TCCU-Net: Transformer and CNN Collaborative Unmixing Network for Hyperspectral Image

Jianfeng Chen, Chen Yang, Lan Zhang, Linzi Yang, Lifeng Bian, Zijiang Luo, and Jihong Wang

***Abstract*—In recent years, deep-learning-based hyperspectral unmixing techniques have garnered increasing attention and made significant advancements. However, relying solely on the use of convolutional neural network (CNN) or transformer approaches is insufficient for effectively capturing both global and fine-grained information, thereby compromising the accuracy of unmixing tasks. In order to fully harness the information contained within hyperspectral images, this article explores a dual-stream collaborative network, referred to as TCCU-Net. It end-to-end learns information in four dimensions: spectral, spatial, global, and local, to achieve more effective unmixing. The network comprises two core encoders: one is a transformer encoder, which includes squeeze-launch modules, DSSCR–vision transformer modules, and stripe pooling modules, while the other one is a CNN encoder, which is composed of two-dimensional (2-D) pyramid convolutions and 3-D pyramid convolutions. By fusing the outputs of these two encoders, the semantic gap between the encoder and decoder is bridged, resulting in improved feature mapping and unmixing outcomes. This article extensively evaluates TCCU-Net and seven hyperspectral unmixing methods on four datasets (Samson, Apex, Jasper Ridge, and Synthetic dataset). The experimental results firmly demonstrate that the proposed approach surpasses others in terms of accuracy, holding the potential to effectively address hyperspectral unmixing tasks.**

***Index Terms*—CNN, global and local information, hyperspectral image unmixing (HSU), spectral and spatial information, transformer.**

I. Introduction

HYPERSPECTRAL imaging (HSI) is a highly regarded remote sensing technology. HSIs [1] combine spectral information reflecting material radiation with spatial information of the terrain. Due to the rich spectral and spatial information within HSIs, they find extensive applications in fields, such as food safety [2], environmental monitoring [3], mineral exploration [4], and so on. However, due to the spatial resolution limitations of HSI instruments [5], pixels in HSIs often consist of mixed spectra [6], [7], representing a combination of various materials, known as mixed pixels. In practical applications, the abundance of mixed pixels can significantly impact the accuracy of pixel-based material classification and area measurement methods, making the development and application of HSI more challenging. To tackle this issue, there are typically two approaches to contemplate: The first involves enhancing the spatial resolution of the spectrometer, which inevitably leads to increased human and financial costs. The second approach, hyperspectral unmixing [8], is often chosen to reduce costs. The primary objective of spectral unmixing is to extract/estimate endmembers and their abundance fractions in each pixel solely based on the observed HSI [9].

Among the numerous methods for hyperspectral unmixing, the linear spectral mixture model (LSMM) [10] stands out for its simplicity, efficiency, and its ability to provide a good description of real spectral mixing processes. Expanding upon the LSMM framework, researchers have introduced several effective unmixing algorithms, including some of the most representative ones, such as geometric, statistical, or sparse methods. In the realm of geometric methods, vertex component analysis (VCA) [11] and fully constrained least squares unmixing (FCLSU) [12] are the most commonly employed techniques. In the field of sparse unmixing methods, on the one hand, collaborative methods, such as the least absolute shrinkage and selection operator [13] consider the spectral variability of endmembers by utilizing a dictionary generated from the data itself in a specific set of sparse unmixing methods. On the other hand, sparse unmixing techniques, like splitting and augmented Lagrangian-based sparse unmixing [14], represent another prominent example due to its exceptional performance, garnering significant attention. The statistical-based approaches also serve as effective alternatives for handling highly mixed HSIs, drawing substantial interest. For instance, the Bayesian framework formulates unmixing as an inference problem, leveraging statistical assumptions and priors to constrain unmixing results [15], [16], [17]. Due to the unique advantages in learning component-based representations, nonnegative matrix factorization (NMF) [18] and L1/2-NMF [19] are two of the most frequently employed algorithms within the statistical methods category for the simultaneous estimation of both endmembers and abundance. Yao et al. [20] leveraged the

Manuscript received 24 October 2023; revised 9 December 2023 and 21 December 2023; accepted 31 December 2023. Date of publication 10 January 2024; date of current version 17 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62065003, in part by the Guizhou Provincial Science and Technology Projects under Grant ZK [2022] Key-020 and Grant General-105, and in part by the Renjihe of Guizhou University (2012). *(Corresponding authors: Chen Yang; Jihong Wang.)*

Jianfeng Chen, Lan Zhang, and Linzi Yang are with the Power Systems Engineering Research Center, Ministry of Education, College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China (e-mail: ylz8097@163.com).

Chen Yang and Jihong Wang are with the Power Systems Engineering Research Center, Ministry of Education, College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China, and also with the China State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China (e-mail: eliot.c.yang@163.com; wjihong20@163.com).

Lifeng Bian is with the Frontier Institute of Chip and System, Fudan University, Shanghai 200433, China.

Zijiang Luo is with the Institute of Intelligent Manufacturing, Shunde Polytechnic, Guangdong Shunde 528300, China.

Digital Object Identifier 10.1109/JSTARS.2024.3352073

insightful properties of natural HSI, specifically the nonlocal smoothness, to propose novel blind hyperspectral unmixing models, NLTV/NLHTV, and logarithm and regularization-based nonnegative matrix factorization (NLTV–LSRNMF/NLHTV–LSRNMF), achieving widespread applications.

In recent years, with the rise of deep learning, various convolutional neural network (CNN)-based methods in the field of hyperspectral unmixing have experienced rapid development. Among these, the EGU-Net proposed by Hong et al. [21] for endmember-guided unmixing has introduced the concept of using endmembers to guide the unmixing network. It represents the first instance of utilizing such techniques in unmixing research, offering new insights for the future development of unmixing studies. Qi et al. [22] introduced the SSCU-Net, which employs spectral–spatial cooperative networks for unmixing tasks. This network is the first to incorporate spectral–spatial cooperation into the unmixing process via dual-branch tasks. Han et al. [23] presented the MU-Net, designed for hyperspectral unmixing with multimodal inputs. This network employs two image inputs to guide the unmixing network, addressing research gaps in unmixing tasks involving different inputs. Rasti et al. [24] proposed the minimum simplex convolutional network (MiSiC-Net), which combines spatial correlations between neighboring pixels and the geometric properties of linear simplex. The recurrent consistency unmixing network [25], introduced by the Dongfeng Hong team, utilizes two convolutional autoencoders that are cascaded and cyclically executed. The proposed loss function includes two terms for spectral reconstruction and one for abundance reconstruction, effectively incorporating high-level semantic information. Yao et al. [26] introduced a novel blind HU model called sparse-enhanced convolutional decomposition (SeCoDe-Net). This model jointly captures the spatial–spectral information of HSI in a tensor-based manner. In SeCoDe-Net, the use of convolutional operations models the spatial relationships between target pixels and their neighboring pixels, providing a robust explanation for the effectiveness of spectral bundles in addressing spectral variability. Simultaneously, it maintains physically continuous spectral components by decomposing invariance and spectral domains. Built upon sparse-enhanced regularization, the network also incorporates an alternative optimization strategy based on the alternating direction method of multipliers to achieve efficient model inference. It can be observed that due to the outstanding generalization capability and accuracy of CNNs, they have made significant strides in unmixing tasks. Other fields besides unmixing, such as, the GNet proposed by Chen et al. [27] has achieved significant success in classification tasks. In addition to representing spectral–spatial features in three classical paradigms (simultaneous, hierarchical, and individual), GNet can learn them in two new processes: multistage and multipath. This approach allows for a comprehensive and balanced exploration of spectral and spatial features. On the other hand, Chen et al. [28]'s work on HSI classification using spectral-induced superpixel segmentation based on local aggregation and global attention network introduces a novel superpixel generation strategy termed spectral-induced alignment superpixel segmentation. This strategy simultaneously leverages segmentation results from HSI with both raw and deeply abstracted spectral features. Chen et al. [29] proposed a temporal difference guided network for HSI change detection. Specifically, the network hierarchically extracts rich spectral features from dual temporal images, generating differences between the two images at various levels using convolutional gated recursive units designed in the spatial dimension. These networks, as the latest research outcomes in deep learning, significantly propel the development of various tasks related to HSI across different domains.

In recent years, development of the transformer [30], [31] has achieved great success in NLP, while the vision transformer (VIT) [32] extends this architecture to the field of computer vision. It has showcased its distinctive capability separate from convolutional operations and has achieved exceptional results in image classification tasks. Such as the recent CASST [33], which explores the use of cross-attention mechanisms for hyperspectral classification tasks. And for example, the ExViT model proposed by the Yao et al. [34] utilizes parallel branches of position-shared VIT extended with separable convolution modules to handle patches in multimodal remote sensing images. This presents an economical solution for leveraging both spatial and modality-specific channel information. The entire VIT model structure consists of several modules: the embedding layer, layer normalization, multihead attention, dropout, MLP block, and MLP head. The primary reason for the widespread adoption of VIT in the computer vision domain is its utilization of the multihead attention mechanism, which facilitates long-range feature association calculations and establishes global feature dependencies. The expression for the multihead attention is as follows:

$$\begin{aligned}&\boldsymbol{Multihead}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)\\&\quad=\boldsymbol{Concat}\left(\boldsymbol{Head}_1,\boldsymbol{Head}_2,\ldots,\boldsymbol{Head}_{\boldsymbol{h}}\right)\cdot\boldsymbol{W}_{\boldsymbol{O}}.\end{aligned}\tag{1}$$

In this equation, Q, K, and V, respectively, denote queries, keys, and values, while Head_i represents the computation result of the ith attention head. The variable "h" signifies the number of attention heads, and "Concatenate" denotes the process of combining their outputs. W_O signifies the final output weight matrix.

Recently, Ghosh et al. [35] introduced the Deep-Trans network, marking the first attempt to apply the transformer architecture to hyperspectral unmixing tasks, achieving remarkable research outcomes. This convincingly demonstrates the viability of transformers in image unmixing tasks. UnDAT [36], led by the team under the leadership of Zhenwei Shi, aims to achieve unmixing tasks by simultaneously harnessing spatial uniformity and spectral correlations within HSIs. The most recent transformer network employed for hyperspectral unmixing is the innovative deep neural U-Net-based model known as UST-Net [37], introduced by Zhiru Yang and colleagues. This network prioritizes discriminative information within the spatial scene and operates on the entire image, eliminating inconsistencies. SpectralGPT [38] is specifically designed for processing hyperspectral remote sensing images using a novel three-dimensional (3-D) generative pretrained transformer. The research presented in this article has also provided us with new avenues for exploration.

