Enhanced Autoencoders With Attention-Embedded Degradation Learning for Unsupervised Hyperspectral Image Super-Resolution

Lianru Gao, *Senior Member, IEEE*, Jiaxin Li, Ke Zheng, and Xiuping Jia, *Fellow, IEEE*

***Abstract*— Recently, unmixing-based networks have shown significant potential in unsupervised multispectral-aided hyperspectral image super-resolution (MS-aided HS-SR) task. Nevertheless, the representation ability of unsupervised networks and the design of loss functions still have not been fully explored, leaving large room for further improvement. To this end, we propose an enhanced unmixing-inspired unsupervised network with attention-embedded degradation learning (EU2ADL) to realize MS-aided HS-SR. First, two coupled autoencoders serve as the backbone of EU2ADL network to simultaneously decompose input modalities into abundances and corresponding endmembers, whose encoder part is composed of a spatial–spectral two-stream subnetwork for modality-salient representation learning and a parameter-shared one-stream subnetwork for modality-interacted representation enhancement. More importantly, a hybrid model-constrained loss containing a perceptual abundance term and a degradation-guided term is introduced to further eliminate the latent distortions. Since the hybrid loss is built on the degradation model, we additionally present an attention-embedded degradation learning network to adaptively estimate the unknown degradation parameters. Extensive experimental results on four datasets demonstrate the effectiveness of our proposed methods when compared with state of the arts.**

***Index Terms*— Hyperspectral image (HSI), spectral unmixing, super-resolution, unsupervised learning.**

I. Introduction

With remarkable developments achieved by remote sensing satellite imaging, numerous multimodal data become readily available nowadays, in which hyperspectral images (HSIs) have been garnering considerable attention in recent years [1]. Since the strong ability in capturing subtle spectral information distinguishes HSIs from traditional multispectral images (MSIs), various applications benefit from this burgeoning technology, such as classification [2], [3], target detection [4], [5], [6], and change detection [7], [8], to name a few.

Abundant spectral details empower HSIs to effectively identify slight variations among different objects [9], [10]. However, under the constraint of imaging principles, the pursuit of high spectral resolution inevitably leads to spatial degradation, therefore limiting the potential in fine-grained tasks to a great extent. To facilitate this problem, HSI super-resolution (HS-SR) is extensively studied to enhance the spatial resolution of HSIs by leveraging one of the following solutions, i.e., spectral SR, spatial SR, and fusion-based SR, as shown in Fig. 1(a). By means of an auxiliary image from the same scenario (e.g., MSI with high resolution), fusion-based SR is more likely to synthesize realistic products compared with the other two techniques [11], [12], [13]. Hence, in this article, we focus on multispectral-aided HS-SR (MS-aided HS-SR) technique and aim to integrate a pair of high-resolution MSI (HrMSI) and low-resolution HSI (LrHSI) to achieve a high-resolution HSI (HrHSI) product.

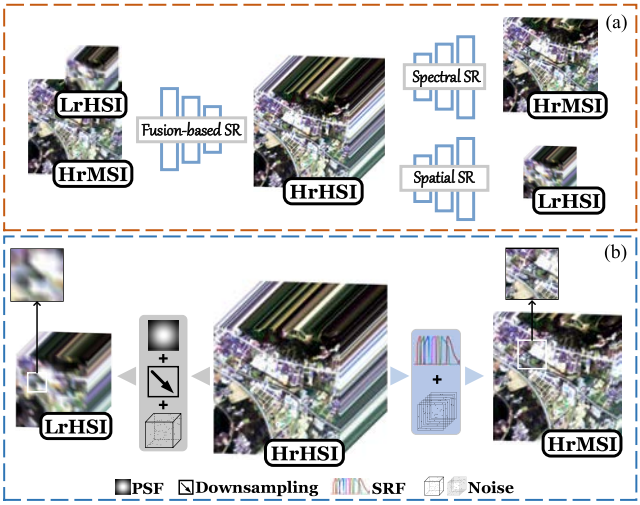

Fig. 1. (a) Illustration for the spectral SR, spatial SR, and fusion-based SR. (b) Degradation model in the MSI-aided HS-SR task.

