Dual Embedding Transformer Network for Hyperspectral Unmixing

Huadong Yang and Chengbi Zhang

***Abstract*—Hyperspectral unmixing is an essential task for achieving accurate perception of hyperspectral remote sensing information, aiming to overcome the limitation of spatial resolution and interpret the distribution of land features. To achieve the spatial and spectral feature representation of hyperspectral images, we propose a dual embedding transformer network (DET-Net) based on an encoder-decoder architecture, which utilizes two transformer modules, including three-view spatial attention (TVA) module with 2-D embedding and multiscale spectral band group feature fusion (BGF) module with 3-D embedding to accomplish the task of hyperspectral unmixing. In TVA module, based on 2-D embedding, we introduce a three-view attention mechanism to extract more comprehensive spatial features. In BGF module, the transformer embedding is extended to band group spatial-spectral 3-D cubed embedding and establishes a series of spectral band groups. A cross-feature fusion mechanism is adopted to achieve multiscale spatial-spectral feature decoupling. With the collaboration of these two embeddings, DET-Net effectively captures complex spatial and spectral dependencies to decouple the tridimensional unmixing feature representation. Experimental results on synthetic and real datasets demonstrates the generalization performance of the proposed method, and the ablation experiments confirm the effectiveness of the TVA and BGF modules.**

***Index Terms*—Abundance map, autoencoder (AE), deep learning, endmember extraction, hyperspectral image (HSI), hyperspectral unmixing (HU), transformer network.**

I. Introduction

HYPERSPECTRAL remote sensing is a multidimensional information acquisition technology that can simultaneously capture 2-D spatial information and 3-D spectral information, resulting a 3-D image cube at hundreds of contiguous bands across the electromagnetic spectrum, providing substantial information of the scene. Accordingly, hyperspectral images (HSIs) have been increasingly applied in many areas including land cover classification [1], [2], [3], [4], precision agriculture [5], food industry [6], biotechnology [5], and medical science [7].

Due to the low spatial resolution of hyperspectral sensors, mixed pixels inevitably exist in the scene, which has become a major obstacle that hinders the quantitative interpretation of hyperspectral remote sensing images. Hyperspectral unmixing (HU) is an effective technique for the mixed pixel decomposition, which aims to decompose the mixed pixel spectrum into a set of "pure" spectra, named endmembers, weighted by the corresponding proportions, named abundances.

Over the last few decades, many HU algorithms have been proposed for HSI interpretation, and most of them are designed based on the linear spectral mixture model since its inherent simplicity from conceptual and implementational points of view. These algorithms can be roughly classified into geometric methods [8], [9], [10], [11], [12], statistical methods [13], [14], [15], sparse regression methods [16], [17], and heuristic intelligent algorithms [18], [19], blind unmixing methods [20], [21], and so on. Compared with other methods, blind unmixing method can extract the endmembers and estimate the corresponding abundances simultaneously.

Recently, deep learning methods have demonstrated astonishing results in many fields [22], [23], [24], such as computer vision, natural language processing, and speech recognition. Naturally, deep neural networks have also been applied to HU task [25], [26], [27]. Hyperspectral deep unmixing networks can be divided into two types: 1) supervised deep unmixing network; and 2) unsupervised deep unmixing network. The supervised unmixing network requires real labels or manually calibrated endmember information. However, acquiring such prior information is often expensive or inaccurate. Therefore, unsupervised blind unmixing network has been attracted by many focuses in recent year. The autoencoder (AE) network is one of the most commonly used unsupervised HU networks, and have achieved a fast development in unmixing applications [21], [28].

Ozkan et al. [29] proposed a two-staged AE network, in which the sparsity of the estimates was improved by incorporating the Kullback–Leibler divergence term with SAD similarity and additional penalty terms. Qu et al. [30] utilized denoising and the l_{21}-norm constraint to propose an untied denoising autoencoder with sparsity. Multiple AE network structures, activation functions, and objective functions were tested for their impact on HU in [31]. Su et al. [32] developed stacked nonnegative sparse autoencoders for hyperspectral data with outliers and low signal-to-noise ratio. A variational autoencoder was employed for blind source separation in [33]. The Wu-net [34] proposed a two-stream network autoencoder architecture, achieving HU by sharing pseudopixel network weights. In addition, superpixels were applied to HU, focusing on local spatial crucial features to achieve better generalization for the network initialization

Received 19 July 2024; revised 7 October 2024 and 29 November 2024; accepted 14 December 2024. Date of publication 30 December 2024; date of current version 20 January 2025. This work was supported in part by the Scientific Research Foundation of Liaoning Educational Committee under Grant LJ212410144018 and in part by the Fundamental Research Funds for the Universities in Liaoning Province under Grant SYLUGXRC202258. *(Corresponding author: Huadong Yang.)*

The authors are with the School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China (e-mail: yanghd8608@163.com; zcb_20222025@163.com).

Digital Object Identifier 10.1109/JSTARS.2024.3523747

task. In TANet [35], superpixel segmentation is adopted as the two-stream network preprocessing to extract the endmember bundles with spatial information. Jin et al. [36] integrated a GAN network with an AE network to further enhance the two-stream network's fitting capability.

The above pixel-based methods cannot explore the global spatial information of HSIs. convolutional neural network (CNN) models are utilized to regard hyperspectral image block or whole hyperspectral image as the inputting of HU network. In [37], hyperspectral image was divided into hyperspectral image blocks as the input of the hyperspectral unmixing network, thus preserving the spatial structure of the HSI. Rasti et al. [38] proposed a minimal simplex convolutional network for deep hyperspectral unmixing. Palsson et al. [39] applied multitask learning to HU. Huang et al. [40] combined spectral information for effective spatial-spectral two-stream unmixing. By addressing unmixing tasks separately for each hyperspectral image block, sharing hidden layers across all tasks can significantly reduce the risk of overfitting. Gao et al. [41] introduced a cycle-consistency unmixing network by training two cascaded AEs in an end-to-end manner, aiming to enhance the unmixing performance more effectively. Zhang et al. [42] applied advanced 3-D CNN for HU.

However, these methods are limited by the finite receptive field of CNN. With the remarkable long-range feature modeling performance of transformer [43] demonstrated in natural language processing, computer vision, and other fields, the transformer networks are utilized to enhance unmixing capabilities. In [44], the significant spatial information in the scene was prioritized using multihead self-attention blocks based on shifted windows. In addition, studies have shown that the application of shifted windows [45] assists transformers in extracting hierarchical features similar to CNN. The authors in[46] and [47] prioritized significant spatial information in the scene using multihead self-attention blocks based on shifted windows. In [48], Duan et al. developed the double-aware transformer for HU by exploiting the region homogeneity and spectral correlation of HSI. TCCU-Net [49] end-to-end learns feature in four dimensions: Spectral, spatial, global, and local, to achieve effective unmixing. Regarding spectral variation, Gao et al. [50] designed Rev-net and demonstrated a theoretical proof for the reversibility of the endmember generation process. Bhakthan et al. [51] used necessary preprocessing, including PCA and 3-D–2-D CNN, to achieve effective unmixing. The DSET-Net [52] adopts an unmixing framework and achieves local and overall feature parameter sharing in the encoder through a "Transformer in Transformer" strategy.

In previous works, spatial embedding has been widely recognized as a crucial factor affecting transformer unmixing performance [44], [46], [47], [48]. However, 2-D ViT-based patch embedding has been adopted in the existing transformer unmixing network, which is suitable for RGB image, but not enough for HSI. Because it is not enough to explore the spatial unmixing information of hyperspectral image cube. Compared with RGB image, HSI is not only with the main view (front view), but also with the left view(side view) and top view, as shown in Fig. 1. The left and top views of HSIs contain abundant spatial unmixing information. However, the 2-D ViT-based patch embedding which only encodes the features of the main view, ignores the spatial feature information of the left view and top view for HSIs. In addition, the 2-D ViT-based patch embedding which encodes only in the spatial dimension, also ignores important feature information in the band dimension. Therefore, it is necessary to establish a 3-D spatial-spectral cubed embedding across spatial and spectral dimensions of HSIs. The band by band embedding requires a large amount of graphics memory, which is not conducive to the convergence of the unmixing network. The band group spatial-spectral cubed embedding can freely generate a suitable number of tokens by setting the embedding scale. Motivated by the above, we propose a dual embedding transformer network (DET-Net) for HU, which mainly includes the three-view spatial attention (TVA) module with 2-D embedding and the multiscale spectral band group feature fusion (BGF) module with 3-D embedding.

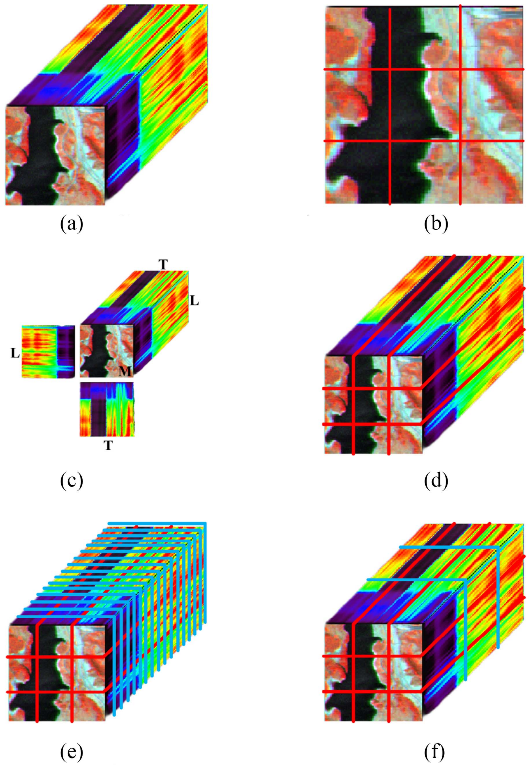


Fig. 1. Embeddings for HSI. (a) HSI. (b) 2-D ViT-based patch embedding. (c) Three views. (d) 2-D ViT-based patch embedding for HSI's main view. (e) Band by band spatial-spectral 3-D cubed embedding for HSI. (f) Band group spatial-spectral 3-D cubed embedding for HSI.

In TVA transformer module, in order to explore spatial feature information in HSIs, we propose a three-view spatial attention mechanism that facilitates the fusion of features from the main view, the top view and the left view. The main view features take the lead in fusing the top and left view features to extract more comprehensive spatial information from the HSI.

In BGF transformer module, we utilize 3-D band group spatial-spectral cubed embedding to extract spectral band feature information of the multiscale spectral band groups. Moreover, a bidirectional cross-fusion mechanism is designed between the multiscale spectral band groups transformer blocks to extract scale-invariant spectral-spatial unmixing features.

The AE DET-Net with aforementioned two modules can effectively capture HSI's spatial and spectral features with unsupervised learning, and accomplish the decoupling of spatial and spectral features. Specifically, the main contributions of this proposed network can be summarized as follows.

1) Based on the transformer hyperspectral unmixing model, We propose a DET-Net with the two feature refinement modules, namely, TVA module with 2-D embedding and BGF module with 3-D embedding. By combining the 2-D embedding and 3-D embedding, the proposed network achieve an improvement in unmixing accuracy.
2) In the TVA module, we introduce a novel spatial attention module based on the three views of the hyperspectral image cube. With the assistance of the three-view spatial attention mechanism, the spatial features of the three views are reweighted to retain more comprehensive spatial information of the HSI.
3) In order to strengthen the spatial-spectral linkage of unmixing feature, we design a 3-D band group spatial-spectral cubed embedding rather than 2-D spatial embedding in the BGF module. Based on 3-D embedding, we establish multiscale spectral band groups through different-scale spatial-spectral band group embedding. In addition, the bidirectional cross-fusion mechanism is established to effectively decouple spatial-spectral unmixing features of multiscale spectral band groups.