In recent years, CNNs have been widely adopted by experts and scholars in the field of hyperspectral unmixing, making significant contributions to the advancement of unmixing tasks. However, due to the complexity of HSIs, the application of CNNs presents considerable challenges. This is primarily because convolution operations are limited to capturing local features determined by kernel sizes, resulting in the loss of a substantial amount of contextual information present in the original HSI. On the one hand, transformers, renowned for their outstanding performance in NLP tasks, have found wide-ranging applications in various domains. Researchers have also harnessed transformers in decomposition tasks, as evidenced by recent developments like the Trans-Net network. Trans-Net effectively applies the transformer to hyperspectral data unmixing, yielding significant results. On the other hand, this article aims to achieve satisfactory performance in the unmixing task by referencing several transformer-based methods. For instance, the CUCaNet proposed by Yao et al. [39] introduces a transformer–DSSCR–VIT with a cross-attention mechanism, anticipating satisfactory performance in unmixing tasks. In addition, the spectral–spatial morphological attention transformer proposed by Roy et al. [40] has been a great source of inspiration. It implements a learnable spectral and spatial morphological network, utilizing spectral and spatial morphological convolution operations (combined with attention mechanisms) to enhance the interaction between structural and shape information of HSI tokens and CLS tokens. However, standalone transformers may focus solely on contextual information, potentially overlooking finer details, and relying solely on neural networks may struggle to capture global information comprehensively. In addressing this issue, we consulted various literature and noted that Danfeng Hong's team proposed a new method called "Global to Local: A Hierarchical Detection Algorithm for Hyperspectral Image Object Detection" [41]. This article introduces a global-to-local hierarchical target detection algorithm for HSI, providing us with a fresh perspective for our upcoming work.

Building on the insights provided by the abovementioned literature, we have designed a novel dual-stream unmixing network that combines transformer and CNN architectures to address the limitations associated with both. On the one hand, this approach employs separate transformer and CNN autoencoders to attend to spectral–spatial information and global–local information, respectively. On the other hand, by merging the outputs of the dual-stream network, it leads to improved quality abundance mapping and overall unmixing results, aiding the decoder in better reconstructing the HSI. The proposed method contributes to this purpose in the following ways.

1) In this article, we introduce a new transformer network, DSSCR_VIT, for the encoder part, which consists of three main components. First, it combines the dual-stream spectral–spatial cross fusion (DSSCF) module with three fusion stages for comprehensive learning of spectral spatial information. Second, channel squeeze–stretch modules are applied at both the input and output ends to facilitate the learning of detailed spectral information, deepening the network to achieve better unmixing effects for the transformer, alleviating interference from excessive irrelevant details. Finally, each layer of DSSCR–VIT incorporates a stripe pooling module, enabling the model to effectively capture spatial information. On the one hand, this logically structured setup provides a more stable configuration for the transformer encoder module for unmixing tasks; on the other hand, it addresses the inherent limitations of the transformer in focusing on image details.
2) Unlike traditional neural networks with limited attention to global information and small receptive fields, we propose a new pyramid convolution network that includes receptive fields of different scales, avoiding the problem of traditional convolution being limited by receptive fields and affecting network performance. Adopting a dual-stream structure, one branch utilizes 3-D pyramid convolution to fully explore the optical frequency spectrum neighborhood information, while the other branch uses 2-D pyramid convolution to learn spatial information. Therefore, compared to traditional CNNs, our network achieves comprehensive learning in both spectral and spatial domains, improving its unmixing performance.
3) Based on the contributions mentioned above, the dual-stream network designed in this article exhibits competitive unmixing performance on four datasets. Compared to three traditional unmixing networks and four recently proposed deep-learning-based unmixing networks, our approach achieves promising unmixing results.

II. Proposed Method

This article proposes a dual-branch codec network for hyperspectral unmixing based on a combination of CNN and transformer. The encoder section of the proposed network is divided into two branches: a transformer encoder branch and a CNN encoder branch, as shown in Fig. 1. The transformer encoder branch employs a channel extrusion structure, allowing subsequent transformers to focus on more relevant information while capitalizing on the transformer's ability to handle long-term tasks with deep networks. In addition, three interlaced cross-VITs are added to the transformer encoder, which are then pooled by two sets of stripes. The cross-VITs utilize the cross-attention mechanism to learn the input pictures of different patches separately. The stripe pooling deploys a long strip of pooled core shape along a spatial dimension, capturing long-distance relationships in isolated regions, thereby helping the cross-VIT learn more about image information. On the other hand, the CNN-based encoder adopts a dual-branch structure with two branches: the 2-D pyramid CNN branch and the 3-D pyramid CNN branch. The 2-D pyramid branch consists of four convolutional layers, each containing different levels of cores with varying sizes and depths to capture varying levels of detail. The 3-D pyramid CNN branch uses three-layer 3-D convolution to learn spectral neighborhood information. Finally, the output of CNN-encoder and transformer-encoder branches are fused and fed into a decoder comprising four convolutional layers for processing. The endmembers are then reconstructed using the output abundance values. In Sections II-B–II-D, we

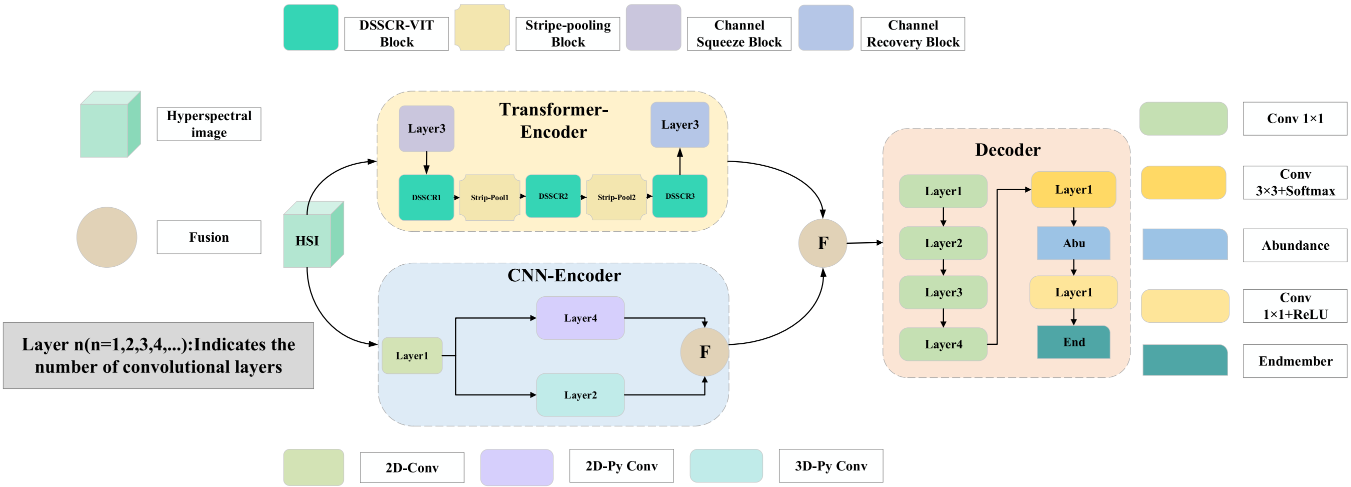

Fig. 1. Proposed network framework.

would provide a detailed discussion of the model components. For Section II-B, we will present the transformer encoder from three perspectives: the extrusion module, the DSSCR module, and the stripe pooling module. In addition, in Section II-C, we will discuss the CNN encoder from two aspects: the 3-D-CNN module and the pyramid convolution module. Finally, in Section II-D, we will primarily focus on the decoder part of this article.

A. *Related Issues*

1) Formulas and Related Symbolic Representations: The symbols involved in this article are represented by the following: Setting up HSIs are indicated by $I \in \mathbb{R}^{B\times H\times W}$, where the spatial dimension is represented by $H \times W$, and the spectral channels are denoted by B. An HSI can be reshaped to produce a matrix $Y = [y_1, y_2, y_3 \ldots y_n] \in \mathbb{R}^{B\times n}$, $n = H \cdot W$ is registered by the number of hyperspectral pixels, and y_i represents the ith observed spectrum. The endmember matrix will be displayed as $E = [e_1, e_2, e_3 \ldots e_R] \in \mathbb{R}^{B\times R}$, e_i is the ith endmember vector, and the number of endmembers present in HSI are represented by R. The corresponding abundance cube is represented by $M \in \mathbb{R}^{R\times H\times W}$. The abundance cube can be reshaped to produce a matrix of $A = [a_1, a_2, a_3 \ldots a_n] \in \mathbb{R}^{R\times n}$, and a_i represents the fractional abundance corresponding to the ith observed pixel.

The linear mixing model (LMM) model has been widely used for unmixing, and the reflectance observed in LMM is

$$\boldsymbol{Y} = \boldsymbol{EA} + \boldsymbol{N}. \tag{2}$$

The $N \in \mathbb{R}^{B\times n}$ is the additive noise present in Y. In addition, three physical constraints should usually be met in the task of unmixing: First, the endmember matrix should be non-negative $E \geq 0$. Second, it is necessary to satisfy the non-negative constraint of abundance ANC($A \geq 0$). Finally, abundance and one constraint ASC($1_R^T A = 1_n^T$), the formula 1_n is indicated N-dimensional column vector of 1.

2) Encoder: In the field of hyperspectral unmixing, autoencoder models have emerged as a representative deep-learning technique due to their strong representation and reconstruction capabilities, which allow them to extract information from given inputs. In this article, autoencoder is composed of two parts: a CNN-encoder and a transformer-encoder. The encoder takes input pixels $y_i \in \mathbb{R}^n$ and transforms them into a hidden low-dimensional representation $v_i \in \mathbb{R}^R$ using the following formula:

$$\boldsymbol{v_i} = \boldsymbol{F_D}\left(\boldsymbol{y_i}\right) = \boldsymbol{F}\left(\boldsymbol{W}^{(d)T}\boldsymbol{y_i} + \boldsymbol{b}^{(d)}\right). \tag{3}$$

Here, $F(.)$ is a nonlinear activation function, such as sigmoid or ReLU, while $W^{(d)}$ and $b^{(d)}$ represent the weight and bias in the dth encoder part, respectively.

3) Decoder: The main task of the decoder is to convert the extracted hidden features into their corresponding original input pixels using LMMs. The reconstructed pixels, denoted by $\hat{y}_i \in \mathbb{R}^n$, are computed as

$$\boldsymbol{\hat{y}_i} = \boldsymbol{f_E}\left(\boldsymbol{v_i}\right) = \boldsymbol{W}^{\boldsymbol{(e)T}}\boldsymbol{v_i}. \tag{4}$$

Here, $W^{(e)}$ is the weight matrix of the decoder part. The extracted endmembers matrix $\hat{e}_i$ and the estimated abundance vector $\hat{a}_i$ correspond to $W^{(e)}$and v_i, respectively.

4) Loss Function: The objective function of the autoencoder unmixing network is to measure reconstruction error (RE) in which y_i and $\hat{y}_i$ are used by different measurement forms. In this article, two loss functions are used that consist of mean square error (MSE) and spectral angular distance (SAD)

$$\boldsymbol{L_{RE}}\left(\boldsymbol{y_i}, \boldsymbol{\hat{y}_i}\right) = \frac{\boldsymbol{1}}{\boldsymbol{H \cdot W}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{H}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{W}} \left(\boldsymbol{\hat{y}_i} - \boldsymbol{y_i}\right)^{\boldsymbol{2}} \tag{5}$$

$$\boldsymbol{L_{SAD}}\left(\boldsymbol{y_i}, \boldsymbol{\hat{y}_i}\right) = \frac{\boldsymbol{1}}{\boldsymbol{R}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{R}} \cos^{\boldsymbol{-1}} \left(\frac{\boldsymbol{\hat{y}_i^T y_i}}{||\boldsymbol{\hat{y}_i}||_{\boldsymbol{2}} ||\boldsymbol{y_i}||_{\boldsymbol{2}}}\right). \tag{6}$$

The MSE objective function is used to calculate the RE loss, which enables the encoder to learn only the fundamental features of the input HSI while discarding irrelevant details.