Manuscript received 26 October 2022; revised 13 February 2023 and 7 March 2023; accepted 14 March 2023. Date of publication 17 April 2023; date of current version 2 May 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB3900502 and in part by the National Natural Science Foundation of China under Grant 42201362. *(Corresponding author: Ke Zheng.)*

Lianru Gao is with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: gaolr@aircas.ac.cn).

Jiaxin Li is with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: lijiaxin203@mails.ucas.ac.cn).

Ke Zheng is with the College of Geography and Environment, Liaocheng University, Liaocheng 252059, China (e-mail: zhengke@lcu.edu.cn).

Xiuping Jia is with the School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2612, Australia (e-mail: x.jia@adfa.edu.au).

Digital Object Identifier 10.1109/TGRS.2023.3267890

1558-0644 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

During the last decade, diverse algorithms featuring different ideas are successively proposed. To be specific, detail injection-based methods, including component substitution and multiresolution analysis, have been adapted from pansharpening to tackle the HS-SR problem. However, these computationally efficient methods are often accompanied by unavoidable distortions. By virtue of the broadly acknowledged degradation model [see Fig. 1(b)], optimization-based algorithms aim to model relationships between observed HrMSI, LrHSI, and target HrHSI under the constraint of point spread function (PSF) and spectral response function (SRF). Compared with these two categories, deep learning (DL) technologies have exhibited more powerful modeling capabilities in recent years, where supervised methods dominate for a long time. Precisely, the performance of the supervised networks is highly dependent on paired triplet data, i.e., HrMSI, LrHSI, and corresponding HrHSI. Since the HrHSI is not available, the downsampling operation is first adopted to generate downsampled training sets so that the original data can be taken as HrHSI. Then, massive training data are fed into deep networks to automatically learn their complicated mapping functions. Eventually, well-trained networks can be utilized to super-resolve the testing data. Therefore, the quality and quantity of training samples are the keys to the supervised methods. However, considering the difficulty of data acquisition and the unavailability of HrHSI, it is hard for supervised methods to apply in real life. Different from the aforementioned supervised manner, unsupervised networks have become a research hotspot in recent years on account of their requirement for only one pair of HSI–MSI. Specifically, unsupervised methods do not distinguish between the training and testing phases as the supervised approach does, that is to say, HrHSI can be directly obtained once the optimization process is completed.

Recently, the strategy of associating the HS-SR task with linear mixing theory has become a practicable approach to achieve unsupervised HR-SR, in which some representative unmixing-based networks have been proposed, including HyCoNet [14] and CUCaNet [15]. However, the representation ability of unsupervised networks and the design of loss functions still lack sufficient investigations, leaving large room for performance improvement. To further alleviate the above problems, we propose an enhanced unmixing-inspired unsupervised network with attention-embedded degradation learning (EU2ADL) for the MS-aided HS-SR task. Precisely, two coupled autoencoders serve as the backbone of our proposed method to automatically learn the latent abundance maps and corresponding endmembers through network optimization. Considering the modality differences and information interaction ignored by HyCoNet and CUCaNet, we specially design an encoder structure to fully extract modality-salient representations and strengthen their information transmission. Though the strategy of cross attention proposed in CUCaNet can facilitate their information flow, we adopt a more straightforward way, i.e., parameter-shared strategy, to effectively enhance their intercourse, successfully deriving abundance maps with strong representation ability. Besides, we propose a hybrid model-constrained loss function to separately guide the generation of abundance maps and the target image, compensating for the insufficient constraints in existing unmixing-based networks. Since the proposed loss functions are built on the degradation model, we additionally design an independent network to estimate unknown degradation parameters, i.e., PSF and SRF. Though some works intend to estimate these parameters through network learning, they either simply treat them as learnable parameters of convolutional layers [16], [17] or need side information (e.g., spectral coverage between HrMSI and LrHSI in HyCoNet and CUCaNet) to assist the estimation. Different from the above methods, we introduce the attention mechanism into the degradation estimation and treat PSF and SRF as the spatial and spectral attention maps, respectively. Due to this subtle design, sum-to-one and nonnegativity constraints can be naturally satisfied without additional operations. From what has been mentioned above, the main contributions of EU2ADL are summarized as follows.