II. Proposed Method

This section presents a comprehensive exposition of the method in our study. First, we provide a brief overview of the mechanism for linear mixture model (LMM). Next, the macroscopic interpretation of our autoencoder network framework is presented. Finally, we provide a detailed exposition of the two core transformer modules, namely the TVA module with 2-D embedding and the BGF module with 3-D embedding, in the encoder.

A. Linear Mixture Model (LMM)

Let HSI of spatial dimensions $H \times W$ with C spectral bands be denoted by $\mathbf{I} \in \mathbb{R}^{C\times H\times W}$. The mathematical expression of the LMM is as follows:

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

where $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_C] \in \mathbb{R}^{C\times Q}$, $Q = H \cdot W$ represent the hyperspectral image matrix, which is reshaped from HSI, with C bands and Q pixels. The endmember matrix $E \in \mathbb{R}^{C\times P}$ contains P endmembers, and $A \in \mathbb{R}^{P\times Q}$ denotes corresponding abundance map. $N \in \mathbb{R}^{C\times Q}$ represent image noise. Two constraints, included abundance nonnegativity constraint (ANC) and the abundance sum-to-one constraint (ASC), must be satisfied to ensure physically meaningful interpretations

$$\begin{gathered}\mathbf{A} \geq 0 \\ \mathbf{1}_P^T\mathbf{A} = \mathbf{1}_Q^T.\end{gathered} \tag{2}$$

B. Overall Architecture

The architecture of the proposed network is shown in Fig. 2. The the main components of DET-Net comprises TVA transformer module and BGF transformer module to achieve end-to-end HU.

In the DET-Net encoder, we first employ CNN to perform coarse feature extraction on the HSI. In the CNN encoder, we employ 2-D convolution with the 1 × 1 kernel to downsample only on the spectral dimension without altering the spatial dimensions. We incorporate batch normalization (BN), dropout, and leaky ReLU operations. These operations have been demonstrated to be effective for hyperspectral unmixing tasks [21]. HSI is transformed by CNN into the feature map $\mathbb{F} \in \mathbb{R}^{T\times H\times W}$, where T represents the reduced number of output bands. Subsequently, to extract the spatial features of the cube-shaped HSI from three different views, we designed a TVA transformer module on the the three views of feature maps. A three-stream network employs TVA mechanism to fuse the features of the top view and left view into the main view features. In BGF transformer module, the 3-D band group spatial-spectral cubed embedding is utilized to split the feature maps of the HSI into multiscale spectral band groups. To extract multiscale spectral band grouping features, we design a bidirectional cross-fusion module to decouple the multiscale spectral unmixing features.

In the DET-Net decoder, the output by two core modules is upsampled and reshaped to the original dimensions of the hyperspectral abundance. Subsequently, softmax is applied to ensure the abundance $\mathbf{A}$ to satisfy ANC and ASC constraints. The reconstructed HSI $\hat{\mathbf{I}} \in \mathbb{R}^{C\times H\times W}$, which can be reshaped to hyperspectral image matrix $\hat{\mathbf{Y}} \in \mathbb{R}^{C\times Q}$,is output through a single convolutional layer without bias. The weights of convolutional layer are the endmember $\mathbf{E}$ [44]. Before training the unmixing network, we initialize the weights of the convolution layer with the endmembers obtained by VCA [10]. For the loss function, DET-Net minimizes the reconstruction loss function, which consists of the reconstruction error (RE) loss and the spectral angle distance (SAD) loss with regularization parameters α and β as follows:

$$L_{\mathrm{RE}}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{Q}\sum_{i=1}^{Q}\left(\hat{\mathbf{Y}}_{\mathbf{i}} - \mathbf{Y}_{\mathbf{i}}\right)^2 \tag{3}$$

$$L_{\mathrm{SAD}}(\mathbf{Y}, \hat{\mathbf{Y}}) = \sum_{i=1}^{Q}\arccos\left(\frac{\left\langle \mathbf{Y}_i, \hat{\mathbf{Y}}_i\right\rangle}{\|\mathbf{Y}_i\|_2 \, \|\hat{\mathbf{Y}_i}\|_2}\right) \tag{4}$$

$$L = \alpha L_{RE} + \beta L_{\mathrm{SAD}}. \tag{5}$$

C. Module 1: Three-View Spatial Attention Module With 2-D Embedding

Some transformer-based unmixing methods [44], [46], [47] only consider the spatiality of the main view of HSI unmixing feature. However, as a HSI cube with three views, HSIs contain not only the main view, but also the top and left views with abundant spatial information. The traditional 2-D ViT patch embedding only on the main view has limitations in unmixing tasks.

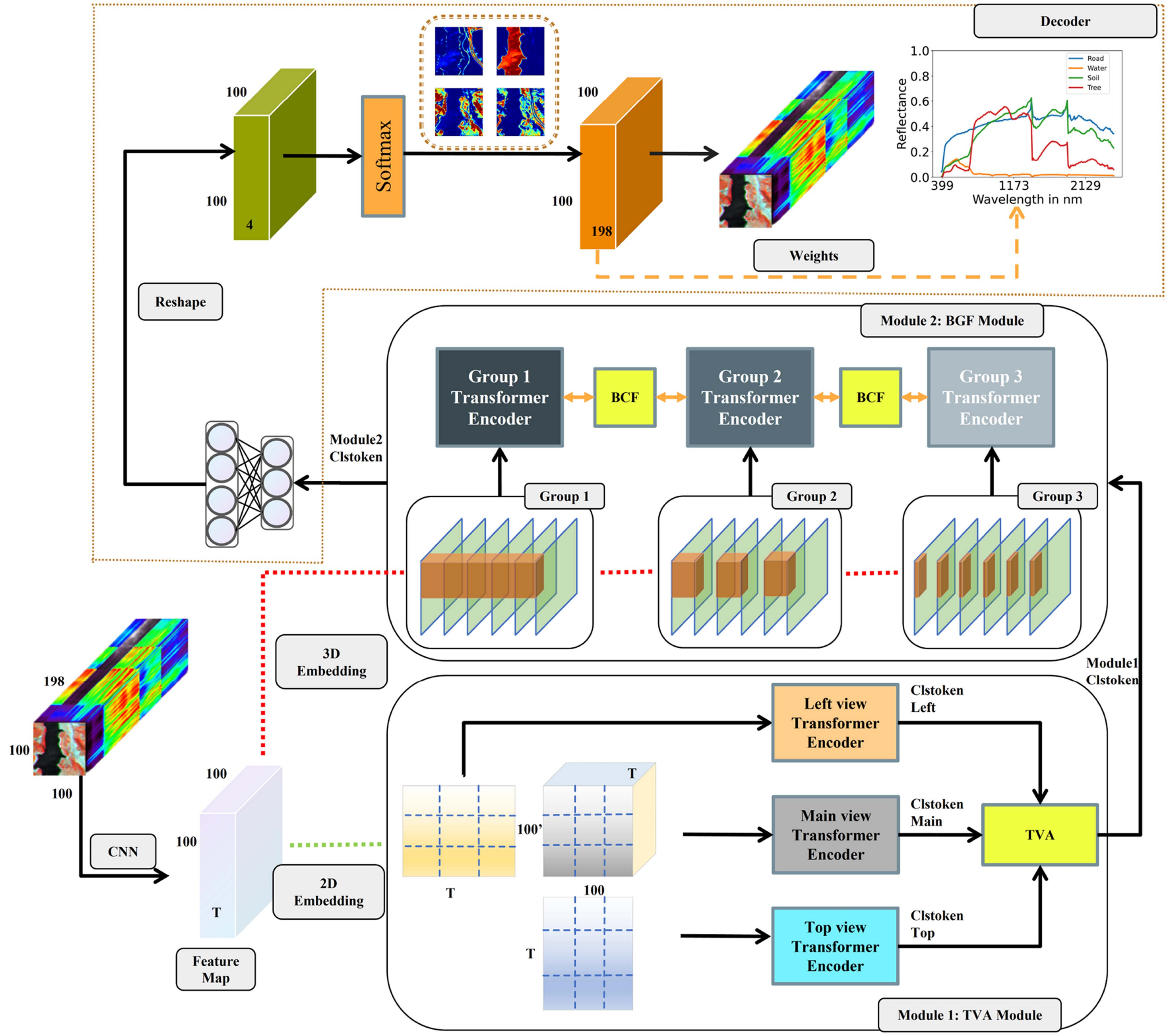


Fig. 2. Proposed DET-Net architecture(as shown in the figure, with the hyperspectral Jasper dataset of 100 × 100 pixels and 198 bands used for demonstration). The input of DET-Net is the original HSI, and the output is the reconstructed HSI with endmembers and abundance maps. TVA Module utilize 2-D ViT-based embedding to extract three-view spatial feature, and 3-D band group spatial-spectral cubed embedding is adopted to decoupling spectral feature in BGF module. For the light yellow blocks, Three-view Spatial Attention (TVA) mechanism and Bidirectional Cross-Fusion (BCF) mechanism are, respectively, introduced in Sections II-C and II-D4.

Therefore, we propose a three-view spatial attention module to emphasize the spatial features of the three views for HU.

We perform 2-D ViT-based patch embedding separately on the feature map of each view. For a more intuitive illustration, we use the hyperspectral Jasper dataset in Section III-E1 as an example to explain TVA module. The feature maps of Jasper dataset is $\mathbb{F}_j \in \mathbb{R}^{T\times100\times100}$. The main view for the feature maps is $\mathbb{F}_{jm} \in \mathbb{R}^{T\times(100\times100')}$. while the dimensions of the top and left views are $\mathbb{F}_{jt} \in \mathbb{R}^{100'\times(100\times T)}$ and $\mathbb{F}_{jl} \in \mathbb{R}^{100\times(T\times100')}$. The spatial feature embedding patch size in the main view is 5 × 5, and it is 5 × 3 for the top and left views. Meanwhile, the patch band(channel) of main view transformer encoder is T. The patch band of top or left view transformer encoder is 100. After the 2-D ViT-based embedding in Section II-D1, three Transformer Blocks with the multihead self-patch attention are, respectively, fed with the resulting sequence of vectors. Further details of multihead self-patch attention are introduced in Section II-D2 [44]. The Clstoken Main, Clstoken Top, and Clstoken Left generated by the three view transformer encoders are input to TVA mechanism block. One of the tasks of HU is abundance estimation, which is reflected by the spatial information of the main view. In other words, the unmixing feature of the main view should be the dominant feature of TVA module. The unmixing features of the top and left views should be fused into the unmixing features of the main view.