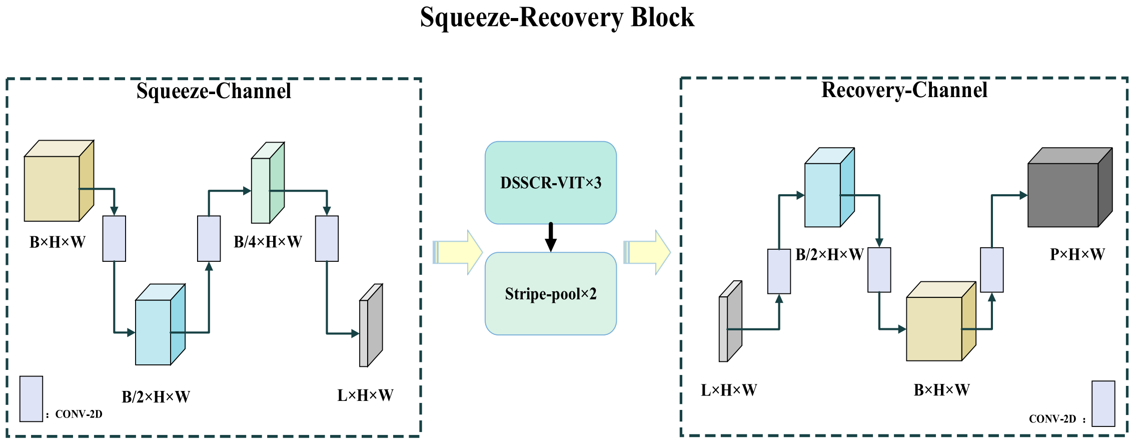

Fig. 2. Squeeze-recovery block.

However, MSE is sensitive to differences in absolute magnitude, which can lead to problems in distinguishing endmembers in HSI unmixing. To address this, the SAD loss, which is a scale-invariant objective function, is added to the model. The SAD loss helps to mitigate the shortcomings of the MSE objective function and accelerate model convergence. The total loss is obtained by adding the RE loss and the SAD loss

$$\boldsymbol{L} = \partial \boldsymbol{L_{RE}} + \varepsilon \boldsymbol{L_{SAD}}. \tag{7}$$

The regularization parameter is ∂ and ε.

B. Transformer Encoder

This article proposes a novel spectral squeezed transformer structure design for the encoding component of the unmixing task, which aims to capitalize on the exceptional capabilities of transformers in processing long-range dependencies to improve learning in unmixing tasks. Then, we will introduce its structure from three components.

1) Squeeze-Recovery Block: The structure of the block is demonstrated in Fig. 2, which has two main functions: the first is to compress the spectral channel so that the next layer transformer module (DSSCR–VIT) can better focus on the global information of the input HSI. The second is to restore the spectral channel, which elevates the original HSI from B channel to P channel $(I \in \mathbb{R}^{B\times H\times W} -> I \in \mathbb{R}^{P\times H\times W})$ for adjusting the output tensor dimension the same as the CNN encoder branch. To achieve the above functions, the proposed structure consists of two parts: a spectral channel compression and restoration module.

The spectral channel compression module consists of three layers of 2-D convolution, which it gradually compresses the features of the input image $I \in R^{B\times H\times W}$ to $I' \in \mathbb{R}^{H\times W\times L}$, where L can be represented as $L = \frac{(\dim * R)}{\text{patch}^2}$. In this expression, dim represents the dimension of the spectral band mapping, R represents the number of endmembers, and p is the patch size. As shown in Fig. 2, it can be seen that after the compression module, the spectral channel is compressed while the spatial scale remains unchanged. The above process can be represented by the following formula:

$$\boldsymbol{I_1 = Conv_1\left(W_1 I + V_1\right), \text{in which } I_1} \in \mathbb{R}^{\boldsymbol{B/2\times H\times W}} \tag{8}$$

$$\boldsymbol{I_2 = Conv_2\left(W_2 I_1 + V_2\right), \text{in which } I_2} \in \mathbb{R}^{\boldsymbol{B/4\times H\times W}} \tag{9}$$

$$\boldsymbol{I_3 = Conv_3\left(W_3 I_2 + V_3\right), \text{in which } I_3} \in \mathbb{R}^{\boldsymbol{L\times H\times W}} \tag{10}$$

$$\boldsymbol{I' = I_3^T, \text{ in which } I'} \in \mathbb{R}^{\boldsymbol{H\times W\times L}}. \tag{11}$$

Among them, $\text{Conv}_1(\cdot)$, $\text{Conv}_2(\cdot)$, and $\text{Conv}_3(\cdot)$ represent a three-layer convolutional layer, W_1, W_2, W_3 and V_1, V_2, V_3 represent the weight and bias of each layer, and the superscript T represents the transpose operation of the matrix.

The spectral channel recovery module is also composed of three layers of 2-D convolutional layers, which perform the opposite operation of the compression module except the output channel is P. The P is an hyperparameter, which used to adjust and optimize network performance.

2) DSSCR–VIT Block: In order to effectively analyze 3-D HSI data structures with spatial spectral joint information, a new dual channel transformer structure with cross-attention mechanism is designed that is called DSSCR–VIT. DSSCR–VIT consists of two modules: a spectral spatial dual branch module for extracting relevant dimensional features and preparing transformer token sequences, and a DSSCF module for spectral and spatial feature interaction. The structure is shown in Fig. 3, and the two modules in detail will be introduced below.

a) Spectral–spatial dual branch module: This module tokenizes the spatial and spectral feature sequences for input into the SSDCA module. The schematic diagram of its structure is shown in Fig. 4. Next, we will describe the specific details of the two branches.

For spectral sequence branches, first of all, a sample pixel $I_s \in \mathbb{R}^{1\times 1\times L}$ is extracted from the input HSI $(I' \in \mathbb{R}^{H\times W\times L})$ and expands into a sample cube $I_s' \in \mathbb{R}^{k\times k\times L}$ in its neighborhood, where k represents the spatial size of the sample cube. Then, shallow features of the input HSI are extracted through

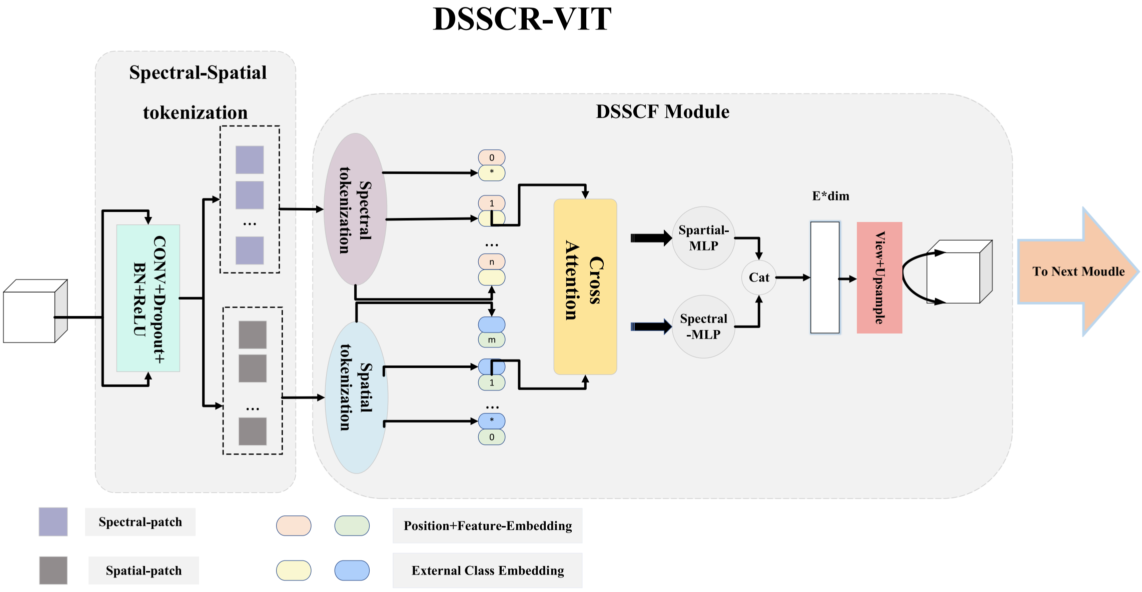

Fig. 3. DSSCR–VIT structure.

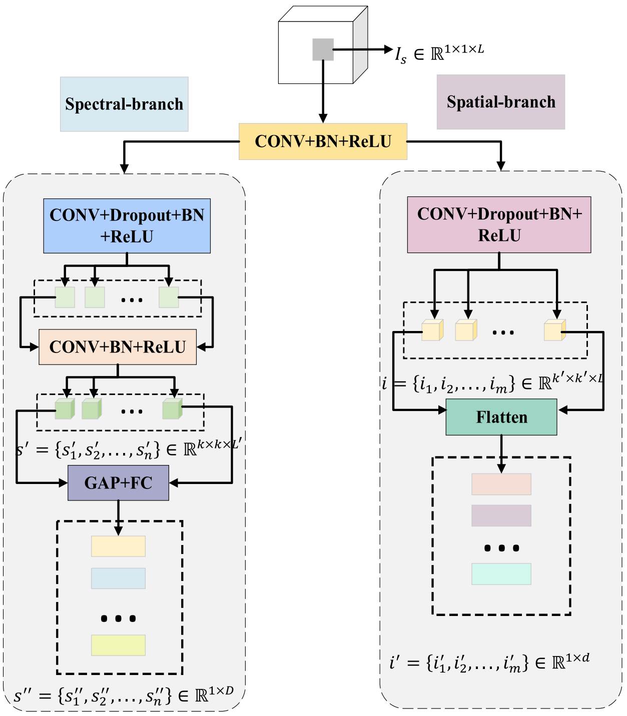
$I_s \in \mathbb{R}^{1\times1\times L}$
Spectral-branch
Spatial-branch
CONV+BN+ReLU
CONV+Dropout+BN +ReLU
CONV+BN+ReLU
$s' = \{s_1', s_2', \ldots, s_n'\} \in \mathbb{R}^{k\times k\times L'}$
GAP+FC
$s'' = \{s_1'', s_2'', \ldots, s_n''\} \in \mathbb{R}^{1\times D}$
CONV+Dropout+BN+ ReLU
$i = \{i_1, i_2, \ldots, i_m\} \in \mathbb{R}^{k'\times k'\times L}$
Flatten
$i' = \{i_1', i_2', \ldots, i_m'\} \in \mathbb{R}^{1\times d}$

Fig. 4. Process of spectral–spatial tokenization.

two convolutional layers with kernels of 1 × 1, the input sample cube is gradually transformed into a spectral vector group $s' = \{s_1', s_2', \ldots, s_n'\} \in \mathbb{R}^{k\times k\times L'}$, where L' represents the size of the spectral channel after convolutional transformation and n represents the number of spectral bands. Finally, spectral sequence characteristics $s'' = \{s_1'', s_2'', \ldots, s_n''\} \in \mathbb{R}^{1\times D}$ are further generated through a pooling layer and a fully connected layer, where D represents the dimension of each spectral sequence feature after tokenization.

