remote sensing

MDPI

Article

A Global Spatial-Spectral Feature Fused Autoencoder for Nonlinear Hyperspectral Unmixing

Mingle Zhang [1,2], Mingyu Yang [1,*], Hongyu Xie [1,2], Pinliang Yue [1,2], Wei Zhang [1,2], Qingbin Jiao [1], Liang Xu [1] and Xin Tan [1]

[1] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China; zhangmingle21@mails.ucas.ac.cn (M.Z.); xiehongyu21@mails.ucas.ac.cn (H.X.); yuepinliang22@mails.ucas.ac.cn (P.Y.); zhangwei20f@mails.ucas.ac.cn (W.Z.); jiaoqingbin@ciomp.ac.cn (Q.J.); xuliang@ciomp.ac.cn (L.X.); tanxin@ciomp.ac.cn (X.T.)

[2] University of the Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: yangmingyu@ciomp.ac.cn

Abstract: Hyperspectral unmixing (HU) aims to decompose mixed pixels into a set of endmembers and corresponding abundances. Deep learning-based HU methods are currently a hot research topic, but most existing unmixing methods still rely on per-pixel training or employ convolutional neural networks (CNNs), which overlook the non-local correlations of materials and spectral characteristics. Furthermore, current research mainly focuses on linear mixing models, which limits the feature extraction capability of deep encoders and further improvement in unmixing accuracy. In this paper, we propose a nonlinear unmixing network capable of extracting global spatial-spectral features. The network is designed based on an autoencoder architecture, where a dual-stream CNNs is employed in the encoder to separately extract spectral and local spatial information. The extracted features are then fused together to form a more complete representation of the input data. Subsequently, a linear projection-based multi-head self-attention mechanism is applied to capture global contextual information, allowing for comprehensive spatial information extraction while maintaining lightweight computation. To achieve better reconstruction performance, a model-free nonlinear mixing approach is adopted to enhance the model's universality, with the mixing model learned entirely from the data. Additionally, an initialization method based on endmember bundles is utilized to reduce interference from outliers and noise. Comparative results on real datasets against several state-of-the-art unmixing methods demonstrate the superior of the proposed approach.

Keywords: hyperspectral unmixing; convolutional neural network; self-attention mechanism; deep learning

Citation: Zhang, M.; Yang, M.; Xie, H.; Yue, P.; Zhang, W.; Jiao, Q.; Xu, L.; Tan, X. A Global Spatial-Spectral Feature Fused Autoencoder for Nonlinear Hyperspectral Unmixing. *Remote Sens.* **2024**, *16*, 3149. https://doi.org/10.3390/rs16173149

Academic Editor: Jon Atli Benediktsson

Received: 7 June 2024
Revised: 7 August 2024
Accepted: 23 August 2024
Published: 26 August 2024

1. Introduction

Hyperspectral imagery (HSI) can be acquired from numerous contiguous spectral bands, enabling the identification of materials that cannot be distinguished in traditional broadband imagery [1]. However, different material substances may contribute to the spectral measurements of individual pixels. For such mixed pixels, we aim to identify the different materials present in the mixture, along with their corresponding proportions. Hyperspectral unmixing (HU) decomposes the spectral measurements of mixed pixels into a set of constituent spectra, or endmembers, and a set of corresponding fractions, or abundances, indicating the proportional presence of each endmember in the pixel [2,3]. Endmembers typically consist of familiar macroscopic substances in the scene, such as soil, trees, water, or any natural or man-made materials. HU provides the capability to identify subpixel details, which has practical value in many scenarios [4–6].

Based on the spectral mixing mechanism, common HU models can be categorized into Linear Mixing Model (LMM) and Nonlinear Mixing Model (NLMM) [7,8]. LMM assumes that each pixel in a HSI is a linear combination of endmembers and their abundances. Due

to its generality, LMM has been the primary model for HU over the past few decades [9]. HU using LMM consists of two steps. The first step involves endmember extraction, with typical methods including Pure Pixel Index (PPI) [10], N-FINDR algorithm [11], and Vertex Component Analysis (VCA) [12]. The second step is abundance estimation based on spectral data and the extracted endmembers, typically achieved through optimization algorithms with abundance nonnegativity constraint (ANC) and abundance sum-to-one constraint (ASC), with the common method being Fully Constrained Least Squares Unmixing (FCLSU) [13]. However, this two-step unmixing approach may lead to error accumulation [14]. To avoid such errors, blind unmixing techniques, which simultaneously perform endmember extraction and abundance estimation, have been widely researched [15–17]. Existing methods often rely on Nonnegative Matrix Factorization (NMF), where many extended NMFs introduce a series of regularization constraints during the matrix factorization process to incorporate prior information on both spectral and spatial domains into the NMF framework, thus enhancing the stability of unmixing [18–20].

However, in practical scenarios, the spectra captured by detectors are not simply the weighted sum of individual endmember spectra [21]. The spectral variability (SV) caused by lighting conditions, terrain, atmospheric effects, and nonlinear effects introduced by complex interactions among materials in the scene limit the ability of LMM to achieve high performance [22]. Many LMM-based methods attempt to introduce additional parameters to model SV, but their modeling capability under complex conditions lacks good generalization [23–26]. To address complex SV and nonlinearity, NLMM is an ideal solution. NLMM can be divided into model-based and model-free methods [27]. Model-based methods assume that the spectral mixing process is known a priori. A popular class of NLMM is the Bilinear Mixing Model (BMM), which simplifies nonlinear theory by assuming that light experiences at most two reflections of the illuminating radiation before reaching the detector. A major variant of this model is the Fan model [28], which performs poorly in scenarios with only linear interactions. To improve model generalization, the Generalized Bilinear Model (GBM) [29], the Linear-Quadratic Model (LQM) [8], and the Polynomial Post Nonlinear Mixing Model (PNMM) [30] are proposed, incorporating a hyperparameter to balance the weights of linear and nonlinear terms in the model. However, the mixing priors are often unknown in practical applications, leading to poor generalization and difficulties in model selection. Therefore, to improve model generalization, the development of model-free unmixing methods is necessary.

In recent years, the powerful learning and data fitting capabilities of deep learning have provided strong support for HU. The network architectures are primarily based on autoencoders and their variants, where HSIs are encoded into corresponding abundance fractions and decoded back into spectra, with the decoder weights representing endmembers. Most deep learning-based HU approaches focus primarily on pixel-wise unmixing, employing various regularizations to constrain the solution space. The mDAE [31] employs a nonnegative sparse autoencoder for unmixing and cascades a marginalized denoising autoencoder to mitigate the effects of noise. Recognizing that cascading introduces additional reconstruction errors, the uDAS [32] incorporates denoising ability as a denoising constraint into the network optimization process. To enhance the sparsity of estimations, EndNet [33] introduces a novel loss function incorporating a Kullback–Leibler divergence term with SAD similarity and several other penalty terms. In contrast to commonly used norm-based sparse priors, OSPAEU [34] observes that different abundance maps are nearly orthogonal, thus proposing an orthogonal sparse prior that achieves better abundance sparsity. Recently, several methods have integrated discriminative networks into their models, where structural distribution similarity is utilized to guide spectral reconstruction [35–37]. However, these methods are limited to pixel-level unmixing, despite ample evidence demonstrating the advantages of incorporating spatial information into the unmixing process.

Leveraging the convenience of neural network frameworks, autoencoder-based methods effectively exploit spatial features through convolutional layers [38]. CNNAEU [39] segments HSIs into a series of patches and extracts spatial information using 2D convolu-

tional neural networks (CNNs). DEAS [40] designs a plug-and-play extended-aggregated convolutional module, which extends the algorithm's spatial receptive field using dilated convolutions at different scales and demonstrates its effectiveness in enhancing the unmixing capabilities of CNNAEU. In fact, the targets exhibit varying scales and sizes, with pure pixels distributed throughout the entire HSI. Networks utilizing local convolutional filters overlook the global material distribution and long-range interdependencies, resulting in the loss of essential spatial feature information during the unmixing process. While MSNet [41] scales the original HSI to expand the receptive field of CNNs, downsampling operations lead to the loss of detailed information, making it challenging to balance the preservation of detailed information and the acquisition of comprehensive information. In contrast to the limited receptive fields of traditional CNNs, considering the non-local spatial correlations between hyperspectral pixels, employing self-attention mechanisms proves to be a viable solution. DeepTrans [42] pioneers the application of transformers [43] in HU, capturing non-local feature dependencies through interactions between image blocks. However, block-based operations introduce inconsistencies associated with patches. UST-Net [44] integrates the advantages of MSNet and DeepTrans, applying a multi-head self-attention mechanism (MHSAM) based on shifted windows to HSIs at different scales, enabling operations on the entire HSI and eliminating inconsistencies between patches. Nevertheless, due to computational constraints, the current non-local spatial correlations are still based on operations between blocks and cannot establish connections between pixels. Additionally, while the introduction of spatial information yields favorable endmember results, it often leads to excessive smoothing of abundance transitions. An ideal approach involves jointly extracting spatial-spectral information from HSI using 3D CNN, albeit at the cost of increased computational burden. Hence, it is common practice to either sequentially extract spatial and spectral information from HSI or employ dual-stream networks for joint extraction of spatial-spectral information. The former is exemplified by SSAE [45], where spatial information is initially utilized for effective endmember extraction, subsequently fixed into the decoder of the abundance estimation network. 1D CNN is then employed to extract the spectral features of HSI, facilitating more accurate abundance estimation. To achieve end-to-end learning, SSANet [46] incorporates an adaptive spectral-spatial attention module, sequentially comprising a spatial attention module and a spectral attention module. The latter typical method is DBA [47], which extracts spatial-spectral information through two branches and adjusts the weighting ratio of both as hyperparameters to regulate their impact on the unmixing results. SSCU-Net [48] and MSSS-Net [49] adopt weight-sharing mechanisms to enable interaction between the information streams, thereby reducing the selection of hyperparameters. Upon summarizing existing unmixing algorithms, it is observed that none fully account for both spatial and spectral information of HSI due to computational constraints, inevitably resulting in a decrease in unmixing performance. Furthermore, the aforementioned unmixing algorithms are all based on the LMM, comprising a meticulously designed encoder and a simple single-layer decoder. DAEU [50] experiments reveal that the simplistic structure of the decoder influences the performance of the autoencoder in reconstructing inputs, indicating that a single-layer decoder fails to fully exploit the robust capabilities of the encoder.

Linear unmixing can be easily addressed using classical methods, while deep learning demonstrates stronger competitiveness in tackling nonlinear problems [51,52]. NAE [53] reconfigures the decoder based on the PNMM [30] and leverages pre-training to enhance unmixing performance. Taking into account the higher degrees of freedom inherent in nonlinear neural networks, AEC [54] designed the encoder as the inverse of the mixing process, thereby enhancing the algorithm's robustness. UHUNA [55] designs three specific nonlinear models for the decoder while retaining the ability for further expansion, thereby improving algorithm versatility. RDAE [56] unfolds the GBM [29] to construct the decoder while extracting endmembers and their second-order scattering interactions. 3DAEU [57] jointly extracts spatial-spectral information of HSI using 3D CNN, with a carefully designed decoder covering several existing artificial models. Compared to linear methods, there are

relatively fewer algorithms developed based on NLMM. On one hand, existing nonlinear unmixing methods often confine themselves to specific mixing models. Developing data-driven model-free unmixing methods can effectively enhance model generalization. On the other hand, due to the inherent non-convexity of blind unmixing methods, the high degrees of freedom in nonlinear unmixing algorithms often generate a set of meaningless endmembers. Addressing the critical issue of setting appropriate initialization and regularization to guide algorithm convergence towards optimal solutions is a key consideration.

1.1. Motivation

While 1D CNN, 2D CNN and self-attention mechanism have been widely employed for feature extraction, none of these methods fully integrate global spatial information and spectral properties of HSI. Therefore, this paper utilizes a dual-stream network to separately extract spatial and spectral information of HSI, and global pixel-level contextual communication is achieved through a MHSAM based on linear projection, reducing the computational complexity from O(N^2) to O(N) without compromising unmixing performance. To fully harness the powerful feature fitting capability of the encoder, a data-driven nonlinear decoder is adopted. The nonlinear type is learned entirely from the data, enabling effective handling of various complex nonlinear scenarios. Considering the challenge of nonlinear decoders easily falling into local optima during model training, a stable initialization method is developed to effectively handle outliers and noise, significantly enhancing unmixing performance.

1.2. Novelty and Contribution

The main contributions of this article are summarized as follows:

- We propose a novel global spatial-spectral unmixing method, which integrates global spatial-spectral information in HSI, achieving pixel-level global spatial information interaction to reduce information loss and improve unmixing performance. Unlike conventional patch-based operations, to the best of our knowledge, this is the first application of pixel-level global attention mechanisms in HU, avoiding discontinuities between pixel blocks.
- We introduce a decoder structure suitable for nonlinear spectral unmixing. Compared to various nonlinear decoders designed for specific models, data-driven nonlinear decoders do not require the application of mixed priors of the scene, enabling adaptive handling of complex mixed images including linear and various nonlinear mixing scenarios.
- We propose a simple and efficient endmember initialization method to mitigate the interference from noise and outliers. Experimental results demonstrate that this method maintains high accuracy across various complex datasets. Moreover, this method can replace the commonly used VCA initialization directly applied to existing autoencoder unmixing algorithms, significantly enhancing unmixing performance.