The structure of TVA mechanism is illustrated in Fig. 3. First, inspired by the different presence of common edges between the main view and the other two views, we restore the Clstoken

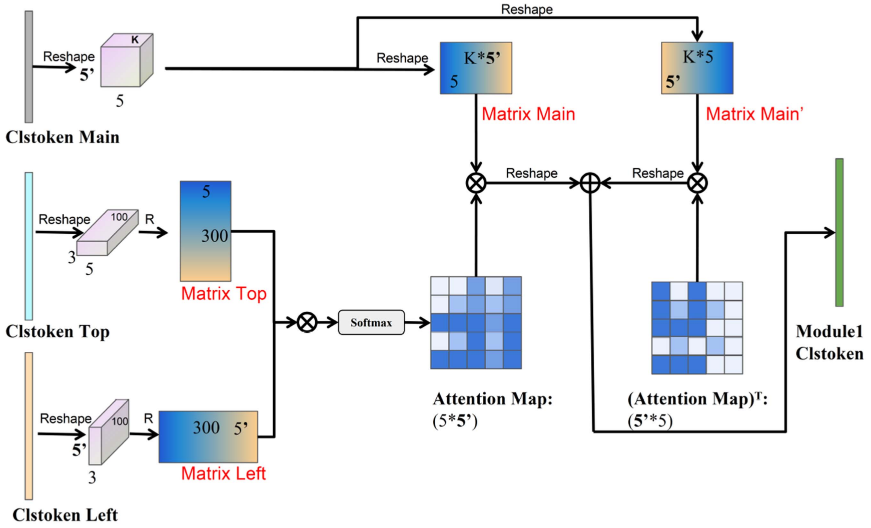


Fig. 3. Proposed TVA mechansim with the hyperspectral Jasper dataset. Module1 Clstoken fused from three-view Clstokens to facilitate feature extraction.

of all views to the original unflattened shape. The edge marked with 5 represents the main view edge corresponding to the top view, and the edge marked with 5' represents the main view edge corresponding to the left view. Next, the Clstoken of the cube shapes for the top view and the left view are reshaped into 2-D feature matrices, namely, Matrix $\mathbf{Top} \in \mathbb{R}^{300\times 5'}$ and Matrix $\mathbf{Left} \in \mathbb{R}^{5\times 300}$. Afterward, we perform a matrix multiplication between Matrix **Top** and Matrix **Left** and apply a softmax layer to calculate the attention map. The main view and the top view share the common edge 5, while the main view and the left view share the common edge 5'. Thus, the Clstoken of the main view is reshaped into two branches. In one branch, Matrix $\mathbf{Main} \in \mathbb{R}^{(K*5')\times 5}$, which is reshaped from Clstoken Main, is chosen to be multiplied by the attention map matrix. In other branch, Matrix $\mathbf{Main}' \in \mathbb{R}^{(K*5)\times 5'}$, which is reshaped from Clstoken Main, is chosen to be multiplied by the transpose of the attention map matrix. Finally, the result of the two branches respectively reshape into a 1-D vector of Clstoken separately. The two 1-D vectors of Clstoken are added element-wise to complete the attention feature of the three views.

As presented above, the Module1 Clstoken is a weighted sum of the features across three views. As a result, it has a global contextual view and selectively aggregates feature information based on spatial attention maps. Moreover, similar semantic features are mutually enhance themselves, improving intra-class compactness and semantic consistency.

D. Module 2: Multiscale Spectral Band Group Feature Fusion Module With 3-D Embedding

To further explore the hyperspectral features in the spectral dimension, we propose a multiscale spectral band group model. As illustrated in the Fig. 4, the lower part demonstrates the 3-D band group spatial-spectral cubed embedding on hyperspectral feature maps. While the upper part depicts the bidirectional cross-fusion mechanism for multispectral band group features. The structure can capture global salient features from long-wave bands and detailed granularity from short-wave bands.

1) 3-D Spatial-Spectral Feature Embedding: For 2-D ViT-based patch embedding of RGB images, ViT extracts tokens by spliting the image into nonoverlapping patches. If ViT is directly applied to the transposed feature maps $\mathbb{F}^T \in \mathbb{R}^{H\times W\times T}$, HSI is splited into the patches $x_1, x_2, \ldots x_N \in \mathbb{R}^{\lfloor \frac{H}{h}\rfloor \times \lfloor \frac{W}{w}\rfloor \times T}$, where $N = h \cdot w$. The above process refers to Fig. 1(d).

Compared with RGB image, HSIs have more bands (channels). Thus, we perform 3-D band group spatial-spectral cubed embedding the hyperspectral transposed CNN feature map by integrating the spatial dimension and the extra spectral dimension. Specifically, tokens are extracted from the hyperspectral feature maps by dividing the image into nonoverlapping cube-shaped patches, $x_1, x_2, \ldots x_N \in \mathbb{R}^{\lfloor \frac{H}{h}\rfloor \times \lfloor \frac{W}{w}\rfloor \times \lfloor \frac{T}{t}\rfloor}$, where $N = h \cdot w \cdot t$. The above process refers to Fig. 1(f).

Each cube-shaped patch, x_i, is then projected into a token. The rearrange operator $\mathbf{G}$ projects each cube into a flattened token $\mathbf{Z}_i = \mathbf{G}x_i$. These tokens are concatenated to form a sequence of Patches, which is then prefixed with a learnable Clstoken $\mathbf{Z}_{cls}$. The positional embedding **Pos** is applied to the sequence. The process can be denoted as

$$\mathbf{Patches}_0 = [\mathbf{G}x_1, \mathbf{G}x_2, \ldots, \mathbf{G}x_N] \tag{6}$$

$$\mathbf{Z}_0 = [\mathbf{Z}_{\text{cls}}, \mathbf{Patches}_0] + \mathbf{Pos} = \left[\mathbf{Z}_{\text{cls}'}, \mathbf{Patches}'_0\right]. \tag{7}$$

2) Preliminary, Transformer Encoder With Multihead Self-Patch Attention: Transformer encoder with multihead self-patch attention mainly consists of multihead self-patch attention (MPA) module and multilayer perceptron (MLP) module. For the convenience of explaining the proposed method, MPA* module, compared with MPA module in [44], is without Layernorm (LN) layer, as shown in the orange box of Fig. 5. Two LN layers are respectively applied before the block. The remaining Layernorm layer is placed on the side of the MPA* module. The residual connections are after every module. The MLP contains two layers with a GELU nonlinearity. The sequence of tokens

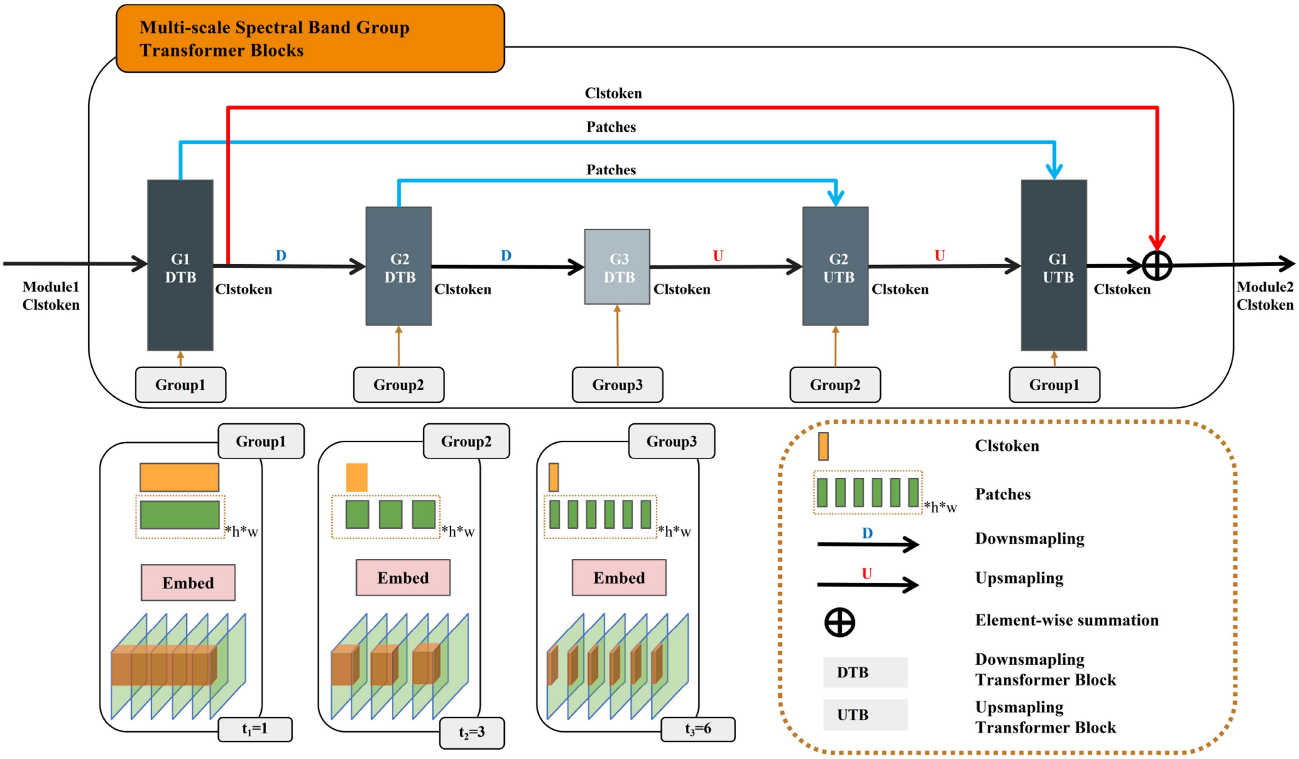


Fig. 4. Crossing structure of the multiscale spectral band group transformer blocks.

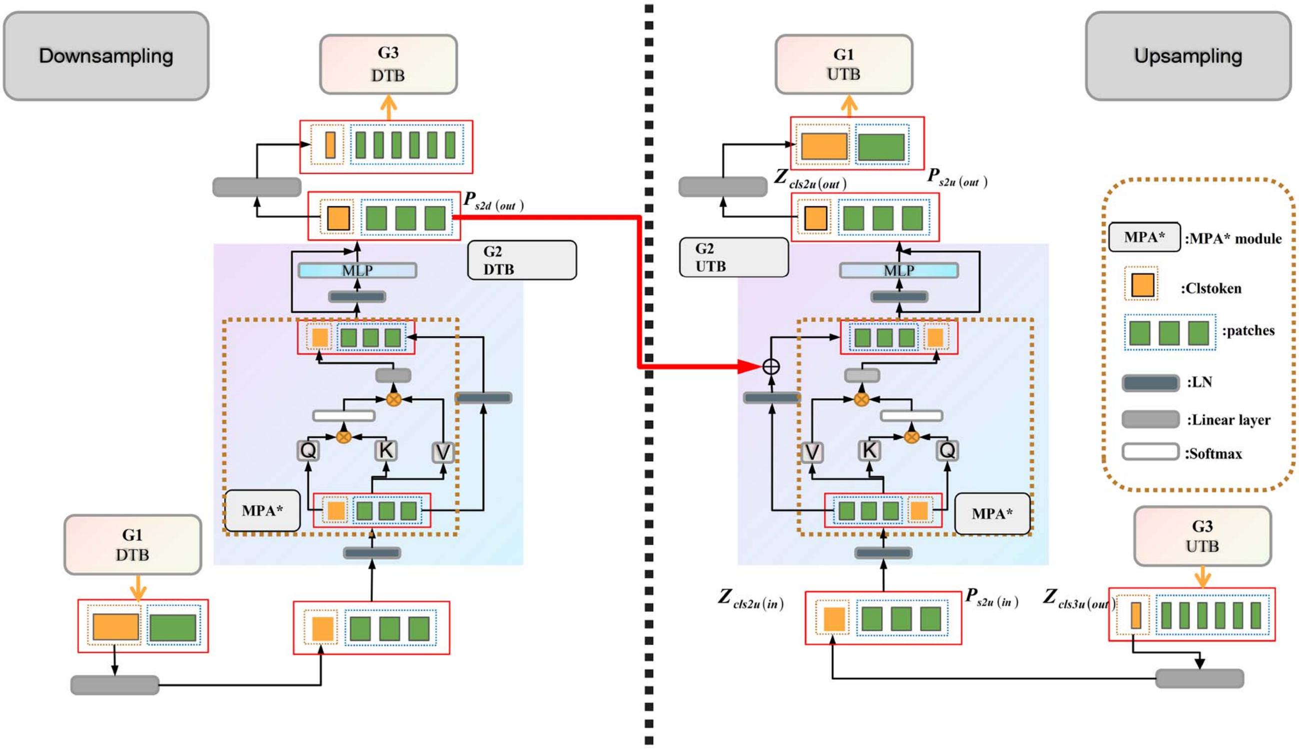


Fig. 5. Horizontal and vertical feature cross-fusion between downsampling and upsampling $Group_2$ transformer blocks.