For the spatial sequence branch, similarly, the cube $I_s' \in \mathbb{R}^{k\times k\times L}$ is first transformed into a space vector group $i = \{i_1, i_2, \ldots, i_m\} \in \mathbb{R}^{k'\times k'\times L}$ through a convolutional layer with a kernel of 3 × 3, where k' represents the size of the space after convolutional transformation and m represents the number of small space blocks. Then, each small block is flattened to generate spatial sequence features $i' = \{i_1', i_2', \ldots, i_m'\} \in \mathbb{R}^{1\times d}$, where d represents the dimensionality of each spatial sequence feature after tokenization.

b) Dual branch spectral–spatial cross fusion module: This article designs a DSSCF module to ensure the adequacy of spectral and spatial feature fusion, as shown in Fig. 5. By constructing a transformer-based cross-attention mechanism, this module performs a three-stage fusion of features from dual channel inputs that is interacting in the early, middle, and late stages.

The early fusion: For spectral space double branch, after embedding the tokenized spatial branch feature sequence x^{spa} and spectral branch feature sequence x^{spe}, the class tokens and feature embeddings of each branch are obtained. The class tokens between the two branches are exchanged to concatenate them with the feature embeddings of the other branch and sent to the multihead self-attention module that could achieve the first fusion between spatial and spectral features.

The intermediate fusion: Taking the spatial branch as an example, the output $x^{\text{spa}'}$ of multihead self-attention is added to the output $x^{\text{spe}'}$ of spectral branch multihead self-attention, as well as the spatial feature sequence x^{spa} to complete the second fusion that further interacts with spectral–spatial information.

The late fusion: The output $x^{\text{spe}''}$ and $x^{\text{spa}''}$ after the second fusion are, respectively, spliced through the MLP layer to complete the third fusion, and thus the spectral and spatial information have been fully learned.

Reshaping block: The output of the third fusion $I_E \in \mathbb{R}^{E\times \text{dim}}$ is reshaped as a 3-D $I_E' \in \mathbb{R}^{E\times H\times W}$, and the I_E' spectral channel

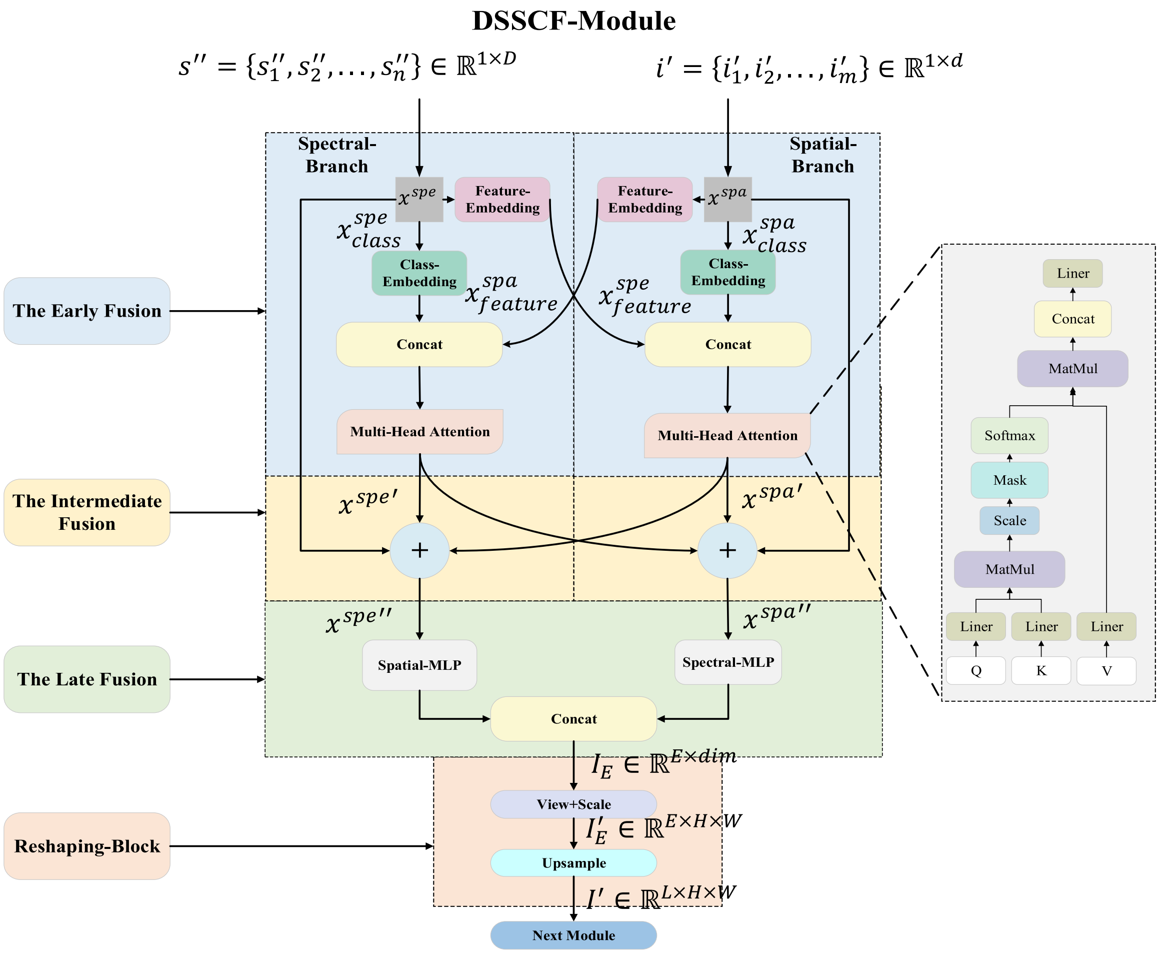

Fig. 5. Process of DSSCF module.

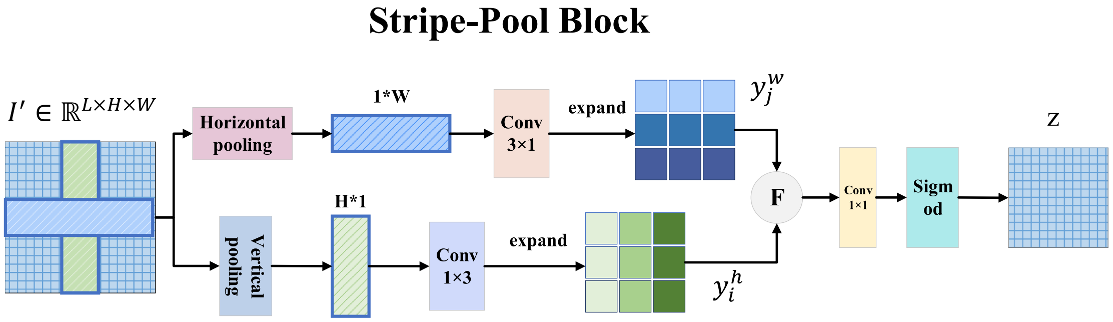

Fig. 6. Process of stripe pooling.

is restored to L (which can be represented as $I' \in \mathbb{R}^{L\times H\times W}$) to match the input of the next stage after the upsampling operation. The above process can be represented by the following formula:

$$x^{spa\prime} = Attention\left[Concat\left(x_{class}^{spa}, x_{feature}{}^{spe}\right)\right] \tag{12}$$

$$x^{spa\prime\prime} = Add\left(x^{spa\prime}, x^{spa}, x^{spe\prime}\right) \tag{13}$$

$$I_E = Concat\{[f_1\left(x^{spe\prime\prime}\right)], [f_2(x^{spa\prime\prime})]\} \tag{14}$$

$$I' = Upsample\left[F\left(I_E\right)\right]. \tag{15}$$

The $x_{\text{class}}^{\text{spa}}$ is a class token for spatial branches, $x_{\text{feature}}^{\text{spe}}$ represents the feature embedding of spectral branches, Attention$(\cdot)$ represents the multihead self-attention module, $f(\cdot)$ represents the MLP processing process, $F(\cdot)$ represents the reshaping operation, and Upsample$(\cdot)$ represents upsampling.

3) Stripe-Pool Block: The stripe-pool block is designed for the output of DSSCR–VIT to expect further fine-grained learning of image features, as shown in Fig. 6. The stripe-pooling module is divided into two parallel branches horizontal pooling and vertical pooling that perform pooling operations along corresponding direction.

Mathematically, the 3-D feature image denoted as $I' \in \mathbb{R}^{L\times H\times W}$undergoes a series of polling and convolution operations to derive the values y_i^h and y_j^w. Notably, the horizontal pooling output y_i^h can be formally defined as

$$\boldsymbol{y_i^h = \frac{1}{W}\sum_{0\le j\le W} I_{i,j}{}'}. \tag{16}$$

Vertical pooling output y_j^w can be represented as

$$\boldsymbol{y_j^w} = \frac{1}{H}\sum_{0\le j\le H} I'_{i,j}. \tag{17}$$

In the formula, the *i* and *j* represent the positions of feature image pixels, and *h* and *w* represent the spatial range of pooling. And then y_i^h and y_j^w are fused as follows:

$$\boldsymbol{Y = y_i^h + y_j^w}. \tag{18}$$

Finally, the output *z* is obtained by a convolution and sigmoid function that can be calculated as

$$\boldsymbol{z = \sigma f\left(Y\right)}. \tag{19}$$

Among them, σ represents the sigmoid function, and $f()$ represents the 1 × 1 convolution.

C. Multiscale Pyramid CNN Encoder

Within the context of the multiscale pyramid CNN encoder branch, this article employs a dual-branch encoding framework that integrates 2-D pyramid convolution and 3-D pyramid convolution. This configuration facilitates the understanding of the input feature image, as depicted in Fig. 7. Notably, the 2-D pyramid convolution branch is dedicated to acquiring spatial features with varying receptive fields, while the 3-D pyramid convolutional branch is strategically designed to delve into the realm of richer spectral contextual information.

1) 2-D-Pyramid Conv: The 2-D pyramid convolutions mentioned in this article serve the purpose of comprehensively capturing the spatial attributes of images. This is achieved through the processing of inputs at various kernel scales, all while avoiding the escalation of computational expenses or model intricacy.

In this branch, the number of convolutional integral groups are set to 1, 4, 4, and 16, and the stride is set to 1. The convolutional kernel size is shown in Fig. 7. The processing of convolution and the I_1'' can be represented as

$$\begin{cases} \boldsymbol{x_1 = Conv}1\left[\boldsymbol{f_1\left(I\right)}\right] \\ \boldsymbol{x_2 = Conv2\left[f_1\left(I\right)\right]} \end{cases} \begin{cases} \boldsymbol{x_3 = Conv3\left[f_1\left(I\right)\right]} \\ \boldsymbol{x_4 = Conv4\left[f_1\left(I\right)\right]} \end{cases} \tag{20}$$

$$\boldsymbol{I_1'' = x_1 + x_2 + x_3 + x_4}. \tag{21}$$

Among them, $f_1()$ represents a convolution with a kernel size of 1 × 1 used to modulate the size of input channel, while *Conv*1, 2, 3, 4 correspond to convolutions with kernel sizes of 3, 5, 7, and 9, respectively.

