) mSAD. (b) RMSE.

TABLE I

SAD, MSAD, AND RMSE OF DIFFERENT UNMIXING METHODS ON THE SAMSON DATASET

	Soil	Tree	Water	mSAD	RMSE
VCA-FCLS	0.0236± 0.00%	0.0426± 0.14%	0.3202± 33.65%	0.1288± 11.23%	0.3262± 0.60%
$L_{1/2}$-NMF	0.0286± 0.23%	0.0596± 0.63%	0.2322± 0.68%	0.1068± 0.10%	0.3214± 0.22%
DAEU	0.0650± 4.21%	0.0461± 1.05%	0.0440± 1.25%	0.0517± 1.88%	0.1510± 5.38%
uDAS	0.0312± 0.07%	0.0534± 0.04%	0.1357± 0.35%	0.0734± 0.10%	0.3161± 0.21%
CNNAEU	0.0588± 0.56%	0.0391± 0.04%	0.1426± 1.81%	0.0802± 0.54%	0.2909± 8.04%
TAEU	**0.0157± 0.46%**	0.0456± 0.58%	0.0693± 2.11%	0.0436± 0.71%	0.1905± 1.21%
SSABN	0.0181± 1.01%	**0.0364± 0.29%**	**0.0452± 0.74%**	**0.0332± 0.20%**	**0.0744± 1.25%**

robustness of the proposed method to noise, additive noises with progressively decreasing signal-to-noise ratios (SNRs) of 30, 20, and 10 dB are introduced into the data.

A learning rate of 0.001 and an epoch of 100 are used for the simulated data. Each algorithm is repeated five times with different SNR data, and its mean and standard deviation are plotted in Fig. 3. The SSABN algorithm is more robust to noise compared to the other algorithms, with the results obtained from the mean SAD (mSAD) more favorable for different noise data. When the SNR value is 10 dB, the RMSE of abundance is optimal and outperforms other methods based on pixel unmixing. At other times, SSABN outperforms networks that consider spatial information.

C. Real Datasets Experiments

1) Samson Dataset: The image size is 95 × 95, with 156 channels per pixel covering wavelengths from 401 to 889 nm. It consists of three ground objects, namely, "#1 soil," "#2 tree," and "#3 water," as illustrated in Fig. 2(c).

For the Samson dataset, the learning rate and epoch are set to 0.002 and 50, respectively. The experimental findings for the Samson dataset are shown in Table I. SSABN distinguishes the water and road EMs most prominently, as illustrated in Fig. 4. Both the mSAD and RMSE values obtained by SSABN are the most optimal. Although the standard deviation

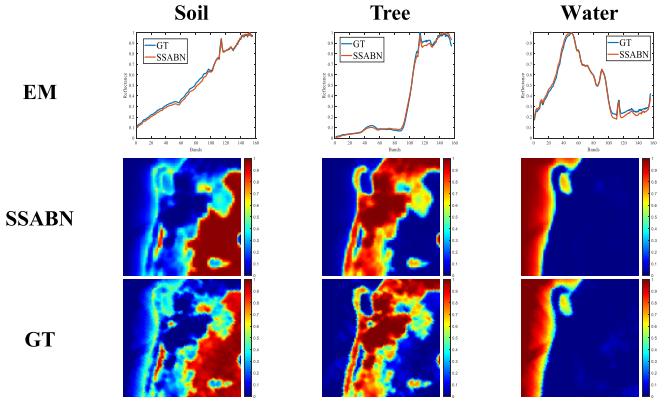

Fig. 4. (Top to Bottom) Real EMs and SSABN extracted EMs, the estimated abundance, and ground truth (GT) abundances.

TABLE II

MSAD AND RMSE OF ABLATION EXPERIMENT ON THREE DATASETS

	Simulated data		Samson dataset		Urban dataset	
	mSAD	RMSE	mSAD	RMSE	mSAD	RMSE
SSMM	0.0620	0.1705	0.2491	0.3100	0.0731	0.1442
SCAM	0.5032	0.3790	0.5112	0.3626	0.2718	0.3160
CBAM	0.7943	0.3527	0.5275	0.3318	0.6340	0.3162
SSMM+SCAM	**0.0060**	**0.0485**	**0.0276**	**0.0827**	**0.0540**	**0.1253**

is slightly poor, it still remains in the suboptimal range. Additionally, the estimated abundance map obtained by the proposed method is noticeably similar to the real abundance values, which confirms that the proposed network accurately identifies ground objects.

2) Urban Dataset: The size of the image is 307 × 307, and it contains 162 bands. The prominent feature categories of these bands include "#1 road," "#2 roof," "#3 grass," and "#4 tree," which are displayed in Fig. 2(d).

The learning rate set for the Urban dataset is 0.006, and 50 epochs are trained. The best results among all the algorithms in the five running trials are selected for comparative analysis. Fig. 5 illustrates various algorithms' estimated abundance plots for the Urban data, revealing that our proposed method yields sharper results when compared to others. As is evidenced in Fig. 5, SSABN could achieve the lowest mSAD and RMSE results simultaneously. This indicates the superiority of our proposed method over other competing methods while showcasing the efficacy of spatial self-attention and spectral self-attention in the bilateral net.