$\mathbf{Z}$ is processed by a transformer encoder of L layers consisting of the following operations:

$$\mathbf{Z}'_\ell = \mathrm{MPA}\left(\mathrm{LN}\left(\mathbf{Z}_{\ell-1}\right)\right) \parallel \mathrm{LN}\left(\mathbf{Patches}'_{\ell-1}\right), \tag{8}$$

$$\mathbf{Z}_\ell = \mathrm{MLP}\left(\mathrm{LN}\left(\mathbf{Z}'_\ell\right)\right) + \mathbf{Z}'_\ell, \quad \ell = 1 \ldots L \tag{9}$$

where $\parallel$ represents the concatenation operation.

3) Multiscale Spectral Band Groups: In order to obtain scale-invariant unmixing features at the spectral band level, we establish multiscale spectral band groups for the hyperspectral unmixing network. The differences among each spectral band group stem from the size of the 3-D embedding scale. In order to focus in extracting the spectral band features, the embedding scales of h and w are consistent for every spectral band group, but the setting of t for each spectral band group is different. With different t setting, tokenization generate different numbers of tokens. A larger embedding scale corresponds to a set of larger patches, generating a smaller number of tokens during the encoding process. Similarly, a smaller embedding scale corresponds to a set of smaller patches, generating a larger number of tokens during the encoding process. We take three spectral band groups as an example and demonstrate the feature embedding process in the lower part of the Fig. 4, where $Group_1$ embedding scale setting:(h, w, t_1), $Group_2$ embedding scale setting:(h, w, t_2), $Group_3$ embedding scale setting:(h, w, t_3) and $1 = t_1 < t_2 < t_3$. Intuitively, fine-grained spectral information can be captured by smaller spectral band groups, while larger spectral band groups capture the slow-changing semantics of the bands. As each spectral band group captures unmixing information at different levels, we utilize a bidirectional cross-fusion mechanism to enhance the correlation of multiscale spectral band group features at different levels, as described in the following section.

4) Bidirectional Cross-Fusion Mechanism of Multiscale Spectral Band Group Features: In the upper part of Fig. 4, multiscale spectral band group transformer blocks are cascaded through upsampling $(Group_3 \rightarrow Group_2 \rightarrow Group_1)$ and downsampling $(Group_1 \rightarrow Group_2 \rightarrow Group_3)$ operations to achieve feature decoupling. In the horizontal direction, the Clstoken realizes feature communication of different levels of spectral band groups. In the vertical direction, we perform vertical cascading of Patches at the same level of spectral band group.

In the horizontal direction, the Clstoken of traditional ViT embedding, which is in the form of empty value or random initialization is concatenated with Patches. In this module, the current Clstoken is from Clstoken of last group. In this way, the umixing features are flowed horizontally among multiscale spectral band group transformer blocks [53]. To ensure multiscale spectral band group transformer blocks can be cascaded adjacent to each other, a set of linear layers are used to unify the dimension of adjacent spectral band groups. Thus, the process achieves a horizontal flow of features from different spectral band groups. It is worth noting that we concatenate the Clstoken of Module1 with the Patches of the first downsampled spectral band group transformer block ($G1$ DTB), ensuring that the feature information extracted by Module1 is passed to Module2 [54]. The embedding scale of Module1 is consistent with $G1$ DTB, where $t_1 = 1$. In the vertical direction, the output Patches of the downsampling transformer block are input into the symmetrical upsampling transformer block at the same level via skip connections. Finally, to prevent attenuation of long-distance feature information, a skip connection is adopted between the output of the first downsampling transformer block and the output of the last upsampling view transformer block.

Algorithm 1: The Bidirectional Cross-Fusion of $Group2$ UTB.

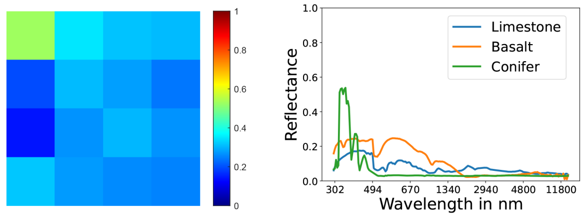
```
INPUT:
    the output Clstoken of $G3$ UTB: $Z_{cls3u}(out)$
    the input Patches of $G2$ UTB: $P_{s2u}(in)$
    the output Patches of $G2$ DTB: $P_{s2d}(out)$
OUTPUT:
    the output Clstoken of $G2$ UTB: $Z_{cls2u}(out)$
    the output Patches of $G2$ UTB: $P_{s2u}(out)$
BEGIN
    $X_1 = P_{s2u}(in) \parallel linear(Z_{cls3u}(out))$
    $X_2 = \mathrm{LN}[X_1]$
    $X_3 = MPA[X_2]$
    $X_4 = X_3 \parallel [\mathrm{LN}(P_{s2u}(in)){+}P_{s2d}(out)]$
    $X_5 = X_4 + MLP(\mathrm{LN}(X_4))$
    $Z_{cls2u}(out) = X_5(1, :)$
    $P_{s2u}(out) = X_5(2, :)$
END
```

The micrograph of the bidirectional cross-fusion is shown in the Fig. 5. We use upsampling $Group_2$ transformer block ($G2$ UTB) as the example to describe the fusion process in detail. In the horizontal direction, the output Clstoken of upsampling $Group_3$ transformer block ($G3$ UTB), $Z_{cls3u}(out) \in \mathbb{R}^{1 \cdot \lfloor \frac{T}{t_3} \rfloor \cdot \lfloor \frac{H}{h} \rfloor \cdot \lfloor \frac{W}{w} \rfloor}$ is dimensionally increased to $Z_{cls2u}(in) \in \mathbb{R}^{1 \cdot \lfloor \frac{T}{t_2} \rfloor \cdot \lfloor \frac{H}{h} \rfloor \cdot \lfloor \frac{W}{w} \rfloor}$ of $G2$ UTB by a linear layer. The above operation ensures that $Z_{cls2u}(in)$ can be concatenated with the Patches of $G2$ UTB, $P_{s2u}(in)$,which is embedded by $Group_2$ embedding scale setting. The concatenated sequence is input into $G2$ UTB. Similarly, the output Clstoken $Z_{cls2u}(out)$of $G2$ UTB transformer block is dimensionally increased and inputted into the transformer block of $G1$ UTB. In the vertical direction, the output Patches $P_{s2d}(out)$ of $G2$ DTB are vertically input into $G2$ UTB. In the $G2$ UTB, $P_{s2d}(out)$ utilize the skip connection to execute element-wise addition in the branch of LN layer, which is beside MPA* module.

Based on bidirectional cross-fusion of Clstoken and Patches, we perform cascading perceptual fusion between spectral band groups. Intuitively, the fusion allows the final spectral band group to aggregate multiscale information from all preceding spectral band groups. For $G2$ UTB, the pseudocode of the bidirectional cascading perceptual fusion is shown in Algorithm 1.

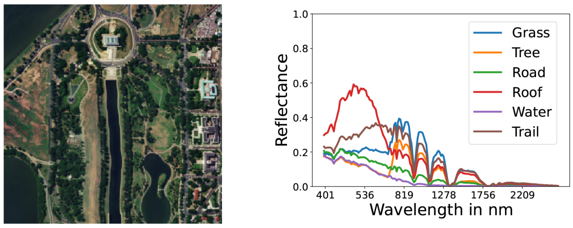
III. Experiments

To evaluate the proposed method, we carry out comprehensive experiments on one synthetic dataset and three representative real datasets, namely Jasper dataset, Washington DC Mall dataset, and Samson dataset. The performance of the proposed



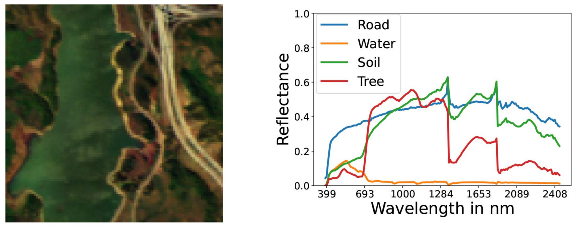
(a) (b)

Fig. 6. Synthetic Dataset. (a) True-color band. (b) Endmembers.



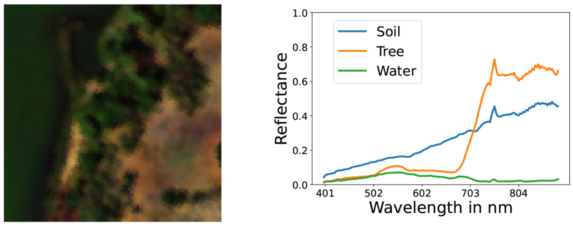
(a) (b)

Fig. 8. Washington DC Mall Dataset. (a) True-color image. (b) Endmembers.



(a) (b)

Fig. 7. Jasper Dataset. (a) True-color image. (b) Endmembers.



(a) (b)

Fig. 9. Samson Dataset. (a) True-color image. (b) Endmembers.

model is compared to six different unmixing methods from different categories: classical unmixing method VCA&FCLS [10], [55], CNN-based unmixing method CNNAUE [37], cycle-consistency unmixing method CyCU-Net [41], deep autoencoder unmixing method DAUE [31], baseline ViT-based unmixing method DeepTrans-Net [44], and Swin Transformer unmixing method UST-Net [46].

In the next subsections, we first introduce the datasets, experimental setup and quantitative performance metrics. Second, the synthetic dataset is experimented to prove the noise robustness of DET-NET. Then, we report our results on the real datasets. Finally, we perform the ablation experiment to verify the performance of two core modules.

A. Datasets

1) Synthetic Dataset: As the Fig. 6 depicted, the synthetic dataset of 80 × 80 pixels is simulated by linear mixtures of three endmembers from the ENVI ASTER spectral library: Limestone, Basalt, and Conifer. These endmembers are shown in Fig. 6 and contain 200 reflection values in the wavelength range [302–13600] nm. The synthetic dataset, whose abundance structure is the same as [44], contains 16 squares, each measuring 20 × 20 pixels, with distinct ternary mixtures.

2) Jasper Ridge Dataset (Jasper Dataset): The Jasper Ridge Dataset was collected over Jasper Ridge by the AVIRIS sensor in central California, USA. The dataset consists of raw data with 512 × 614 pixels, each recorded in 224 channels ranging from 380 to 2500 nm. The spectral resolution is up to 9.46 nm. Since this HSI is too complex to get the ground truth, we consider a subimage of 100 × 100 pixels. Due to dense water vapor and atmospheric effects, 198 channels are retained. The endmembers in this dataset are Soil, Water, Tree, and Road, as shown in Fig. 7. The GT endmembers and abundance maps are obtained from rslab.[1]

3) Washington DC Mall Dataset (DC Dataset): The hyperspectral digital image acquisition experiment (HYDICE) sensor was used to collect the HSI over the Washington DC mall. The original dataset have 1280 × 307 pixels. The cropped image is an area of 290 × 290 pixels. The dataset is with 191 bands, covering wavelengths from 400 to 2400 nm. Six endmembers are obtained from this image, namely, Grass, Tree, Roof, Road, Water, and Trail, as shown in Fig. 8. The GT endmembers and abundance maps are obtained from [44].