1) A tailor-designed encoder structure is incorporated into our coupled autoencoders backbone, which consists of a spatial–spectral two-stream subnetwork (SSTS-Net) and a parameter-shared one-stream subnetwork (PSOS-Net), for comprehensive feature extraction and information interaction, leading to an enhanced representation of abundance maps.
2) To break through existing performance bottlenecks, a hybrid model-constrained loss function is presented to guide our network toward a more accurate fusion. Specifically, we first introduce the perceptual loss to measure the similarity between extracted abundance maps from two modalities, ensuring their consistency in deep feature spaces. Then, a degradation-guided loss is proposed to further improve the quality of the target product from spatial and spectral domains by virtue of estimated PSF and SRF.
3) Inspired by the attention mechanism and the degradation model, we design an independent network, called ADLnet, for degradation adaptive learning and hence successfully get rid of dependence on degradation functions.

The rest of this article is organized as follows. Related works are detailed in Section II, followed by Sections III and IV, which elaborates on our proposed EU2ADL. Extensive experiments and analysis are presented in Section V. Finally, Section VI summarizes this article.

II. Related Work

In this section, we classify existing MS-aided HS-SR methods into two groups, i.e., traditional methods and DL-based methods, and give a brief review along each research line.

A. Traditional Methods

Detail injection- and optimization-based techniques constitute the main parts of traditional methods. The former originates from pansharpening domain and then is modified to tackle the HS-SR task by adopting band selection or synthesization strategy. Precisely, Selva et al. [18] proposed a novel framework by synthesizing high-resolution bands for each

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

low-resolution band and hence successfully divided the HS-SR task into several pansharpening problems. Sylla et al. [19] calculated the cross correlation between the band set to select proper high-resolution bands for HSIs.

In contrast, the latter is designed to realize spatial enhancement on the strength of matrix factorization or tensor representation. Specifically, the SR task is reformulated as an ill-posed inverse problem with a data-fidelity term and a regularization term. Inspired by spectral unmixing, Kawakami et al. [20] separately estimated reflectance spectra and associated coefficients to reconstruct target HrHSI through a two-stage process. Following the same line, Huang et al. [21] made an improvement to predict SRF in advance from input HSI–MSI correspondence, hence realizing a semiblind fusion. Motivated by the methods mentioned above, Akhtar et al. [22] introduced two more prior assumptions, i.e., spatial smooth and nonnegativity, to constrain unknown coefficients and obtain plausible fusion results. Instead of independently calculating these matrices, Lanaras et al. [23] attempted to jointly unmix input data through an alternation way, where complete unmixing constraints are exploited to regularize this coupled unmixing problem. Instead of utilizing one dictionary to represent abundant information, Han et al. [24] added a spatial dictionary to sufficiently preserve spatial information. Considering the negative influence caused by the difference in acquisition time, Fu et al. [25] imposed the group sparsity on the inputs to account for the localized changes between two images.

Rather than destroy the inherent structure of HSIs, HS-SR has extended its branch from matrix factorization into tensor representation. Essentially, tensor representation treats HSIs as high-dimensional cubes and hence naturally maintains their spatial correlations with adjacent pixels. To be specific, Li et al. [26] incorporated the sparse prior into the core tensor and formulated a coupled sparse tensor factorization. Bu et al. [27] proposed graph Laplacian regularizations to model the local submainfold structures. Dian et al. [28] exploited local and nonlocal correlation by clustering similar patches. Apart from Tucker decomposition-based methods, Kanatsoulis et al. [29] proposed a coupled tensor method based on canonical decomposition to realize a semiblind fusion task. Ding et al. [30] made the first attempt to introduce block term decomposition into the HS-SR domain and naturally endow latent factors with physical interpretations. Xu et al. [31] incorporated nonlocal similarity and tensor sparse coding together via a novel tensor–tensor product. To capture high-order correlations, they further adopt tensor ring representation to model the multiscale structures and spectral features [32]. Dian et al. [33] employed tensor train rank to constrain grouped 4-D tensors and hence successfully learn the cluster structures. Jin et al. [34] adopted the tensor network decomposition to describe the multiscale correlation.

