A Spectral-Spatial Attention Network for Hyperspectral Unmixing

Xuanwen Tao, *Member, IEEE*, Bikram Koirala, *Member, IEEE*, Behnood Rasti, *Senior Member, IEEE*, Antonio Plaza, *Fellow, IEEE*, and Paul Scheunders, *Senior Member, IEEE*.

***Abstract*—Hyperspectral unmixing, an essential and fundamental task in remote sensing, focuses on estimating endmembers (spectrally pure components) and their fractional abundances within each mixed pixel of a hyperspectral image. With the advent of deep learning (DL), the field of hyperspectral unmixing has made significant progress. Among DL approaches, autoencoder-based models have shown promising results. However, most unmixing methods estimate the endmembers by the weights of the linear layers in the decoder of their networks, making their performance highly dependent on weight initialization. Moreover, noise is not explicitly accounted for in most recent methods that use spectral angle distance (SAD) loss. To avoid the initialization problems, we developed an innovative inversion strategy to directly estimate the endmembers. Moreover, to optimally account for noise, an end-to-end network is proposed that integrates both denoising and unmixing. Finally, for an improved feature extraction, a novel spectral-spatial attention module is integrated in the network. Extensive experiments on a synthetic and three real datasets show that the proposed method significantly and consistently outperforms the compared state-of-the-art methods. The full code is available at https://github.com/xuanwentao for public evaluation.**

***Index Terms*—Hyperspectral unmixing, autonomous unmixing module, autoencoder, deep denoising module, attention module.**

I. Introduction

HYPERSPECTRAL images provide rich spectral information and intricate spatial characteristics of target scenes, making them indispensable in a wide range of applications, including environmental monitoring, mineral exploration, and precision agriculture [1]–[4]. These images capture detailed spectral signatures comprising hundreds of contiguous spectral bands, allowing an accurate identification of materials and substances in the scene.

However, one of the inherent challenges with hyperspectral images is their low spatial resolution. A single pixel in a hyperspectral image can cover a large physical area, often ranging from tens to thousands of square meters. This large footprint leads to mixed pixels, where multiple materials are captured within a single pixel, resulting in a blurred representation of the actual scene. This spatial ambiguity, where distinct materials are blended within individual pixels, poses a significant limitation for many applications that require high spatial precision. To overcome this critical obstacle, hyperspectral unmixing techniques [5] have been developed. These methods aim to solve the mixed pixel problem by decomposing each pixel into a set of spectrally pure components, known as endmembers, and estimating their corresponding proportions, referred to as abundances.

The most widely used mixing model is the linear mixing model (LMM) [6]–[9]. LMM operates under the assumption that each incident light ray interacts with a single pure material within the pixel's instantaneous field of view before reaching the sensor, allowing the spectrum of each pixel to be expressed as a linear combination of a set of endmembers and their corresponding abundances.

A. Traditional Unmixing Methods

Traditional hyperspectral unmixing techniques that solve the LMM can be broadly categorized into three groups: geometry-based, statistical-based, and sparse regression-based methods.

Geometry-based unmixing methods operate under the assumption that the observed spectral vectors lie within a simplex, with vertices corresponding to the endmembers, ensuring that the resulting abundances are non-negative and sum to one to provide a physically interpretable solution. Two main categories exist: pure pixel-based methods [10]–[12] and minimum volume-based methods [13]–[15]. Pure pixel-based methods such as N-FINDR, vertex component analysis (VCA), and maximum distance analysis (MDA) [10], [11], [16] assume the existence of at least one pure pixel for each endmember in the image. When no pure pixels are present, minimum volume-based methods such as robust collaborative nonnegative matrix factorization (R-CoNMF) [17] and minimum volume simplex analysis (MVSA) [13] aim to estimate the endmembers as vertices of a simplex with minimal volume. Statistical methods describe the unmixing problem as an inference task, often within a Bayesian framework [18]–[20]. These techniques model the uncertainty of both endmembers and abundances, allowing for a more flexible and probabilistic interpretation of the results. Sparse regression-based methods tackle hyperspectral unmixing by selecting endmember candidates from a spectral library [21]–[24]. These methods exploit the fact that, in many cases, only a small subset of materials from the spectral library are present in each pixel, leading to sparse abundance representations.

Xuanwen Tao (corresponding author), Bikram Koirala and Paul Scheunders are with the Imec-VisionLab, Department of Physics, University of Antwerp, Belgium (email: taoxuanwenupc@gmail.com; bikram.koirala@uantwerpen.be; paul.scheunders@uantwerpen.be).

Behnood Rasti is with the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, 10587 Berlin, Germany (e-mail: behnood.rasti@gmail.com).

Antonio Plaza is with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, University of Extremadura, 10071 Cáceres, Spain (e-mail: aplaza@unex.es).

B. DL-based Unmixing Methods

Traditional unmixing algorithms require manual parameter setting or rely on strict mathematical assumptions. Recently, neural networks [25]–[29] have been widely applied to hyperspectral image analysis due to their powerful generalization and modeling capabilities, which have been significantly improved by advances in deep learning (DL) technology. DL leverages the ability of neural networks to learn complex representations, making them particularly suitable for tasks such as hyperspectral unmixing. DL-based unmixing methods have increasingly become the mainstream approach for addressing the mixed pixel problem in hyperspectral remote sensing [30]–[32].

A notable unsupervised approach that has attracted attention is the autoencoder framework. Autoencoders are designed with an encoder-decoder architecture, where the encoder reduces the dimensionality of the input data and the decoder reconstructs the original data from the learned representations. This has proven successful in hyperspectral unmixing by automatically learning low-dimensional representations (i.e., the abundances) and reconstructing the original data using a corresponding basis (i.e., the endmembers), and many autoencoder-based unmixing networks have been proposed in the literature, such as deep autoencoder unmixing (DAEU) [33], multitask autoencoder unmixing (MTAEU) [34], spectral information divergence autoencoder unmixing (SIDAEU) [35], adaptive abundance smoothing (AAS) [36], and deep autoencoder network (DAN) [37]. However, most autoencoder-based approaches tend to focus solely on capturing spectral information and overlook the spatial relationships between different pixels. This limitation may hinder the full exploitation of spatial–contextual information, which is crucial for accurate unmixing in scenarios where spatial continuity or homogeneity plays a significant role.

Recently, convolutional neural networks (CNNs) have shown significant progress in hyperspectral image processing by incorporating both 2D or 3D convolutions to effectively learn spatial and spectral information simultaneously. The ability of CNNs to capture both local spatial patterns and spectral features has stimulated the development of many innovative unmixing methods [30], [31], [38]–[42]. In particular, convolutional neural network autoencoder unmixing (CNNAEU) [39] exploited the inherent spatial and spectral structures of hyperspectral images to estimate both endmembers and abundances. Unlike traditional CNN architectures that rely on pooling or upsampling layers, CNNAEU processes image patches directly, preserving the spatial structure throughout the network. This allows abundance maps to be generated as feature maps from hidden convolutional layers, preserving detailed spatial relationships during unmixing. The cross-convolution unmixing network (CrossCUN) [30] is a hybrid approach, combining both 3D and 2D convolutions for a more efficient feature extraction. The multi-stage convolutional autoencoder network (MSNet) [31] employs a multi-stage approach to acquire long-range dependencies and high-level contextual information without losing fine spatial details. The network consists of three sub-stages of CNN autoencoders. The first two stages extract broad information and capture long-range dependencies through multi-stage downsampling operations, while the final stage operates at the original resolution to preserve fine spatial details. Although CNN-based unmixing methods have shown significant improvements, the limited receptive field of the traditional CNN limits its ability to capture the global distribution of features across the entire scene.

Very recently, inspired by the remarkable ability of the attention mechanism to capture long-range dependencies in images [43]–[45], several techniques have been proposed that use attention to improve the extraction of spatial and spectral features from hyperspectral data for hyperspectral unmixing [46]–[52]. In [46], a new attention mechanism called multi-head self-patch attention was proposed that employs the long-range dependencies between patches for performing hyperspectral unmixing. In [51], a unidirectional local-attention autoencoder network (ULA-Net) was introduced to effectively capture and utilize discriminative local spatial–spectral information while addressing spectral variability. ULA-Net successfully addresses the challenges of high computational complexity and slow convergence often encountered when dealing with global information. In [53], the authors proposed a graph attention convolutional autoencoder (GACAE), which employs graph attention convolution to learn the global spatial relationships within hyperspectral data for unsupervised unmixing.

Besides linear unmixing, much work has been devoted to the problem of nonlinear unmixing. Nonlinear spectral unmixing assumes that the incident light ray undergoes multiple scattering before reaching the sensor. Traditional [54]–[56] as well as deep learning methods [57]–[60] have been proposed. Because the scope of the current work is on linear unmixing, we will not further elaborate on the problem of nonliear unmixing.

C. Motivations and Contributions

Despite the advances made by autoencoders, CNNs, and attention-based unmixing methods, they face two primary challenges: (1) Most decoders estimate the endmembers as weights of a linear layer. To initialize these weights, many methods still rely on techniques such as VCA. However, the endmembers estimated by VCA are non-unique, as each run produces a different set of endmembers. Because the networks initialize their decoder weights with VCA's output, this randomness impacts the final unmixing performance. (2) The use of SAD loss rather than mean squared error loss has become increasingly common in recent deep-learning-based unmixing methods because it can address scaling spectral variability. However, SAD loss is ineffective in handling Gaussian noise and previous deep-learning-based unmixing methods have not explicitly accounted for noise. Although some mature hyperspectral image denoising algorithms have been proposed [61], [62], applying denoising as a preprocessing step may constrain unmixing performance. Optimal unmixing performance requires an end-to-end network that integrates both denoising and unmixing.

To tackle these two challenges, we propose the spectral-spatial attention unmixing network (SSAUN) for hyperspectral

unmixing. The key innovative contributions of this work can be summarized as follows:

- We propose an autonomous unmixing module (AUM) that includes a novel approach to estimate the endmembers using an inversion strategy, unlike existing methods that obtain endmembers from the weights of the final linear layer of the network. This direct estimation of the endmembers from the network avoids the randomness associated with other methods.
- We propose a deep denoising module (DDM) to generate a clean hyperspectral image as input to the AUM. This is necessary for the inversion strategy and to justify the use of SAD loss. To the best of our knowledge, this is the first attempt in the literature for an end-to-end network that integrates both denoising and unmixing.
- We propose a spectral-spatial attention module (SSAM) within the AUM that improves the quality of the estimated fractional abundances, which requires capturing both long-distance vertical and horizontal spatial dependencies, as well as dependencies across spectral bands to further enhance the unmixing performance.

The remainder of this article is structured as follows. Section II describes the related work. Section III elaborates the proposed SSAUN method. Section IV evaluates the proposed method on synthetic and real hyperspectral datasets. Section V provides relevant discussions, evaluating the effect of the proposed strategy. Finally, the conclusion is given in Section VI.