4) Samson Dataset: The Samson dataset, acquired through the SAMSON sensor and illustrated in Fig. 9, consists of a region of interest (ROI) with three endmembers: Soil, Water, and Tree. The original dataset comprises 952 × 952 pixels and 156 channels, with a high spectral resolution of 3.13 nm. Due to computational constraints, the analysis focuses on a reduced ROI of 95 × 95 pixels. The GT endmembers and abundance maps are obtained from [44].

B. Experiment Setup

The experimental results of the deep unmixing network depend on the settings of hyperparameters. The hyperparameter settings of the deep unmixing network proposed in this article are as follows: The number of self-attention heads is set to 4. DET-Net is a multiscale network that contains multiple Transformer blocks. In TVA transformer module, the transformer blocks of the main view are set to 2 layers, and the transformer blocks of the top view and left view are set to 1 layer. In BGF transformer module, the number of band groups is set to 5. The computational complexity of proposed method is shown in the Table I. For

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

TABLE I
COMPUTATIONAL COMPLEXITY ON JASPER DATASET

Proposed Method		Module1		Module2	
FLOPs	Param	FLOPs	Param	FLOPs	Param
3.750G	18.062M	2.446G	13.512M	1.678G	5.736M

floating point operation (FLOPs) and Param, Module 1 is with a relatively high computational complexity, while Module 2 is with a relatively low computational complexity. Compared to six unmixing methods in this article, DET-Net is with the higher complexity, and still has potential improvement on Flops and Params. For the embedding scale of *Group*, five different t values are set as:$[t_1, t_2, t_3, t_4, t_5] = [1, 2, 3, 4, 6]$. To implement dynamic learning rates, the StepLr scheduler is adopted for the deep unmixing network. In the loss function, two parameters need to be set. The specific hyperparameter settings are provided in Table II.

C. Quantitative Performance Metrics

Two quantitative performance metrics, namely, the root MSE (RMSE) of the abundance and the SAD of the endmember, are utilized to evaluate unmixing method. The formulas of these indicators are as follows:

$$\mathrm{RMSE}(\mathbf{A}, \hat{\mathbf{A}}) = \sqrt{\frac{1}{P \cdot Q} \sum_{i=1}^{P} \sum_{j=1}^{Q} \left(\hat{\mathbf{A}}_{ij} - \mathbf{A}_{ij} \right)^2} \tag{10}$$

$$\mathrm{SAD}(\mathbf{E}, \hat{\mathbf{E}}) = \frac{1}{P} \sum_{i=1}^{P} \arccos \left(\frac{\left\langle \mathbf{E}_{(i)}, \hat{\mathbf{E}}_{(i)} \right\rangle}{\left\| \mathbf{E}_{(i)} \right\|_2 \left\| \hat{\mathbf{E}_{(i)}} \right\|_2} \right) \tag{11}$$

where A and $\hat{A}$ denote the GT abundance and the estimated abundance, respectively, and $\hat{E}$ and E denote the extracted endmember and GT endmember, respectively.

D. Noise Robustness Experiment

In this experiment, we analyzed the algorithm's noise robustness on synthetic dataset with different levels of noise pollution, namely, 10 dB, 20 dB, and 30 dB signal-to-noise ratio (SNR) noise. Quantitative experimental result are shown in Table III. As the noise level decreases, the unmixing performance of the algorithm is more precise. To demonstrate the effectiveness of our proposed network, we compare the endmember extraction results with other six unmixing methods under the SNR of 20 dB, as shown in Fig. 10. Our experiment results achieve the best performance of endmember extraction and anundance estimation, demonstrating the powerful unmixing capabilities of the algorithm. It is worth noting that the structured abundance maps may affect the RMSE evaluation criterion of the CNNAUE. The visualization of abundance estimation is shown in Fig. 11. As the noise decreases, the noise spots of the abundance map become less and less. The visualization of endmember extraction is shown in Fig. 12. As expected, the endmember signatures obtained by the proposed algorithm are almost the same as the ones provided by GT under various noise levels.

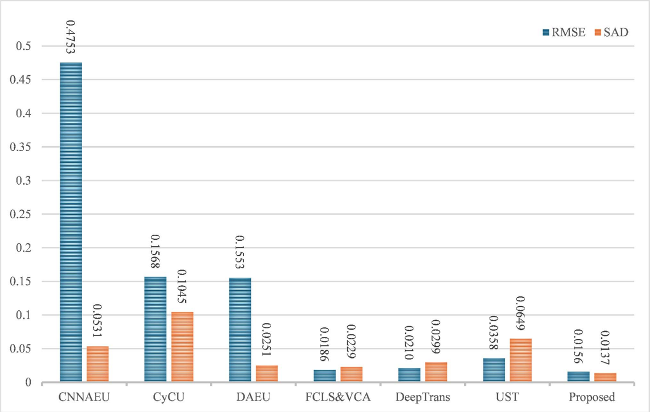


Fig. 10. Under 20-db noise, endmember extraction and abundance estimation results of seven method on the synthetic data.

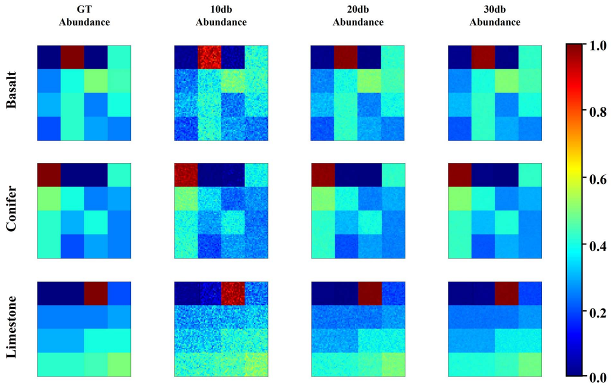


Fig. 11. Abundance estimation of the synthetic dataset. The first column is the GT of abundance maps.

E. Experiment With Real Datasets

1) Jasper Dataset: The results of Jasper dataset are presented in Tables IV and V. The proposed method achieved the best performance on the overall criterion for endmember extraction and abundance estimation. CNNAUE and DAUE excessively focus on endmember extraction, leading to significant error propagation in abundance estimation. While UST-Net achieves a balance between endmember extraction and abundance estimation, the experimental results do not reach the optimal performance. Due to poor performance on the Water endmember, the overall unmixing effect of the CyCU-Net is affected. The Road endmember of the Jasper dataset is found to be a challenge for the other methods, however, our proposed method obtains the best results. Figs. 13 and 14 further illustrate the competitiveness of our method.

2) Washington DC Mall Dataset: This dataset comprises 290 × 290 pixels with 191 bands and contains a great number of endmembers. Consequently, the DC dataset poses higher challenges for deep unmixing methods. Tables VI and VII display the quantification results. In the comprehensive assessment, the

TABLE II
HYPERPARAMETERS FOR THE PROPOSED METHOD

Hyperparameters	Synthetic dataset			Jasper dataset	Washington DC Mall dataset
	10db	20db	30db		
Embedding Scale	$(5 \times 5 \times 3)$	$(5 \times 5 \times 3)$	$(5 \times 5 \times 3)$	$(5 \times 5 \times 3)$	$(10 \times 10 \times 3)$
T	24	24	24	24	24
α	1e4	2e4	3e4	4.3e3	5e3
β	5e-2	5e-2	5e-2	8.2e-2	2e-4
Epoch	100	150	200	230	140
Learning rate	4e-3	4e-3	4e-3	9e-3	6e-3
Step Lr	(15, 0.8)	(15, 0.8)	(15, 0.8)	(15, 0.8)	(15, 0.8)

TABLE III
QUANTITATIVE RESULTS OF NOISE ROBUSTNESS EXPERIMENTS

Synthetic Dataset	10db		20db		30db	
	RMSE	SAD	RMSE	SAD	RMSE	SAD
Basalt	0.0435	0.0245	0.0158	0.0065	**0.0101**	**0.0059**
Conifer	0.0333	0.0275	0.0108	0.0141	**0.0089**	**0.0179**
Limestone	0.0562	0.0265	0.0191	0.0206	**0.0141**	**0.0035**
Overall	0.0453	0.0261	0.0156	0.0137	**0.0113**	**0.0091**

The best performances are shown in bold.

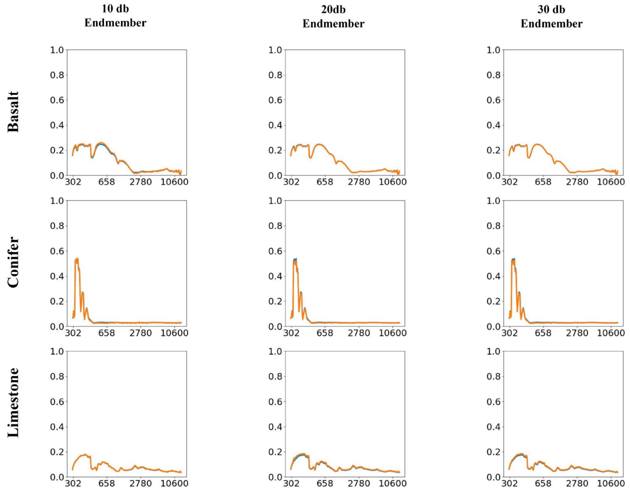


Fig. 12. Endmember extraction visual results of the synthetic dataset, where the orange lines are the endmembers extraction result and GT is in blue, wavelength in nm.

TABLE IV
RMSE (JASPER DATASET)

RMSE	CNNAUE	CyCU	DAUE	FCLS	DeepTrans	UST	Proposed
Road	0.7216	0.1254	0.6729	0.1589	0.1162	0.0958	**0.0648**
Soil	0.7544	0.1339	0.6528	0.2045	0.1296	0.1358	**0.0925**
Tree	0.5653	0.0929	0.5203	0.1302	0.0859	**0.0794**	0.0850
Water	0.2582	0.1049	0.2053	0.1143	0.0675	**0.0312**	0.0635
Overall	0.6075	0.1154	0.5458	0.1558	0.1028	0.0934	**0.0775**

The best performances are shown in bold.

TABLE V
SAD (JASPER DATASET)

SAD	CNNAUE	CyCU	DAUE	VCA	DeepTrans	UST	Proposed
Road	0.1442	0.0494	0.1302	0.0901	0.0521	0.0409	**0.0248**
Soil	0.1246	**0.0284**	0.0769	0.0893	0.0937	0.0986	0.0533
Tree	0.0734	**0.0374**	0.0251	0.1166	0.0255	0.0483	0.0446
Water	0.0607	0.1459	**0.0285**	0.1481	0.0455	0.0346	0.0312
Overall	0.1007	0.0653	0.0651	0.1110	0.0542	0.0556	**0.0385**

The best performances are shown in bold.