B. DL-Based Methods

Though convincing products can be achieved, traditional methods are criticized for weak representation capacity and undesirable time consumption. Recently, driven by the successful applications of DL techniques in computer vision, HS-SR has extended its branch into the DL domain [35], such as convolutional neural networks (CNNs), generative adversarial networks (GANs), and transformer [36]. In the early stage, the one-stream network is adopted as the standard structure by directly stacking the upsampled HSI and MSI as inputs [37], [38], [39]. This intuitive operation not only ignores inherent features in each modality but also produces extra computational pressure, and hence, the multistream network becomes the mainstream architecture of existing methods. For instance, Xu et al. [40] first extracted spatial–spectral features in different scales and blended them under the guidance of a specially designed loss function. To further preserve structural information, Hu et al. [41] designed a novel architecture to learn the residual map by virtue of attention mechanism and realize a competitive result. Different from the above methods, Yang et al. [42] proposed a deep network with two CNN branches to construct the target image pixel by pixel, where one branch is devoted to deriving spectral features of each pixel and another branch is to extract spatial features from nearby patches. Gao et al. [43] proposed two novel loss functions to generate discriminative samples and exploit physical characteristics, enabling an improvement in the fusion results. Wang et al. [44] devised a U-shaped fusion network to capture multipath and multiscale features, in which dense skip connections are embedded to realize more flexible feature fusion. Liu et al. [45] also built their U-shaped network to preserve the original spatial and spectral information, and a specially designed attention module is placed at the end of U-net to extract discriminative features. Besides, deep unrolling technique is also successfully applied to HS-SR, in which traditional optimization algorithms are embedded into deep unfolding networks to realize parameter updating. Based on this cutting-edge solution, a series of unfolding networks is proposed, leading to a transparent and explainable network [46], [47], [48]. To break through the bottleneck of CNN architectures, Hu et al. [49] pioneered the first transformer-based method to fully represent global information. Xiao et al. [16] proposed a physics-based GAN while considering the model degradation process. Very recently, Li et al. [50] attempted to jointly fuse HSI, MSI, and panchromatic images through one network, realizing a more realistic product.

Unfortunately, the superior performance of the methods above largely depends on synthesized training datasets, which greatly hampers their application in real life. Therefore, developing unsupervised methods, which only rely on one pair of HSI-MSI, has become significant in recent years [35]. Generally, there exist two ways to realize the unsupervised HS-SR task, including generation- and unmixing-based methods. The former aims to produce the target image through well-designed generators under the constraint of the degradation model. Specifically, Fu et al. [51] designed a special layer to automatically select or learn the optimal spectral response and hence effectively assist MSI to reconstruct the target image. Han et al. [52] utilized the clustering information to realize a multibranch reconstruction. Similarly, Li et al. [53] proposed a three-stage training strategy to better generate spatial details from input MSI. Besides, some works [54], [55] attempt to recover the target image from the noise map by

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

resorting to an encoder–decoder feature extracting network. Zhang et al. [17], [56] first generated an encoded image with abundant image-specific information and then fed it into the generation network for the following fusion. The latter makes assumptions that each pixel is the linear combination of several distinct materials. To be specific, Qu et al. [57] pioneered the first unsupervised network to achieve an end-to-end training process with the help of a given SRF. However, the fully connected structures severely restrict the spatial representability and hence hinder performance improvements. Following this milestone work, Wang et al. [58], [59] proposed a variational network based on nonlinear unmixing but ignored the importance of PSF, which builds the bridge between HrHSI and LrHSI. To solve the aforementioned problems, Zheng et al. [14] and Yao et al. [15] proposed a coupled autoencoder framework to realize degradation adaptive learning. Nevertheless, the lack of sufficient constraints and plain design of the encoder part inevitably degrade the fusion performance. Recently, Liu et al. [60] designed a model-inspired deep network based on nonnegative matrix factorization and fulfill a pixelwise fusion.

III. Problem Formulation

The goal of MS-aided HS-SR is to reconstruct unobservable HrHSI $\mathcal{X} \in \mathbb{R}^{H\times W\times C}$ by fusing observable LrHSI $\mathcal{Y} \in \mathbb{R}^{h\times w\times C}$ and HrMSI $\mathcal{Z} \in \mathbb{R}^{H\times W\times c}$ over the under-studying scene, where (H, W, C) and (h, w, c) are the (height, width, and number of bands) of product in optimal and degraded resolution, respectively. In particular, $w \ll W, h \ll H$, and $c \ll C$ are satisfied due to the unavoidable imaging constraint. To simplify the notation, 3-D tensors are reformulated as matrices along the spectral mode, with each row indicating a spectrum at a given pixel, i.e., $\mathbf{X} \in \mathbb{R}^{HW\times C}$, $\mathbf{Y} \in \mathbb{R}^{hw\times C}$, and $\mathbf{Z} \in \mathbb{R}^{HW\times c}$.