This article is organized as follows. Section 2 provides a brief introduction to the mixing model and autoencoder structures. Section 3 elaborates on the proposed unmixing autoencoder framework, including the specific network architecture, component modules, and loss functions. Section 4 presents the experimental section, where performance comparisons are made with existing state-of-the-art unmixing methods. Finally, the conclusions of our work are presented in Section 5.

2. Problem Formulation

2.1. Data Model

Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N] \in \mathbb{R}^{D \times N}$ represent a HSI, where D is the number of spectral bands and N is the number of pixels. The LMM assumes that each pixel is a linear combination of pure materials, which can be formulated as:

$$\mathbf{Y} = \mathbf{AX} + \mathbf{R} \tag{1}$$

where $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_P] \in \mathbb{R}^{D \times P}$ represents an endmember matrix with P pure materials, $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{P \times N}$ is the corresponding abundance matrix, and $\mathbf{R} \in \mathbb{R}^{D \times N}$ is an additive noise matrix. Typically, both the endmember matrix and the abundance matrix are non-negative, and the columns of the abundance matrix sum to one, that is

$$\begin{aligned} \mathbf{A} \geq 0, \mathbf{X} \geq 0 \\ 1_R^{\mathrm{T}} \mathbf{X} = 1_N^{\mathrm{T}} \end{aligned} \tag{2}$$

where 1_R and 1_N are column vectors of ones with lengths R and N, respectively.

Although LMM is popular, it struggles to handle the nonlinearity occurring among different materials in real HSIs. In such cases, NLMM can better represent the unmixing process, with its general form as follows:

$$\mathbf{Y} = \phi(\mathbf{A}, \mathbf{X}) + \mathbf{R} \tag{3}$$

where ϕ defines the nonlinear interactions between endmembers. To extract interpretable endmembers and abundances from the model, most NLMMs decompose the nonlinear process into a linear term and a nonlinear perturbation term:

$$\mathbf{Y} = \mathbf{AX} + \psi(\mathbf{AX}) + \mathbf{R} \tag{4}$$

Moreover, the mixture model defined in Equation (4) offers better control over the degree of nonlinearity. PNMM [30] and MLM [58] are representative methods of such models.

2.2. Autoencoder

Deep learning-based unsupervised HU methods commonly employ an autoencoder structure, comprising an encoder and a decoder. The encoder compresses input data into a low-dimensional representation, representing abundance information in HSIs for unmixing tasks. Subsequently, the decoder reconstructs the original HSI based on the abundance information, with the decoder weights representing endmembers. The unsupervised training of the entire network is achieved by selecting appropriate objective functions. An overview of the overall network architecture is depicted in Figure 1a.

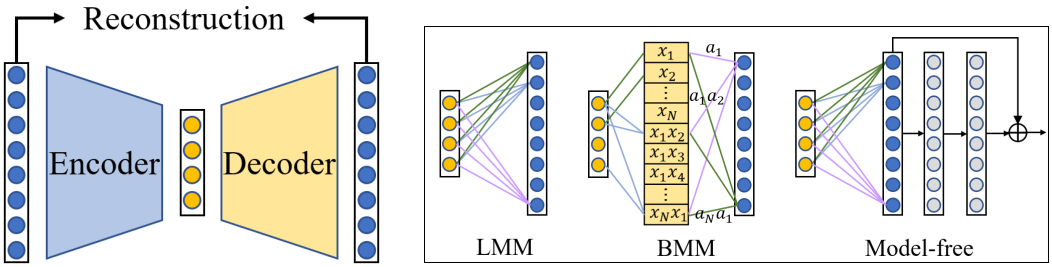


Figure 1. Schematic diagram of autoencoder architecture: (**a**) Autoencoder architecture. (**b**) Several common decoder architectures.

Figure 1b summarizes several commonly used decoding models in existing unmixing algorithms. Most unmixing algorithms are built upon the LMM, employing a single-layer linear decoder. The simplistic linear mixing assumption constrains the robust feature fitting capability of autoencoder networks. Consequently, several decoder architectures based on NLMM have been proposed, primarily categorized into model-based and model-free approaches. Model-based decoders strictly adhere to the underlying NLMM; methods like UHUNA [55] design decoders rigorously based on the BMM, yet such methods can only handle single mixing scenarios. In contrast, model-free approaches can simultaneously address multiple mixing scenarios, autonomously learning mixing forms from data via

autoencoder networks without requiring predefined mixing priors, thereby better handling various complex nonlinear scenarios.

3. Proposed Method

In this section, we elaborate on the proposed method. Similar to other blind unmixing methods, the proposed approach follows the structure of an autoencoder, comprising an encoder and a decoder, as illustrated in Figure 2. The entire HSI is input into the network, where spatial and spectral features of the HSI are initially extracted separately through a dual-stream network consisting of a Spatial Feature Extraction Module (SAFEM) and a Spectral Feature Extraction Module (SEFEM). Detailed explanations regarding these modules are provided in Section 3.1.

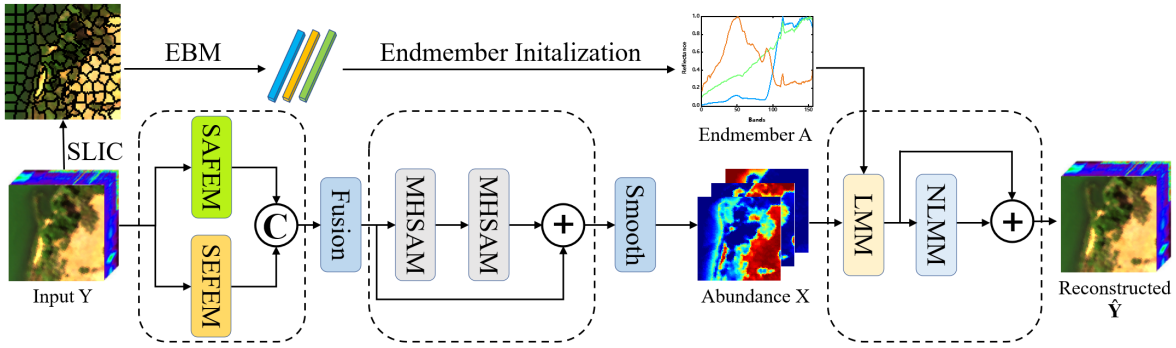


Figure 2. The architecture of the proposed AE network for hyperspectral unmixing.

Subsequently, the spatial and spectral features of HSI are fused together and then subjected to two consecutive MHSAMs for global spatial information interaction. The module has been redesigned to significantly reduce computational complexity, enabling pixel-wise feature interaction without the need for patch operations. Residual connections are utilized to accelerate the training process of deep neural networks, and finally, an abundance matrix is generated through a smoothing module. Specific implementation details are provided in Section 3.2.

Finally, considering that a simple LMM-based decoder underutilizes the powerful feature fitting capability of the encoder, the decoder is designed to consist of a linear term and a nonlinear perturbation term. The linear term based on LMM enables rapid fitting of the network, with the weights of the linear term corresponding to the endmember matrix. The nonlinear term is regarded as a perturbation of linear reconstruction, automatically learning nonlinear mixing patterns in HSI through a data-driven approach, thus enhancing the network's generalization across various mixing scenarios. More details on the decoder are provided in Section 3.3.

3.1. Spatial-Spectral Feature Extraction Module

HSIs typically exhibit high similarity between adjacent pixels, which can be effectively extracted using 2D CNN to capture local spatial information. However, this approach overlooks the rich spectral information in HSIs. While 1D CNN operates convolutions along the spectral dimension, it fails to consider the spatial properties of the image. On the other hand, 3D CNN can simultaneously extract local spectral-spatial features from HSIs, yet high parameter size in practical applications leads to increased computational costs and a risk of overfitting [59]. The direct application of 3D CNN for feature extraction incurs a time complexity of $O(M^2 \cdot K^2 \cdot L \cdot C_{in} \cdot C_{out})$, where M is the side length of each convolutional kernel output feature map, and K and L represent the kernel's spatial dimensions and depth, respectively. C_{in} and C_{out} denote the number of input and output channels for the convolutional layer. To balance computational costs and feature extraction capability, a dual-stream network is proposed, leveraging the strengths of 1D CNN and 2D CNN to separately extract spatial and spectral features from HSI. The time complexity of this

approach is O($M^2 \cdot K^2 \cdot C_{in} \cdot C_{out} + M^2 \cdot L \cdot C_{in} \cdot C_{out}$), representing approximately a L-fold reduction in computational complexity compared to direct application of 3D CNN.

The specific structure of the network is presented in Figure 3, with both streams adopting a four-layer network architecture. Dropout is employed in the first layer of the dual-stream network to prevent overfitting. Batch normalization is applied to each layer to alleviate the vanishing gradient problem and accelerate network training. In the SEFEM, max-pooling is used to enhance significant features in the HSIs and eliminate redundant spectral information. LeakyReLU is chosen as the activation function for the network, addressing the issue of "neuron death" associated with ReLU. Notably, in the spectral feature extraction module, both convolutional and max-pooling strides are typically set to 2. Minor adjustments may be applied to ensure consistent dimensions across different datasets.

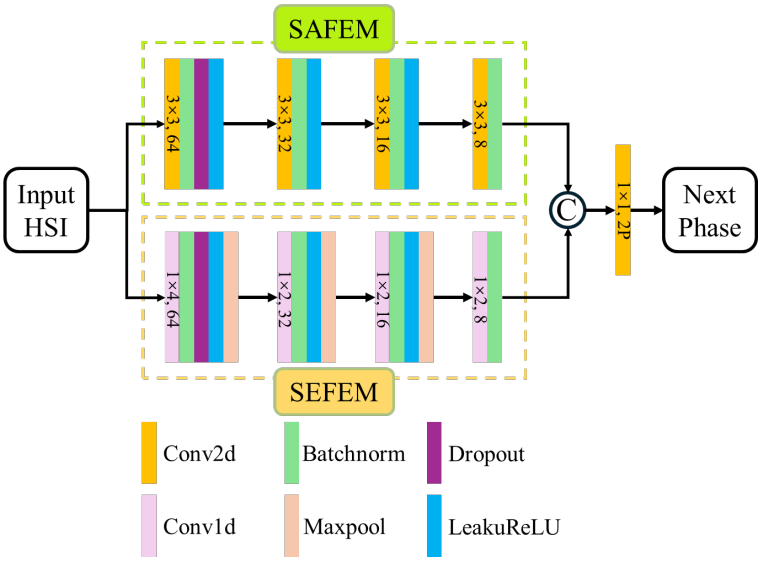


Figure 3. The architecture of the Spatial-Spectral Feature Extraction Module.

Finally, the extracted spectral-spatial features are concatenated together and fused using a 2D CNN. The output dimension of the network is twice the number of endmembers.

3.2. Global Spatial Information Interaction Module

The pure pixel distribution of HSI spans the entire image. Leveraging self-attention mechanisms facilitates substantial enhancement in the accuracy of unmixing algorithms by enabling global spatial information interaction. Inspired by Linformer [60], we adopt linear projection to reduce both the space and time complexity of network operations to O(N), achieving pixel-level information interaction within reasonable computational time.

The output matrix from the previous module is denoted as $\mathbf{X} \in \mathbb{R}^{N \times 2P}$. It is worth noting that we did not add positional encodings to the embedded representation $\mathbf{X}$ because the local spatial similarity of the pixels has already been extracted in the previous module. Adding positional information at this stage would lead to a decrease in performance. In this step, after the layer normalization process, $\mathbf{X}$ is first passed through three different linear layers to obtain embedding matrices $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V} \in \mathbb{R}^{N \times 2P}$:

$$\mathbf{Q} = \mathbf{X}\mathbf{W_q}, \quad \mathbf{K} = \mathbf{X}\mathbf{W_k}, \quad \mathbf{V} = \mathbf{X}\mathbf{W_v} \tag{5}$$

The scaled dot-product attention computation performed between them enables the model to capture the dependency relationships between any two pixels in the HSI, regardless of their spatial distance within the image:

$$\text{Attention} = \text{softmax}\left[\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{2P}}\right]\mathbf{V} \tag{6}$$

Scale factor $1/\sqrt{2P}$ is utilized to alleviate the gradient vanishing caused by the softmax function. And the time and space complexity resulting from the multiplication of several $(N \times 2P)$ matrices is $O(N^2)$. When implementing pixel-level attention mechanisms for the task of unmixing, the sequence length N can reach magnitudes of tens of thousands. To avoid costly computations, patch operations must be employed.

Linformer demonstrated that the self-attention matrix is low rank. By compressing the dimensions of inputs $\mathbf{K}$ and $\mathbf{V}$, significant reductions in computational complexity can be achieved. We employ the same linear projection matrix $\mathbf{L} \in \mathbb{R}^{P \times N}$ to act on both $\mathbf{K}$ and $\mathbf{V}$. After the projection is applied, the attention mechanism is recalculated as follows:

$$\text{Attention} = \text{softmax}\left[\frac{\mathbf{Q}(\mathbf{LK})^{\mathrm{T}}}{\sqrt{2P}}\right]\mathbf{LV} \tag{7}$$

Our projection dimension is only 128, which implies that our computational complexity decreases to $O(N)$, making it feasible to compute the global spatial correlation among pixels in HSI within a reasonable time. The justification of projection dimension compression will be validated in Section 4.4.