TABLE VI
RMSE (DC DATASET)

RMSE	CNNAUE	CyCU	DAUE	FCLS	DeepTrans	UST	Proposed
Grass	0.4085	0.2036	0.2718	0.2651	0.1770	0.3083	**0.1601**
Road	0.3021	0.2731	**0.1274**	0.3354	0.1597	0.1619	0.1290
Roof	0.1039	0.0829	0.3103	**0.0446**	0.0844	0.1477	0.0780
Tree	0.3440	0.3250	0.3267	0.1712	0.0983	0.2603	**0.0947**
Trail	0.2224	0.1544	0.1106	**0.1099**	0.1631	0.1270	0.1233
Water	0.1794	0.3012	**0.1176**	0.2626	0.1590	0.1224	0.1258
Overall	0.2795	0.2391	0.2307	0.2218	0.1446	0.2008	**0.1214**

The best performances are shown in bold.

TABLE VII
SAD (DC DATASET)

SAD	CNNAUE	CyCU	DAUE	VCA	DeepTrans	UST	Proposed
Grass	0.0835	0.5548	**0.0555**	0.3221	0.2645	0.1756	0.2055
Road	0.2328	0.4676	0.1747	0.5682	0.1743	0.1450	**0.0437**
Roof	0.7856	0.4342	0.5204	**0.0326**	0.3071	0.6754	0.3351
Tree	0.1327	0.6639	0.2109	0.2070	0.1418	0.2555	**0.0884**
Trail	0.2254	0.1404	**0.0517**	0.1073	0.1347	0.3719	0.0585
Water	0.0263	0.6488	**0.0157**	0.0338	0.0923	0.0367	0.0215
Overall	0.2477	0.4849	0.1715	0.2118	0.1858	0.2767	**0.1254**

The best performances are shown in bold.

proposed algorithm achieves the best overall performance both in endmember extraction and abundance estimation. The visual results of endmember extraction and abundance estimation are

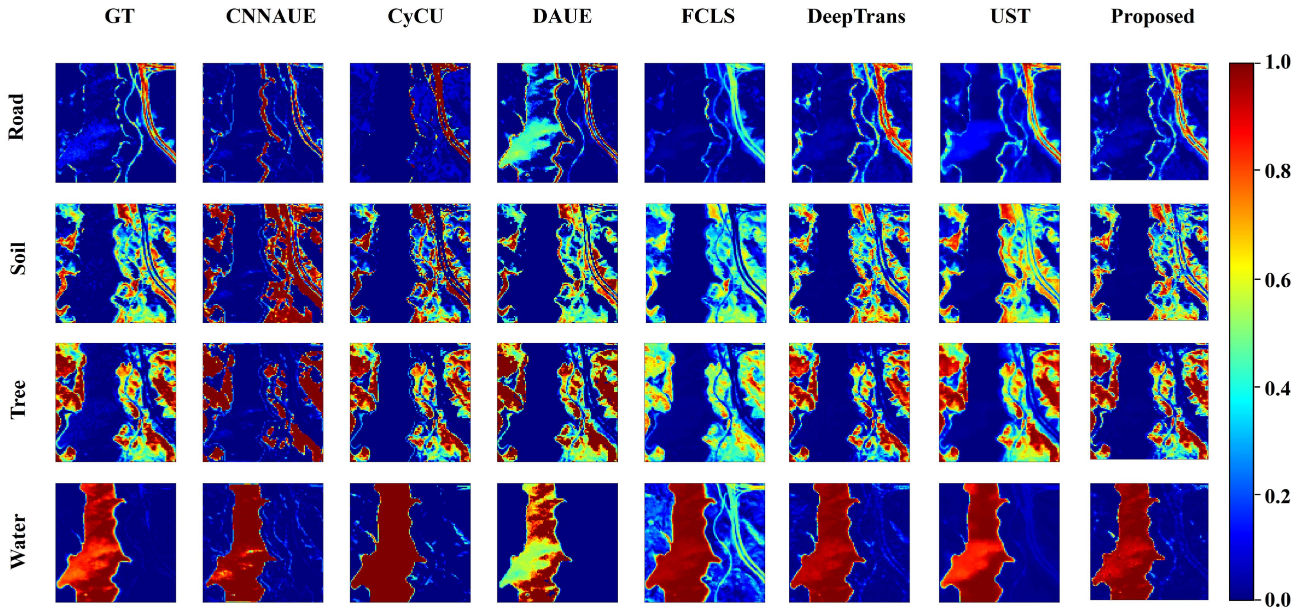
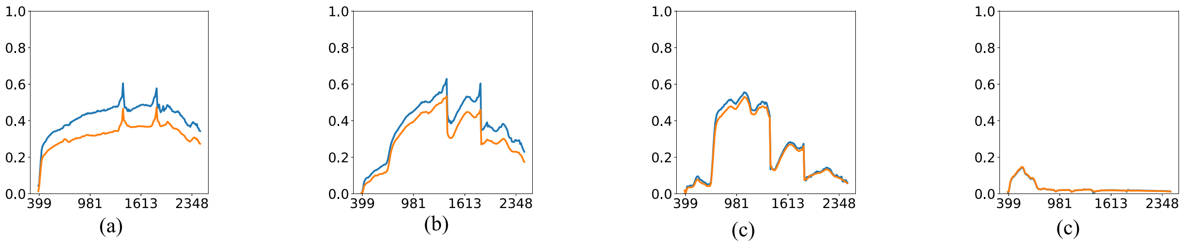


Fig. 13. Abundance estimation of the Jasper dataset is compared with six methods. The first column is the GT of abundance maps.



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

Fig. 14. Endmember extraction visual results of the Jasper dataset, where the orange lines are the endmembers extraction result and GT is in blue, wavelength in nm. (a) Road. (b) Soil. (c) Tree. (d) Water.

TABLE VIII
RMSE (SAMSON DATASET)

RMSE	CNNAUE	CyCU	DAUE	VCA	DeepTrans	UST	Proposed
Soil	0.3175	0.2873	0.2258	0.1758	0.0826	0.1250	**0.0672**
Tree	0.1899	0.1618	0.2061	0.0772	0.0755	0.1469	**0.0753**
Water	0.2649	0.1847	0.3164	0.1588	0.0995	0.2073	**0.0812**
Overall	0.2627	0.2182	0.2540	0.1438	0.0864	0.1635	**0.0748**

The best performances are shown in bold.

TABLE IX
SAD (SAMSON DATASET)

SAD	CNNAUE	CyCU	DAUE	VCA	DeepTrans	UST	Proposed
Soil	0.0371	0.0597	0.0380	0.0141	0.0259	**0.0102**	0.0146
Tree	0.0734	0.0508	0.0634	0.0958	0.0757	0.0770	**0.0436**
Water	0.0753	0.1008	**0.0423**	0.1554	0.1037	0.0489	0.0634
Overall	0.0619	0.0705	0.0479	0.0924	0.0645	0.0454	**0.0405**

The best performances are shown in bold.

depicted in Figs. 15 and 16. As for the Roof abundance estimation result, all the algorithms fail to achieve the results at the level of 0.01 except for the VCA&FCLS. Comparing the abundance estimation visual results, we found that some of the Trail have been recognized as the Roof. This may be due to the fact that the main constituents of Trail and Roof are mainly made of civil materials, such as concrete.

3) Samson Dataset: The quantitative experimental results of Samson dataset are presented in Tables VIII and IX. Most methods have achieved good results in endmember extraction. Among them, the proposed method outperforms the other unmixing methods with a mean RMSE value of 0.0748 and a mean SAD value of 0.0405. The visualization of endmember extraction and abundance estimation is depicted in Figs. 17 and 18. These results attest to the effectiveness of the proposed deep unmixing method.

F. Module Ablation Experiment

The proposed deep unmixing method relies on two core modules, included Module1: TVA module with 2-D embedding and Module2: BGF module with 3-D embedding. Ablation

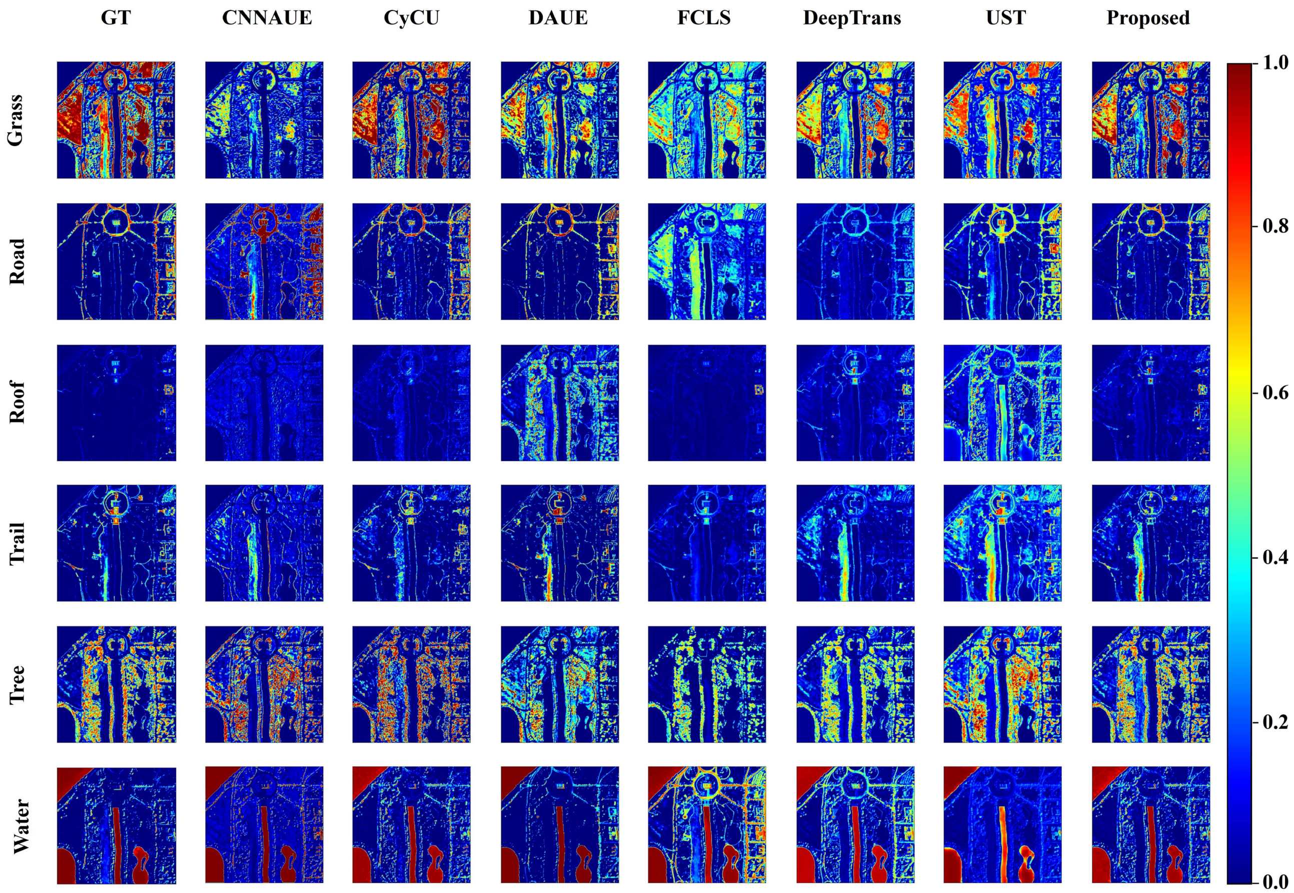


Fig. 15. Abundance estimation of the Washington DC Mall dataset is compared with six methods. The first column is the GT of abundance maps.