II. Related Work

A. Linear Mixture Model (LMM)

In linear hyperspectral unmixing, the observed spectral reflectance at each pixel of a hyperspectral image is assumed to be a linear combination of pure spectral signatures, known as endmembers, and their corresponding proportions, referred to as abundances. This relationship can be formulated as:

$$\mathbf{X} = \mathbf{MA} + \mathbf{N}, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{Nb \times Np}$ represents the hyperspectral image with *Nb* bands and *Np* pixels, $\mathbf{M} \in \mathbb{R}^{Nb \times c}$ is the endmember matrix, in which each column corresponds to the spectral signature of an endmember, and c is the number of endmembers. $\mathbf{A} \in \mathbb{R}^{c \times Np}$ denotes the abundance matrix, in which each element a_{ij} represents the proportion (or abundance) of the *i*-th endmember in the *j*-th pixel. $\mathbf{N}$ is additive noise. Generally, the abundance matrix satisfies two physical constraints, i.e., the abundance nonnegative constraint (ANC) and the abundance sum-to-one constraint (ASC), which can be expressed as:

$$\text{ANC} : a_{ij} \geq 0; \forall i, j, \tag{2}$$

and

$$\text{ASC} : \sum_{i=1}^{c} a_{ij} = 1; \forall j. \tag{3}$$

B. Autoencoder

An autoencoder is a type of unsupervised neural network that aims to learn compact, low-dimensional representations of input data and that is commonly applied to tasks such as dimensionality reduction, feature extraction, and data reconstruction. The fundamental architecture of an autoencoder consists of two primary components: an encoder and a decoder.

The encoder compresses the input data $\mathbf{X}$ into a latent space representation $\mathbf{Y}$, effectively reducing the dimensionality while preserving the most salient information. This process can be mathematically represented as:

$$\mathbf{Y} = f(\mathbf{X}; \mathbf{W}, \mathbf{b}), \tag{4}$$

where f is the activation function of the encoder, typically a non-linear function such as rectified linear unit (ReLU) or sigmoid, and $\mathbf{W}$ and $\mathbf{b}$ represent the learnable weights and biases of the encoder, respectively. The decoder is used to reconstruct the input from the latent representation:

$$\hat{\mathbf{X}} = g(\mathbf{Y}; \tilde{\mathbf{W}}, \tilde{\mathbf{b}}), \tag{5}$$

where g is the activation function of the decoder, and $\tilde{\mathbf{W}}$ and $\tilde{\mathbf{b}}$ are its weights and biases. The objective of the decoder is to generate a reconstruction $\hat{\mathbf{X}}$ that closely approximates the original input $\mathbf{X}$, capturing the underlying structure of the data. The overall learning process of the autoencoder is driven by the minimization of the reconstruction loss, which measures the difference between the input $\mathbf{X}$ and its reconstruction $\hat{\mathbf{X}}$. The loss is typically formulated by the mean squared error (MSE) or by other loss functions such as binary cross-entropy and spectral angle distance, depending on the nature of the data. In summary, the autoencoder learns to map the input data to a compressed representation and to reconstruct it effectively, thus capturing key patterns and dependencies in the data.

C. Self-Attention Mechanism

The self-attention mechanism has emerged as a crucial concept in DL, especially in tasks involving sequential or structured data, such as natural language processing and image analysis. This attention mechanism allows models to focus on specific parts of the input data, effectively capturing both local and global dependencies.

Typically, the input to the self-attention mechanism consists of three main components, i.e., query *Q*, key *K*, and value *V*. These components are derived from the same input sequence, allowing the model to dynamically assess the relationships between different parts of the data. The attention scores are calculated by taking the matrix multiplication of the query and the transpose of the key:

$$\mathbf{Att} = \mathbf{QK}^{\mathrm{T}}, \tag{6}$$

This operation produces a scoring matrix that quantifies how much attention each element of the input data should receive relative to others. To ensure that the attention scores are interpretable and stable, they are typically normalized using the *Softmax* function. This step converts the raw scores into

a probability distribution, effectively emphasizing the most relevant parts of the input:

$$\begin{aligned}\mathbf{Att}_{\text{norm}} &= \mathit{Softmax}(\frac{\mathbf{Att}}{\sqrt{d}}) \\ &= \mathit{Softmax}(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d}}),\end{aligned} \tag{7}$$

where d is the dimension of $\mathbf{Q}$ and $\mathbf{K}$. The final output of the self-attention mechanism is obtained by weighting the value vectors $\mathbf{V}$ with the normalized attention scores:

$$\mathbf{Z} = \mathbf{Att}_{\text{norm}}\mathbf{V} = \mathit{Softmax}(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d}})\mathbf{V}, \tag{8}$$

where $\mathbf{Z}$ represents the final output of the attention layer, capturing the important features of the input while filtering out irrelevant information.

III. Proposed Method

The proposed approach (SSAUN) is an end-to-end network, integrating a novel DDM to filter the noise in the hyperspectral images, and an AUM that automatically estimates the abundances and endmembers and simultaneously completes the image reconstruction. To improve performance, a novel SSAM is embedded in the AUM. Fig. 1 shows the architecture of the proposed SSAUN. The three modules are further elaborated below.

A. Deep Denoising Module

In general, most existing unmixing methods consider the autoencoder (subsection II-B) as the main backbone of the networks to perform abundance estimation and endmember extraction. Noise seriously hinders network robustness and unmixing accuracy, but is usually not taken into account. To this end, we design a DMM to pre-filter noise in hyperspectral images, ensuring cleaner data for further analysis and improving overall unmixing performance. The proposed DDM uses an autoencoder as the core framework, including several encoder-decoder layers that gradually compress and reconstruct the input image while skip connections combine the output of different encoders and decoders to preserve high-level spatial and spectral features across the various network stages. The main advantage of using multiple encoders and decoders is that the module can extract multi-scale features, progressively denoise input images, and enhance robustness by combining spatial and spectral features from multiple encoder-decoder stages. This is particularly important in practical applications, where noise varies in magnitude and characteristics across different datasets.

The input of DDM is a 3D hyperspectral image $\mathbb{X} \in \mathbb{R}^{Nb\times Nx\times Ny}$ with spatial dimensions *Nx* and *Ny* (width and height), and the number of spectral bands *Nb*. A first encoder block transforms the input image to $\mathbb{E}_1 \in \mathbb{R}^{dim\times Nx\times Ny}$:

$$\mathbb{E}_1 = \mathit{ReLU}(\mathit{BN}(\mathit{Conv}(\mathbb{X}))), \tag{9}$$

where Conv denotes the 2D convolutional operation with *dim* kernels of size 3×3 (in our experiment, *dim* = 32), BN stands for batch normalization, and ReLU is the activation function.

Then, a two-times iteration of a combination of a downsampling and an encoder block follow. For $i = 1, 2$:

$$\begin{aligned}\mathbb{F}_i^{(E)} &= \mathit{Downsampling}(\mathit{Conv}(\mathbb{E}_i)) \\ \mathbb{E}_{i+1} &= \mathit{ReLU}(\mathit{BN}(\mathit{Conv}(\mathbb{F}_i^{(E)})))\end{aligned} \tag{10}$$

where downsampling by max-pooling is performed, hereby reducing the spatial dimensions by a factor of 2, and doubling the number of channels (i.e., $2 * i * dim$ filters are applied).

The decoder reverses the operations performed by the encoder, gradually reconstructing the denoised image while restoring the original spatial dimensions. Starting from the deepest encoder output $E_3 = D_3$, a combination of an upsampling and a decoder block is performed to obtain:

$$\begin{aligned}\mathbb{F}_2^{(D)} &= \mathit{Upsampling}(\mathit{Conv}(\mathbb{D}_3)) \\ \mathbb{D}_2 &= \mathit{ReLU}(\mathit{BN}(\mathit{Conv}(\mathbb{E}_2 + \mathbb{F}_2^{(D)}))).\end{aligned} \tag{11}$$

In this case, upsampling restores the spatial dimensions, each time enlarging the spatial dimensions by a factor of 2, and halving the number of channels (i.e., $i{*}dim$ filters are applied). To ensure that fine-grained features from the encoder are preserved during reconstruction, skip connections are applied between the corresponding encoder and decoder layers to obtain $\mathbb{E}_2 + \mathbb{F}_2^{(D)}$. Next, we use similar operations to obtain $\mathbb{F}_1^{(D)}$ and $\mathbb{D}_1$ by:

$$\begin{aligned}\mathbb{F}_1^{(D)} &= \mathit{Upsampling}(\mathit{Conv}(\mathbb{D}_2)) \\ \mathbb{D}_1 &= \mathit{Conv}(\mathbb{E}_1 + \mathbb{F}_1^{(D)}).\end{aligned} \tag{12}$$

After the decoding process, a final 2D convolutional layer is applied to map the features back to the original number of spectral bands *Nb*, and the output is combined with the input $\mathbb{X}$ to produce the final denoised hyperspectral image:

$$\hat{\mathbb{X}}_1 = \mathit{Conv}(\mathbb{D}_1) + \mathbb{X}, \tag{13}$$

where $\hat{\mathbb{X}}_1 \in \mathbb{R}^{Nb\times Nx\times Ny}$. This additive operation ensures noise removal while preserving both the spectral and spatial details from the input hyperspectral data.

B. Autonomous Unmixing Module

The denoised image $\mathbb{A}_0 = \hat{\mathbb{X}}_1$ is the input of the AUM. AUM performs a sequence of blocks, each of them containing our newly developed spectral-spatial attention module SSAM (see subsection III-C) along with a convolutional layer with a decreasing number of kernels. In this work, 4 blocks are applied with respectively 128, 64, 32, and c kernels. The convolution layers also contain batch normalization and ReLU. The output $\mathbb{A}_l$ of the l-th block is given by:

$$\mathbb{A}_l = \mathit{ReLU}(\mathit{BN}(\mathit{Conv}(\mathit{SSAM}(\mathbb{A}_{l-1})))), \tag{14}$$

The final estimated abundance maps $\hat{\mathbb{A}}$ is given by the output $\mathbb{A}_4 \in \mathbb{R}^{c\times Nx\times Ny}$ of layer 4 (with c nodes).