According to the degradation model showing the relationship between observations and the target, $\mathbf{Y}$ and $\mathbf{Z}$ can be treated as degraded versions of unknown $\mathbf{X}$ in the spatial and spectral domains, respectively, which is modeled as

$$\begin{aligned}\mathbf{Y} &= \mathbf{PX} + \mathbf{N}_y\\ \mathbf{Z} &= \mathbf{XS} + \mathbf{N}_z.\end{aligned} \tag{1}$$

Here, $\mathbf{P} \in \mathbb{R}^{\mathrm{wh}\times \mathrm{WH}}$ includes the blurring matrix derived from PSF, followed by a decimation operation along the spatial domain, and $\mathbf{S} \in \mathbb{R}^{C\times c}$ serves as SRF to map each pixel in HrHSI to a specific multispectral sensor. $\mathbf{N}_y$ and $\mathbf{N}_z$ include modeling errors and sensor noises, respectively.

By virtue of the linear mixing theory, each pixel can be linearly expressed by the product of two matrices with additional residuals [61], [62], [63], [64]. Therefore, $\mathbf{X}$ is formulated as

$$\mathbf{X} = \mathbf{AE} + \mathbf{N} \tag{2}$$

where $\mathbf{E} \in \mathbb{R}^{P\times C}$ is the endmembers matrix, with each row denoting a pure spectral signature, $\mathbf{A} \in \mathbb{R}^{\mathrm{WH}\times P}$ is the abundance matrix, with each row representing the proportion of each endmember at a specific pixel, and $\mathbf{N}$ is the residuals.

By introducing (2) into (1), $\mathbf{Y}$ and $\mathbf{Z}$ can be approximated by

$$\begin{aligned}\mathbf{Y} &= \widetilde{\mathbf{A}}\mathbf{E}, \quad \widetilde{\mathbf{A}} = \mathbf{PA}\\ \mathbf{Z} &= \mathbf{A}\widetilde{\mathbf{E}}, \quad \widetilde{\mathbf{E}} = \mathbf{ES}.\end{aligned} \tag{3}$$

Here, $\widetilde{\mathbf{A}} \in \mathbb{R}^{\mathrm{wh}\times P}$ and $\widetilde{\mathbf{E}} \in \mathbb{R}^{P\times c}$ are spatially and spectrally degraded version of abundance matrix $\mathbf{A}$ and endmember matrix $\mathbf{E}$, respectively. In this way, the goal of MS-aided HS-SR task is transformed to estimate the abundance matrix $\mathbf{A}$ and endmember matrix $\mathbf{E}$ under the constraint of (3). Hence, the SR formulation can be rewritten as

$$\begin{aligned}&\min_{\mathbf{A},\mathbf{E}} \left\|\mathbf{Y} - \widetilde{\mathbf{A}}\mathbf{E}\right\|_F^2 + \left\|\mathbf{Z} - \mathbf{A}\widetilde{\mathbf{E}}\right\|_F^2\\ &\text{s.t. } \mathbf{A}, \widetilde{\mathbf{A}}, \mathbf{E}, \widetilde{\mathbf{E}} \succeq \mathbf{0}, \quad \mathbf{A}\mathbf{1}_P = \mathbf{1}_{\mathrm{WH}}, \ \widetilde{\mathbf{A}}\mathbf{1}_P = \mathbf{1}_{\mathrm{wh}}\end{aligned} \tag{4}$$

where $\succeq$ and $||\cdot||_F$ indicate the componentwise inequality and the Frobenius norm, respectively, and $\mathbf{1}_P$ denotes the all-one vector of size P. However, the limited observations render the SR problem ill-posed, and hence, no stable solution can be obtained without reasonable constraints. Therefore, diverse prior assumptions are contrived to regularize this inverse problem, such as nonlocal [32], sparsity [65], and low rank [66]. However, these handcrafted priors largely depend on domain-specific knowledge and only fit for specific scenarios. Hence, we attempt to extract the latent abundance maps and corresponding endmembers from input images themselves by virtue of a coupled autoencoder network. We will elaborate on our network architecture as follows.