The specific module workflow is illustrated in Figure 4. In practical applications, the concept of multi-head self-attention is adopted, with a fixed number of heads denoted as P. This allows the model to simultaneously focus on different elemental information for individual pixels. The final output is as follows:

$$\text{head}_i = \text{Attention}_i = \text{softmax}\left[\frac{\mathbf{Q}_i(\mathbf{LK}_i)^{\mathrm{T}}}{\sqrt{2P}}\right]\mathbf{LV}_i \tag{8}$$

$$\mathbf{X'} = \text{Concat}(\text{head}_1, \text{head}_2, ..., \text{head}_P)\mathbf{W_o} \tag{9}$$

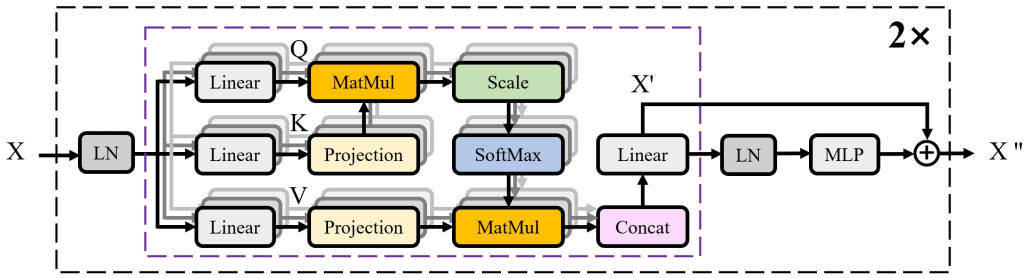


Figure 4. Module of Multi-Head Self-Attention Modules based on Linear Projection.

The output $\mathbf{X'}$ from the multi-head self-attention module is fed into a layer normalization layer and then passed through a multi-layer perceptron (MLP) block. The final output is obtained through residual connections:

$$\mathbf{X''} = \mathbf{X'} + \text{MLP}\big(\text{LN}(\mathbf{X'})\big) \tag{10}$$

Multiple preceding modules can be cascaded, where the cascade number is set as 2. The overall output dimension of the network is represented as follows:

$$\mathbf{X'''} = \mathbf{X} + \text{MHSAM}(\text{MHSAM}(\mathbf{X})) \tag{11}$$

$\mathbf{X'''} \in \mathbb{R}^{W \times H \times 2P}$, where W and H are the width and height of the HSI, respectively. After passing through the entire encoder, the global spatial-spectral information of the HSI is sufficiently extracted. And then, a 3×3 convolutional smoothing module is used to ensure a smooth overall distribution of the abundance map output and to compress the output dimension from $2P$ to P. A softmax layer is applied to the P dimension to ensure nonnegativity constraints as well as sum-to-one constraints, as described in Equation (2).

3.3. Unmixing with Decoder

The decoder consists of a linear component and a nonlinear perturbation component. The linear layer represents a single reflection of the illuminating solar radiation, which can be obtained through a simple linear combination of endmembers and abundances. The output of the linear component is as follows:

$$\mathbf{Y}_{linear} = \mathbf{E}^{(1)}\mathbf{X} \tag{12}$$

where $\mathbf{X} \in \mathbb{R}^{W\times H\times P}$ represents the abundance obtained through the encoder, $\mathbf{E}^{(1)}$ is defined as the weights of the first layer of the decoder, which are obtained using a 1×1 2D CNN without bias. Typically, the weights are initialized using methods such as VCA to accelerate the training of the network. $\mathbf{Y}_{linear} \in \mathbb{R}^{W\times H\times D}$ represents the linearly reconstructed HSI, which forms the main part of the network reconstruction.

The nonlinear component does not rely on existing mixture models but instead learns parameters automatically through the network. Research has shown that neural networks with two hidden layers can represent any arbitrary nonlinear relationship between inputs [61]. Therefore, we use a two-layer unbiased 2D CNN to construct the nonlinear component of the network. The spatial information of HSI can effectively improve the accuracy of abundance extraction in the encoder, but the network reconstruction in the decoder is limited to the pixel level. Among the literature reviewed, only CNNAEU assumes a new spectral-spatial model that considers the influence of neighboring pixels on the reconstruction in the decoder, but this model is still limited to LMM.

In this article, we assume that the influence of neighboring pixels should be included in the nonlinear reconstruction of any given pixel. Therefore, the first 2D CNN receptive field in the nonlinear module is set to 3×3, while the second remains at 1×1. The final reconstruction process of $\hat{\mathbf{Y}}$ is as follows:

$$\mathbf{Y}_{nonlinear} = \mathbf{E}^{(2)}\mathbf{Y}_{linear} \tag{13}$$

$$\hat{\mathbf{Y}} = \mathbf{Y}_{linear} + \mathbf{Y}_{nonlinear} \tag{14}$$

where $\mathbf{E}^{(2)}$ represents the weight of the nonlinear perturbation term. To avoid overfitting, the nonlinear network is pretrained with the same input and output before training begins. Therefore, during training, the nonlinear part can be regarded as a perturbation of the HSI.

3.4. Overall Loss Function

The Spectral Angle Distance (SAD), which exhibits spectral scale invariance, is chosen as the primary reconstruction function. It evaluates the similarity between two spectral curves by calculating the angle between the target spectrum and the reference spectrum. A smaller angle between two spectral indicates a greater similarity. The formula for calculating SAD is as follows:

$$\mathcal{L}_{\text{SAD}} = \frac{1}{W \cdot H}\sum_{i=1}^{W}\sum_{j=1}^{H}\arccos\left(\frac{\langle \mathbf{Y}_{ij}, \hat{\mathbf{Y}}_{ij}\rangle}{||\mathbf{Y}_{ij}||_2||\hat{\mathbf{Y}}_{ij}||_2}\right) \tag{15}$$

Considering only scale invariance may lead to significant errors, hence Mean Square Error (MSE) is introduced into the network to ensure the stability of unmixing.

$$\mathcal{L}_{\text{MSE}} = \frac{1}{W \cdot H}\sum_{i=1}^{W}\sum_{j=1}^{H}\left(\hat{\mathbf{Y}}_{ij} - \mathbf{Y}_{ij}\right)^2 \tag{16}$$

Given that adjacent pixels have similar abundance vectors, Total Variation (TV) loss is added to the abundance results to impose spatial smoothness. The TV loss function is defined as:

$$\mathcal{L}_{\mathrm{TV}} = \frac{1}{W \cdot H \cdot P} \sum_{i=1}^{W} \sum_{j=1}^{H} \sum_{k=1}^{P} \sqrt{\left(\mathbf{X}_{i,j+1,k} - \mathbf{X}_{i,j,k}\right)^2 + \left(\mathbf{X}_{i+1,j,k} - \mathbf{X}_{i,j,k}\right)^2} \tag{17}$$

To mitigate overfitting caused by nonlinear functions, L_2-norm is employed to constrain the weights of the nonlinear term:

$$\mathcal{L}_{\mathrm{nl}} = ||\mathbf{E}^{(2)}||^2 \tag{18}$$

In our model, the overall training loss function comprises the following four components:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{SAD}} + \lambda_1 \mathcal{L}_{\mathrm{MSE}} + \lambda_2 \mathcal{L}_{\mathrm{TV}} + \lambda_3 \mathcal{L}_{\mathrm{nl}} \tag{19}$$

λ_1, λ_2 and λ_3 represent the coefficients of the regularization terms, which are uniformly set to 1×10^{-3} in the Samson dataset and Jasper Ridge dataset, and uniformly set to 1×10^{-4} in the Urban dataset. It is worth noting that, for a fair comparison with competitive methods in experiments, we did not fine-tune the regularization coefficients based on different datasets. Therefore, fine-tuning of hyperparameters for different datasets may lead to better results.

To sum up, the pseudocode of the proposed method is shown in Algorithm 1.

Algorithm 1 Pseudocode of the Proposed Method.

Input: HSI **Y**;
Output: Endmembers **A**;
Abundance **X**;

Initialize **A** by the VCA-bundles algorithm;
Initialize $\mathbf{Y}_{nonlinear}$;
Training stage:
for *Epochs* **do**
  Update X using Equation (11)
  Update A using Equation (14)
  Compute Loss $\mathcal{L}_{\mathrm{total}}$ using Equation (19)
  Back propagation
end for
Test stage:
Forward propagation: Feed **Y** into the network;
Obtain **A** and **X**

4. Experiments

In order to better evaluate the proposed method, detailed ablation experiments were conducted on the dual-branch spatial-spectral feature extraction module and the nonlinear decoder within the network. The proposed method was compared with several representative existing methods on three real datasets.

The mean spectral angle distance (mSAD) was used to evaluate the quality of the extracted endmembers, while the mean root mean square error (mRMSE) was utilized to assess the accuracy of the abundance estimation, which are defined as follows:

$$\mathrm{mSAD} = \frac{1}{P} \sum_{i=1}^{P} \arccos\left(\frac{\langle \mathbf{A}_i, \hat{\mathbf{A}}_i \rangle}{||\mathbf{A}_i||_2 ||\hat{\mathbf{A}}_i||_2}\right) \tag{20}$$

$$\mathrm{mRMSE} = \sqrt{\frac{1}{W \cdot H \cdot P} \sum_{i=1}^{W} \sum_{j=1}^{H} \sum_{k=1}^{P} \left(\hat{\mathbf{X}}_{ijk} - \mathbf{X}_{ijk}\right)^2} \tag{21}$$

The proposed method is implemented in Python 3.9 using the PyTorch framework. The network's learning rate is initialized to 1×10^{-3}, while the linear decoder is initialized to 5×10^{-4}. Every 30 epochs of training, the learning rate is decayed by a factor of 0.8. The training epochs for different datasets were adjusted, and the network was run for 200, 400, and 90 epochs respectively for the Samson dataset, Jasper Ridge dataset, and Urban dataset. The decoder weights are initialized based on the endmembers extracted in Section 4.2.

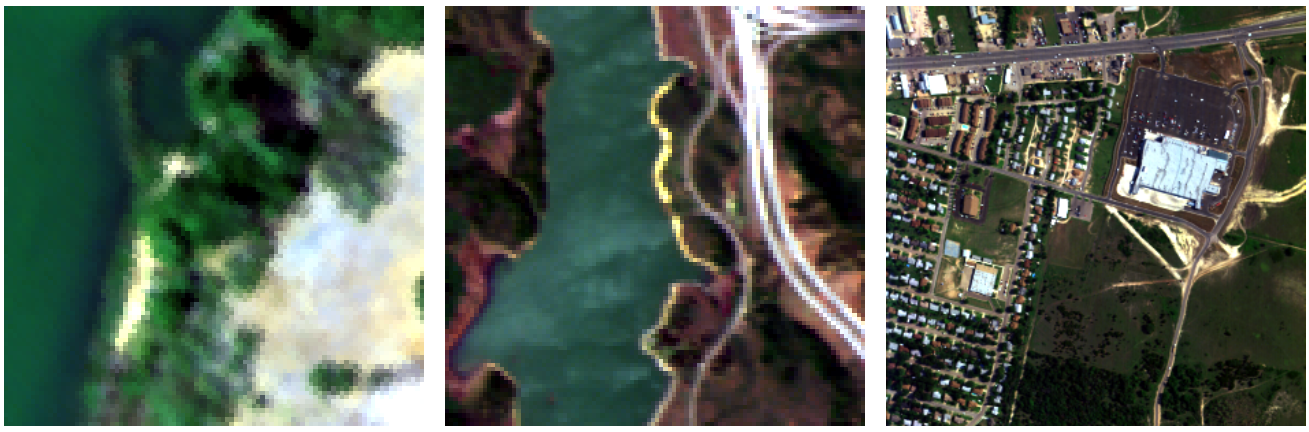
4.1. Data Description

Due to the challenge of synthetic datasets in reflecting the complex nonlinear interactions in real-world scenarios, only three widely used real datasets were employed to validate the unmixing results of different algorithms, as illustrated in Figure 5.

Samson contains 95×95 pixels, with each pixel encompassing 156 spectral channels ranging from 401 nm to 889 nm. This dataset is not affected by bad pixels or severe noise contamination and consists of three endmembers: Soil, Tree, and Water.

Jasper Ridge comprises 100×100 pixels, with each pixel containing 198 spectral channels ranging from 380 nm to 2500 nm after the removal of channels 1–3, 108–112, 154–166, and 220–224 due to dense water vapor and atmospheric effects. Four endmembers are included in this dataset: Road, Soil, Tree, and Water.

Urban is a large dataset consisting of 307×307 pixels, with each pixel containing 162 spectral channels spanning from 400 nm to 2500 nm after the removal of channels 1–4, 76, 87, 101–111, 136–153, and 198–210 due to dense water vapor and atmospheric effects. This dataset contains a significant number of outliers and encompasses four endmembers: Asphalt, Grass, Tree, and Roof.



(**a**) (**b**) (**c**)

Figure 5. Dataset: (**a**) Samson dataset. (**b**) Jasper Ridge dataset. (**c**) Urban dataset.