The abundance maps are then reshaped into matrices of size $(c \times (Nx \times Ny))$:

$$\hat{\mathbf{A}} = \mathit{Reshape}(\hat{\mathbb{A}}) \tag{15}$$

where $\hat{\mathbf{A}} \in \mathbb{R}^{c\times Np}$, with $Np = Nx \times Ny$. Similarly, $\hat{\mathbf{X}}_1$ is reshaped:

$$\hat{\mathbf{X}}_1 = \mathit{Reshape}(\hat{\mathbb{X}}_1), \tag{16}$$

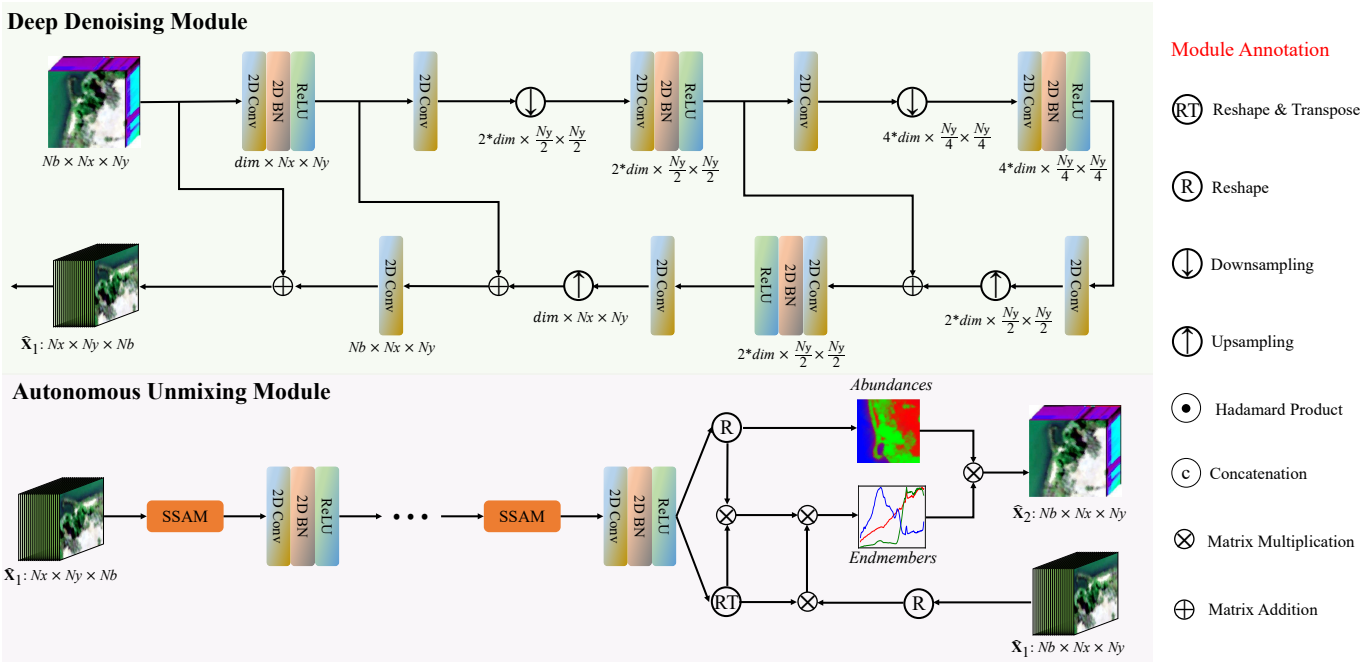

Fig. 1. Architecture of the proposed spectral-spatial attention unmixing network (SSAUN), containing a deep denoising module (DDM) and an autonomous unmixing module (AUM). The designed spectral-spatial attention module (SSAM) is embedded into AUM to enhance vital spectral-spatial features.

where $\hat{\mathbf{X}}_1 \in \mathbb{R}^{Nb \times Np}$.

Under the assumption that the reconstructed image $\hat{\mathbf{X}}_1$ is clean, Eq. (1) becomes:

$$\hat{\mathbf{X}}_1 = \hat{\mathbf{M}}\hat{\mathbf{A}}, \tag{17}$$

where $\hat{\mathbf{M}} \in \mathbb{R}^{Nb \times c}$ is the endmember matrix that needs to be estimated. A direct estimation is obtained by inverting Eq. (17):

$$\hat{\mathbf{M}} = \hat{\mathbf{X}}_1\hat{\mathbf{A}}^{\mathrm{T}}(\hat{\mathbf{A}}\hat{\mathbf{A}}^{\mathrm{T}})^{-1}, \tag{18}$$

where the symbol '-1' denotes a matrix inversion. From the estimated $\hat{\mathbf{A}}$ and $\hat{\mathbf{M}}$, a final reconstructed image $\hat{\mathbf{X}}_2$ can be obtained as:

$$\begin{aligned}\hat{\mathbf{X}}_2 &= \hat{\mathbf{M}}\hat{\mathbf{A}} \\ &= \hat{\mathbf{X}}_1\hat{\mathbf{A}}^{\mathrm{T}}(\hat{\mathbf{A}}\hat{\mathbf{A}}^{\mathrm{T}})^{-1}\hat{\mathbf{A}}.\end{aligned} \tag{19}$$

C. Spectral-Spatial Attention Module

The performance of the unmixing process can be enhanced by improving on the spectral-spatial feature extraction. For this, a spectral-spatial attention module is designed and embedded in the AUM. The module contains a spectral attention module, a spatial attention module and a fusion of both. The architecture of the proposed SSAM is presented in Fig. 2.

1) Spectral Attention Module: The spectral attention focuses on capturing the dependencies between different feature channels. This attention mechanism re-calibrates the input feature maps by focusing on the most relevant channels. Assuming that the feature map of one hidden layer is $\mathbb{F} \in \mathbb{R}^{B \times W \times H}$ with B channels, where W is the width and H is the height. Query, key and value matrices are initially identical and computed as follows:

$$\begin{aligned}\mathbf{Q} &= Reshape(\mathbb{F}) \in \mathbb{R}^{B \times WH}; \\ \mathbf{K} &= Reshape(\mathbb{F}) \in \mathbb{R}^{B \times WH}; \\ \mathbf{V} &= Reshape(\mathbb{F}) \in \mathbb{R}^{B \times WH},\end{aligned} \tag{20}$$

The matrices $\mathbf{Q}$ and $\mathbf{K}$ are used to compute an attention map $\mathbf{A}_{Spe} \in \mathbb{R}^{B \times B}$ as follows:

$$\mathbf{A}_{Spe} = Softmax(\mathbf{Q}\mathbf{K}^{\mathrm{T}}), \tag{21}$$

where T denotes the transpose operation. $\mathbf{A}_{Spe}$ represents the attention weights across channels, capturing the relationships between different feature channels in the input. Next, channel attention is applied to the attention map $\mathbf{A}_{Spe}$ by multiplying it with the value matrix V to produce the feature map $\mathbf{F}_{Spe} \in \mathbb{R}^{B \times WH}$:

$$\mathbf{F}_{Spe} = \mathbf{A}_{Spe}\mathbf{V}, \tag{22}$$

Finally, the output of the spectral attention module is obtained by reshaping $\mathbf{F}_{Spe}$ back to its original spatial dimensions and adding it to the original input feature map:

$$\mathbb{F}_{Spe} = \alpha Reshape(\mathbf{F}_{Spe}) + \mathbb{F}, \tag{23}$$

where $\mathbb{F}_{Spe} \in \mathbb{R}^{B \times W \times H}$ and α is a learnable scaling parameter, initialized to zero.

2) Dual Spatial Attention Module: Spatial attention mechanisms play a crucial role in enhancing the performance of neural networks by focusing on the most informative parts of an input feature map. To capture a comprehensive representation of the input features, in this work, we introduce a dual spatial attention module that effectively utilizes both maxpooling and meanpooling, identifying the most salient features and the

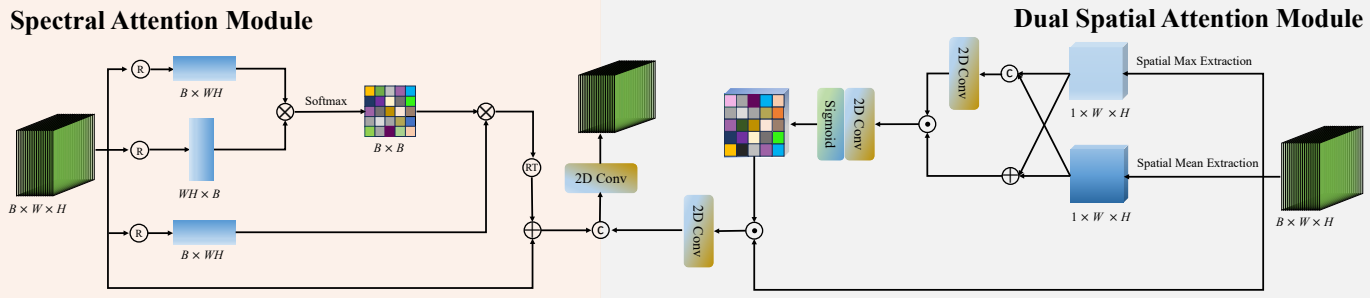

Fig. 2. Architecture of the SSAM, containing a spectral attention module and an dual spatial attention module.

average contextual information, respectively across the spatial domain.

Let $\mathbb{F} \in \mathbb{R}^{B\times W\times H}$ be the input feature map from a hidden layer. Maxpooling and meanpooling are performed along the channel dimension to distill the feature map into two distinct forms: $\mathbb{F}_{max}$, capturing the most pronounced features by computing the maximum value across the channels and $\mathbb{F}_{mean}$, providing an average overview of the channels, thereby preserving the contextual information:

$$\begin{aligned}\mathbb{F}_{max}(1,W,H) &= \underset{B}{Maxpooling}(\mathbb{F}(B,W,H));\\ \mathbb{F}_{mean}(1,W,H) &= \underset{B}{Meanpooling}(\mathbb{F}(B,W,H));\end{aligned} \tag{24}$$

The module then synthesizes $\mathbb{F}_{max} \in \mathbb{R}^{1\times W\times H}$ and $\mathbb{F}_{mean} \in \mathbb{R}^{1\times W\times H}$, using both concatenation and summation:

$$\begin{aligned}\mathbb{A}_{cat} &= Conv(Concatenate(\mathbb{F}_{max},\mathbb{F}_{mean}));\\ \mathbb{A}_{sum} &= Sum(\mathbb{F}_{max},\mathbb{F}_{mean}),\end{aligned} \tag{25}$$

with $\mathbb{A}_{cat} \in \mathbb{R}^{1\times W\times H}$ (a 1×1 convolution was applied to retain the dimensionality) and $\mathbb{A}_{sum} \in \mathbb{R}^{1\times W\times H}$. This dual approach forms robust spatial attention signals that integrate the distinct characteristics of the max and mean pooled features. These maps are then multiplied in element-wise fashion, and a 1×1 convolution followed by a sigmoid function is applied to bound the attention weights between 0 and 1:

$$\mathbb{A}_{Spa} = Sigmoid(Conv(\mathbb{A}_{cat}\odot\mathbb{A}_{sum})), \tag{26}$$

where $\odot$ denotes the Hadamard Product. The refined attention map $\mathbb{A}_{Spa} \in \mathbb{R}^{1\times W\times H}$ is then used to modulate the input feature map, enhancing crucial areas while subduing less relevant regions:

$$\mathbb{F}_{Spa} = \mathbb{A}_{Spa}\odot\mathbb{F}. \tag{27}$$

where $\mathbb{F}_{Spa} \in \mathbb{R}^{B\times W\times H}$ has the same dimensions as the input. This representation not only highlights significant features but also maintains a balance with the contextual background.