IV. Methodology

A. Overview

The limitations of existing methods discussed in Section I drive us to design an effective network for HS-SR task. As shown in Fig. 2, two coupled autoencoders are built as the backbone network of our proposed EU2ADL model, which aims at deriving the high-resolution abundance maps $\mathbf{A}$ and corresponding endmembers $\mathbf{E}$ of the desired HrHSI. In this section, we mainly describe the architecture of the proposed EU2ADL model, whose specific components are detailed in the following.

Given HrMSI $\mathbf{Z}$ and LrHSI $\mathbf{Y}$, we intend to explicitly learn their abundance maps by sending them into our tailor-designed encoder part. Precisely, $\mathbf{Z}$ and $\mathbf{Y}$ are first fed to different streams in SSTS-Net for modality-salient representation learning. By the aid of cascaded multiscale spatial residual blocks (MSRBs) and spectral residual blocks (SRBs), SSTS-Net is able to separately extract their spatial and spectral features. Then, these extracted high-level features pass through the parameter-shared PSOS-Net for modality-interacted representation enhancement. The whole procedure of abundance maps learning can be expressed as

$$\mathbf{A} = f_{\mathrm{en}}\left(\mathbf{Z}; \mathbf{W}_{f,\mathrm{en}}\right) \tag{5}$$

$$\widetilde{\mathbf{A}} = g_{\mathrm{en}}\left(\mathbf{Y}; \mathbf{W}_{g,\mathrm{en}}\right) \tag{6}$$

where $f_{\mathrm{en}}(\cdot)$ and $g_{\mathrm{en}}(\cdot)$ denote the mapping function from inputs to corresponding abundance maps with learnable parameters $\mathbf{W}$.