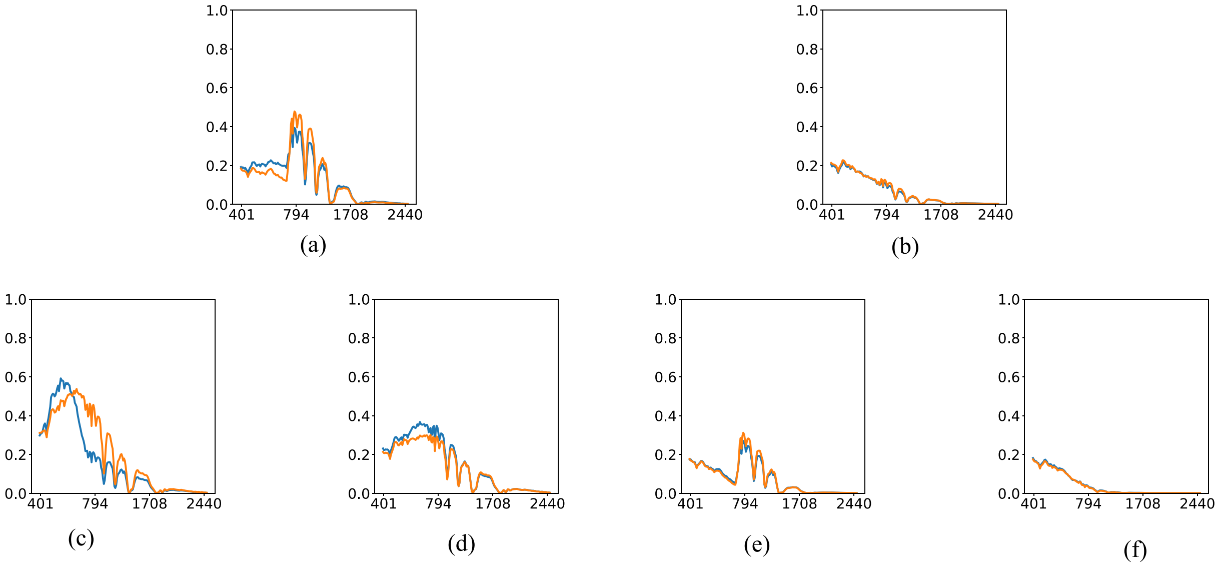


Fig. 16. Endmember extraction visual results of the DC dataset, where the orange lines are the endmembers extraction result and GT is in blue, wavelength in nm. (a) Grass. (b) Road. (c) Roof. (d) Trail. (e) Tree. (f) Water.

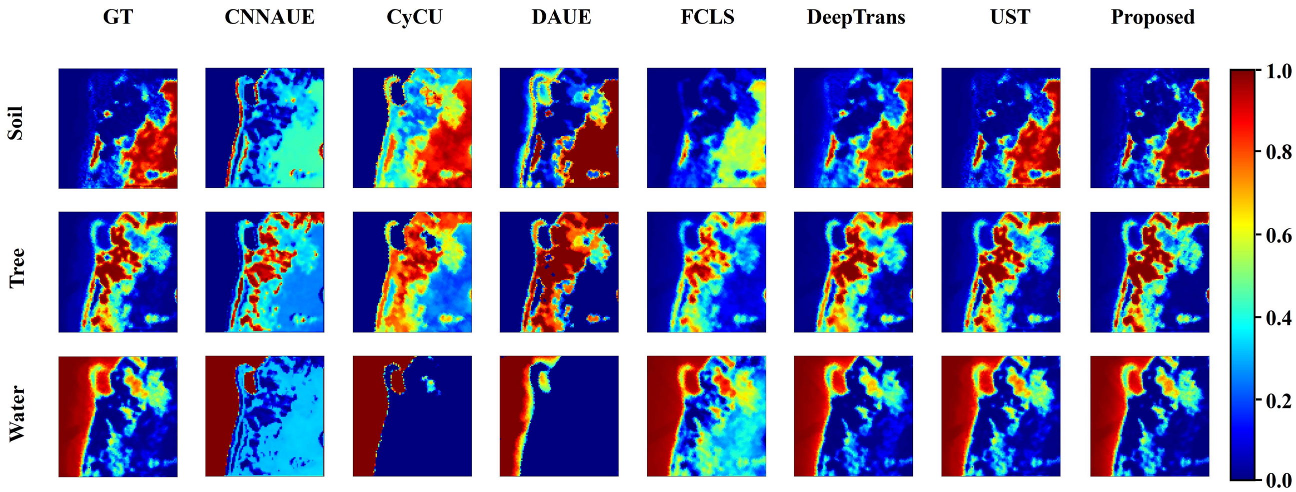
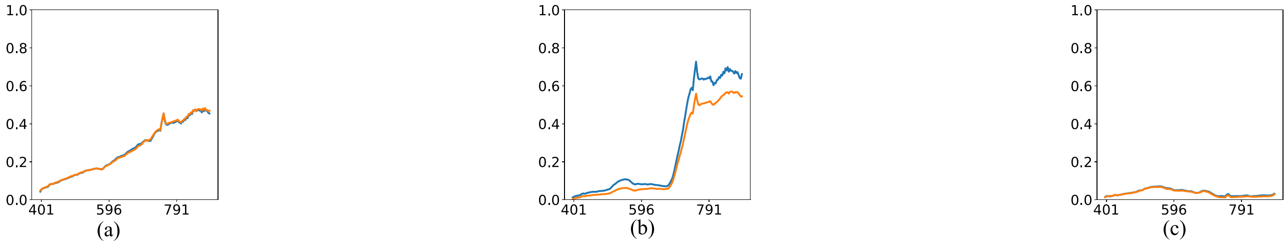


Fig. 17. Abundance estimation of the Samson dataset is compared with six methods. The first column is the GT of abundance maps.



(a) (b) (c)

Fig. 18. Endmember extraction visual results of the Samson dataset, where the orange lines are the endmembers extraction result and GT is in blue, wavelength in nm. (a) Soil. (b) Tree. (c) Water.

TABLE X
Quantitative Results of Module Ablation Experiment

Module1	Module2	Synthetic data(20db)		Jasper		DC	
		RMSE	SAD	RMSE	SAD	RMSE	SAD
✓	✗	0.0171	0.0077	0.0686	0.0507	0.1539	0.1309
✗	✓	0.0139	0.0169	0.0690	0.0628	0.1138	0.1380
✓	✓	0.0156	0.0137	0.0775	0.0385	0.1214	0.1254

The best performances are shown in red, and the suboptimal performances are shown in blue.

experiments are designed to validate the contributions of the two core modules for the DET-Net. The experimental results of the comparative analysis are shown in Table X. Although the single module performs well in one evaluation criterion, it is built upon sacrificing the other evaluation criterion. The collaboration of the two core modules achieved excellent and balanced performance in the comprehensive experimental results. Table XI presents the quantitative endmembers subresults of module ablation experiment on the Washington Mall dataset. As shown in Table XI, except for the SAD of Tree endmember, the proposed method quantization results of each endmember are covered by the optimal and suboptimal performances. For single module, the optimal performance is covered by staggered distribution on Module1 and Module2. Among them, The optimal experimental performance of Module2 is mostly on RMSE performance metric. SAD performance metric is more suitable for the optimal experimental results of Module1. But Module2 is more outstanding compared to Module1. Either optimal experimental performance or the optimal and suboptimal experimental performances, Module2 occupies a higher proportion. Meanwhile, Table XI presents the applicability for individual endmembers. On the Road and Trail endmembers, Module2's RMSE and SAD are ahead of Module1. On the Roof endmember, the RMSE and SAD of Module1 are ahead of Module2.

In Fig. 19, we utilize the Washington DC mall dataset as an example to visually analyze the ablation experiment. In the endmember extraction results, the visualization results of Roof

TABLE XI
QUANTITATIVE ENDMEMBER SUB-RESULTS OF MODULE ABLATION EXPERIMENT ON WASHINGTON DC MALL DATASET

Module1	Module2	Grass		Road		Roof		Tree		Trail		Water	
		RMSE	SAD	RMSE	SAD	RMSE	SAD	RMSE	SAD	RMSE	SAD	RMSE	SAD
✓	✗	0.1761	0.1491	0.1775	0.1435	0.0764	0.3476	0.1849	0.0581	0.1413	0.0623	0.1400	0.0249
✗	✓	0.1508	0.2078	0.1073	0.0953	0.0810	0.3767	0.1266	0.0702	0.1004	0.0522	0.1034	0.0260
✓	✓	0.1601	0.2055	0.1290	0.0437	0.0780	0.3351	0.0947	0.0884	0.1233	0.0585	0.1258	0.0215

The best performances are shown in red, and the suboptimal performances are shown in blue.

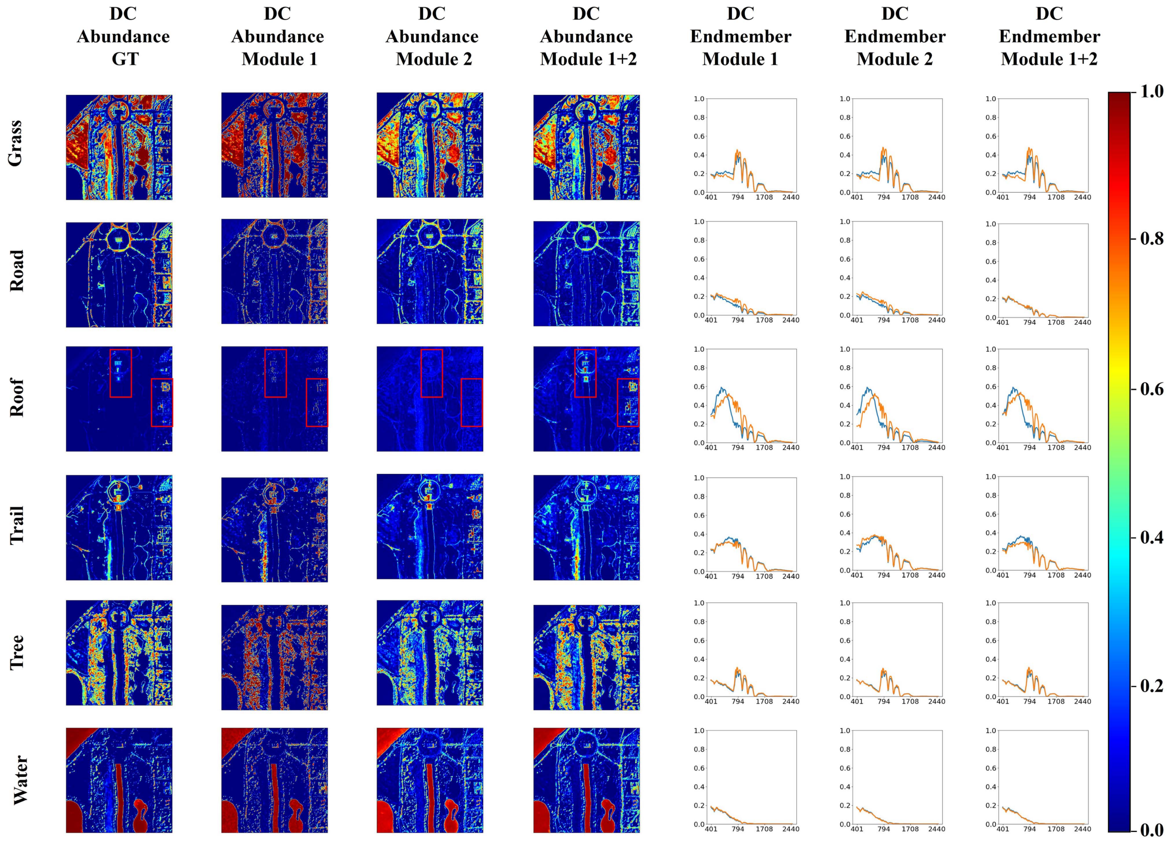


Fig. 19. Module ablation visual results for the Washington DC Mall dataset, wavelength in nm.

endmember and Trail endmember are significantly different. In visual abundance estimation, the two modules exhibit complementary effects on the detailed information of abundance maps. Specifically, in the abundance estimation of Grass and Tree, Module2: BGF module gifts DET-Net with the advantage of vegetation detection. In the abundance estimation of Roof, Module1: TVA module gifts DET-Net with the ability to detect small texture features in Roof abundance map, as shown in the red boxes of Fig. 19.