3) Attention Fusion: Since both spectral and spatial information play vital roles in capturing the characteristics of a scene, the spectral-attention feature map $\mathbb{F}_{Spe} \in \mathbb{R}^{B\times W\times H}$ and the spatial-attention feature map $\mathbb{F}_{Spa} \in \mathbb{R}^{B\times W\times H}$ are concatenated along the channel dimension. To reduce the number of channels back to B, a 1×1 convolution is applied to the concatenated feature map:

$$\mathbb{F}_{SSAM} = Conv(Concatenate(\mathbb{F}_{Spe},\mathbb{F}_{Spa})), \tag{28}$$

where $\mathbb{F}_{SSAM} \in \mathbb{R}^{B\times W\times H}$ is the final feature map, enriched with information from both spectral and spatial attention mechanisms. This SSAM module is embedded in each of the layers of the AUM to enhance the spectral-spatial feature representation.

D. Loss Function

The proposed SSAUN is an end-to-end unmixing network that combines a DDM for noise reduction and an AUM for unmixing and reconstruction of the hyperspectral image. To optimize the network, a loss function is designed that combines a reconstruction loss, a denoising loss, and an abundance loss. These losses guide the network in the tasks of learning to denoise the input data, accurately unmixing the spectral signatures, and reconstructing the hyperspectral images with improved precision. The reconstruction loss is computed by:

$$L_R = \frac{1}{Np}\sum_{j=1}^{Np}\arccos\frac{\mathbf{x}_j\cdot\hat{\mathbf{x}}_{2j}}{||\mathbf{x}_j||_2||\hat{\mathbf{x}}_{2j}||_2}, \tag{29}$$

where $\mathbf{x}_j \in \mathbb{R}^{Nb\times1}$ denotes the original hyperspectral pixel, $\hat{\mathbf{x}}_{2j} \in \mathbb{R}^{Nb\times1}$ is the reconstructed pixel from the AUM, and $||\cdot||_2$ denotes the L_2 norm. To guide the denoising module DDM, the denoising loss is defined as:

$$L_D = \frac{1}{Np}\sum_{j=1}^{Np}\arccos\frac{\mathbf{x}_j\cdot\hat{\mathbf{x}}_{1j}}{||\mathbf{x}_j||_2||\hat{\mathbf{x}}_{1j}||_2}, \tag{30}$$

where $\hat{\mathbf{x}}_{1j} \in \mathbb{R}^{Nb\times1}$ represents the denoising pixel in the denoised image $\hat{\mathbf{X}}_1$ from the DDM. The abundance loss includes the ASC and ANC constraints:

$$\begin{aligned}L_{ASC} &= \frac{1}{Np}\sum_{j=1}^{Np}[(\sum_{m=1}^{c}\hat{a}_{mj})-1]^2,\\ L_{ANC} &= \frac{1}{c\times Np}\sum_{i=1}^{c}\sum_{j=1}^{Np}\max(0,-\hat{a}_{mj}),\\ L_A &= L_{ASC}+L_{ANC}.\end{aligned} \tag{31}$$

where $\hat{a}_{mj}$ is the abundance of the m-th endmember at the j-th pixel, and c is the number of endmembers.

Finally, the total loss function is defined as:

$$L = L_R + \beta L_D + \gamma L_A, \tag{32}$$

where β and γ are regularization parameters, introduced to balance the contributions of the different loss components in

the optimization process. These parameters adjust the relative importance of each of the terms, allowing the network to find an optimal trade-off between accurately reconstructing the hyperspectral image and estimating the endmember signatures and the abundances.

IV. Experiments

To evaluate the performance of the proposed SSAUN, experiments were performed on a synthetic dataset (with different noise levels) and three real hyperspectral datasets. Three classical and four state-of-the-art DL-based unmixing methods were chosen for comparison: VCA followed by fully constrained least squares unmixing (VCA-FCLS) [16], [63], the minimum-volume based methods R-CoNMF [17] and MVSA [13], the autoencoder-based methods DAEU [33], MTAEU [34] and SIDAEU [35], and the CNN-based MSNet [31].

A. Data Descriptions

1) Synthetic Dataset: The proposed method assumes that the observed spectra are linear combinations of pure material spectra. To demonstrate its effectiveness, a synthetic dataset is generated that fully complies with the LMM. This dataset consists of 60×60 pixels and is created by linearly combining five randomly selected endmembers from the United States Geological Survey (USGS) spectral library. The spectral data contains 188 bands with a resolution of 10 nm, covering a wavelength range from 0.38 μm to 2.5 μm. A 64×64 pixel image is divided into 8×8 blocks, with each block filled with one of the five randomly chosen endmembers. The abundance maps are created by applying a 5×5 mean low-pass filter to ensure smooth transitions between adjacent endmembers. The outer 2×2 rows and columns of pixels are removed, resulting in a final image of 60×60 pixels. To test the robustness of the method under noisy conditions, Gaussian noise with different signal-to-noise ratios (SNRs) is added to the spectral data. For each noise level, different sets of endmembers are used.

2) Samson Dataset: The Samson dataset consists of 952×952 pixels with 156 spectral bands, covering a wavelength range from 401 nm to 889 nm. To reduce the computational cost for the analysis, a 95×95 pixel area is extracted, starting from the pixel at coordinates (252, 332). This subset of the dataset contains three endmembers: soil, tree, and water.

3) Jasper Ridge Dataset: The Jasper Ridge dataset contains 224 spectral bands with wavelengths from 380 to 2500 nm and consists of 512×614 pixels. For computational efficiency, a subset of 100×100 pixels is extracted, starting from pixel coordinates (105, 269). Due to atmospheric interference and dense water vapor, spectral bands 1–3, 108–112, 154–166, and 220–224 are removed, leaving 198 bands for further analysis. The final Jasper Ridge dataset includes four endmembers: water, soil, tree, and road.

4) Urban Dataset: The Urban dataset contains 307×307 pixels with 210 spectral bands covering wavelengths from 400 to 2500 nm. To minimize the impact of atmospheric interference and dense water vapor, bands 1–4, 76, 87, 101–111, 136–153, and 198–210 are removed, leaving 162 bands for further analysis. The final Urban dataset includes four endmembers: asphalt, grass, tree, and roof. The three real datasets and ground-truth for endmembers and abundances can be found in remote sensing lab[1].

B. Evaluation Metrics

In our experiments, we employ three widely-used evaluation metrics to assess the performance of the algorithms: spectral angle distance (SAD), root mean square error (RMSE), and reconstruction error (RE) with respect to the ground truth. These metrics are used to evaluate the accuracy of the estimated endmember signatures, the abundances, and the reconstructed hyperspectral images, respectively. SAD is defined as follows:

$$\mathrm{SAD}(\mathbf{m}_k, \hat{\mathbf{m}}_k) = \arccos \frac{\mathbf{m}_k \cdot \hat{\mathbf{m}}_k}{||\mathbf{m}_k||_2 ||\hat{\mathbf{m}}_k||_2}, \tag{33}$$

where $\hat{\mathbf{m}}_k$ and $\mathbf{m}_k$ represent the estimated and the ground truth endmember signatures, respectively. RMSE is computed by:

$$\mathrm{RMSE}(\hat{\mathbf{a}}_j, \mathbf{a}_j) = \sqrt{\frac{1}{Np} \sum_{j=1}^{Np} (\hat{\mathbf{a}}_j - \mathbf{a}_j)^2}, \tag{34}$$

where $\hat{\mathbf{a}}_j$ and $\mathbf{a}_j$ represent the estimated and ground-truth abundances, and *Np* is the number of pixels. Finally, RE is defined as:

$$\begin{aligned} \mathrm{RE}(\mathbf{X}, \hat{\mathbf{X}}_2) &= \frac{1}{Np} \sum_{j=1}^{Np} \mathrm{RE}(\mathbf{x}_j, \hat{\mathbf{x}}_{2j}) \\ &= \frac{1}{Np} \sum_{j=1}^{Np} \arccos \frac{\mathbf{x}_j \cdot \hat{\mathbf{x}}_{2j}}{||\mathbf{x}_j||_2 ||\hat{\mathbf{x}}_{2j}||_2}, \end{aligned} \tag{35}$$

where $\hat{\mathbf{x}}_{2j}$ and $\mathbf{x}_j$ are the reconstructed and ground truth spectra of pixel *j*.

C. Experiments on the Synthetic Dataset

Table I presents the mean SAD and mean RMSE obtained by the different unmixing methods on the synthetic dataset with different noise levels. The proposed SSAUN consistently outperformed the other methods in terms of both SAD and RMSE. This indicates that SSAUN is highly effective in extracting endmembers and estimating their abundances, even when noise is introduced into the synthetic datasets. To further assess the noise robustness of the proposed ASSUN, RE values are shown with the original noisy and clean image as ground truth, respectively (note that the noisy image was always the input for all unmixing methods). It can be observed that the DL-based unmixing methods DAEU, MTAEU, SIDAEU, and MSNet, exhibited superior performance in terms of image reconstruction compared to the traditional techniques VCA-FCLS, R-CoNMF, and MVSA. Notably, the reconstruction results of SSAUN were closer to the original clean image compared to other techniques, demonstrating that it is immune to noise, highlighting its capability to effectively deal with noise, and leading to more accurate unmixing results (even in challenging conditions).

[1] https://rslab.ut.ac.ir/data

Fig. 3 and Fig. 4 show the endmembers and abundance maps estimated by all unmixing methods on synthetic data with 20dB noise. From the figures, one can observe that the endmembers obtained by our proposed SSAUN have good consistency with the true endmembers, and the estimated abundance maps are closer to the ground truth compared to other unmixing methods.

D. Experiments on the Samson Dataset

Table II shows a quantitative performance assessment of all algorithms on the Samson dataset. The results show that VCA-FCLS, MSNet, and SSAUN achieved the best performance in extracting the tree, water, and soil endmembers, respectively. However, SSAUN standed out overall, providing the best results in terms of mean SAD, mean RMSE, and RE, demonstrating that SSAUN is highly effective for endmember extraction, abundance estimation, and image reconstruction. Fig. 5 provides a visual comparison of the endmembers estimated by different unmixing methods. Clearly, the DL-based methods outperformed traditional methods such as R-CoNMF and MVSA, which struggle to effectively extract the endmembers. The abundance maps estimated by different unmixing methods are presented in Fig. 6. It can be observed that DAEU, MTAEU, SIDAEU, and SSAUN successfully separated the three endmembers more clearly compared to VCA-FCLS, R-CoNMF, and MVSA, demonstrating the superiority of the DL-based approaches for this task.

E. Experiments on the Jasper Ridge Dataset

Table III provides a quantitative comparison of the different algorithms. The table indicates that DAEU accurately estimated the endmembers for trees and water, while MSNet and SSAUN were the most successful in estimating soil and road, respectively. Overall, SSAUN standed out in terms of mean SAD, as well as in RMSE and RE. Fig. 7 presents the endmember signatures estimated by different unmixing methods on the Jasper Ridge dataset. The results show that the endmember signatures from SSAUN aligned closely with the ground truth. Fig. 8 displays the visual results of abundance maps estimated by different unmixing methods. It has been noted that traditional unmixing methods, such as VCA-FCLS, R-CoNMF, and MVSA, did not accurately estimate the abundances of endmembers. In contrast, our proposed SSAUN effectively estimated the abundances of trees, water, and roads, whereas MSNet excelled in generating abundance maps for soil.