4.2. Endmember Initialization

The unmixing problem can be formulated as a non-convex minimization problem, and reasonable initialization can significantly improve the unmixing performance. Most unmixing algorithms use geometric extraction algorithms such as VCA for initialization, but the existence of outliers and noise may lead to the extraction of meaningless endmembers, which strongly hinder the unmixing process. DAEN [62] combines stacked autoencoders and VCA to generate well-initialized endmembers, eliminating the influence of outliers, but it is computationally expensive. OSPAEU [34] removes outliers by measuring the uniformity of neighboring pixels over the entire image. MAAENet [51] proposes an SLIC-VCA algorithm, which generates spatial groups through image segmentation [63]. The spectral within the same group are averaged to alleviate the impact of outliers and noise. We further optimize the SLIC-VCA algorithm by following the concept of endmember bundles.

The specific workflow is illustrated in Figure 6. HSI exhibits similar spectral characteristics within compact spatial neighborhoods. Under the assumption that pure pixels

do not exist independently, SLIC [63] is employed to segment the HSI. Notably, SLIC clusters pixels by considering both spatial Euclidean distances and spectral similarity. The formulation is given by:

$$\mathbf{SLIC}_{\text{feature}} = \sqrt{\frac{\Delta x_{ij}^2 + \Delta y_{ij}^2}{S^2} + \frac{||\mathbf{y}_i - \mathbf{y}_j||^2}{m^2}} \tag{22}$$

where $\Delta x_{ij}^2 + \Delta y_{ij}^2$ is the squared Euclidean distance between two pixels, S is the search size of SLIC, and m is a hyperparameter balancing pixel distance and spectral similarity strength.

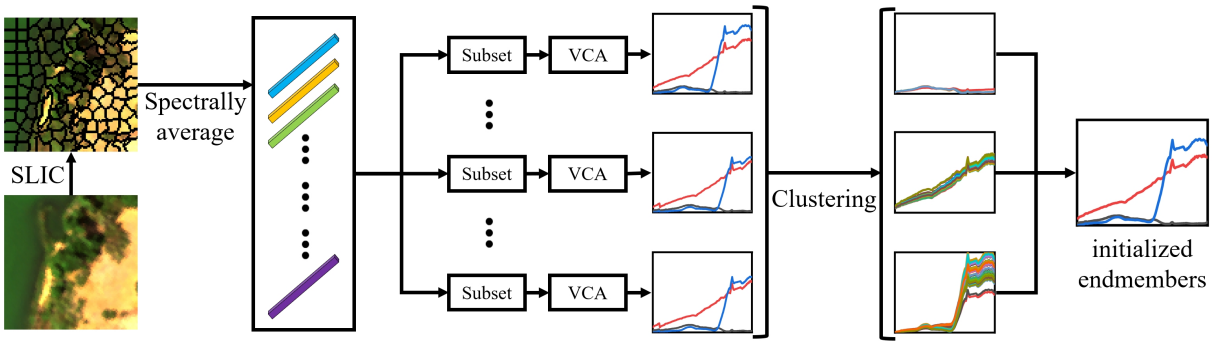


Figure 6. The flowchart of the proposed endmember initialization method.

After SLIC segmentation, the HSI is divided into highly correlated sub-pixel blocks. Each sub-pixel block undergoes averaging, effectively eliminating outliers and reducing noise. Following the averaging of spectral collections, based on the concept of endmember bundles, a subset of the spectral collection is randomly selected to run VCA. This approach assumes that a small percentage of image pixels can approximate the original image statistics. Each run of the endmember extraction algorithm yields a different set of endmember spectra.

Next, the extracted spectral library undergoes k-means clustering based on Euclidean distance as a similarity metric. This partitions the spectral library into independent endmember bundles for each ground component, characterizing each endmember with a set of spectra exhibiting spectral variability. Finally, averaging is applied to each group of endmember bundles to obtain the desired initial endmembers.

Utilizing the concept of endmember bundles instead of directly applying the endmember extraction algorithm on the spectral collection offers several advantages. Firstly, running VCA across the entire image typically yields inconsistent results, whereas averaging over multiple samplings minimizes uncertainties. This ensures that even if some VCAs extract incorrect spectra, the final endmembers remain unaffected. Secondly, averaging a set of endmember bundles containing various spectral variabilities further mitigates the impact of outliers and noise, achieving accurate and reliable endmember initialization.

The challenging Urban dataset is employed for comparative experiments in endmember extraction, with VCA [12], NFINDR [11], and SLIC-VCA [51] selected as the benchmark methods. Figure 7 illustrates the visual results of endmember extraction by different algorithms, while Table 1 quantitatively lists the performance of all compared methods. Conventional geometric extraction methods are adversely affected by outliers, resulting in the poorest outcomes. SLIC-VCA, by locally averaging spectral signatures, effectively mitigates the impact of outliers, significantly enhancing the accuracy of endmember extraction. However, it fails to accurately differentiate highly similar endmember spectra such as “Grass” and “Tree”. In comparison, VCA-bundles accurately extracts all endmembers, demonstrating robust adaptability to various complex natural scenes compared to other methods.

To further investigate the robustness of VCA-bundles, Figure 8 visualizes the results of endmember bundle extraction. It can be observed that for each reference endmember,

a set of spectral subsets with varying spectral variabilities is generated, and averaging operations help reduce the influence of spectral variability. Moreover, for endmembers that are difficult to extract, such as "Grass", partial erroneously extracted endmembers are eliminated through averaging, thereby yielding robust results.

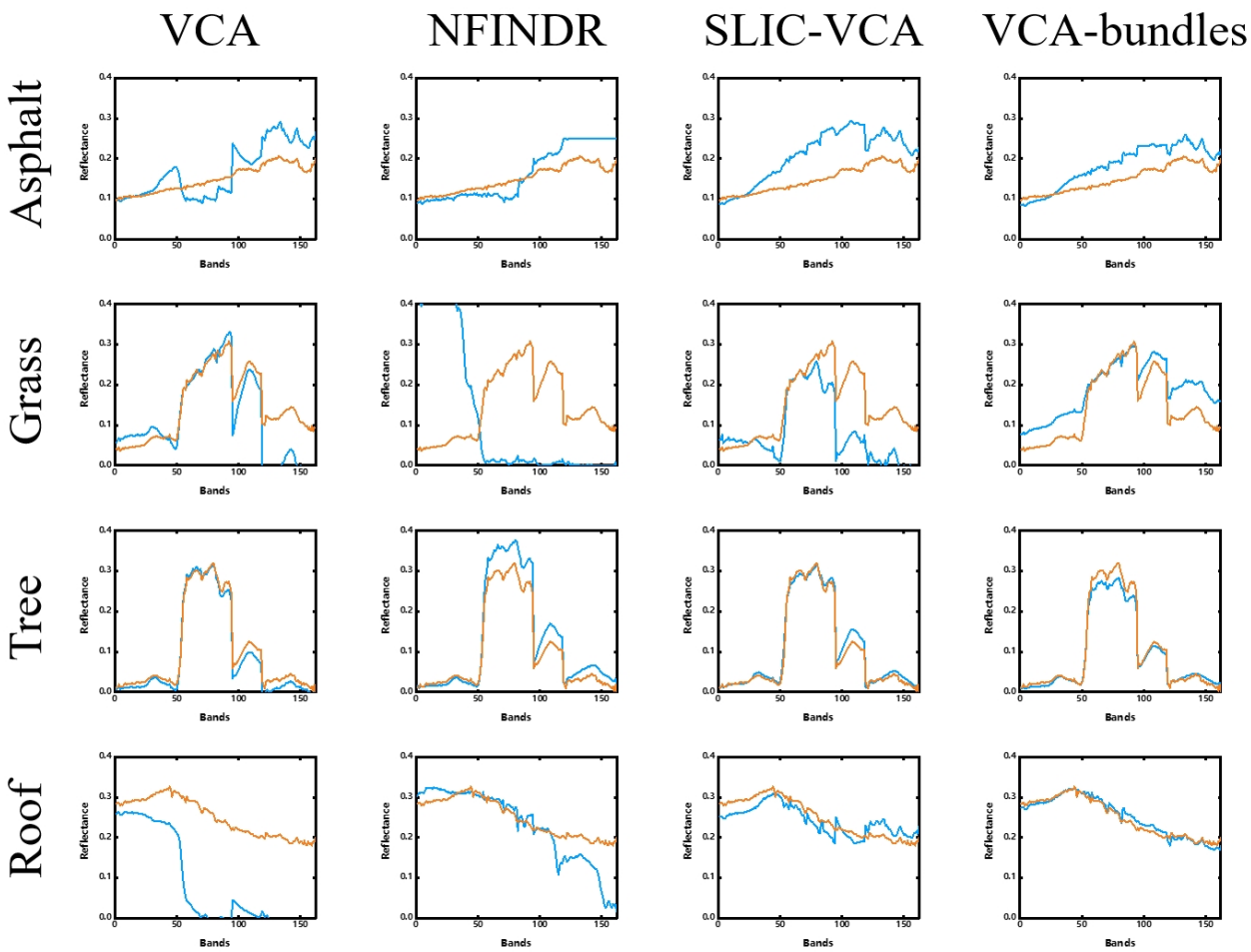


Figure 7. The results of endmember extraction (Urban dataset): extracted endmembers (blue) and actual endmembers (orange).

Table 1. The SAD results of each comparative algorithm on the Urban dataset.

Urban		VCA	NFINDR	SLIC-VCA	VCA-Bundles
SAD	Asphalt	20.95	19.10	16.29	9.98
	Grass	41.33	136.92	54.02	22.41
	Tree	10.32	7.44	8.07	4.81
	Roof	82.66	21.74	8.89	4.04
Mean SAD		38.82	46.30	21.98	10.31

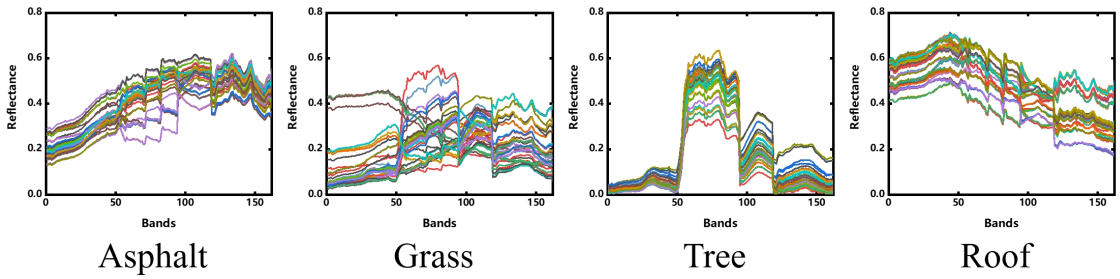


Figure 8. Visualization results of endmember bundle extraction (Urban dataset).

4.3. Ablation Experiments

The effectiveness of the proposed branch modules was analyzed by systematically removing the SAFEM, SEFEM, or NLMM decoder module individually from three datasets until all were removed. When the Spatial-Spectral Feature Extraction Module was entirely removed, the encoding module was substituted with a 2D CNN incorporating 1×1 spatial channel attention. Table 2 summarizes the outcomes of endmember extraction and abundance estimation under varied conditions. As anticipated, complete removal of all modules resulted in the poorest performance. When only one module was reintroduced, augmenting the encoder with spatial or spectral information proved more advantageous. The independent use of the nonlinear decoder integrating local spatial information yielded only marginal improvements; however, when paired with SAFEM, which also extracts local spatial information, a significant enhancement in unmixing accuracy was observed. Ultimately, simultaneous utilization of all three modules yielded optimal results, underscoring the efficacy of the proposed dual-branch encoder and nonlinear decoder.

Table 2. The mSAD and mRMSE results after conducting ablation experiments on the branch modules. Best result are bold.

Ablation Modules			Samson		Jasper Ridge		Urban	
SAFEM	**SEFEM**	**NLMM**	**mSAD**	**mRMSE**	**mSAD**	**mRMSE**	**mSAD**	**mRMSE**
✗	✗	✗	4.04	13.67	4.12	10.94	8.01	11.87
✔	✗	✗	2.83	11.19	4.01	10.85	7.92	11.69
✗	✔	✗	3.49	11.56	4.02	10.82	7.82	11.79
✗	✗	✔	3.72	13.10	4.19	10.98	7.90	11.74
✔	✔	✗	2.57	8.60	3.95	9.24	**7.74**	11.56
✔	✗	✔	2.62	8.91	3.99	9.66	7.79	11.62
✗	✔	✔	3.20	12.65	4.13	10.60	7.96	12.26
✔	✔	✔	**2.25**	**7.57**	**3.93**	**8.99**	7.75	**11.46**

In addition, a highlight of this study is the proposed MHSAM based on linear projection, facilitating pixel-level global spatial interaction. Ablation experiments were conducted using linear layers, CNN layers, and a 5×5 MHSAM based on patch-based operations as substitutes for this module on the Samson and Jasper Ridge datasets, as illustrated in Table 3. The results indicate that using solely linear layers yielded the poorest performance, while CNN layers considering local spatial information offered moderate improvement. In contrast, employing MHSAM with both methods comprehensively integrating global spatial information resulted in significant enhancements in endmember accuracy. Figure 9 illustrates abundance visualizations of tree, water, dirt, and road on the Jasper Ridge dataset, demonstrating that MHSAM without patch-based processing achieved the smoothest pixel transitions and best visual performance.