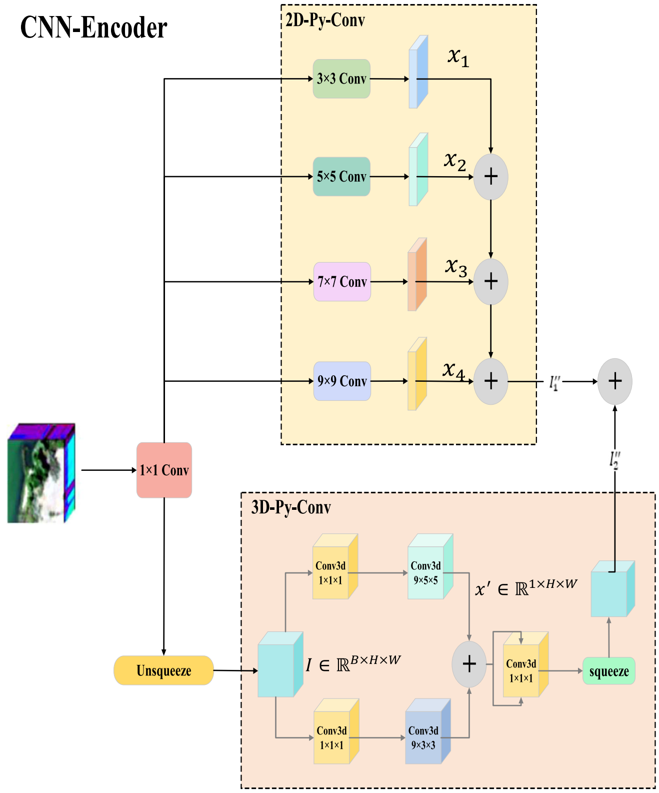

Fig. 7. CNN-encoder structure.

2) 3-D-Pyramid Conv: In this branch, first, a nonextrusion operation is performed to channel, which can be represented as

$$\boldsymbol{x' = unsequeeze\left(I, 1\right)}. \tag{22}$$

Second, $x' \in \mathbb{R}^{1\times H\times W}$ is fed into two sub-branches with different receptive fields, and then further fused to obtain x_3'. The process can be expressed as follows:

$$\boldsymbol{x_1' = Conv3d_2\left(Conv3d_1\right)} \tag{23}$$

$$\boldsymbol{x_2' = Conv3d_4\left(Conv3d_3\right)} \tag{24}$$

$$\boldsymbol{x_3' = x_1' + x_2'}. \tag{25}$$

Among them $Conv3d_1, Conv3d_2, Conv3d_3, Conv3d_4$ are 3-D convolutions of kernel sizes are 1 × 1 × 1, 9 × 5 × 5, 1 × 1 × 1, 9 × 5 × 5, respectively.

Finally, after sequentially passing through a 1 × 1 × 1 convolution and a sequence operation, the $I_2'' \in \mathbb{R}^{P\times H\times W}$ is restored to the same channel size of the 2-D pyramid convolution branch. It can be represented as

$$\boldsymbol{I_2'' = squeeze\left(Conv3d_5\left(x_3'\right), P\right)}. \tag{26}$$

Among them Conv3d_5 is a 3-D convolution with kernel size of 1 × 1.

D. Decoder

After fusing the output of the dual stream network, we will send the output of the automatic encoder to the decoder module for decoding. The specific operating steps are as follows:

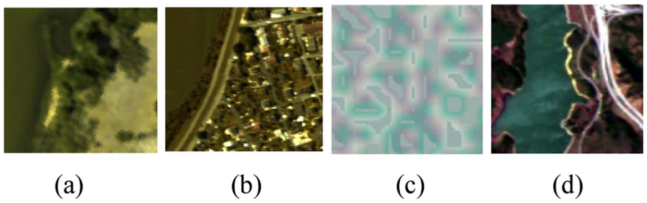

Fig. 8. (a) Samson true-color image. (b) Apex true-color image. (c) Synthetic RGB image. (d) Jasper ridge RGB image.

First, the decoder in this article compresses the channel to the corresponding number of end elements in the dataset through a four-layer convolutional layer with a kernel of 1 × 1. Second, after convolutional layer processing, the accuracy of endmember reconstruction and abundance estimation is improved by sharing weights. The estimated abundance is obtained through a convolution and softmax function with a kernel of 3 × 3. Finally, the estimated abundance obtained from the output is passed through a convolutional layer with a kernel of 1 × 1 to obtain the reconstructed endmember.

III. Experimental Results

This article compares the results of the proposed network with seven different unmixing techniques, specifically FCLSU [12], $L_{1/2}$-NMF [19], Coolab [13], EGU-Net [21], MiSiC-Net [24], Trans-Net [35], and UST-Net [37], applied individually to three real datasets (Samson, Apex, and Jasper Ridge) and one synthetic dataset.

A. Hyperspectral Dataset Description

Samson dataset: The Samson hyperspectral dataset was employed in this article, which included 156 different spectral channels in the wavelength range of [401–889] nm, as well as 95 × 95 pixels. The true color image of the Samson dataset, as shown in Fig. 8(a), contained three endmembers: soil, trees, and water.

Apex dataset: The cropped images of the Apex dataset used in this article were shown in Fig. 8(b). This image contained 110 × 110 pixels and covered 285 different bands spanning spectral channels of [413–2420] nm. There were four endmembers in this HSI, namely water, trees, roads, and roofs.

Synthetic dataset: We created the simulated data based on the method followed by [42], and its RGB images were shown in Fig. 8(c). Its size was 104 × 104 pixels, distributed over 200 spectral bands, with four endmembers. We generated these mixed pixels by multiplying four endmembers and four abundance maps based on LMM.

Jasper ridge dataset: The original data comprises 512 × 614 pixels, with each pixel recorded across 224 channels spanning from 380 to 2500 nm. Due to the complexity of the original image, which is less conducive to the specific research at hand, we considered and adopted 100 × 100 pixel subimages in this article, retaining a total of 198 channels, as illustrated in Fig. 8(d).

B. Description of Experimental Equipment and Parameters

The proposed model has been researched on a PC equipped with an i7 8750H core processor and NVIDA GTX 1050Ti GPU, using a Python 3.8.0 interpreter. This article explores several hyperparameters across different datasets, which are summarized in Table I. As shown in the table, the regularization parameters are δ and ε, which are used to adjust contribution of RE and SAD loss. Furthermore, detailed of training parameters of epoch, learning rate, and weight decay are also illustrated in the table.

C. Experimental Results and Comparison

This article evaluates the performance of the proposed model and compares it with six different approaches on above three datasets. These include three traditional unmixing methods: the FCLSU utilizing VCA for endmember extraction, the $L_{1/2}$-NMF employing abundance sparsity for unmixing, and the collab considering spatial–spectral information for joint unmixing. In addition, three state-of-the-art deep neural network based methods are considered: the EGU-Net, an endmember-guided unmixing network; the MiSiC-Net, a spectral-spatial collaborative network; and the Trans-Net, the first application of a transformer network for unmixing.

Samson dataset: The quantitative results on the Samson dataset can be found in Tables II and III. The results indicate that the proposed model outperforms other techniques in terms of abundance and endmember estimation in most cases. Among them, the average RMSE of our network is 0.05501, which is 30.90% higher than the suboptimal method; and its average SAD is 0.05762, which is 22.58% higher than the suboptimal method. The experimental results reveal the competitiveness of the proposed network on Samson dataset, and it proves the feasibility and superiority of the cross-fusion network between CNN and transformer in the task of unmixing.

Apex dataset: The quantitative results of the Apex dataset are shown in Tables IV and V. From the table, it can be seen that the endmembers "Road" and "Water" in the Apex dataset pose a considerable challenge compared to most other methods. However, the proposed method greatly improves the estimation of these two endmembers, mainly due to the proposed network's collaborative learning strategy for spectral spatial information, which takes into account the geometric information of the endmembers and fully utilizes effective spectral bands. The average RMSE value of this method is 34.26% higher than that of the suboptimal method, and the average SAD is 71.10% higher than that of the suboptimal method. Furthermore, based on the SAD value, the proposed method provides the best endmember estimation for all endmembers.

Synthetic dataset: To evaluate the robustness of the proposed method to noise, we added Gaussian white noise with different noise powers to the simulated dataset, and obtained data with signal-to-noise ratios of 20, 30, 40, and 50 dB, respectively. The endmember and abundance estimation results under different noise conditions are shown in Tables VI and VII, including the results of all alternative unmixing methods. Overall, on the one

TABLE I
INFORMATION ON HYPERPARAMETER SETTINGS

Hyperparameters	Samson	Apex	Synthetic
δ	$5e^{3}$	$1e^{6}$	$1e^{5}$
ε	$3e^{-3}$	$9e^{-1}$	$1e^{-4}$
Epoch	100	100	50
Learning rate	$1e^{-3}$	$1e^{-3}$	$4e^{-5}$
Weight decay	$2e^{-5}$	$4e^{-5}$	$5e^{-5}$

TABLE II
MEAN RMSE RESULTS OF THE SAMSON DATASET

	FCLSU	$L_{1/2}$-NMF	Collab	EGU-Net	MiSiC-Net	Trans-Net	UST-Net	proposed
Soil	0.16712	0.15694	0.17232	0.19219	0.14861	0.07561	0.08095	**0.06115**
Tree	0.12563	0.16788	0.07326	0.13112	0.11232	**0.06135**	0.13245	0.06277
Water	0.09634	0.16988	0.10261	0.23225	0.07167	0.09313	0.07641	**0.03740**
Mean	0.12969	0.16490	0.11209	0.18518	0.11086	0.07962	0.10099	**0.05501**

The best one is shown in bold.

TABLE III
MEAN SAD RESULTS OF THE SAMSON DATASET

	VCA	$L_{1/2}$-NMF	Collab	EGU-Net	MiSiC-Net	Trans-Net	UST-Net	proposed
Soil	0.13615	0.16763	0.09556	0.13115	0.09191	0.06891	0.03443	**0.02059**
Tree	0.16552	0.16611	0.12553	0.12189	0.10295	0.08113	0.08919	**0.06584**
Water	0.09164	1.27861	0.26954	0.27234	0.09632	0.07326	**0.06728**	0.07644
Mean	0.13110	0.53745	0.16354	0.17513	0.09706	0.07443	0.07963	**0.05762**

The best one is shown in bold.

TABLE IV
MEAN RMSE RESULTS OF THE APEX DATASET

	FCLSU	$L_{1/2}$-NMF	Collab	EGU-Net	MiSiC-Net	Trans-Net	UST-Net	proposed
Road	0.16164	0.68812	0.45631	0.25112	0.13166	0.12718	**0.10771**	0.12591
Roof	0.27631	0.28631	0.21664	0.15019	0.13568	0.12957	**0.06418**	0.07249
Tree	0.15996	0.17112	0.10450	0.12223	0.12125	**0.07132**	0.09134	0.08345
Water	0.52369	1.82363	0.52011	0.22614	0.12317	0.05399	0.06830	**0.04841**
Mean	0.28040	0.74229	0.32439	0.18742	0.12794	0.093115	0.08539	**0.08256**

The best one is shown in bold.