F. Experiments on the Urban Dataset

Table IV quantitatively assesses the performance of the different methods on the Urban dataset. SSAUN was not only superior in estimating each of the endmembers but also achieved the best results in terms of mean mean SAD and RMSE. Although SIDAEU obtained the best RE values, SSAUN was competitive, showing only a small gap. Overall, the results underline SSAUN's effectiveness in hyperspectral unmixing and image reconstruction, confirming its effectiveness across multiple performance metrics. Fig. 9 provides a visual comparison of the reference signatures and the endmembers estimated by different methods. The figure highlights that SSAUN exhibited high accuracy and consistency compared to other approaches, indicating its stable ability in extracting precise endmembers. Fig. 10 shows the estimated abundance maps for the Urban dataset, generated by different methods. Clearly, MSNet had the best result in estimating abundances of roof, and SSAUN had clearer abundance maps in estimating asphalt, grass, and tree than other methods.

G. Computational Cost

We conducted all experiments on a computer equipped with a 2.6 GHz Intel Core i7 CPU, 16 GB of memory, and an NVIDIA GeForce RTX 2060 GPU. The running time of the different unmixing methods across all datasets are documented in Table V. The table shows that traditional methods such as VCA-FCLS, R-CoNMF, and MVSA achieved higher efficiency compared to DL-based unmixing methods. Nevertheless, when considering a comprehensive evaluation, the efficiency of the proposed SSAUN method remained acceptable. This balance between performance and computational efficiency highlights the practical utility of SSAUN in hyperspectral unmixing applications. To assess the resource consumption of each of the 2 modules DDM and AUM separately, we investigated the number of parameters in each module, which can be regarded as an indirect indicator of resource consumption. It was observed that AUM contains up to three times less parameters than DMM and therefore requires fewer computing resources.

H. Hyperparameter Settings

We employ an Adam optimizer alongside Eq. (32) as the loss function to guide the optimization process. The regularization parameters β and γ in Eq. (32) play a crucial role in balancing the trade-off between the image reconstruction, denoising, and abundance estimation tasks. The size of the applied convolution kernel across the different layers of the network is another hyperparameter. Fig. 11 provides a detailed analysis of the impact of these hyperparameters. The analysis shows that setting the regularization parameter β to 0.01 and keeping the kernel size at 3×3 consistently yielded the best performance for all datasets. The optimal settings for the regularization parameter γ varied per dataset: 0.02 for the Samson dataset, 0.015 for the Jasper Ridge dataset, and 0.1 for the Urban dataset. These specific settings significantly improve performance in terms of mean SAD and mean RMSE, thereby confirming their effectiveness in hyperspectral unmixing tasks.

V. Discussion

A. Ablation Study for the Spectral-Spatial Attention Module

In the proposed approach, we embedded a novel SSAM into the AUM to enhance the performance of hyperspectral unmixing. SSAM contains a spatial attention module and a spectral attention module that are fused together. We performed an ablation study, in which either or both of the attention modules are applied, and the obtained results are compared. Table

TABLE I
PERFORMANCE EVALUATION OF DIFFERENT UNMIXING METHODS ON THE SYNTHETIC DATASET. BEST RESULTS ARE SHOWN IN BOLD.

SNR	Metrics	VCA-FCLS	R-CoNMF	MVSA	DAEU	MTAEU	SIDAEU	MSNet	SSAUN
10dB	Mean SAD	0.0741	0.0602	0.1620	0.1277	0.1332	0.1179	0.0654	**0.0499**
	Mean RMSE	0.1154	0.1284	0.2367	0.2513	0.2154	0.1965	0.1194	**0.1129**
	RE (Noisy GT)	0.4050	0.3989	0.4137	0.3592	0.3577	0.3956	0.3577	**0.3571**
	RE (Clean GT)	0.1901	0.1770	0.2144	0.0743	0.0579	0.1745	0.0578	**0.0543**
20dB	Mean SAD	0.1379	0.1457	0.1870	0.1430	0.1094	0.2630	0.1579	**0.0364**
	Mean RMSE	0.0662	0.1067	0.1765	0.2174	0.1372	0.2211	0.1020	**0.0603**
	RE (Noisy GT)	0.2118	0.1897	0.1617	0.1286	0.1219	0.1276	**0.1200**	0.1204
	RE (Clean GT)	0.1628	0.1398	0.1044	0.0462	0.0289	0.0478	0.0217	**0.0239**
30dB	Mean SAD	0.0141	0.0265	0.0469	0.0796	0.0938	0.1223	0.0264	**0.0033**
	Mean RMSE	0.0397	0.0697	0.1070	0.2170	0.2320	0.2340	0.1154	**0.0394**
	RE (Noisy GT)	0.1585	0.1686	0.1785	0.0419	0.0412	0.0470	0.0377	**0.0343**
	RE (Clean GT)	0.1530	0.1637	0.1741	0.0230	0.0216	0.0311	0.0149	**0.0051**

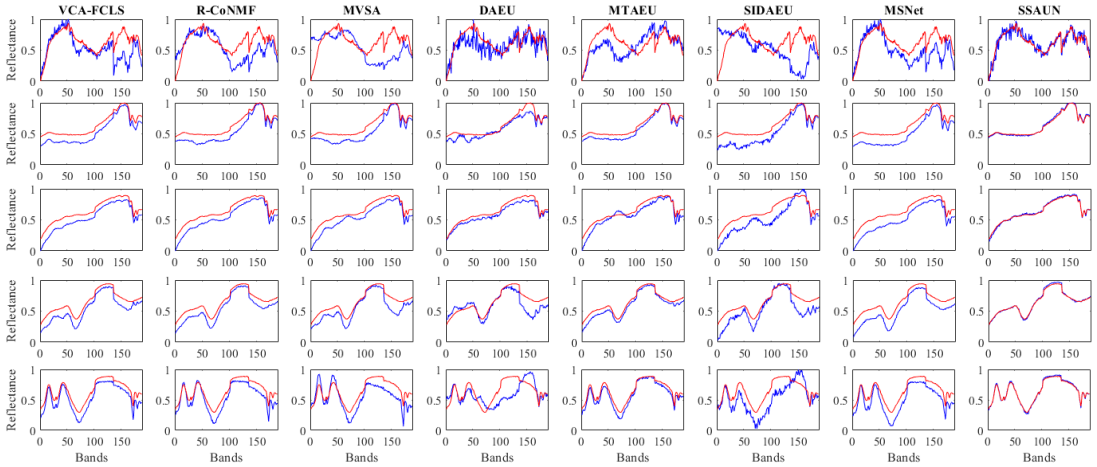

Fig. 3. Reference (red) and estimated (blue) endmember signatures by the different unmixing methods on synthetic data with 20dB.

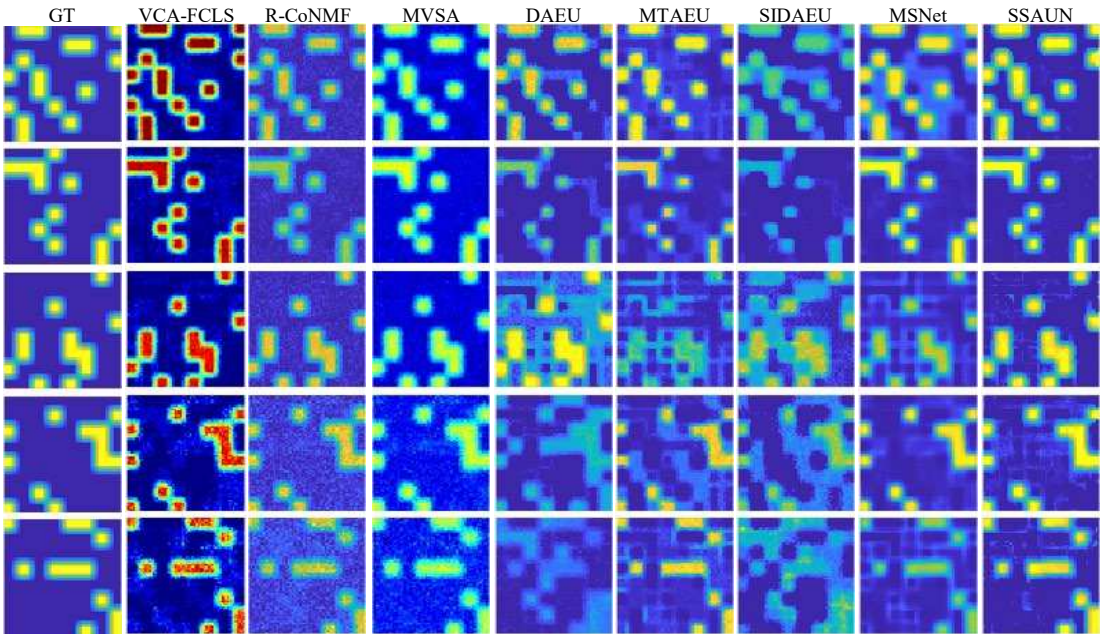

Fig. 4. Abundance maps estimated by the different unmixing methods on synthetic data with 20dB. The GT column shows the ground truth abundance maps.

TABLE II
PERFORMANCE EVALUATION OF DIFFERENT UNMIXING METHODS ON THE SAMSON DATASET. BEST RESULTS ARE SHOWN IN BOLD.

Methods		VCA-FCLS	R-CoNMF	MVSA	DAEU	MTAEU	SIDAEU	MSNet	SSAUN
SAD	Soil	0.0232	0.0534	0.0455	0.0271	0.0393	0.0236	0.0163	**0.0112**
	Tree	**0.0206**	0.0290	0.0247	0.0322	0.0430	0.0243	0.0218	0.0219
	Water	0.1058	0.1861	0.3724	0.0467	0.0439	0.0542	**0.0377**	0.0405
Mean SAD		0.0499	0.0895	0.1475	0.0353	0.0421	0.0340	0.0253	**0.0245**
Mean RMSE		0.2748	0.3145	0.2747	0.0700	0.0786	0.0614	0.0457	**0.0269**
RE		0.4009	0.4791	0.4578	0.0501	0.0544	0.0590	0.0406	**0.0362**

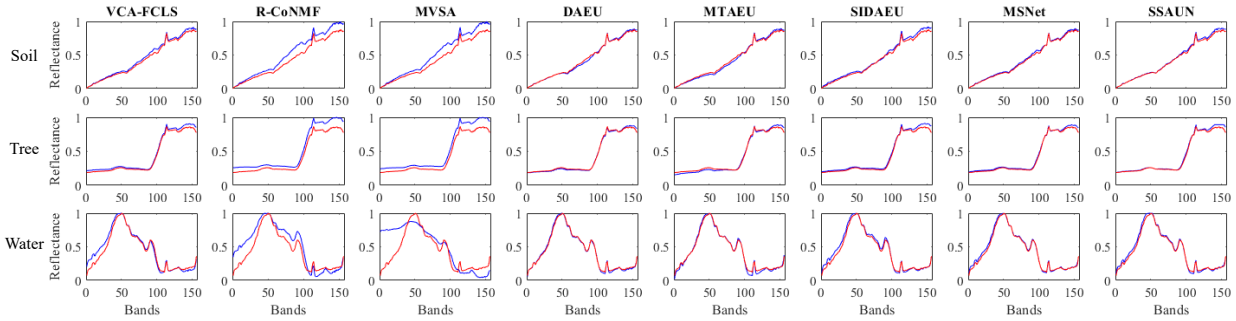

Fig. 5. Reference (red) and endmember signatures (blue) estimated by the different unmixing methods on the Samson dataset.