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

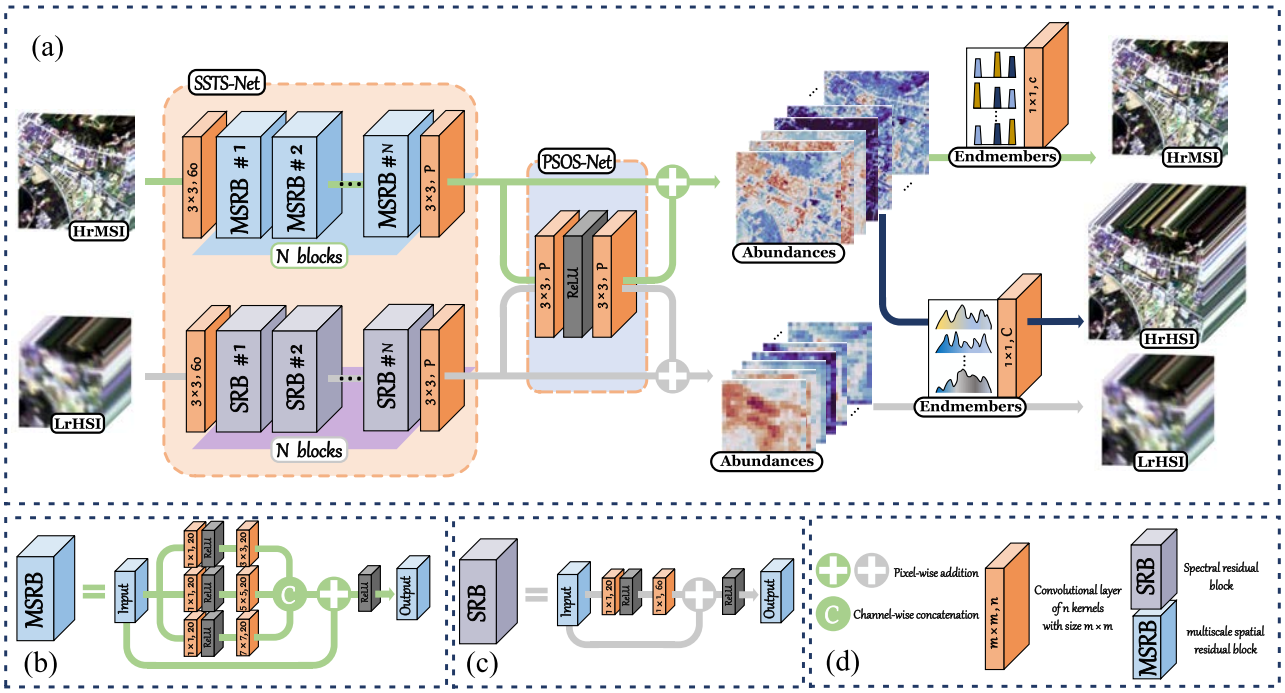

Fig. 2. Overall structures of the proposed EU2ADL and its detailed components. (a) Overview of EU2ADL that contains two coupled autoencoders, (b) and (c) details of MSRB and SRB in SSTS-Net, and (d) corresponding legend.

Many works have demonstrated the potential of linear operators in modeling linear mixing problems. Therefore, 1×1 convolution without any bias is applied here to constitute our decoder part, aiming at recovering the original inputs from their own abundance maps. This can be calculated as

$$\widehat{\mathbf{Z}} = f_{\text{de}}\left(\mathbf{A}; \mathbf{W}_{f,\text{de}}\right) = f_{\text{de}}\left(f_{\text{en}}\left(\mathbf{Z}; \mathbf{W}_{f,\text{en}}\right); \mathbf{W}_{f,\text{de}}\right) \quad (7)$$

$$\widehat{\mathbf{Y}} = g_{\text{de}}\left(\widetilde{\mathbf{A}}; \mathbf{W}_{g,\text{de}}\right) = g_{\text{de}}\left(g_{\text{en}}\left(\mathbf{Y}; \mathbf{W}_{g,\text{en}}\right); \mathbf{W}_{g,\text{de}}\right). \quad (8)$$

Here, $f_{\text{de}}(\cdot)$ and $g_{\text{de}}(\cdot)$ act as the decoders of the proposed EU2ADL, and hence, $\mathbf{W}$ can be naturally interpreted as endmembers, i.e., $\widetilde{\mathbf{E}}$ and $\mathbf{E}$. Once unknown parameters $\mathbf{W}_{f,\text{en}}$ and $\mathbf{W}_{g,\text{de}}$ are obtained, we can reconstruct the desired HrHSI by resorting to (2), i.e.,

$$\widehat{\mathbf{X}} = g_{\text{de}}\left(f_{\text{en}}\left(\mathbf{Z}; \mathbf{W}_{f,\text{en}}\right); \mathbf{W}_{g,\text{de}}\right). \quad (9)$$

B. Encoder for Salient Representation Learning and Interacted Enhancement

In the previous works, simple network architectures without any subtle designs are usually employed to derive abundance maps, which largely weakens the feature representation and leads to quality degradation. To solve this problem, an elaborately designed encoder structure, containing SSTS-Net and PSOS-Net, is embedded into our coupled auntoencoders to realize thorough feature extraction and interaction.

1) SSTS-Net for Modality-Salient Representation Learning: SSTS-Net contains two independent and parallel streams, which aims to differentially exploit the modality-salient features from two diverse modalities, i.e., HrMSI-stream for spatial content feature extraction and LrHSI-stream for spectral feature extraction. Initially, a 3×3 convolutional layer is deployed at the beginning of each stream to simply detect the shallow features of $\mathbf{Z}$ and $\mathbf{Y}$, which can be expressed as

$$\mathbf{Z}_1 = \text{Conv}(\mathbf{Z}), \quad \mathbf{Y}_1 = \text{Conv}(\mathbf{Y}). \quad (10)$$