TABLE V
MEAN SAD RESULTS OF THE APEX DATASET

	VCA	$L_{1/2}$-NMF	Collab	EGU-Net	MiSiC-Net	Trans-Net	UST-Net	proposed
Road	0.24193	0.19234	0.31652	0.20913	0.17631	0.17891	0.15601	**0.02418**
Roof	0.13625	0.24991	0.20916	0.15136	0.12615	0.10256	0.11544	**0.01848**
Tree	0.13022	0.23909	0.16538	0.24316	0.26139	0.12694	0.13540	**0.01540**
Water	0.14067	0.38965	0.17526	0.29936	0.42123	0.09135	0.23636	**0.04411**
Mean	0.16226	0.26774	0.19158	0.22575	0.27127	0.12494	0.13581	**0.03562**

The best one is shown in bold.

TABLE VI
MEAN RMSE RESULTS OF THE SYNTHETIC DATASET

	FCLSU	$L_{1/2}$-NMF	Collab	EGU-Net	MiSiC-Net	Trans-Net	UST-Net	proposed
20dB	0.24631	0.21686	0.22329	0.15791	0.10125	0.15314	0.20566	**0.03258**
30dB	0.19657	0.22691	0.18691	0.14963	0.09631	0.12311	0.13517	**0.02861**
40dB	0.15639	0.18614	0.17631	0.14031	0.08912	0.10167	0.10542	**0.02239**
50dB	0.14998	0.15691	0.17665	0.13667	0.08011	0.09963	0.09208	**0.01542**
Mean	0.18731	0.19670	0.1760	0.14613	0.09169	0.11938	0.14158	**0.02475**

The best one is shown in bold.

TABLE VII
MEAN SAD RESULTS OF THE SYNTHETIC DATASET

	VCA	$L_{1/2}$-NMF	Collab	EGU-Net	MiSiC-Net	Trans-Net	UST-Net	proposed
20dB	0.28646	0.31191	0.29167	0.14125	0.05361	0.07326	0.03837	**0.02791**
30dB	0.24926	0.20165	0.21855	0.13517	0.04310	0.08112	0.02388	**0.00455**
40dB	0.23169	0.18715	0.14161	0.09364	0.02102	0.06132	0.02057	**0.00306**
50dB	0.16139	0.15189	0.08846	0.08710	0.01131	0.03962	0.00495	**0.00274**
Mean	0.22322	0.21315	0.18072	0.11429	0.03226	0.06383	0.02195	**0.00956**

The best one is shown in bold.

TABLE VIII
MEAN RMSE RESULTS OF THE JASPER DATASET

	FCLSU	$L_{1/2}$-NMF	Collab	EGU-Net	MiSiC-Net	Trans-Net	UST-Net	proposed
Tree	0.76169	0.68812	0.16234	0.17939	0.13655	0.10718	0.11771	**0.11037**
Water	0.88731	0.91324	0.11369	0.26974	0.11794	0.10957	0.12037	**0.08761**
Soil	0.23791	0.30136	0.19367	0.31796	0.13125	0.12132	0.11961	**0.10122**
Road	0.16645	028252	0.17922	0.23039	0.14074	0.09617	0.11279	**0.08324**
Mean	0.51334	0.54631	0.16223	0.24937	0.12963	0.10856	0.11762	**0.09561**

The best one is shown in bold.

TABLE IX
MEAN SAD RESULTS OF THE JASPER DATASET

	VCA	$L_{1/2}$-NMF	Collab	EGU-Net	MiSiC-Net	Trans-Net	UST-Net	proposed
Tree	0.94316	0.86923	0.23615	0.28933	0.15671	0.12916	0.15636	**0.08105**
Water	0.67133	0.67922	0.12975	0.20397	0.11652	**0.09364**	0.11544	0.10385
Soil	0.76712	0.61334	0.26377	0.48637	0.12937	0.10763	0.08671	**0.07991**
Road	0.71435	072369	0.14651	0.49111	0.11648	0.09133	0.09063	**0.08867**
Mean	0.77399	0.72137	0.19404	0.36769	0.12976	0.10544	0.11228	**0.08837**

The best one is shown in bold.

hand, as the signal-to-noise ratio increases, the abundance of each network and the estimation results of endmembers have improved. On the other hand, under different signal-to-noise ratios, most deep-learning-based unmixing networks have better effects than traditional methods that also reflects the superiority of deep learning in the unmixing tasks. In addition, among several deep-learning-based methods, there are notable distinctions. One point to consider, the first transformer-based unmixing task, Trans-Net, outperforms the classic CNN network, EGU-Net. However, it still falls slightly short of MiSiC-Net, which leverages spectral–spatial information for unmixing. This underscores the vast potential for further exploration of transformers in unmixing tasks. On the flip side, in comparison to alternative networks, the network proposed in this article yields superior unmixing results, exhibiting significant advantages in terms of RMSE and SAD across all signal-to-noise ratios. This is primarily attributed to the proposed method's amalgamation of CNN and transformer strengths, coupled with its comprehensive and meticulous learning of both spatial and spectral information.

Jasper ridge dataset: Quantitative results in the Jasper dataset are presented in Tables VIII and IX. As evident from the tables, our approach outperforms most other methods in estimating the abundances of the four endmembers in the Jasper dataset, delivering competitive results. The method exhibits an average RMSE value that is 11.93% higher than the second-best method,

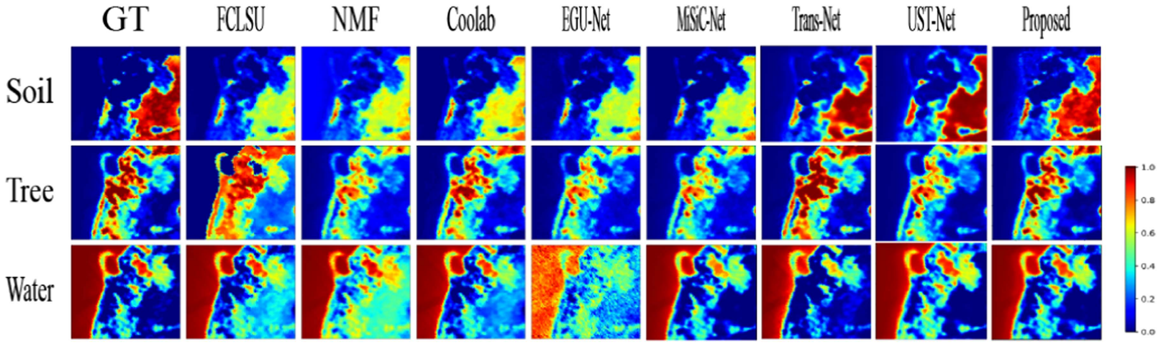

Fig. 9. Abundance map of the Samson dataset.

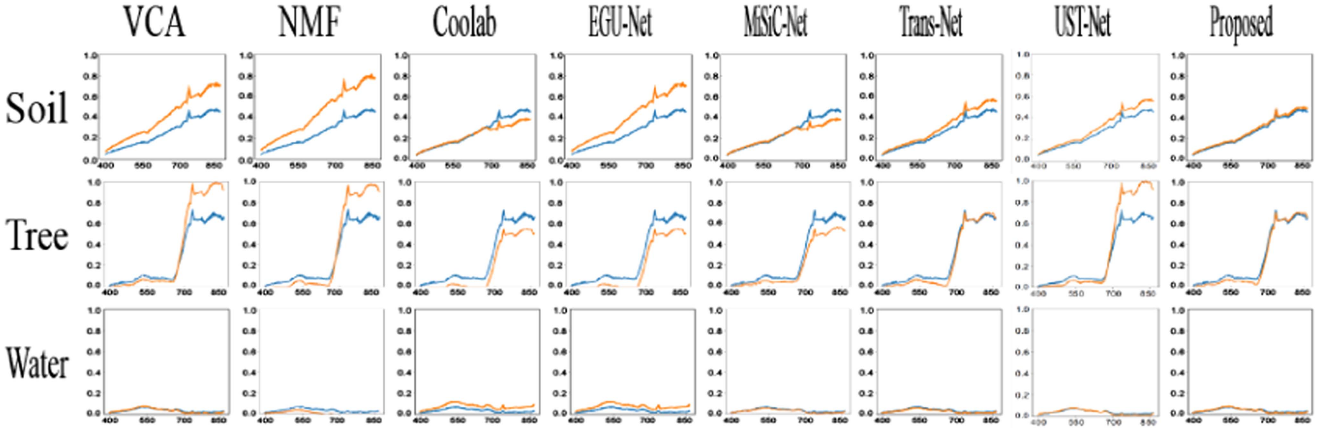

Fig. 10. Endmember graph of the Samson dataset.

and an average SAD value that is 16.19% higher than the second-best method. Moreover, the proposed method provides the best estimates for the abundances of all endmembers. Across the four datasets mentioned above, our method consistently achieves the best average SAD and RMSE values, reaffirming the superiority of the approach introduced in this article.

D. Visual Analysis

After experimental research on different datasets and comparative networks, we have decided to visually analyze the abundance and endmember maps generated from each dataset. From the abundance maps and endpoint maps, it can be seen that the abundance maps and endpoint maps obtained by the proposed method in this article are visually closest to ground truths (GTs) compared to other comparison networks. Next, the results on each dataset will be discussed one by one.

For the Samson dataset as shown in Figs. 9 and 10, the traditional unmixing methods, such as FCLSU, NMF, and Coolab, have shown inadequate results on the abundance and endmember plots. This is because their performance is affected by the differences between end elements, which limits their accuracy in unmixing tasks and leads to overall performance losses for these models. The overall effect of the EGU-Net method is relatively good, but its performance on the end element "Water" is less satisfactory. It is probably due to the network that mainly targets pure endmembers for unmixing, and the performance of this type of task in mixed pixel unmixing is not outstanding. The MiSiC-Net, as one of the representative neural networks used in hyperspectral unmixing tasks in recent years, has also shown competitive unmixing results on Samson. Nevertheless, owing to the inherent constraints of CNN, it fails to effectively capture global information, thereby leaving room for enhancement in the final reconstructed abundance and endmember maps. As the first network based on transformer is to handle unmixing tasks, Trans-Net has improved the accuracy of unmixing tasks. However, due to the fact that transformers mainly focus on global information, their unmixing performance is usually not as good as the proposed method.