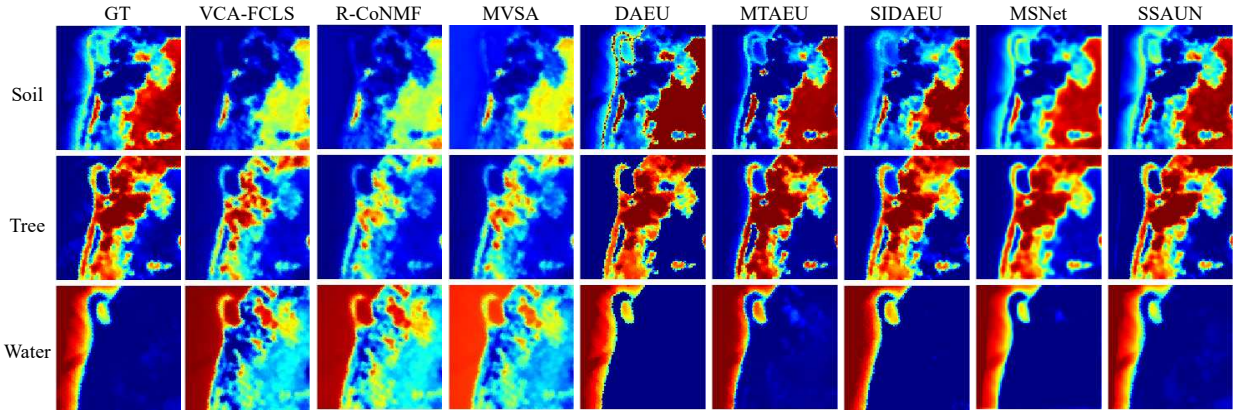

Fig. 6. Abundance maps estimated by the different unmixing methods on the Samson dataset. The GT column shows the ground truth abundance maps.

VI presents the results. The combination of both spectral and spatial attention led to the best results. In some cases, only applying the spatial or the spectral module matches the results of applying both. In other cases, the combination is significantly superior, underlining the advantage of attentively focusing on both spectral and spatial information.

B. Ablation Study for the Spatial Attention Module

The dual spatial attention module computes the maximum and mean values along the channel dimension to enhance spatial features. This attention mechanism allows the model to focus on critical spatial regions within the input data, improving its ability to capture key features in the image. Fig. 12 shows a visual analysis of this dual spatial attention module. As can be observed, the mean-based operation smoothens the image by reducing high-frequency variations, highlighting broader, less detailed features, while the max-based operation retains and enhances higher intensity values, highlighting sharper features in the image. Moreover, in the dual spatial attention module, both concatenation and summation steps are applied to merge the feature maps $\mathbb{F}_{max}$ and $\mathbb{F}_{min}$, as detailed in Eq. (24). We performed an ablation study in which either or both of the steps are applied, and the results are compared in Table VII. The combined use of concatenation and summation produced superior results.

C. Future direction

Existing deep-learning-based unmixing methods cannot easily be generalized to nonlinear cases. This limitation arises mainly since endmembers are typically extracted from the final

TABLE III
PERFORMANCE EVALUATION OF DIFFERENT UNMIXING METHODS ON THE JASPER RIDGE DATASET. BEST RESULTS ARE SHOWN IN BOLD.

Methods		VCA-FCLS	R-CoNMF	MVSA	DAEU	MTAEU	SIDAEU	MSNet	SSAUN
SAD	Tree	0.3176	0.0330	0.0490	**0.0185**	0.0961	0.0300	0.0442	0.0310
	Water	0.2956	0.0253	0.1227	**0.0098**	0.0955	0.0261	0.0427	0.0146
	Soil	0.6569	0.0786	0.1137	0.0506	0.1048	0.0796	**0.0479**	0.0606
	Road	0.5195	0.0473	0.0550	0.2056	0.2172	0.2433	0.1014	**0.0240**
Mean SAD		0.4474	0.0460	0.0851	0.0711	0.1284	0.0948	0.0591	**0.0326**
Mean RMSE		0.3545	0.1641	0.2007	0.1319	0.2700	0.1309	0.0977	**0.0913**
RE		0.3276	0.3951	0.4380	0.0768	0.0888	0.0718	0.0728	**0.0627**

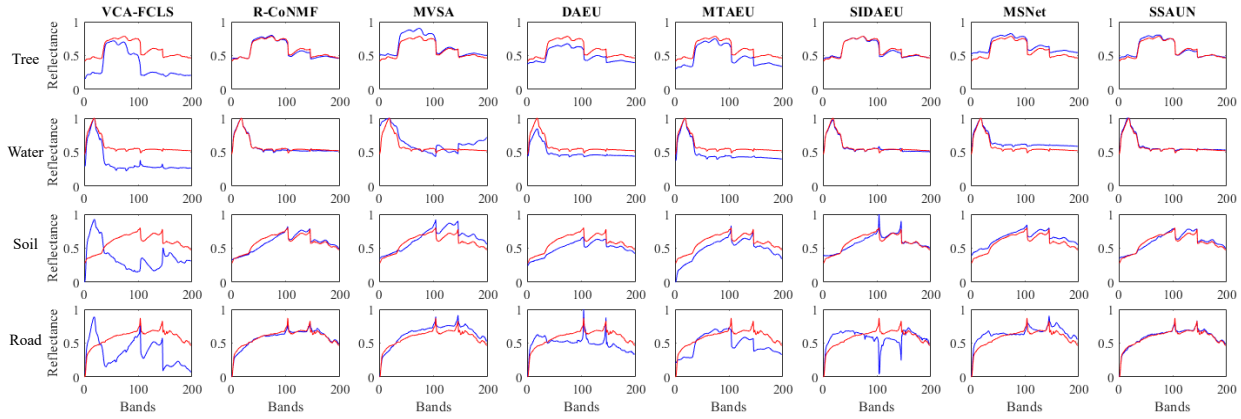

Fig. 7. Reference (red) and endmember signatures (blue) estimated by different unmixing methods on the Jasper Ridge dataset.

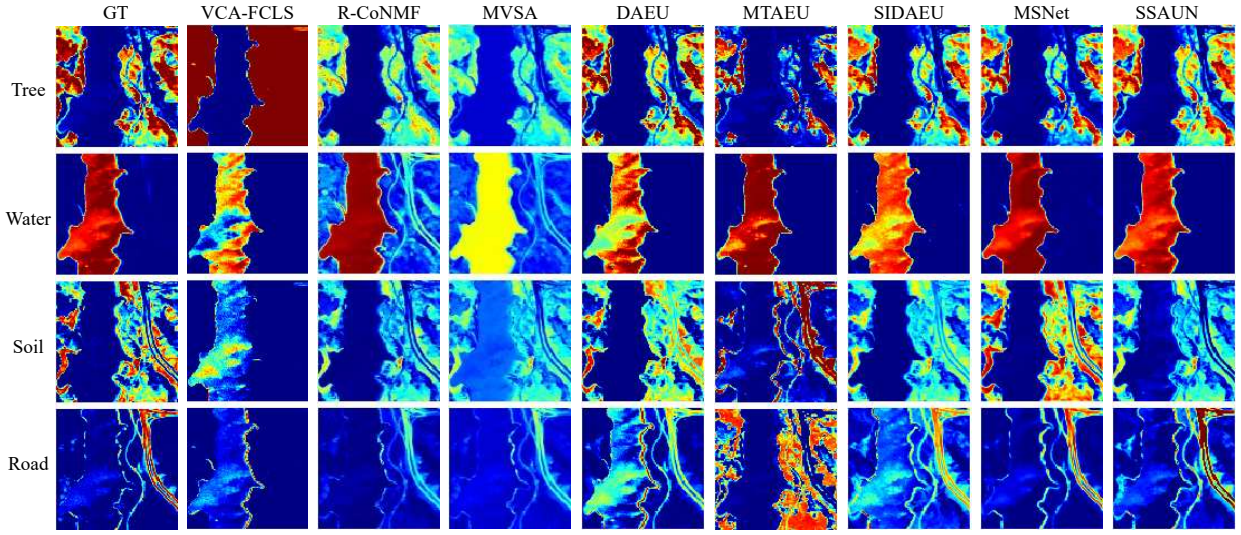

Fig. 8. Abundance maps estimated by different unmixing methods on the Jasper Ridge dataset. The GT column shows the ground truth abundance maps.

linear layer of the decoder, and the data is reconstructed by combining endmembers and fractional abundances. Although these methods can be extended to bilinear models by incorporating both linear (**MA**) and nonlinear terms (**MA** $\odot$ **MA**), it is not easy to generalize them to more complex models, such as multilinear mixing models and the Hapke model. Instead, a dedicated nonlinear layer is required that explicitly represents the mixing model, making the task significantly

TABLE IV
PERFORMANCE EVALUATION OF DIFFERENT UNMIXING METHODS ON THE URBAN DATASET. BEST RESULTS ARE SHOWN IN BOLD.

Methods		VCA-FCLS	R-CoNMF	MVSA	DAEU	MTAEU	SIDAEU	MSNet	SSAUN
SAD	Asphalt	0.3459	0.2091	0.1923	0.2106	0.1411	0.2901	0.1511	**0.0773**
	Grass	0.1170	1.0611	1.0430	0.2171	0.3427	0.3471	0.1022	**0.0659**
	Tree	0.2314	0.0728	0.0825	0.0806	0.0773	0.1338	0.0948	**0.0222**
	Roof	0.3608	0.1341	0.1314	0.2436	0.1473	0.2555	0.1736	**0.1235**
Mean SAD		0.2638	0.3693	0.3623	0.1880	0.1771	0.2566	0.1304	**0.0722**
Mean RMSE		0.3184	0.3080	0.2887	0.1454	0.1560	0.1670	0.1259	**0.1103**
RE		0.2835	0.4925	0.4454	0.0548	0.0713	**0.0509**	0.0837	0.0521

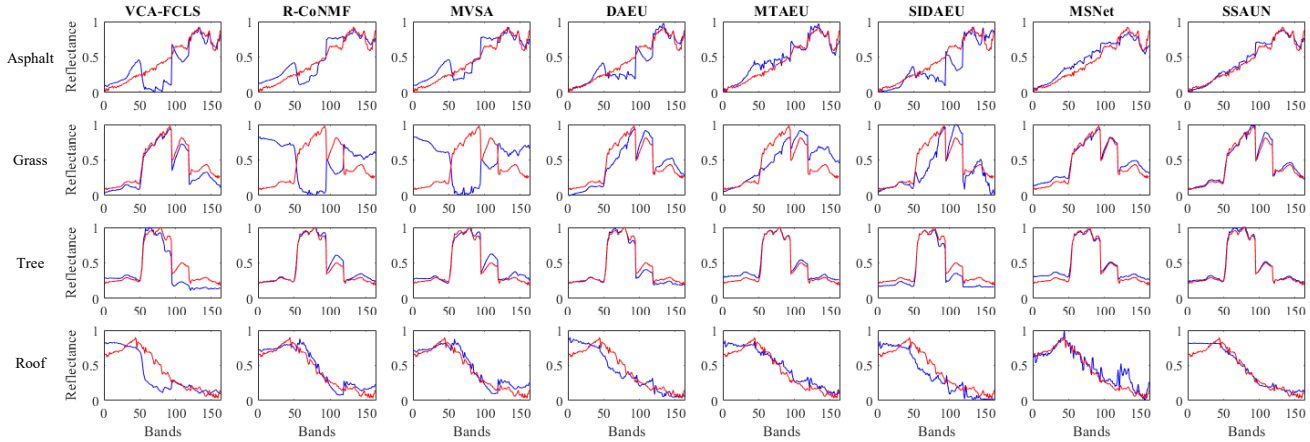

Fig. 9. Reference (red) and endmember signatures (blue) estimated by different unmixing methods on the Urban dataset.