D. Ablation Experiment

The convolutional block attention module (CBAM) [21] is a module that incorporates a convolutional attention mechanism. We conduct separate experiments on the single-stream network and compared it with both the dual-stream network and the network incorporating CBAM. The experimental results for the three datasets are presented in Table II. Among these, the simulated data refers to data with an SNR of 30 dB. The table clearly demonstrates that the SCAM module network, which solely considers spectral information, exhibits poor unmixing results. Conversely, the spatial self-attention mechanism module (SSMM), which primarily considers spatial information,

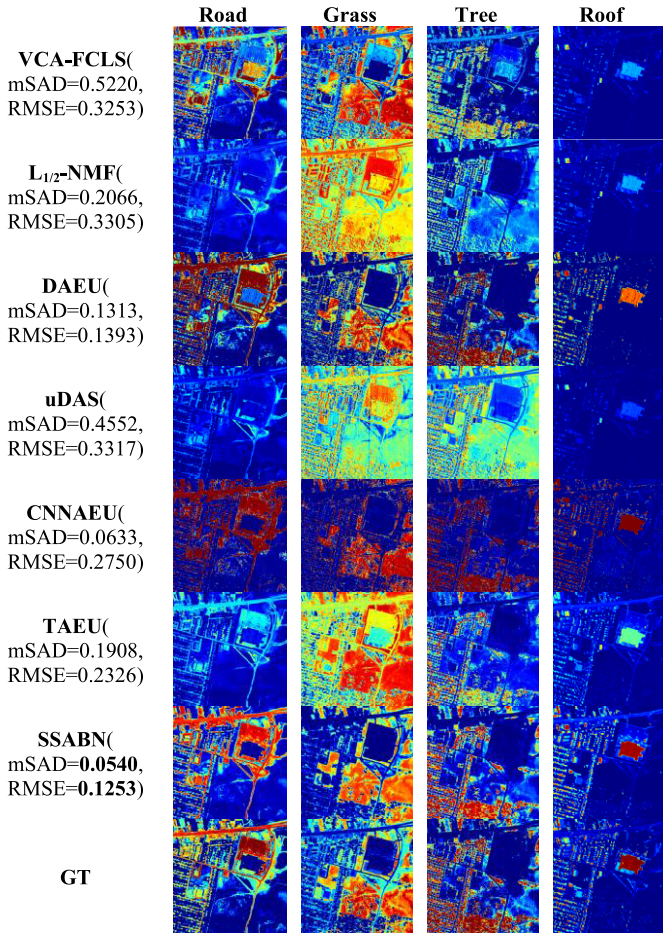

Fig. 5. Abundance maps of Urban dataset. (Top to Bottom) VCA-FCLS, $L_{1/2}$-NMF, DAEU, uDAS, CNNAEU, TAEU, and the proposed network.

achieves relatively good unmixing performance. Combining these two modules yields an optimal outcome as the SCAM module compensates for the limitations of SSMM in capturing weak spectral information. Although the CBAM network can simultaneously learn the spatial and spectral features of an image, its ability to effectively fuse these two types of information is inadequate, resulting in a subpar unmixing effect.

IV. Conclusion

In this letter, a spatial-spectral attention-based bilateral AE network for unmixing was proposed. By utilizing the self-attention mechanism, the network can obtain crucial non-local data features. The spatial attention is extracted from a multi-head self-attention module based on shifted windows, while the spectral attention is obtained using a simplified channel attention mechanism. This network simultaneously extracts both spatial and spectral features and effectively fuses them. Our proposed method's validity was confirmed through experimentation using simulated and real data as well as ablation experiment.

References

[1] J. Chen, M. Zhao, X. Wang, C. Richard, and S. Rahardja, "Integration of physics-based and data-driven models for hyperspectral image unmixing," 2022, *arXiv:2206.05508*.

[2] W.-K. Ma et al., "A signal processing perspective on hyperspectral unmixing: Insights from remote sensing," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 67–81, Jan. 2014.

[3] J. M. P. Nascimento and J. M. B. 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.

[4] D. Shah, T. Zaveri, Y. N. Trivedi, and A. Plaza, "Entropy-based convex set optimization for spatial–spectral endmember extraction from hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4200–4213, 2020.

[5] D. C. Heinz and C.-I. Chang, "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 3, pp. 529–545, Mar. 2001.

[6] R. Feng, Y. Zhong, and L. Zhang, "An improved nonlocal sparse unmixing algorithm for hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 915–919, Apr. 2015.

[7] Z. Li, R. Feng, L. Wang, and T. Zeng, "Spectral-spatial-sparse unmixing with superpixel-oriented graph Laplacian," *Int. J. Remote Sens.*, vol. 44, no. 8, pp. 2573–2589, Apr. 2023.

[8] 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.

[9] Y. Su, A. Marinoni, J. Li, J. Plaza, and P. Gamba, "Stacked nonnegative sparse autoencoders for robust hyperspectral unmixing," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 9, pp. 1427–1431, Sep. 2018.

[10] 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.

[11] 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.

[12] M. Zhao, S. Shi, J. Chen, and N. Dobigeon, "A 3-D-CNN framework for hyperspectral unmixing with spectral variability," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5521914.

[13] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[14] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[15] D. He, Q. Shi, X. Liu, Y. Zhong, and X. Zhang, "Deep subpixel mapping based on semantic information modulated network for urban land use mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10628–10646, Dec. 2021.

[16] 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.

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

[19] Q. Jin, Y. Ma, X. Mei, and J. Ma, "TANet: An unsupervised two-stream autoencoder network for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5506215.

[20] 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, Jan. 2002.

[21] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," 2018, *arXiv:1807.06521*.