For the Apex dataset as shown in Figs. 11 and 12, we observed that the proposed method on the "Road" endmember is closest to the original abundance map, and relatively speaking, other methods have lower accuracy in estimating endmembers and abundance compared to the proposed method on the roof and tree endmembers, and the Trans-Net and the proposed method are significantly superior to the previous methods; and on the water endmember, the proposed method is also more competitive. The significant enhancement in overall network performance primarily stems from the collaborative training of CNN and transformer architectures in this article. This collaborative approach

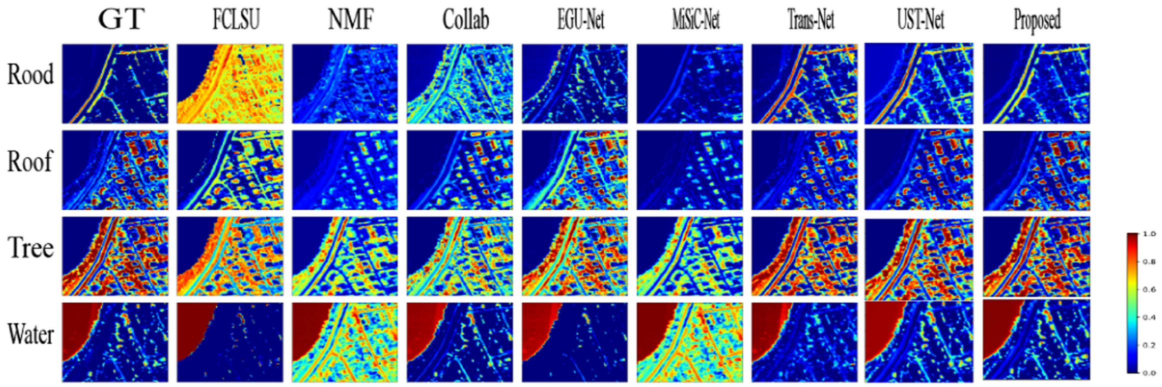

Fig. 11. Abundance map of the Apex dataset.

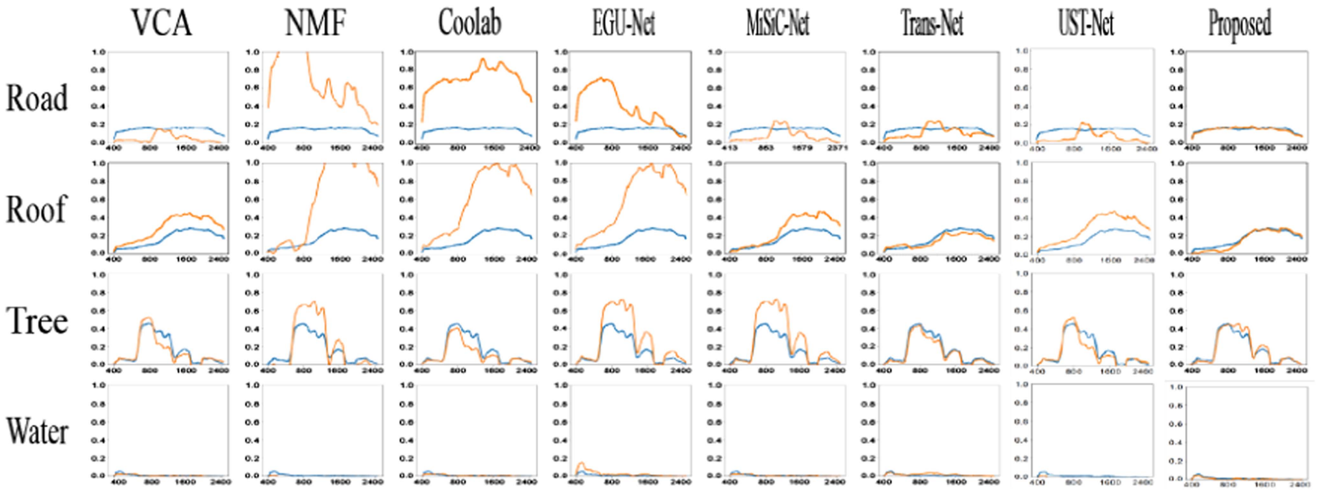

Fig. 12. Endmember graph of the Apex dataset.

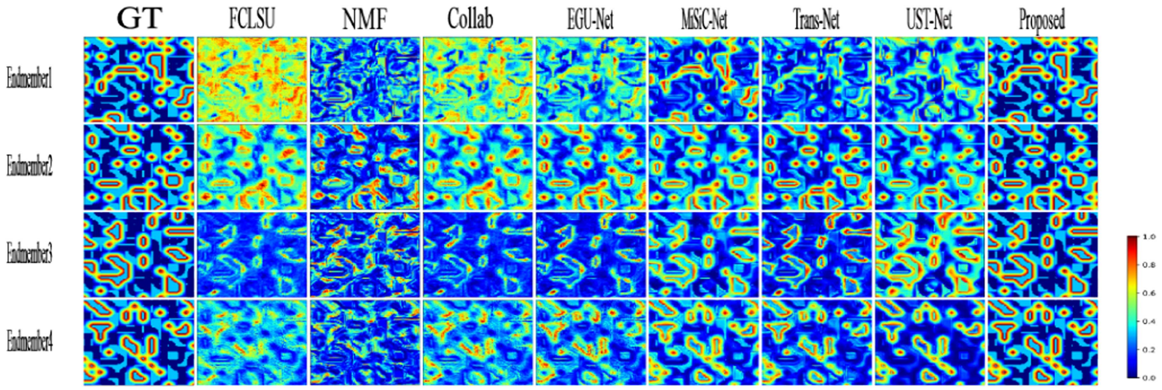

Fig. 13. Abundance map of the synthetic dataset.

empowers the network to acquire both global and intricate details, leading to a substantial improvement in network performance.

For the synthetic dataset as shown in Figs. 13 and 14, it can be seen that the deep-learning methods proposed in recent years have achieved better results compared to the classical unmixing methods. The MiSiC-Net, Trans-Net, as well as the abundance maps and endmember maps generated by the proposed methods in this article exhibit a visual resemblance that is closer to the GTs. This is mainly due to the stronger learning ability of deep-learning methods, as well as their wider coverage and better adaptability.

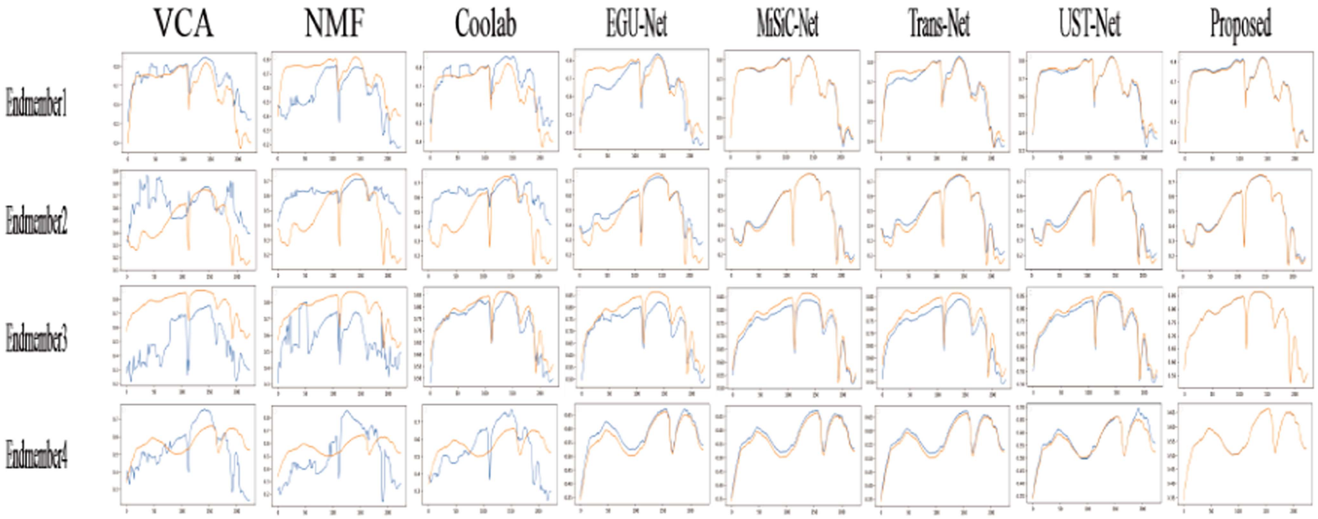

Fig. 14. Endmember graph of the synthetic dataset.

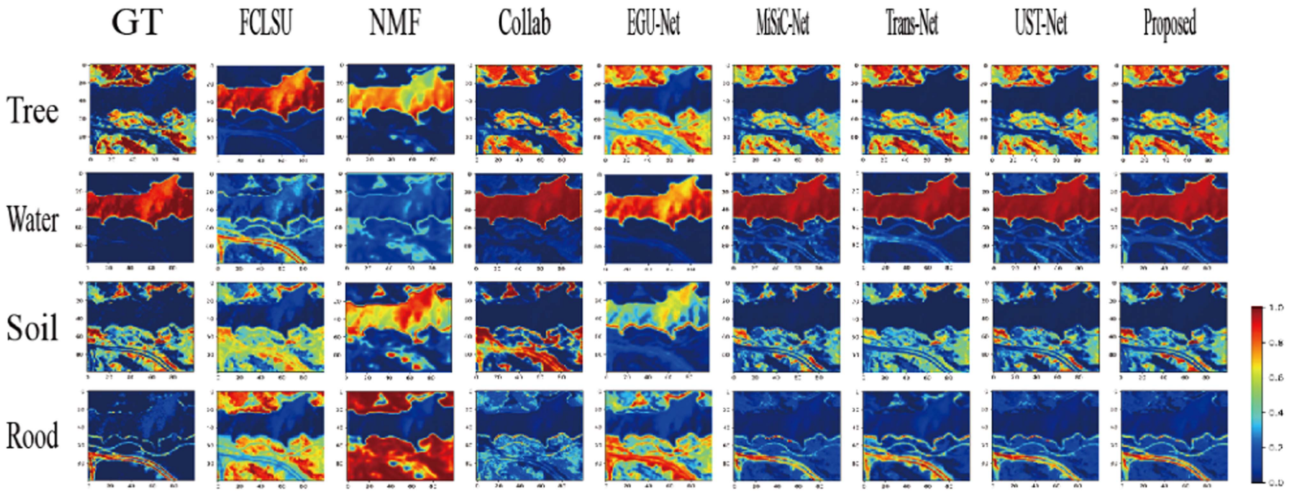

Fig. 15. Abundance map of the Jasper dataset.

For the Jasper dataset, as depicted in Figs. 15 and 16, we observe that traditional methods like VCA and NMF fail to correctly estimate “Tree” and “Water.” This is due to the limitations of traditional methods in capturing fine details in images. In contrast, deep-learning-based approaches perform significantly better, underscoring the superiority of deep learning over traditional methods. Our proposed method also excels in estimating the “Soil” endmember, further demonstrating that our neural network is better at learning image details and contextual information compared to other deep-learning methods. This highlights the effectiveness of our proposed approach.

E. Ablation Study

This article employs a dual-branch encoder–decoder structure to tackle unmixing tasks across three different datasets. To validate the roles of various encoder components within the proposed network, we conducted ablation experiments on the dual-branch encoder.

From Table X, it is evident that using only a single-branch encoder in the proposed network yields inferior unmixing results compared to the fusion achieved by the dual-branch approach. This highlights a substantial enhancement in source separation when utilizing the dual-branch fusion. Notably, across all three datasets, the CNN branch demonstrates higher effective separation compared to the transformer branch. We attribute this to the nature of the transformer as a network that excels in tasks with long-range dependencies, typically performing optimally in deeper networks.

Table XI presents experimental results on three datasets using a single module from the CNN encoder. The results indicate that the individual use of either the 2-D pyramid module or the 3-D pyramid module is less effective than their combination.