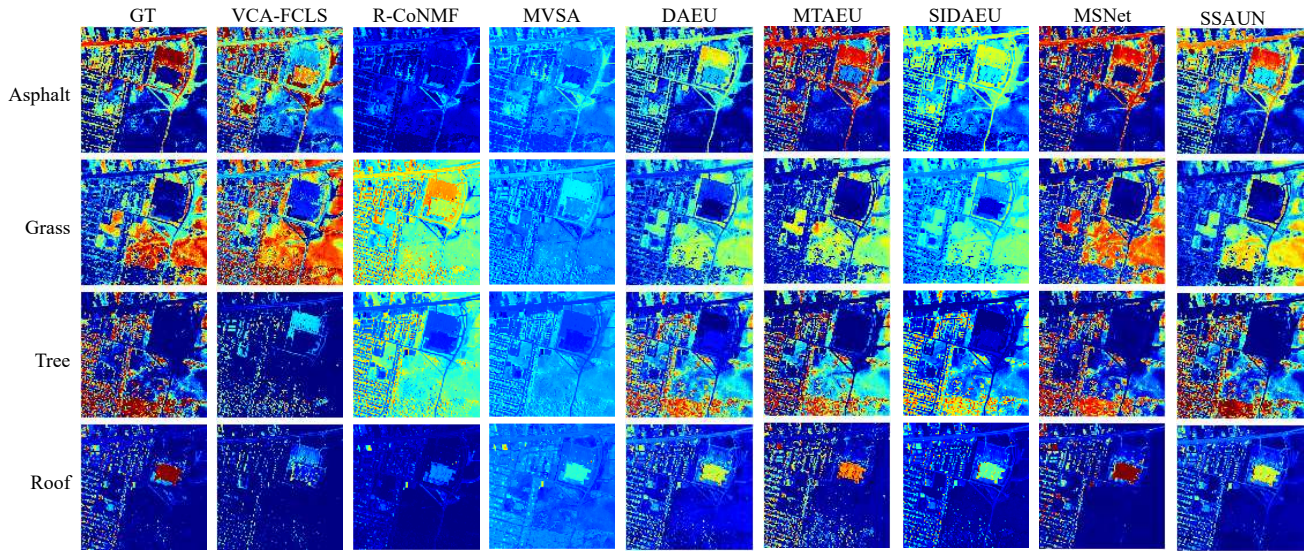

Fig. 10. Abundance maps estimated by different unmixing methods on the Urban dataset. The GT column shows the ground truth abundance maps.

more challenging.

Our method addresses this challenge by estimating endmembers through the inversion of the mixing model, enabling their integration into any mixing model. In future work, we

TABLE V
COMPUTATIONAL COST OF ALL METHODS ON DIFFERENT DATASETS (IN SECONDS).

Methods	Synthetic (20dB)	Samson	Jasper Ridge	Urban
VCA-FCLS	0.2909	0.4276	0.4401	4.2662
R-CoNMF	2.2753	1.1273	1.2587	20.9045
MVSA	0.7224	0.8615	0.4283	3.3006
DAEU	18.4865	43.8835	47.9196	52.8121
MTAEU	35.5518	85.9920	99.3320	107.5883
SIDAEU	4.9552	41.2642	10.7187	48.4249
MSNet	10.7441	11.2745	11.6601	49.6279
SSAUN	16.8025	23.3912	25.1283	172.1192

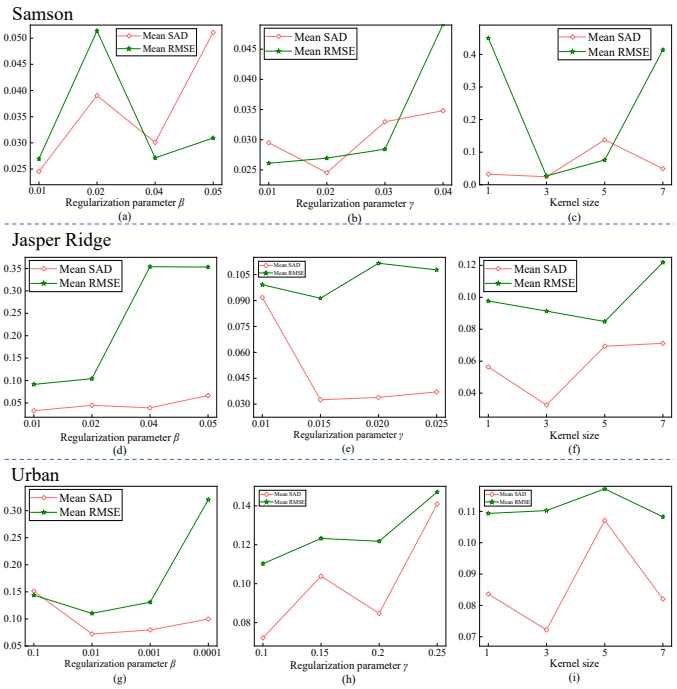

Fig. 11. Hyperparameter analysis conducted by varying: regularization parameter β, regularization parameter γ, and kernel size for different datasets: (a)-(c) Samson dataset, (d)-(f) Jasper Ridge dataset, and (g)-(i) Urban dataset.

TABLE VI
ABLATION STUDY FOR OUR PROPOSED SSAUN WITH VARYING ATTENTION MODULES ON DIFFERENT DATASETS. BEST RESULTS ARE SHOWN IN BOLD.

Datasets	Spectral attention	Spatial attention	Mean SAD	Mean RMSE
Synthetic data (20dB)	✓	✗	0.1671	0.1563
	✗	✓	0.1759	0.0854
	✓	✓	**0.0364**	**0.0603**
Samson	✓	✗	0.0385	0.0506
	✗	✓	0.0556	0.0489
	✓	✓	**0.0245**	**0.0269**
Jasper Ridge	✓	✗	0.0332	0.1023
	✗	✓	0.0364	0.0974
	✓	✓	**0.0326**	**0.0913**
Urban	✓	✗	0.0849	0.1168
	✗	✓	0.0726	0.1140
	✓	✓	**0.0722**	**0.1103**

TABLE VII
ABLATION STUDIES FOR OUR PROPOSED DUAL SPATIAL ATTENTION MODULE USING DIFFERENT COMBINATIONS FOR DIFFERENT KINDS OF SPATIAL EXTRACTION ON DIFFERENT DATASETS. BEST RESULTS ARE SHOWN IN BOLD.

Datasets	Concatenation	Summation	Mean SAD	Mean RMSE
Synthetic data (20dB)	✓	✗	0.1514	0.1229
	✗	✓	0.3097	0.3491
	✓	✓	**0.0364**	**0.0603**
Samson	✓	✗	0.0460	0.0361
	✗	✓	0.0537	0.0409
	✓	✓	**0.0245**	**0.0269**
Jasper Ridge	✓	✗	0.0468	0.1476
	✗	✓	0.0928	0.1156
	✓	✓	**0.0326**	**0.0913**
Urban	✓	✗	0.0985	0.3279
	✗	✓	0.0989	0.1429
	✓	✓	**0.0722**	**0.1103**

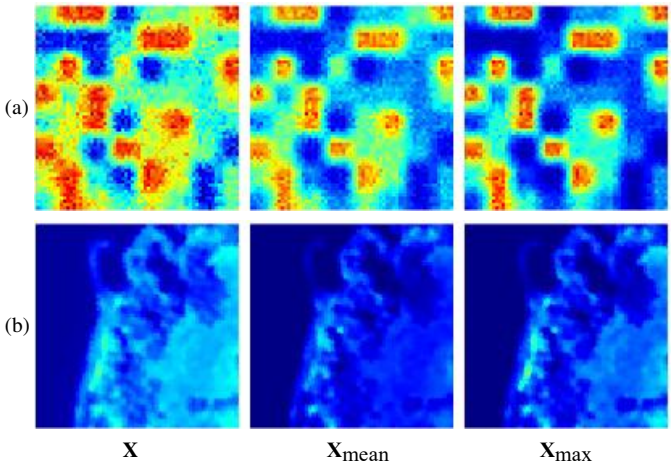

Fig. 12. Visual analysis of dual spatial attention module on: (a) Synthetic dataset (20dB) and (b) Samson dataset. $\mathbf{X}$: original hyperspectral image (band 100); $\mathbf{X}_{\text{mean}}$: the Hadamard Product between $\mathbf{X}$ and $\mathbb{F}_{mean}$; $\mathbf{X}_{\max}$: the Hadamard Product between $\mathbf{X}$ and $\mathbb{F}_{max}$.

plan to extend the proposed approach to nonlinear cases.

VI. CONCLUSION

In this work, we proposed a spectral-spatial attention unmixing network (SSAUN) for effective hyperspectral unmixing and image reconstruction. We addressed the challenges common to conventional DL-based unmixing methods and improved performance through several key innovations. SSAUN includes a DDM that preemptively filters out noise, strengthening the robustness and accuracy of the network. The core unmixing tasks are handled by AUM, which directly estimates the abundance maps from the output. Furthermore, instead of using weights to estimate the endmembers, an innovative inversion strategy is used. A SSAM is carefully designed to identify and highlight critical features across spectral channels while capturing extensive vertical and horizontal spatial dependencies. This dual attention to spectral and spatial details allows our newly developed ASSUN to distinguish and enhance relevant features, significantly improving unmixing accuracy. Extensive evaluations on both synthetic and real-world datasets

have validated the effectiveness and robustness of ASSUN, confirming its superior performance in hyperspectral unmixing and image reconstruction tasks. Our future efforts will focus on expanding ASSUN's capabilities to address nonlinear unmixing problems. This initiative aims to adapt the network to more complex scenarios, increasing its applicability and its effectiveness in various hyperspectral unmixing applications.