Table 3. The mSAD and mRMSE results after conducting ablation experiments on the MHSAM modules. Best result are bold.

	Samson		Jasper Ridge	
	mSAD	**mRMSE**	**mSAD**	**mRMSE**
Linear	5.64	13.58	12.63	10.48
CNN	5.35	11.07	8.68	10.24
MHSAM (5×5)	2.95	11.70	4.40	9.53
Proposed	**2.25**	**7.57**	**3.93**	**8.99**

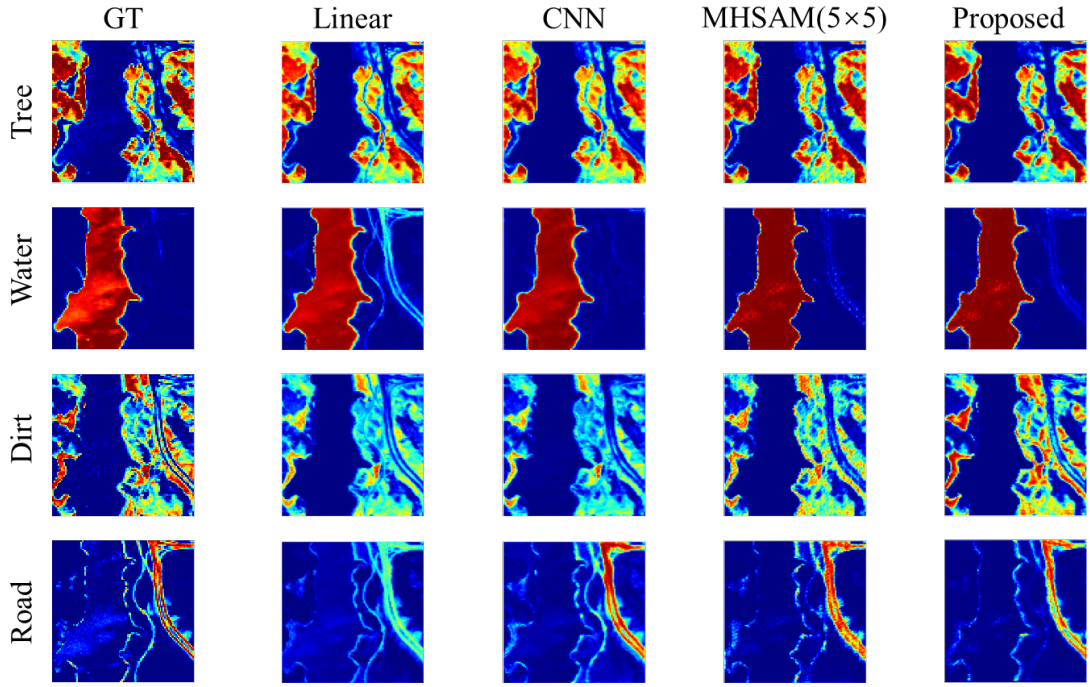


Figure 9. Abundance maps of tree, water, dirt, and road on the Jasper Ridge dataset obtained by different modules.

4.4. Projection Dimension Analysis

The projection dimension in the global spatial information interaction module is an optional hyperparameter, as demonstrated in Linformer [60] where a lower projection dimension leads to faster network training speed. However, it may also result in a decrease in network performance. Tests were conducted on different projection dimensions using the Samson dataset and Jasper Ridge dataset, with specific results shown in Figure 10. Worth noting is that the network performance does not significantly deteriorate as the projection dimension decreases, which could be attributed to the relatively simple nature of the unmixing task. On the other hand, a more noticeable decrease is observed in computation time, which converges when the projection dimension is below 128. Therefore, the projection dimension is set to 128 to minimize information loss while maintaining computational efficiency.

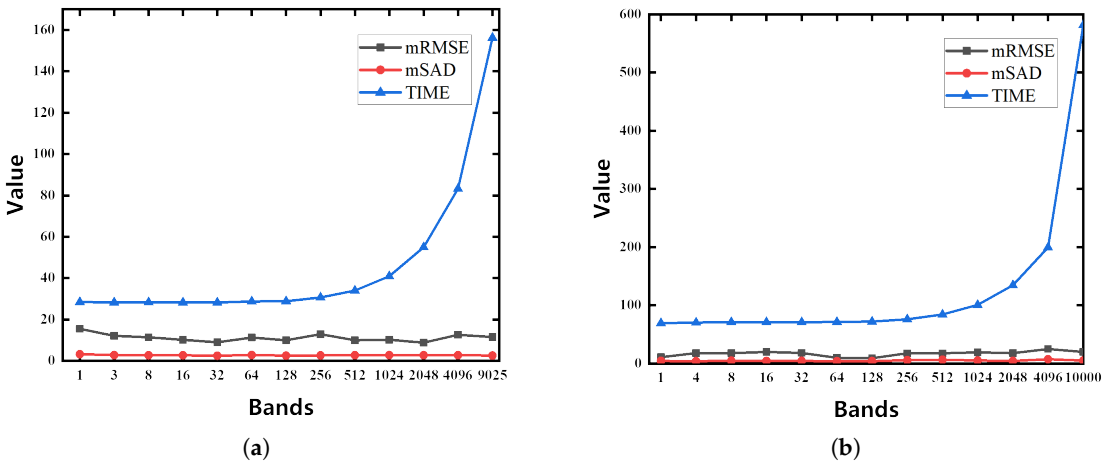


Figure 10. The results of mRMSE and mSAD under different projection dimensions, along with the corresponding computation times (measured in seconds). (**a**) Samson dataset. (**b**) Jasper Ridge dataset.

4.5. Experiments

In this section, we conducted comparative experiments with other methods. We took into consideration both linear and nonlinear methods for method selection, as the three chosen datasets are widely used in various linear unmixing algorithms. The comparative methods are as follows:

(1) FCLSU [13]: The most commonly used abundance estimation method. In our experiments, VCA-bundles were used as the endmember extraction method in conjunction with FCLSU.
(2) DeepTrans [42]: A linear unmixing network based on deep learning, which captures nonlocal feature dependencies through operations between image patches.
(3) uDAS [32]: A linear unmixing network based on deep learning, with denoising capability incorporated into network optimization in the form of denoising constraints.
(4) SGSNMF [19]: A linear unmixing network based on NMF, where the group-structured prior information of HSI is integrated into nonnegative matrix factorization optimization, with data organized into spatial groups.
(5) NAE [53]: A nonlinear unmixing network based on deep learning, trained through pixel-wise network.
(6) rNMF [64]: A nonlinear unmixing network based on NMF, with an additional term introduced in the model to consider nonlinear effects.
(7) 3DAEU [57]: A nonlinear unmixing network based on deep learning, capturing spatial-spectral information of HSI through 3DCNN, with the design of the nonlinear model encompassing several existing artificial models.
(8) A2SAN [65]: A linear unmixing network based on deep learning, utilizing spectral and spatial modules to extract spatial-spectral information of HSI, and employing attention mechanisms for direct reconstruction.
(9) USTNet [44]: A linear unmixing network based on deep learning, employing multi-head self-attention blocks based on shifted windows to extract HSI feature maps at different scales, minimizing loss of detailed information.

All comparative methods were independently run ten times on each dataset. The subsequent evaluation calculated the mean and standard deviation for each method.

4.5.1. Samson Dataset

The Table 4 presents the abundance RMSE and endmember SAD obtained by different unmixing methods on the Samson dataset. It is observed that all unmixing methods can accurately extract the endmembers "Soil" and "Tree", but most encounter difficulties in extracting "Water", possibly due to its low reflectance which makes it difficult to distinguish subtle differences in the loss function. In comparative experiments, only USTNet and the proposed method accurately extract the "Water" endmember, suggesting that the introduction of global spatial information aids in endmember extraction, a finding reinforced by [45]. FCLSU employing VCA-bundles as the endmember extraction method achieves suboptimal mean SAD results, further demonstrating the advantage of the proposed initialization method. Regarding abundance estimation, it is evident that the proposed method and A2SAN significantly outperform other methods in accuracy, highlighting the advantage of using scale-invariant SAD as a loss function in conjunction with spectral information extraction. Visual results of abundances and endmembers are shown in Figures 11 and 12, respectively. Comparisons indicate that deep learning methods exhibit significant advantages over traditional methods in both endmember extraction and abundance estimation, with recent approaches yielding abundance maps closest to ground truth.

4.5.2. Jasper Ridge Dataset

In this section, all the compared unmixing methods are applied to the Jasper Ridge dataset. Table 5 quantitatively demonstrates the performance of all competing methods, while the visualization results of abundance and endmembers are respectively presented in

Figures 13 and 14. NAE and rNMF confused "Dirt" and "Road", resulting in the poorest performance. Although 3DAEU successfully extracted all endmembers, it completely ignored "Road" in abundance estimation. This indicates that nonlinear methods with more parameters are prone to falling into local minima, and may require more prior information and careful hyperparameter tuning. Many linear methods performed better in endmember extraction on this dataset, while DeepTrans, A2SAN and USTNet were affected by initialization and interfered with the extraction of "Road". FCLSU initialized by VCA-bundles achieved suboptimal performance, further highlighting that more accurate endmember initialization might bring greater benefits compared to sophisticated algorithm design. In the observation of the abundance map, it is evident that the majority of comparative algorithms did not accurately separate the roads, whereas the proposed method fully considered the spatial and spectral information in HSI, achieving better road separation. Demonstrating its stable processing capabilities across different datasets, our method consistently achieved the best mSAD and mRMSE, although it may not have outperformed others in individual comparisons.

4.5.3. Urban Dataset

Table 6 presents the abundance RMSE and endmember SAD achieved by different unmixing methods on the Urban dataset. The Urban dataset is highly complex, with a considerable number of outliers. Most methods face challenges distinguishing between "Grass" and "Tree", with uDAS and SGSNMF completely unable to differentiate between the two. Methods incorporating spatial information such as DeepTrans, 3DAEU, and A2SAN perform well in extracting most endmembers, while the proposed method and USTNet achieve optimal mSAD in two specific endmembers. Influenced by endmember extraction challenges, Figure 15's abundance visualization results show suboptimal performance for most methods, whereas ASAN, USTNet, and the proposed method, proposed in the last two years, integrate spatial-spectral information extraction and self-attention mechanisms, achieving the best results. Visual results of endmembers are shown in Figure 16.

4.6. Processing Time

Table 7 presents the running times of all methods on three datasets. Specifically, FCLSU, uDAS, SGSNMF and rNMF were implemented in Matlab (2022a), while the remaining methods were implemented in PyCharm (2022). The experiments were conducted on a computer equipped with an Intel Core i5-13600KF processor, 32 GB of memory, and an NVIDIA GeForce RTX 2080 Ti graphics processing unit. On the Samson and Jasper Ridge datasets, the runtime of all methods remained within approximately one hundred seconds, except for USTNet, which did not utilize GPU acceleration. The Urban dataset, being larger, particularly saw a sharp increase in computation time for the 3D CNN-based method 3DAEU. The proposed method, leveraging global spatial-spectral information extraction, achieved a runtime lower than that of pixel-wise methods like uDAS, thus maintaining acceptable computational costs.

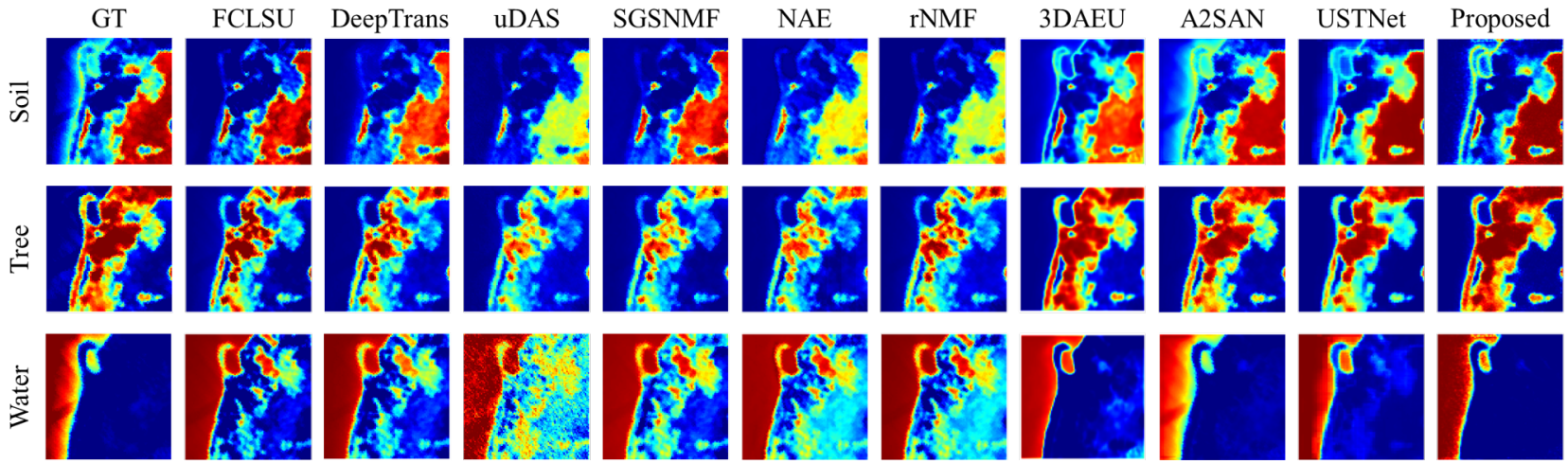


Figure 11. Abundance maps of soil, tree, water on the Samson dataset obtained by different methods.