1) Module1 Ablation Experiment: To evaluate the effectiveness of the TVA module, we compare the TVA module with three-view spatial attention and the ViT-based method with only 2-D embedding on the main view in Table XII. The depth of the Transformer Block for the method being compared is set to 2. Except for RMSE metric for the Washington DC mall dataset, TVA module with three-view spatial attention dominates all optimal performances. These results indicate the effectiveness of TVA module.

TABLE XII
QUANTITATIVE RESULTS OF THREE VIEW ATTENTION ABLATION EXPERIMENT

	Synthetic data(20db)		Jasper		DC	
	RMSE	SAD	RMSE	SAD	RMSE	SAD
Main View	0.0216	0.0295	0.0990	0.0510	**0.1336**	0.1610
Three Views	**0.0171**	**0.0077**	**0.0686**	**0.0507**	0.1539	**0.1309**

The best performances are shown in bold.

2) Module2 Ablation Experiment: The number of spectral band group in BGF module affects the experimental performance of the deep unmixing network. In order to eliminate the influence of TVA module, the network for the ablation

TABLE XIII
QUANTITATIVE RESULTS OF SPECTRAL BAND GROUP ABLATION EXPERIMENT

Groups	Synthetic data(20db)		Jasper		DC	
	RMSE	SAD	RMSE	SAD	RMSE	SAD
2	0.0449	0.0189	0.0840	**0.0408**	0.1420	0.1424
3	0.0176	0.0181	0.0784	0.0412	0.1246	0.1425
4	0.0156	**0.0147**	0.0753	0.0688	0.1186	**0.1328**
5	**0.0139**	0.0169	**0.0690**	0.0628	**0.1138**	0.1380

The best performances are shown in bold.

experiment only included Module2: BGF module. The ablation experiment results are presented in Table XIII. For the Jasper dataset, as the number of band groups increases, the balanced performance of RMSE and SAD gradually is emerged. For the synthetic dataset and the Washington Mall dataset dataset, the best unmixing quantitative experimental results are performed in the large number of spectral band groups, which roughly conform to the characteristics of the pyramid feature structure.

IV. CONCLUSION

In this article, we introduce a novel hyperspectral deep unmixing network, DET-Net, consisting of two core modules, namely TVA transformer module with 2-D embedding and BGF transformer module with 3-D embedding. In TVA transformer module, the main view features are primary to establish an attention mechanism that integrates top view and left view feature information. In BGF transformer module, we utilize the 3-D band group spatial-spectral cubed embedding and the bidirectional cross-fusion to achieve multiscale spectral band group feature fusion. Based on two core transformer modules, the proposed unmixing network can extract spatial-spectral features for the cube-shaped HSIs. Experimental results on the synthetic dataset and real datasets indicate the effectiveness of DET-Net. The ablation experiment proves that both core modules play the complementary and effective roles in the proposed network. In order to explore band by band hyperspectral unmixing, an important future work will be on how to enable spatial-spectral band by band embedding to select fewer and more critical tokens via sparse representation.

ACKNOWLEDGMENT

The authors would like to thank the authors of the methods compared in this article for the code available on GitHub.

REFERENCES

[1] B. Rasti et al., "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, Dec. 2020.

[2] J. Feng et al., "Class-aligned and class-balancing generative domain adaptation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5509617.

[3] Z. Gao, B. Pan, X. Xu, T. Li, and Z. Shi, "LiCa: Label-indicate-conditional-alignment domain generalization for pixel-wise hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5519011.

[4] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2019.

[5] T. D. McRae, D. Oleksyn, J. Miller, and Y.-R. Gao, "Robust blind spectral unmixing for fluorescence microscopy using unsupervised learning," *PLoS One*, vol. 14, no. 12, 2019, Art. no. e0225410.

[6] A. T. Badaró et al., "Near infrared hyperspectral imaging and spectral unmixing methods for evaluation of fiber distribution in enriched pasta," *Food Chem.*, vol. 343, 2021, Art. no. 128517.

[7] G. J. Edelman, E. Gaston, T. G. Van Leeuwen, P. Cullen, and M. C. Aalders, "Hyperspectral imaging for non-contact analysis of forensic traces," *Forensic Sci. Int.*, vol. 223, no. 1-3, pp. 28–39, 2012.

[8] M. E. Winter, "N-FINDR: An algorithm for fast autonomous spectral endmember determination in hyperspectral data," in *Proc. Imag. Spectrometry V*, SPIE, 1999, pp. 266–275.

[9] J. W. Boardman, "Automating spectral unmixing of AVIRIS data using convex geometry concepts," in *Proc. JPL, Summaries 4th Annu. JPL Airborne Geosci. Workshop*, 1993, pp. 11–14.

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

[11] C.-I. Chang, C.-C. Wu, W. Liu, and Y.-C. Ouyang, "A new growing method for simplex-based endmember extraction algorithm," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 10, pp. 2804–2819, Oct. 2006.

[12] J. H. Gruninger, A. J. Ratkowski, and M. L. Hoke, "The sequential maximum angle convex cone (SMACC) endmember model," in *Proc. Algorithms Technol. Multispectral, Hyperspectral, Ultraspectral Imagery X*, SPIE, 2004, pp. 1–14.

[13] S. Moussaoui, C. Carteret, D. Brie, and A. Mohammad-Djafari, "Bayesian analysis of spectral mixture data using Markov chain Monte Carlo methods," *Chemometrics Intell. Lab. Syst.*, vol. 81, no. 2, pp. 137–148, 2006.

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

[15] S. Moussaoui et al., "On the decomposition of Mars hyperspectral data by ICA and Bayesian positive source separation," *Neurocomputing*, vol. 71, no. 10-12, pp. 2194–2208, 2008.

[16] A. Zare and P. Gader, "Sparsity promoting iterated constrained endmember detection in hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 3, pp. 446–450, Jul. 2007.

[17] J. M. Bioucas-Dias and M. A. Figueiredo, "Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing," in *Proc. 2nd Workshop Hyperspectral Image Signal Process.: Evol. Remote Sens.*, 2010, pp. 1–4.

[18] L. Gao et al., "Multiple algorithm integration based on ant colony optimization for endmember extraction from hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2569–2582, Jun. 2015.

[19] L. Gao, L. Zhuang, Y. Wu, X. Sun, and B. Zhang, "A quantitative and comparative analysis of different preprocessing implementations of DPSO: A robust endmember extraction algorithm," *Soft Comput.*, vol. 20, pp. 4669–4683, 2016.

[20] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Process.*, vol. 24, no. 1, pp. 1–10, 1991.

[21] B. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Blind hyperspectral unmixing using autoencoders: A critical comparison," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1340–1372, 2022.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[24] B. Mann et al., "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[25] Q. Zhang, Q. Yuan, M. Song, H. Yu, and L. Zhang, "Cooperated spectral low-rankness prior and deep spatial prior for HSI unsupervised denoising," *IEEE Trans. Image Process.*, vol. 31, pp. 6356–6368, 2022.

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

[27] Y. Wang et al., "Spectral–spatial–temporal transformers for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536814.

[28] M. Sewak, S. K. Sahay, and H. Rathore, "An overview of deep learning architecture of deep neural networks and autoencoders," *J. Comput. Theor. Nanosci.*, vol. 17, no. 1, pp. 182–188, 2020.

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

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

[31] 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.

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

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

[34] D. Hong, J. Chanussot, N. Yokoya, U. Heiden, W. Heldens, and X. X. Zhu, "WU-Net: A weakly-supervised unmixing network for remotely sensed hyperspectral imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 373–376.

[35] Q. Jin, Y. Ma, X. Mei, and J. Ma, "TANet: An unsupervised two-stream autoencoder network for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5506215.

[36] Q. Jin, Y. Ma, F. Fan, J. Huang, X. Mei, and J. Ma, "Adversarial autoencoder network for hyperspectral unmixing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 8, pp. 4555–4569, Aug. 2023.

[37] B. Palsson, M. O. Ulfarsson, and J. R. Sveinsson, "Convolutional autoencoder for spatial-spectral hyperspectral unmixing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 357–360.

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

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

[40] Y. Huang, J. Li, L. Qi, Y. Wang, and X. Gao, "Spatial-spectral autoencoder networks for hyperspectral unmixing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 2396–2399.

[41] L. Gao, Z. Han, D. Hong, B. Zhang, and J. Chanussot, "CyCu-Net: Cycle-consistency unmixing network by learning cascaded autoencoders," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5503914.

[42] X. Zhang, Y. Sun, J. Zhang, P. Wu, and L. Jiao, "Hyperspectral unmixing via deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1755–1759, Nov. 2018.

[43] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

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

[45] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[46] 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 Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5528815.

[47] Y. Wang, S. Shi, and J. Chen, "Efficient blind hyperspectral unmixing with non-local spatial information based on swin transformer," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 5898–5901.

[48] Y. Duan, X. Xu, T. Li, B. Pan, and Z. Shi, "UnDAT: Double-aware transformer for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5522012.

[49] J. Chen et al., "TCCU-Net: Transformer and CNN collaborative unmixing network for hyperspectral image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 8073–8089, 2024.

[50] Y. Gao, B. Pan, X. Xu, X. Song, and Z. Shi, "A reversible generative network for hyperspectral unmixing with spectral variability," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5519115.

[51] S. M. Bhakthan and A. Loganathan, "A hyperspectral unmixing model using convolutional vision transformer," *Earth Sci. Informat.*, vol. 17, no. 3, pp. 2255–2273, 2024.

[52] X. You, Y. Su, L. Gao, X. Sun, P. Li, and D. Liu, "Dset-Net: Deep self-embedded transformer network for hyperspectral unmixing," in *Proc. 4th Int. Conf. Geol., Mapping, Remote Sens.*, SPIE, 2024, pp. 110–115.

[53] W. Wang et al., "CrossFormer : A versatile vision transformer hinging on cross-scale attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3123–3136, May 2024.

[54] C.-F. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 357–366.

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

Huadong Yang received the M.S. and Ph.D. degrees in computer science from Dalian Maritime University, Dalian, China, in 2010 and 2015, respectively.

He is currently an Associate Professor with the School of Information Science and Engineering, Shenyang Ligong University, Shenyang, China. His research interests include hyperspectral image processing and machine/deep learning.

Chengbi Zhang received the bachelor's degree in optical information science and technology from the Harbin University of Science and Technology, Harbin, China, in 2014. He is currently working toward the master's degree in computer science and technology with the School of Information Science and Engineering, Shenyang Ligong University, Shenyang, China.

His research interests include image processing and hyperspectral unmixing.