ACKNOWLEDGEMENT

The research presented in this paper is funded by the Research Foundation-Flanders (project G031921N). Bikram Koirala is a postdoctoral fellow of the Research Foundation Flanders, Belgium (FWO: 1250824N-7028)

REFERENCES

[1] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and remote sensing magazine*, vol. 1, no. 2, pp. 6–36, 2013.

[2] M. B. Stuart, A. J. McGonigle, and J. R. Willmott, "Hyperspectral imaging in environmental monitoring: A review of recent developments and technological advances in compact field deployable systems," *Sensors*, vol. 19, no. 14, p. 3071, 2019.

[3] S. Peyghambari and Y. Zhang, "Hyperspectral remote sensing in lithological mapping, mineral exploration, and environmental geology: an updated review," *Journal of Applied Remote Sensing*, vol. 15, no. 3, pp. 031501–031501, 2021.

[4] P. K. Sethy, C. Pandey, Y. K. Sahu, and S. K. Behera, "Hyperspectral imagery applications for precision agriculture-a systemic survey," *Multimedia Tools and Applications*, vol. 81, no. 2, pp. 3005–3038, 2022.

[5] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[6] W.-K. Ma, J. M. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A. J. Plaza, A. Ambikapathi, and C.-Y. Chi, "A signal processing perspective on hyperspectral unmixing: Insights from remote sensing," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 67–81, 2013.

[7] P.-A. Thouvenin, N. Dobigeon, and J.-Y. Tourneret, "Hyperspectral unmixing with spectral variability using a perturbed linear mixing model," *IEEE Transactions on Signal Processing*, vol. 64, no. 2, pp. 525–538, 2015.

[8] X. Tao, M. E. Paoletti, L. Han, J. M. Haut, P. Ren, J. Plaza, and A. Plaza, "Fast orthogonal projection for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[9] J. Wei and X. Wang, "An overview on linear unmixing of hyperspectral data," *Mathematical Problems in Engineering*, vol. 2020, no. 1, p. 3735403, 2020.

[10] M. E. Winter, "N-findr: An algorithm for fast autonomous spectral endmember determination in hyperspectral data," in *Imaging spectrometry V*, vol. 3753, pp. 266–275, SPIE, 1999.

[11] X. Tao, M. E. Paoletti, J. M. Haut, P. Ren, J. Plaza, and A. Plaza, "Endmember estimation with maximum distance analysis," *Remote Sensing*, vol. 13, no. 4, p. 713, 2021.

[12] C.-I. Chang and A. Plaza, "A fast iterative algorithm for implementation of pixel purity index," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 1, pp. 63–67, 2006.

[13] J. Li and J. M. Bioucas-Dias, "Minimum volume simplex analysis: A fast algorithm to unmix hyperspectral data," in *IGARSS 2008-2008 IEEE International Geoscience and Remote Sensing Symposium*, vol. 3, pp. III–250, IEEE, 2008.

[14] L. Miao and H. Qi, "Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 3, pp. 765–777, 2007.

[15] T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma, "A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4418–4432, 2009.

[16] J. M. Nascimento and J. M. Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE transactions on Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 898–910, 2005.

[17] J. Li, J. M. Bioucas-Dias, A. Plaza, and L. Liu, "Robust collaborative nonnegative matrix factorization for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6076–6090, 2016.

[18] K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "A novel hierarchical bayesian approach for sparse semisupervised hyperspectral unmixing," *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 585–599, 2011.

[19] H. Liu, Y. Lu, Z. Wu, Q. Du, J. Chanussot, and Z. Wei, "Bayesian unmixing of hyperspectral image sequence with composite priors for abundance and endmember variability," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[20] N. Dobigeon, S. Moussaoui, M. Coulon, J.-Y. Tourneret, and A. O. Hero, "Joint bayesian endmember extraction and linear unmixing for hyperspectral imagery," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4355–4368, 2009.

[21] M.-D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Collaborative sparse regression for hyperspectral unmixing," *IEEE Transactions on geoscience and remote sensing*, vol. 52, no. 1, pp. 341–354, 2013.

[22] S. Zhang, J. Li, K. Liu, C. Deng, L. Liu, and A. Plaza, "Hyperspectral unmixing based on local collaborative sparse regression," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 5, pp. 631–635, 2016.

[23] S. Zhang, J. Li, H.-C. Li, C. Deng, and A. Plaza, "Spectral–spatial weighted sparse regression for hyperspectral image unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 6, pp. 3265–3276, 2018.

[24] X. Shen, H. Liu, X. Zhang, K. Qin, and X. Zhou, "Superpixel-guided local sparsity prior for hyperspectral sparse regression unmixing," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[25] K. Golhani, S. K. Balasundram, G. Vadamalai, and B. Pradhan, "A review of neural networks in plant disease detection using hyperspectral data," *Information Processing in Agriculture*, vol. 5, no. 3, pp. 354–371, 2018.

[26] L. Liu, T. Miteva, G. Delnevo, S. Mirri, P. Walter, L. de Viguerie, and E. Pouyet, "Neural networks for hyperspectral imaging of historical paintings: a practical review," *Sensors*, vol. 23, no. 5, p. 2419, 2023.

[27] Z. Wu, H. Lu, M. E. Paoletti, H. Su, W. Jing, and J. M. Haut, "Kacnet: Kolmogorov-arnold convolution network for hyperspectral anomaly detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[28] Y. Fu, Z. Liu, and J. Lyu, "Transferable adversarial attacks for remote sensing object recognition via spatial-frequency co-transformation," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[29] Y. Xu, H. Wang, F. Zhou, C. Luo, X. Sun, S. Rahardja, and P. Ren, "Mambahsisr: Mamba hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[30] X. Tao, M. E. Paoletti, L. Han, Z. Wu, P. Ren, J. Plaza, A. Plaza, and J. M. Haut, "A new deep convolutional network for effective hyperspectral unmixing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 6999–7012, 2022.

[31] Y. Yu, Y. Ma, X. Mei, F. Fan, J. Huang, and H. Li, "Multi-stage convolutional autoencoder network for hyperspectral unmixing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 113, p. 102981, 2022.

[32] X. Tao, B. Koirala, A. Plaza, and P. Scheunders, "A new dual-feature fusion network for enhanced hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[33] 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.

[34] B. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Spectral-spatial hyperspectral unmixing using multitask learning," *IEEE Access*, vol. 7, pp. 148861–148872, 2019.

[35] F. Palsson, J. Sigurdsson, J. R. Sveinsson, and M. O. Ulfarsson, "Neural network hyperspectral unmixing with spectral information divergence objective," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 755–758, IEEE, 2017.

[36] Z. Hua, X. Li, Q. Qiu, and L. Zhao, "Autoencoder network for hyperspectral unmixing with adaptive abundance smoothing," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 9, pp. 1640–1644, 2020.

[37] Y. Su, J. Li, A. Plaza, A. Marinoni, P. Gamba, and S. Chakravortty, "Daen: Deep autoencoder networks for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4309–4321, 2019.

[38] B. Rasti and B. Koirala, "Suncnn: Sparse unmixing using unsupervised convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[39] B. Palsson, M. O. Ulfarsson, and J. R. Sveinsson, "Convolutional autoencoder for spectral–spatial hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 535–549, 2020.

[40] B. Rasti, B. Koirala, P. Scheunders, and P. Ghamisi, "UnDIP: Hyperspectral unmixing using deep image prior," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–15, 2021.

[41] Y. Fang, Y. Wang, L. Xu, R. Zhuo, A. Wong, and D. A. Clausi, "Bcun: Bayesian fully convolutional neural network for hyperspectral spectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[42] B. Rasti, B. Koirala, P. Scheunders, and J. Chanussot, "Misicnet: Minimum simplex convolutional network for deep hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[43] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[44] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational visual media*, vol. 8, no. 3, pp. 331–368, 2022.

[45] D. Soydaner, "Attention mechanism in neural networks: where it comes and where it goes," *Neural Computing and Applications*, vol. 34, no. 16, pp. 13371–13385, 2022.

[46] P. Ghosh, S. K. Roy, B. Koirala, B. Rasti, and P. Scheunders, "Hyperspectral unmixing using transformer network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[47] Z. Yang, M. Xu, S. Liu, H. Sheng, and J. Wan, "Ust-net: A u-shaped transformer network using shifted windows for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[48] J. Chen, C. Yang, L. Zhang, L. Yang, L. Bian, Z. Luo, and J. Wang, "Tccu-net: Transformer and cnn collaborative unmixing network for hyperspectral image," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[49] Y. Duan, X. Xu, T. Li, B. Pan, and Z. Shi, "Undat: Double-aware transformer for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[50] L. Qi, X. Qin, F. Gao, J. Dong, and X. Gao, "Sawu-net: Spatial attention weighted unmixing network for hyperspectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[51] S. Xiang, X. Li, J. Ding, S. Chen, and Z. Hua, "Unidirectional local-attention autoencoder network for spectral variability unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[52] X. Tao, M. E. Paoletti, Z. Wu, J. M. Haut, P. Ren, and A. Plaza, "An abundance-guided attention network for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[53] D. Jin and B. Yang, "Graph attention convolutional autoencoder-based unsupervised nonlinear unmixing for hyperspectral images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[54] R. Heylen, M. Parente, and P. Gader, "A review of nonlinear hyperspectral unmixing methods," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 1844–1868, 2014.

[55] N. Dobigeon, J.-Y. Tourneret, C. Richard, J. C. M. Bermudez, S. McLaughlin, and A. O. Hero, "Nonlinear unmixing of hyperspectral images: Models and algorithms," *IEEE Signal processing magazine*, vol. 31, no. 1, pp. 82–94, 2013.

[56] Q. Wei, M. Chen, J.-Y. Tourneret, and S. Godsill, "Unsupervised nonlinear spectral unmixing based on a multilinear mixing model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4534–4544, 2017.

[57] M. Li, B. Yang, and B. Wang, "Emlm-net: An extended multilinear mixing model-inspired dual-stream network for unsupervised nonlinear hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[58] Y. Su, Z. Zhu, L. Gao, A. Plaza, P. Li, X. Sun, and X. Xu, "Daan: A deep autoencoder-based augmented network for blind multilinear hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[59] T. Fang, F. Zhu, and J. Chen, "Hyperspecral unmixing based on multilinear mixing model using convolutional autoencoders," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[60] Y. Su, X. Xu, J. Li, H. Qi, P. Gamba, and A. Plaza, "Deep autoencoders with multitask learning for bilinear hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 10, pp. 8615–8629, 2020.

[61] L. Zhuang and M. K. Ng, "Fasthymix: Fast and parameter-free hyperspectral image mixed noise removal," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 8, pp. 4702–4716, 2021.

[62] B. Rasti, J. R. Sveinsson, M. O. Ulfarsson, and J. A. Benediktsson, "Hyperspectral image denoising using first order spectral roughness penalty in wavelet domain," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2458–2467, 2013.

[63] D. C. Heinz *et al.*, "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *IEEE transactions on geoscience and remote sensing*, vol. 39, no. 3, pp. 529–545, 2001.