Table 4. Valuation metrics SAD and RMSE results of Samson dataset ($\times 10^{-2}$). Best results are bold.

Samson		FCLSU	DeepTrans	uDAS	SGSNMF	NAE	rNMF	3DAEU	A2SAN	USTNet	Proposed
SAD	Soil	1.67 ± 0.06	2.49 ± 0.40	3.12 ± 0.09	1.63 ± 0.11	2.08 ± 0.03	3.52 ± 0.01	1.18 ± 0.06	2.26 ± 0.21	**1.05 ± 0.02**	1.46 ± 0.03
	Tree	3.71 ± 0.09	4.93 ± 0.24	5.44 ± 0.36	6.04 ± 0.38	4.96 ± 0.03	8.26 ± 0.01	**2.97 ± 0.02**	4.07 ± 0.04	3.35 ± 0.03	3.15 ± 0.02
	Water	9.85 ± 0.12	8.81 ± 0.56	13.93 ± 1.15	23.29 ± 0.30	13.31 ± 0.14	23.73 ± 0.02	24.26 ± 0.43	13.36 ± 0.37	2.47 ± 0.03	**2.15 ± 0.13**
Mean SAD		5.08 ± 0.06	5.41 ± 0.34	7.49 ± 0.49	10.32 ± 0.03	6.78 ± 0.05	11.84 ± 0.01	9.49 ± 0.17	6.56 ± 0.08	2.29 ± 0.02	**2.25 ± 0.04**
RMSE	Soil	17.52 ± 0.04	16.40 ± 0.33	25.29 ± 0.84	20.12 ± 0.27	23.01 ± 1.95	25.87 ± 0.00	11.65 ± 0.05	**7.63 ± 0.15**	8.54 ± 0.13	9.20 ± 0.10
	Tree	16.28 ± 0.14	17.35 ± 1.03	25.29 ± 0.82	25.56 ± 0.55	22.19 ± 1.60	20.47 ± 0.00	6.45 ± 0.06	**4.77 ± 0.08**	11.11 ± 0.10	7.45 ± 0.07
	Water	28.29 ± 0.10	28.28 ± 0.96	41.37 ± 0.68	37.62 ± 0.27	36.07 ± 1.92	37.34 ± 0.00	10.03 ± 0.11	6.07 ± 0.18	9.37 ± 0.08	**5.64 ± 0.16**
Mean RMSE		21.34 ± 0.07	20.68 ± 0.77	30.65 ± 0.57	27.76 ± 0.25	27.85 ± 1.43	28.77 ± 0.00	9.62 ± 0.04	**6.16 ± 0.09**	9.73 ± 0.09	7.57 ± 0.09

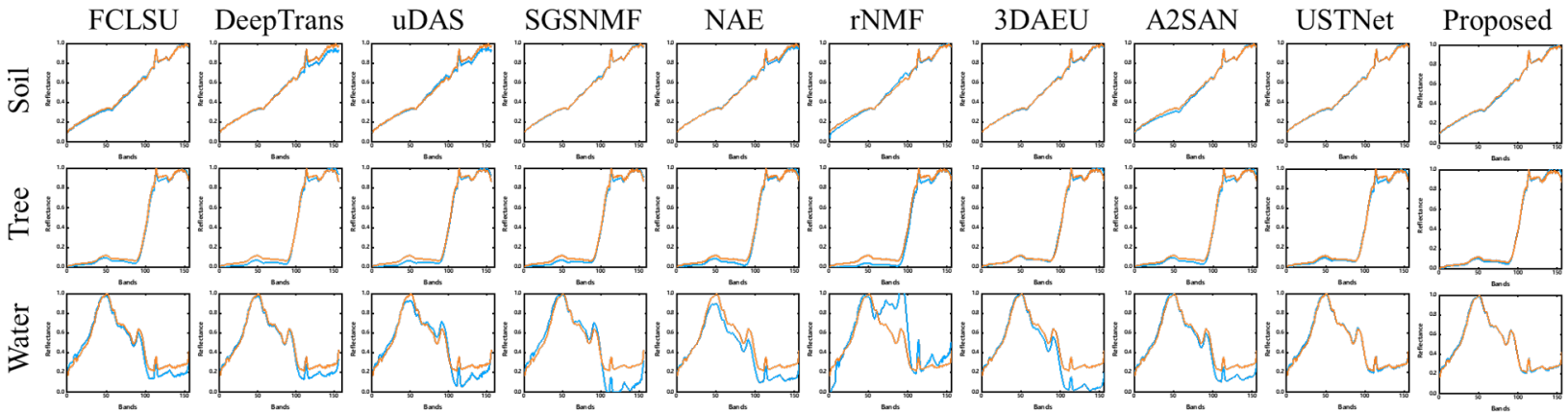


Figure 12. Extracted endmember comparison between the different algorithms and the corresponding GTs in the Samson dataset.

Table 5. Valuation metrics SAD and RMSE results of Jasper Ridge dataset ($\times 10^{-2}$). Best results are bold.

Jasper Ridge		FCLSU	DeepTrans	uDAS	SGSNMF	NAE	rNMF	3DAEU	A2SAN	USTNet	Proposed
SAD	Tree	9.40 ± 0.20	5.42 ± 2.01	14.90 ± 1.76	13.71 ± 0.38	26.32 ± 0.06	24.94 ± 0.01	7.77 ± 0.93	11.41 ± 1.88	4.86 ± 0.17	**4.03 ± 0.01**
	Water	18.08 ± 1.12	11.18 ± 2.76	9.58 ± 1.63	21.27 ± 1.95	29.60 ± 0.28	28.69 ± 0.01	22.59 ± 2.68	12.45 ± 3.31	**3.76 ± 0.03**	5.15 ± 0.04
	Dirt	5.52 ± 0.28	6.24 ± 0.64	13.98 ± 5.00	13.94 ± 3.71	22.40 ± 0.02	5.49 ± 0.00	7.31 ± 0.50	11.30 ± 1.12	18.35 ± 0.29	**2.42 ± 0.03**
	Road	4.77 ± 0.30	16.75 ± 1.39	5.85 ± 0.20	**4.07 ± 0.52**	54.30 ± 0.24	70.20 ± 0.02	4.62 ± 0.19	17.70 ± 1.76	10.00 ± 0.24	4.14 ± 0.12
Mean SAD		9.45 ± 0.31	9.90 ± 1.55	11.08 ± 0.72	13.25 ± 0.91	33.16 ± 0.09	32.33 ± 0.01	10.57 ± 0.83	13.21 ± 1.23	9.24 ± 0.14	**3.93 ± 0.03**
RMSE	Tree	9.29 ± 0.17	8.21 ± 0.55	16.16 ± 0.07	13.36 ± 0.44	18.89 ± 1.71	14.21 ± 0.01	**7.06 ± 0.43**	14.28 ± 1.94	12.05 ± 0.23	7.68 ± 0.09
	Water	9.03 ± 0.03	6.3 ± 0.52	19.91 ± 0.58	17.34 ± 0.67	8.42 ± 0.87	8.01 ± 0.00	6.64 ± 0.98	8.17 ± 1.05	**6.20 ± 0.05**	6.40 ± 0.03
	Dirt	**11.09 ± 0.15**	20.41 ± 0.90	12.58 ± 0.24	12.15 ± 0.31	25.61 ± 1.43	26.03 ± 0.01	23.34 ± 2.69	14.31 ± 1.17	23.51 ± 0.35	11.41 ± 0.08
	Road	**6.46 ± 0.12**	19.85 ± 1.00	12.45 ± 0.56	12.04 ± 0.32	21.70 ± 0.60	24.92 ± 0.00	22.76 ± 0.00	10.70 ± 0.77	31.60 ± 0.32	9.65 ± 0.04
Mean RMSE		9.12 ± 0.07	15.15 ± 0.51	15.59 ± 0.15	13.89 ± 0.34	19.74 ± 0.81	19.78 ± 0.00	17.05 ± 0.82	17.05 ± 1.08	20.82 ± 0.20	**8.99 ± 0.04**

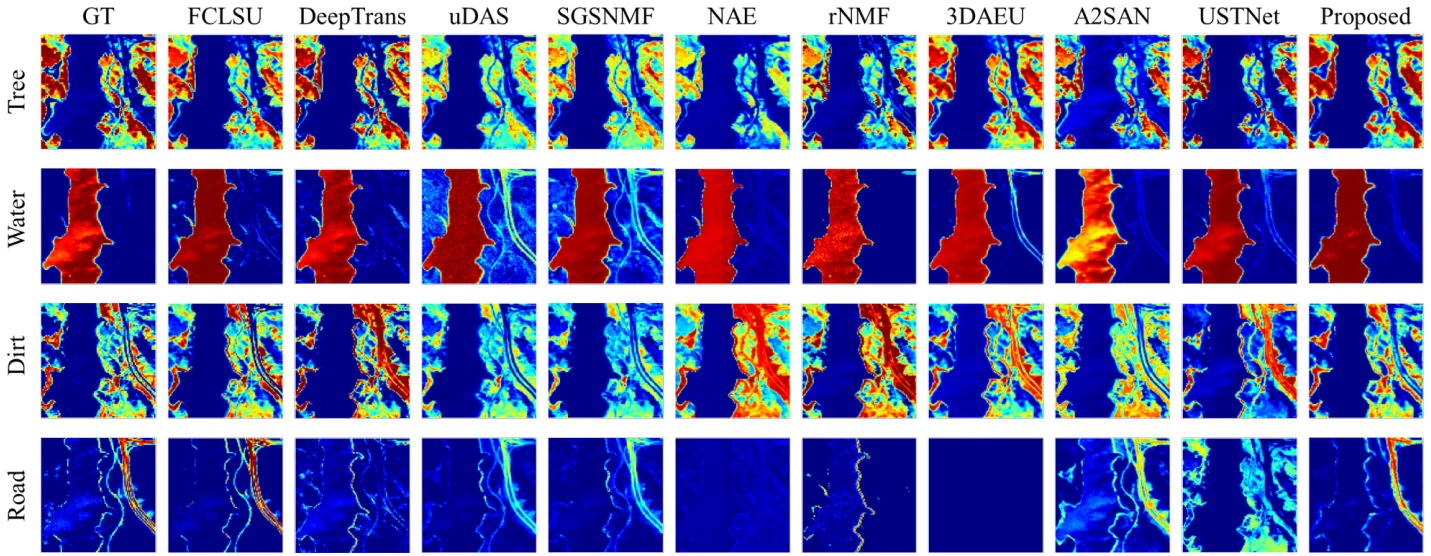


Figure 13. Abundance maps of tree, water, dirt, road on the Jasper Ridge dataset obtained by different methods.

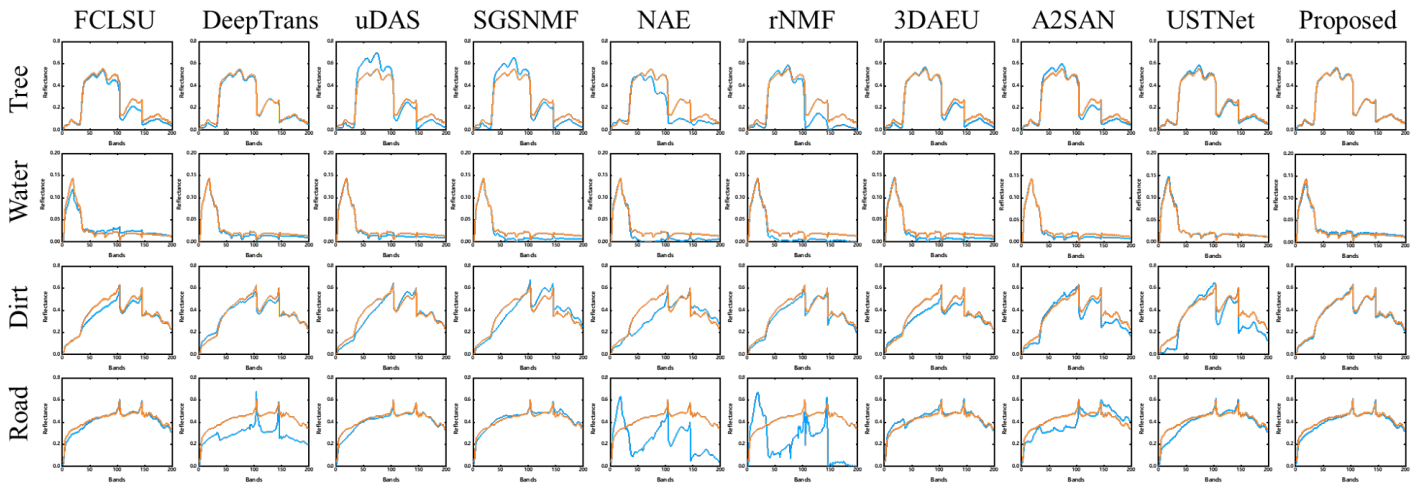


Figure 14. Extracted endmember comparison between the different algorithms and the corresponding GTs in the Jasper Ridge dataset.