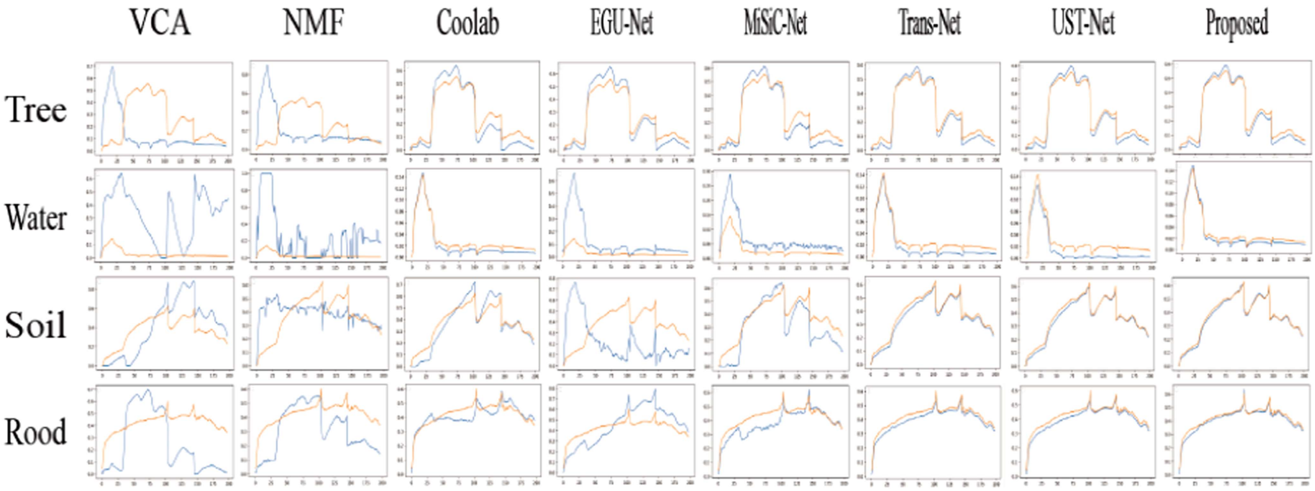

Fig. 16. Endmember graph of the Jasper dataset.

TABLE X
PERFORMANCE OF DIFFERENT ENCODER MODULES IN THE PROPOSED NETWORK ON THREE DATASETS

	Encoder-CNN		Encoder-transformer		Proposed	
	SAD	RMSE	SAD	RMSE	SAD	RMSE
Samson	0.09632	0.10339	0.11325	0.12931	**0.05762**	**0.05501**
Apex	0.10329	0.11361	0.09997	0.08622	**0.06935**	**0.07302**
Synthetic	0.05336	0.06562	0.06913	0.07931	**0.00865**	**0.02434**

The bold values represent the best effect.

TABLE XI
PERFORMANCE OF THE PROPOSED NETWORK USING ONLY A SINGLE MODULE OF CNN ENCODER ON THREE DATASETS

	2-D-pyramid-CNN+transformer-encoder		3-D-pyramid-CNN+transformer-encoder	
	SAD	RMSE	SAD	RMSE
Samson	0.06133	0.05991	0.08726	0.07411
Apex	0.09364	0.08972	0.09765	0.09963
Synthetic	0.02366	0.04115	0.02061	0.03697

TABLE XII
PERFORMANCE OF THE PROPOSED NETWORK USING ONLY TRANSFORMER ENCODER SINGLE OR TWO COMBINED MODULES ON THREE DATASETS

	DSSCR–VIT+CNN-encoder		DSSCR–VIT+squeeze +CNN-encoder		DSSCR–VIT+stripe-pool +CNN-encoder	
	SAD	RMSE	SAD	RMSE	SAD	RMSE
Samson	0.07131	0.06933	0.065141	0.06639	0.06983	0.07016
Apex	0.07539	0.08562	0.07128	0.07633	0.07496	0.08235
Synthetic	0.03697	0.04014	0.01326	0.03658	0.01063	0.03366

This substantiates the rationality of the designed CNN encoder modules. In addition, it is observed that the 2-D pyramid convolution outperforms the 3-D pyramid convolution slightly on all three datasets, likely due to the 3-D pyramid convolution compressing spectral channels, leading to the loss of some spectral information.

Table XII illustrates the combinations of various components within the transformer encoder across the three datasets that the performance remains subpar when compared to the collaborative effect of all three modules. Furthermore, the performance of the DSSCR–VIT module alone is the poorest among the three combinations, indicating that the inclusion of the Squeeze-launch module and stripe-pooling module enhances the unmixing performance of the transformer encoder.

IV. Conclusion

In this article, we introduced a hyperspectral unmixing network based on VIT and pyramid CNN, which exhibits competitive performance in unmixing accuracy compared to the benchmark networks. The design of the squeeze–stretch module in the transformer encoder of the proposed TCCU-Net helps deepen the network while mitigating the interference of overly detailed information. The inclusion of the DSSCF module enables the transformer encoder to simultaneously focus on spectral and spatial information. In addition, the use of the stripe-pooling module ensures that the transformer encoder pays attention to local fine-grained details, further enhancing the network's performance. In the CNN encoder, the 2-D pyramid convolution focuses on spatial information by employing convolution kernels of varying scales, while the 3-D pyramid convolution module prioritizes spectral neighborhood information by configuring the spectral channels as one, thus allowing the CNN encoder to attend to both spatial and spectral information simultaneously. To validate the effectiveness of the proposed TCCU-Net, extensive ablation studies were conducted on the Samson, Apex, Jasper Ridge, and synthetic datasets. The experimental results demonstrate that the TCCU-Net is an effective unmixing method, consistently delivering robust unmixing performance across all four datasets. We look forward to the application of the TCCU-Net in relevant domains, contributing to the advancement of hyperspectral unmixing.

Acknowledgment

The authors give thanks for the computing support of the State Key Laboratory of Public Big Data, Guizhou University.

References

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

[2] M. O. Smith, P. E. Johnson, and J. B. Adams, "Quantitative determination of mineral types and abundances from reflectance spectra using principal component analysis," in *Proc. Lunar Planet. Sci. Conf.*, 1985, pp. 797–904.

[3] Z. Lv et al., "Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective," *Proc. IEEE*, vol. 110, no. 12, pp. 1976–1991, Dec. 2022.

[4] Z. Lv, P. Zhong, W. Wang, Z. You, and N. Falco, "Multiscale attention network guided with change gradient image for land cover change detection using remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 2501805.

[5] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 44–57, Jan. 2002.

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

[7] P. Ghamisi et al., "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, Dec. 2017.

[8] R. O. Green et al., "Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (A VIRIS)," *Remote Sens. Environ.*, vol. 65, no. 3, pp. 227–248, 1998.

[9] G. Shaw and D. Manolakis, "Signal processing for hyperspectral image exploitation," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 12–16, Jan. 2002.

[10] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 44–57, Jan. 2002.

[11] D. Manolakis and G. Shaw, "Detection algorithms for hyperspectral imaging applications," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 29–43, Jan. 2002.

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

[13] L. Drumetz, T. R. Meyer, J. Chanussot, A. L. Bertozzi, and C. Jutten, "Hyperspectral image unmixing with endmember bundles and group sparsity inducing mixed norms," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3435–3450, Jul. 2019, doi: 10.1109/TIP.2019.2897254.

[14] M.-D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2014–2039, Jun. 2011.

[15] 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.

[16] O. Eches, N. Dobigeon, C. Mailhes, and J.-Y. Tourneret, "Bayesian estimation of linear mixtures using the normal compositional model: Application to hyperspectral imagery," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1403–1413, Jun. 2010.

[17] J. M. Nascimento and J. M. Bioucas-Dias, "Hyperspectral unmixing based on mixtures of Dirichlet components," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 863–878, Mar. 2012.

[18] J. Li, J. M. Bioucas-Dias, A. Plaza, and L. Liu, "Robust collaborative nonnegative matrix factorization for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6076–6090, Oct. 2016.

[19] Y. Qian, S. Jia, J. Zhou, and A. Robles-Kelly, "Hyperspectral unmixing via 1 {1/2} sparsity-constrained nonnegative matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4282–4297, Nov. 2011.

[20] J. Yao, D. Meng, Q. Zhao, W. Cao, and Z. Xu, "Nonconvex sparsity and nonlocal-smoothness based blind hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2991–3006, Jan. 2019, doi: 10.1109/TIP.2019.2893068.

[21] D. Hong et al., "Endmember-guided unmixing network (EGU-Net): A general deep learning framework for self-supervised hyperspectral unmixing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6518–6531, Nov. 2022, doi: 10.1109/TNNLS.2021.3082289.

[22] L. Qi, F. Gao, J. Dong, X. Gao, and Q. Du, "SSCU-Net: Spatial–spectral collaborative unmixing network for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5407515, doi: 10.1109/TGRS.2022.3150970.

[23] Z. Han, D. Hong, L. Gao, J. Yao, B. Zhang, and J. Chanussot, "Multimodal hyperspectral unmixing: Insights from attention networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5524913, doi: 10.1109/TGRS.2022.3155794.

[24] 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, doi: 10.1109/TGRS.2022.3146904.

[25] 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.

[26] J. Yao, D. Hong, L. Xu, D. Meng, J. Chanussot, and Z. Xu, "Sparsity-enhanced convolutional decomposition: A novel tensor-based paradigm for blind hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5505014, doi: 10.1109/TGRS.2021.3069845.

[27] Z. Chen, D. Hong, and H. Gao, "Grid network: Feature extraction in anisotropic perspective for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5507105, doi: 10.1109/LGRS.2023.3297612.

[28] Z. Chen, G. Wu, H. Gao, Y. Ding, D. Hong, and B. Zhang, "Local aggregation and global attention network for hyperspectral image classification with spectral-induced aligned superpixel segmentation," *Expert Syst. Appl.*, vol. 232, 2023, Art. no. 120828.

[29] Z. Chen et al., "Temporal difference guided network for hyperspectral image change detection," *Int. J. Remote Sens.*, vol. 44, no. 19, 2023, Art. no. 60336059, doi: 10.1080/01431161.2023.2258563.

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

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[32] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” 2020, *arXiv:2010.11929*.

[33] Y. Peng, Y. Zhang, B. Tu, Q. Li, and W. Li, “Spatial–spectral transformer with cross-attention for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5537415, doi: 10.1109/TGRS.2022.3203476.

[34] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, “Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415, doi: 10.1109/TGRS.2023.3284671.

[35] 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, doi: 10.1109/TGRS.2022.3196057.

[36] 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, doi: 10.1109/TGRS.2023.3310155.

[37] 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, doi: 10.1109/TGRS.2023.3321839.

[38] D. Hong et al., “SpectralGPT: Spectral remote sensing foundation model,” *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2024.3362475.

[39] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, “Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 208–224, doi: 10.1007/978-3-030-58526-6_13.

[40] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza, “Spectral–spatial morphological attention transformer for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5503615, doi: 10.1109/TGRS.2023.3242346.

[41] Z. Chen et al., “Global to local: A hierarchical detection algorithm for hyperspectral image target detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5544915, doi: 10.1109/TGRS.2022.3225902.

[42] C. Debes et al., “Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014, doi: 10.1109/JSTARS.2014.2305441.