Table 6. Valuation metrics SAD and RMSE results of Urban dataset ($\times 10^{-2}$). Best results are bold.

Urban		FCLSU	DeepTrans	uDAS	SGSNMF	NAE	rNMF	3DAEU	A2SAN	USTNet	Proposed
SAD	Asphalt	9.98 ± 0.24	15.36 ± 0.71	15.89 ± 2.88	27.64 ± 1.09	20.40 ± 0.02	19.30 ± 0.43	17.01 ± 0.45	14.86 ± 0.65	**6.28 ± 0.04**	7.90 ± 0.02
	Grass	22.41 ± 2.25	16.43 ± 1.54	114.0 ± 1.60	120.4 ± 7.80	63.40 ± 3.51	49.88 ± 3.87	20.03 ± 0.83	12.85 ± 0.80	**9.79 ± 0.04**	17.36 ± 0.03
	Tree	4.81 ± 0.21	10.36 ± 0.71	11.86 ± 2.60	8.56 ± 0.10	11.46 ± 0.13	11.78 ± 0.10	9.94 ± 0.12	9.29 ± 0.23	2.88 ± 0.01	**2.60 ± 0.01**
	Roof	4.04 ± 0.16	28.03 ± 3.19	27.31 ± 0.85	19.71 ± 3.24	81.15 ± 0.38	79.13 ± 0.58	47.44 ± 0.99	9.08 ± 0.51	3.46 ± 0.03	**3.14 ± 0.03**
Mean SAD		10.31 ± 0.53	17.54 ± 1.42	42.27 ± 1.58	44.08 ± 1.00	44.20 ± 0.89	44.02 ± 0.70	23.60 ± 0.31	11.52 ± 0.41	**5.60 ± 0.02**	7.75 ± 0.01
RMSE	Asphalt	38.26 ± 0.14	13.09 ± 0.46	32.32 ± 1.26	30.08 ± 0.65	28.90 ± 0.05	26.91 ± 0.14	32.18 ± 0.87	**12.79 ± 0.57**	15.29 ± 0.12	13.07 ± 0.31
	Grass	54.05 ± 0.46	14.00 ± 0.71	45.31 ± 1.71	45.08 ± 0.36	25.76 ± 1.02	47.40 ± 0.20	27.50 ± 0.77	14.50 ± 0.65	13.53 ± 0.13	**13.11 ± 0.53**
	Tree	22.54 ± 0.77	11.65 ± 0.49	28.50 ± 2.42	26.15 ± 0.40	24.79 ± 0.22	40.15 ± 0.63	23.78 ± 0.21	13.83 ± 0.20	**7.44 ± 0.06**	10.84 ± 2.86
	Roof	21.83 ± 0.16	11.67 ± 0.32	20.73 ± 0.41	19.50 ± 0.26	19.97 ± 1.14	15.94 ± 0.02	14.86 ± 0.65	11.13 ± 0.44	8.47 ± 0.07	**8.08 ± 0.20**
Mean RMSE		36.64 ± 0.24	12.60 ± 0.43	31.71 ± 0.59	30.20 ± 0.41	25.07 ± 0.10	34.77 ± 0.21	25.39 ± 0.44	13.13 ± 0.40	11.66 ± 0.08	**11.46 ± 0.98**

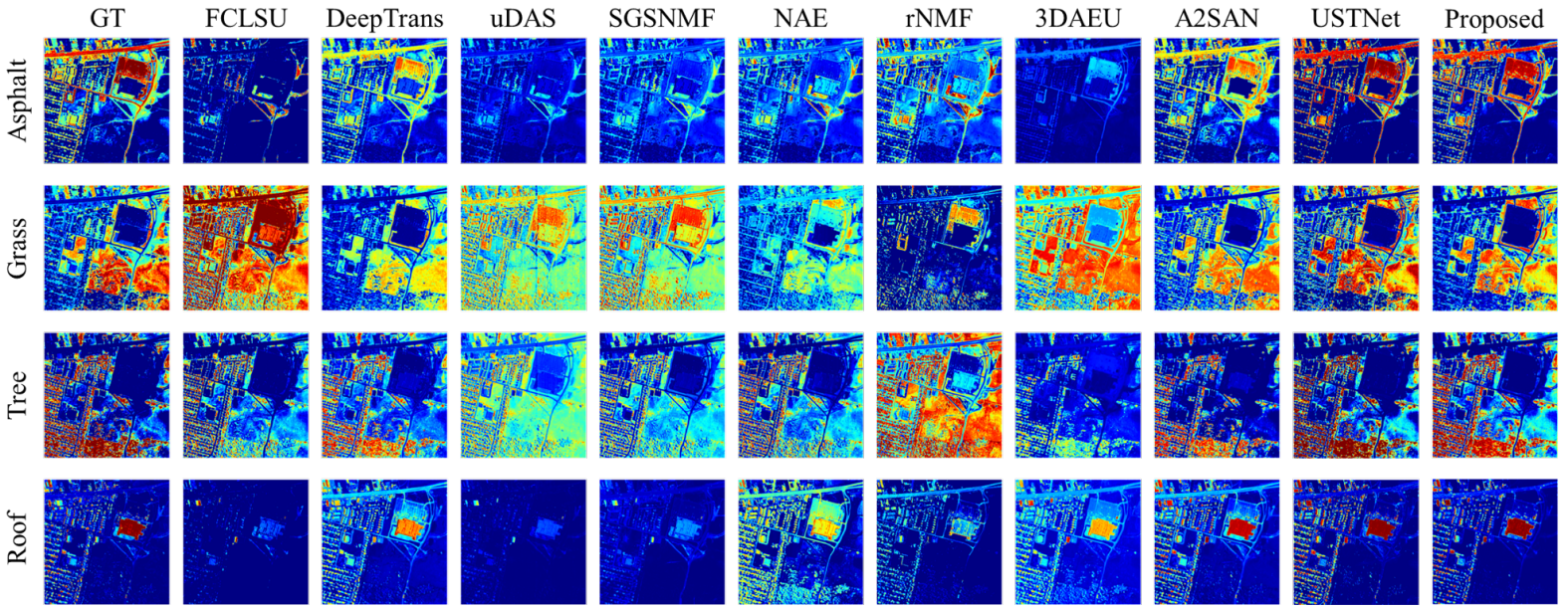


Figure 15. Abundance maps of asphalt, grass, tree, roof on the Urban dataset obtained by different methods.

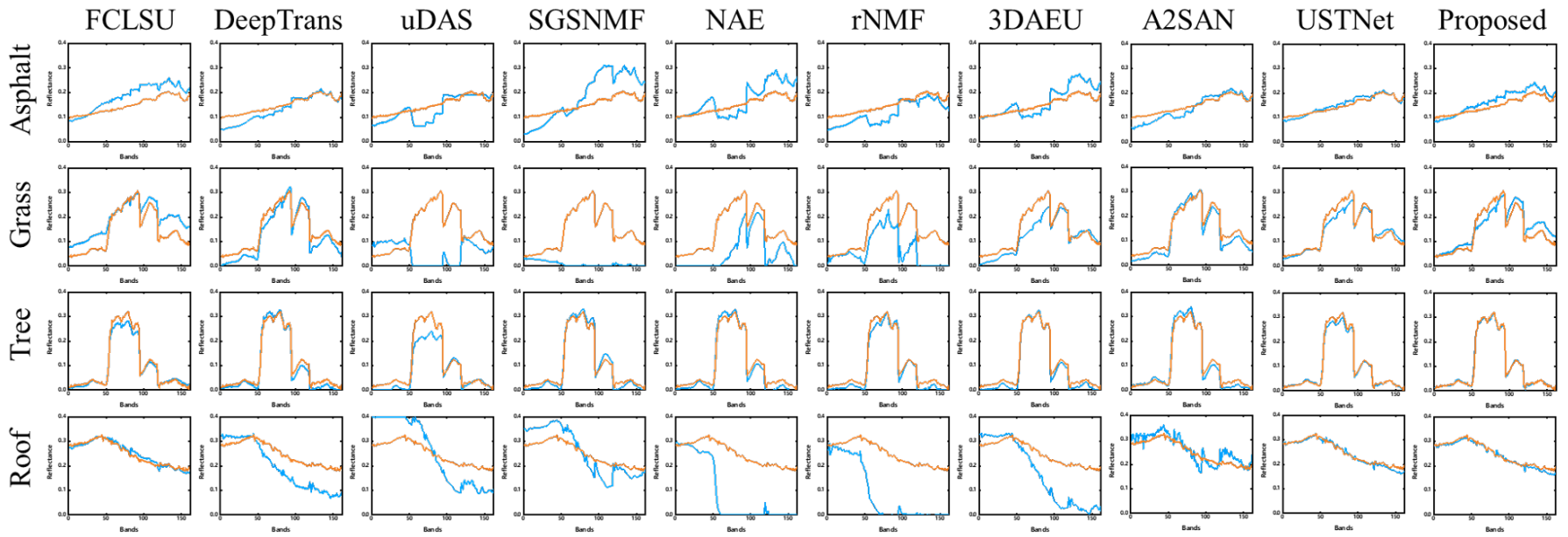


Figure 16. Extracted endmember comparison between the different algorithms and the corresponding GTs in the Urban dataset.

Table 7. The processing time (in seconds) of each method on three datasets. Best results are bold.

	FCLSU	DeepTrans	uDAS	SGSNMF	NAE	rNMF	3DAEU	A2SAN	USTNet	Proposed
Samson	**1.21**	6.56	13.24	12.83	4.95	10.65	44.77	11.83	75.22	27.39
Jasper Ridge	**1.93**	12.86	62.95	16.63	6.62	19.34	98.89	9.12	227.43	71.01
Urban	**9.20**	78.04	429.65	192.83	41.00	113.02	7855.08	79.44	1261.15	331.91

5. Conclusions

In this paper, we propose a global spatial-spectral feature fused autoencoder for nonlinear hyperspectral unmixing. The network extracts spatial and spectral structural information of HSI separately through dual-stream networks and captures global feature dependencies via two consecutive MHSAMs. Pixel-wise correlations are achieved through linear projection, enabling the network to learn more comprehensive spatial information. The data-driven learning approach ensures that the decoder does not rely on existing models, and the meticulously designed decoder module can handle various complex nonlinear scenarios simultaneously. Additionally, a stable and fast endmember initialization method is employed, which enables robust endmember extraction even in the presence of outliers and noise interference. The proposed method achieves the first-ever global spatial-spectral information extraction, albeit with increased computational costs. Developing simpler and more efficient global attention mechanism modules, or selectively extracting more valuable spatial information instead of global extraction, could be potential directions for future research.

Author Contributions: Conceptualization, M.Z. and H.X.; methodology, M.Z.; software, M.Z.; validation, Q.J.; investigation, M.Z. and M.Y.; writing—original draft preparation, M.Z.; writing—review and editing, M.Y. and X.T.; visualization, W.Z.; supervision, P.Y. and L.X.; funding acquisition, X.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Changchun science and technology development plan project (22SH03); Jilin province and Chinese Academy of Sciences Science and Technology Cooperation High Tech Special Fund project (2023SYHZ0020, 2022SYHZ0008, 2023SYHZ0047); Jilin Province Science and Technology Development Plan Project (20220201060GX, 20230204095YY, 20230508038RC, 20230201045GX); Scientific and Technological Innovation Project of Black Land Protection and Utilization (XDA28050201); National Natural Science Foundation of China (NSFC) (42377037).

Data Availability Statement: The data presented in this study are available in the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Goetz, A.F.H.; Vane, G.; Solomon, J.E.; Rock, B.N. Imaging Spectrometry for Earth Remote Sensing. *Science* **1985**, *228*, 1147–1153. [CrossRef] [PubMed]
2. Ma, W.K.; Bioucas-Dias, J.M.; Chan, T.H.; Gillis, N.; Gader, P.; Plaza, A.J.; Ambikapathi, A.; Chi, C.Y. A Signal Processing Perspective on Hyperspectral Unmixing: Insights from Remote Sensing. *IEEE Signal Process. Mag.* **2014**, *31*, 67–81. [CrossRef]
3. Keshava, N. A survey of spectral unmixing algorithms. *Linc. Lab. J.* **2003**, *14*, 55–78.
4. Poulet, F.; Ehlmann, B.L.; Mustard, J.F.; Vincendon, M.; Langevin, Y. Modal mineralogy of planetary surfaces from visible and near-infrared spectral data. In Proceedings of the 2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, Reykjavik, Iceland, 14–16 June 2010; pp. 1–4. [CrossRef]
5. Yang, C.; Everitt, J.H.; Du, Q.; Luo, B.; Chanussot, J. Using High-Resolution Airborne and Satellite Imagery to Assess Crop Growth and Yield Variability for Precision Agriculture. *Proc. IEEE* **2013**, *101*, 582–592. [CrossRef]
6. Teke, M.; Deveci, H.S.; Haliloğlu, O.; Gürbüz, S.Z.; Sakarya, U. A short survey of hyperspectral remote sensing applications in agriculture. In Proceedings of the 2013 6th International Conference on Recent Advances in Space Technologies (RAST), Istanbul, Turkey, 12–14 June 2013; pp. 171–176. [CrossRef]
7. Keshava, N.; Mustard, J. Spectral unmixing. *IEEE Signal Process. Mag.* **2002**, *19*, 44–57. [CrossRef]
8. Heylen, R.; Parente, M.; Gader, P. A Review of Nonlinear Hyperspectral Unmixing Methods. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 1844–1868. [CrossRef]

9. Palsson, B.; Sveinsson, J.R.; Ulfarsson, M.O. Blind Hyperspectral Unmixing Using Autoencoders: A Critical Comparison. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2022**, *15*, 1340–1372. [CrossRef]
10. Boardman, J.W. Automating spectral unmixing of AVIRIS data using convex geometry concepts. In Proceedings of the JPL, Summaries of the 4th Annual JPL Airborne Geoscience Workshop, Washington, DC, USA, 25–29 October 1993. Available online: https://api.semanticscholar.org/CorpusID:140591692 (accessed on 5 June 2024).
11. Winter, M.E. N-FINDR: An algorithm for fast autonomous spectral end-member determination in hyperspectral data. In *Proceedings of the Imaging Spectrometry V. SPIE*; SPIE: Bellingham, WA, USA, 1999; Volume 3753, pp. 266–275.
12. Nascimento, J.; Dias, J. Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 898–910. [CrossRef]
13. Heinz, D.; Chang, C.I. Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. *IEEE Trans. Geosci. Remote Sens.* **2001**, *39*, 529–545. [CrossRef]
14. Bioucas-Dias, J.M.; Plaza, A.; Dobigeon, N.; Parente, M.; Du, Q.; Gader, P.; Chanussot, J. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 354–379. [CrossRef]
15. Qian, Y.; Jia, S.; Zhou, J.; Robles-Kelly, A. Hyperspectral Unmixing via $L_{1/2}$ Sparsity-Constrained Nonnegative Matrix Factorization. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 4282–4297. [CrossRef]
16. Feng, X.R.; Li, H.C.; Li, J.; Du, Q.; Plaza, A.; Emery, W.J. Hyperspectral Unmixing Using Sparsity-Constrained Deep Nonnegative Matrix Factorization with Total Variation. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6245–6257. [CrossRef]
17. Peng, J.; Zhou, Y.; Sun, W.; Du, Q.; Xia, L. Self-Paced Nonnegative Matrix Factorization for Hyperspectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1501–1515. [CrossRef]
18. Zhu, F.; Wang, Y.; Fan, B.; Xiang, S.; Meng, G.; Pan, C. Spectral Unmixing via Data-Guided Sparsity. *IEEE Trans. Image Process.* **2014**, *23*, 5412–5427. [CrossRef] [PubMed]
19. Wang, X.; Zhong, Y.; Zhang, L.; Xu, Y. Spatial Group Sparsity Regularized Nonnegative Matrix Factorization for Hyperspectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 6287–6304. [CrossRef]
20. Huang, R.; Li, X.; Zhao, L. Spectral–Spatial Robust Nonnegative Matrix Factorization for Hyperspectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8235–8254. [CrossRef]
21. Somers, B.; Zortea, M.; Plaza, A.; Asner, G.P. Automated Extraction of Image-Based Endmember Bundles for Improved Spectral Unmixing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 396–408. [CrossRef]
22. Hong, D.; Gao, L.; Yao, J.; Yokoya, N.; Chanussot, J.; Heiden, U.; Zhang, B. Endmember-Guided Unmixing Network (EGU-Net): A General Deep Learning Framework for Self-Supervised Hyperspectral Unmixing. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 6518–6531. [CrossRef]
23. Veganzones, M.; Drumetz, L.; Tochon, G.; Dalla Mura, M.; Plaza, A.; Bioucas-Dias, J.; Chanussot, J. A new extended linear mixing model to address spectral variability. In Proceedings of the 2014 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Lausanne, Switzerland, 24–27 June 2014; pp. 1–4. [CrossRef]
24. Drumetz, L.; Henrot, S.; Veganzones, M.A.; Chanussot, J.; Jutten, C. Blind hyperspectral unmixing using an extended linear mixing model to address spectral variability. In Proceedings of the 2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Tokyo, Japan, 2–5 June 2015; pp. 1–4. [CrossRef]
25. Thouvenin, P.A.; Dobigeon, N.; Tourneret, J.Y. Hyperspectral Unmixing with Spectral Variability Using a Perturbed Linear Mixing Model. *IEEE Trans. Signal Process.* **2016**, *64*, 525–538. [CrossRef]
26. Hong, D.; Yokoya, N.; Chanussot, J.; Zhu, X.X. An Augmented Linear Mixing Model to Address Spectral Variability for Hyperspectral Unmixing. *IEEE Trans. Signal Process.* **2019**, *28*, 1923–1938. [CrossRef]
27. Chen, J.; Zhao, M.; Wang, X.; Richard, C.; Rahardja, S. Integration of Physics-Based and Data-Driven Models for Hyperspectral Image Unmixing: A summary of current methods. *IEEE Signal Process. Mag.* **2023**, *40*, 61–74. [CrossRef]
28. Dobigeon, N.; Tourneret, J.Y.; Richard, C.; Bermudez, J.C.M.; McLaughlin, S.; Hero, A.O. Nonlinear Unmixing of Hyperspectral Images: Models and Algorithms. *IEEE Signal Process. Mag.* **2014**, *31*, 82–94. [CrossRef]
29. Halimi, A.; Altmann, Y.; Dobigeon, N.; Tourneret, J.Y. Nonlinear unmixing of hyperspectral images using a generalized bilinear model. In Proceedings of the 2011 IEEE Statistical Signal Processing Workshop (SSP), Nice, France, 28–30 June 2011; pp. 413–416. [CrossRef]
30. Altmann, Y.; Halimi, A.; Dobigeon, N.; Tourneret, J.Y. Supervised Nonlinear Spectral Unmixing Using a Postnonlinear Mixing Model for Hyperspectral Imagery. *IEEE Trans. Image Process.* **2012**, *21*, 3017–3025. [CrossRef] [PubMed]
31. Guo, R.; Wang, W.; Qi, H. Hyperspectral image unmixing using autoencoder cascade. In Proceedings of the 2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Tokyo, Japan, 2–5 June 2015; pp. 1–4. [CrossRef]
32. Qu, Y.; Qi, H. uDAS: An Untied Denoising Autoencoder with Sparsity for Spectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 1698–1712. [CrossRef]
33. Ozkan, S.; Kaya, B.; Akar, G.B. EndNet: Sparse AutoEncoder Network for Endmember Extraction and Hyperspectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 482–496. [CrossRef]
34. Dou, Z.; Gao, K.; Zhang, X.; Wang, H.; Wang, J. Hyperspectral Unmixing Using Orthogonal Sparse Prior-Based Autoencoder with Hyper-Laplacian Loss and Data-Driven Outlier Detection. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 6550–6564. [CrossRef]

35. Min, A.; Guo, Z.; Li, H.; Peng, J. JMnet: Joint Metric Neural Network for Hyperspectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5505412. [CrossRef]
36. Zhao, M.; Wang, M.; Chen, J.; Rahardja, S. Perceptual Loss-Constrained Adversarial Autoencoder Networks for Hyperspectral Unmixing. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6006505. [CrossRef]
37. Jin, Q.; Ma, Y.; Fan, F.; Huang, J.; Mei, X.; Ma, J. Adversarial Autoencoder Network for Hyperspectral Unmixing. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 4555–4569. [CrossRef]
38. Sun, L.; Chen, Y.; Li, B. SISLU-Net: Spatial Information-Assisted Spectral Information Learning Unmixing Network for Hyperspectral Images. *Remote Sens.* **2023**, *15*, 817. [CrossRef]
39. Palsson, B.; Ulfarsson, M.O.; Sveinsson, J.R. Convolutional Autoencoder for Spectral–Spatial Hyperspectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 535–549. [CrossRef]
40. Gao, Y.; Pan, B.; Song, X.; Xu, X. Extended-Aggregated Strategy for Hyperspectral Unmixing Based on Dilated Convolution. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 5507005. [CrossRef]
41. Yu, Y.; Ma, Y.; Mei, X.; Fan, F.; Huang, J.; Li, H. Multi-stage convolutional autoencoder network for hyperspectral unmixing. *Int. J. Appl. Earth Observ. Geoinf.* **2022**, *113*, 102981. [CrossRef]
42. Ghosh, P.; Roy, S.K.; Koirala, B.; Rasti, B.; Scheunders, P. Hyperspectral Unmixing Using Transformer Network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5535116. [CrossRef]
43. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762 [CrossRef]
44. Yang, Z.; Xu, M.; Liu, S.; Sheng, H.; Wan, J. UST-Net: A U-Shaped Transformer Network Using Shifted Windows for Hyperspectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5528815. [CrossRef]
45. Huang, Y.; Li, J.; Qi, L.; Wang, Y.; Gao, X. Spatial-Spectral Autoencoder Networks for Hyperspectral Unmixing. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 2396–2399. [CrossRef]
46. Wang, J.; Xu, J.; Chong, Q.; Liu, Z.; Yan, W.; Xing, H.; Xing, Q.; Ni, M. SSANet: An Adaptive Spectra-Spatial Attention Autoencoder Network for Hyperspectral Unmixing. *Remote Sens.* **2023**, *15*, 2070. [CrossRef]
47. Hua, Z.; Li, X.; Feng, Y.; Zhao, L. Dual Branch Autoencoder Network for Spectral-Spatial Hyperspectral Unmixing. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5507305. [CrossRef]
48. Qi, L.; Gao, F.; Dong, J.; Gao, X.; Du, Q. SSCU-Net: Spatial–Spectral Collaborative Unmixing Network for Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5407515. [CrossRef]
49. Qi, L.; Chen, Z.; Gao, F.; Dong, J.; Gao, X.; Du, Q. Multiview Spatial–Spectral Two-Stream Network for Hyperspectral Image Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5502016. [CrossRef]
50. Palsson, B.; Sigurdsson, J.; Sveinsson, J.R.; Ulfarsson, M.O. Hyperspectral Unmixing Using a Neural Network Autoencoder. *IEEE Access* **2018**, *6*, 25646–25656. [CrossRef]
51. Su, L.; Liu, J.; Yuan, Y.; Chen, Q. A Multi-Attention Autoencoder for Hyperspectral Unmixing Based on the Extended Linear Mixing Model. *Remote Sens.* **2023**, *15*, 2898. [CrossRef]
52. Cheng, Y.; Zhao, L.; Chen, S.; Li, X. Hyperspectral Unmixing Network Accounting for Spectral Variability Based on a Modified Scaled and a Perturbed Linear Mixing Model. *Remote Sens.* **2023**, *15*, 3890. [CrossRef]
53. Wang, M.; Zhao, M.; Chen, J.; Rahardja, S. Nonlinear Unmixing of Hyperspectral Data via Deep Autoencoder Networks. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1467–1471. [CrossRef]
54. Li, H.; Borsoi, R.A.; Imbiriba, T.; Closas, P.; Bermudez, J.C.M.; Erdoğmuş, D. Model-Based Deep Autoencoder Networks for Nonlinear Hyperspectral Unmixing. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5506105. [CrossRef]
55. Shahid, K.T.; Schizas, I.D. Unsupervised Hyperspectral Unmixing via Nonlinear Autoencoders. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5506513. [CrossRef]
56. Yang, X.; Chen, J.; Wang, C.; Chen, Z. Residual Dense Autoencoder Network for Nonlinear Hyperspectral Unmixing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 5580–5595. [CrossRef]
57. Zhao, M.; Wang, M.; Chen, J.; Rahardja, S. Hyperspectral Unmixing for Additive Nonlinear Models with a 3-D-CNN Autoencoder Network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5509415. [CrossRef]
58. Heylen, R.; Scheunders, P. A Multilinear Mixing Model for Nonlinear Spectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 240–251. [CrossRef]
59. Kong, F.; Chen, M.; Li, Y.; Li, D. A Global Spectral–Spatial Feature Learning Network for Semisupervised Hyperspectral Unmixing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 3190–3203. [CrossRef]
60. Wang, S.; Li, B.Z.; Khabsa, M.; Fang, H.; Ma, H. Linformer: Self-Attention with Linear Complexity. *arXiv* **2020**, arXiv:2006.04768.
61. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
62. Su, Y.; Li, J.; Plaza, A.; Marinoni, A.; Gamba, P.; Chakravortty, S. DAEN: Deep Autoencoder Networks for Hyperspectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 4309–4321. [CrossRef]
63. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef] [PubMed]

64. Févotte, C.; Dobigeon, N. Nonlinear Hyperspectral Unmixing with Robust Nonnegative Matrix Factorization. *IEEE Trans. Image Process.* **2015**, *24*, 4810–4819. [CrossRef] [PubMed]
65. Tao, X.; Paoletti, M.E.; Wu, Z.; Haut, J.M.; Ren, P.; Plaza, A. An Abundance-Guided Attention Network for Hyperspectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5505414. [CrossRef]