Then, these initial features are fed into cascaded blocks for the subsequent process. Specifically, considering the abundant structure and texture information in HrMSI, we purposefully contrive MSRB for HrMSI-stream to capture spatial information and jointly exploit their multiscale features. First, feature maps are separately sent to three parallel branches, each of which contains a 1×1 convolution for dimension reduction, a rectified linear unit (ReLU), and a convolutional layer for multiscale feature extraction. Subsequently, channelwise concatenation is followed to integrate primary features extracted from different scales. Finally, to further utilize low- and high-level features and promote information flow, a skip connection is established between input and output via pixelwise addition. Different from HrMSI-branch which mainly focuses on spatial domains, SRB is implemented in the LrHSI branch to make full use of rich spectral information at each pixel. Similar to MSRB, residual learning is employed to build the bridge between input and output. Inside the residual structure, 1×1 kernel size is employed in all convolutional operators because it exhibits great ability in capturing spectral features without involving adjacent pixels. Finally, a set of MSRBs and SRBs are intentionally cascaded to progressively learn the hierarchical features, which can be described as

$$\begin{aligned} \mathbf{Z}_2 &= f_{\text{MSRB},N}\left(f_{\text{MSRB},N-1}\left(\cdots f_{\text{MSRB},1}(\mathbf{Z}_1)\cdots\right)\right) \\ \mathbf{Y}_2 &= f_{\text{SRB},N}\left(f_{\text{SRB},N-1}\left(\cdots f_{\text{SRB},1}(\mathbf{Y}_1)\cdots\right)\right) \end{aligned} \quad (11)$$

where $f_{\text{MSRB},N}$ and $f_{\text{SRB},N}$ denote the operation of the Nth MSRN and SRN, respectively. At the end of SSTS-Net,

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

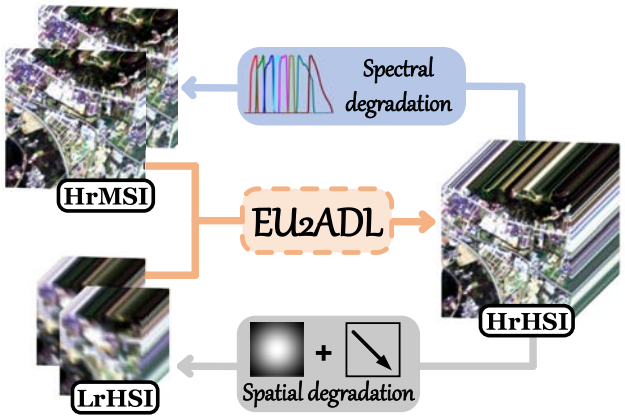

Fig. 3. Illustration for the proposed degradation-guided loss.

we apply a 3 × 3 convolution to generate the final modality-salient representation, which can be expressed as

$$\begin{aligned}\mathbf{Z}_{\text{salient}} &= \text{Conv}(\mathbf{Z}_2)\\ \mathbf{Y}_{\text{salient}} &= \text{Conv}(\mathbf{Y}_2).\end{aligned} \tag{12}$$

2) PSOS-Net for Modality-Interacted Representation Enhancement: Unlike previous works that derive abundance maps by resorting to two separate branches with mutually independent parameters, we specially design PSOS-Net to enhance the feature representation by virtue of the parameter-shared coupling strategy. Concretely, two consecutive 3 × 3 convolutional layers in tandem with a ReLU activation function constitute the main part of PSOS-Net. Different from isolated networks whose parameters are optimized independently, the advantages of such coupling design are two folds. First, it can effectively shield computation to a certain extent. Second, this design will greatly promote information interaction since different modalities will jointly influence the parameter learning of PSOS-Net. By sending the modality-salient representations into PSOS-Net, the final enhanced representations can be obtained by virtue of a weighted summation strategy with the following formulation:

$$\begin{aligned}\mathbf{A} &= \omega f_{\text{PSOS-Net}}(\mathbf{Z}_{\text{salient}}) + \mu \mathbf{Z}_{\text{salient}}\\ \widetilde{\mathbf{A}} &= \omega f_{\text{PSOS-Net}}(\mathbf{Y}_{\text{salient}}) + \mu \mathbf{Y}_{\text{salient}}\end{aligned} \tag{13}$$

where $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ are the enhanced representations, i.e., abundance maps, along with the weight coefficients ω and μ.

C. Network Training

1) Loss Functions: The loss function of EU2ADL is a weighted combination of a basic reconstruction loss, abundance sum-to-one constraint (ASC) loss, and our purposefully designed model-constrained loss, which can be expressed as

$$\begin{aligned}\mathcal{L}_{\text{overall}} = \underbrace{\lambda_{\text{rec}}\mathcal{L}_{\text{rec}}}_{\text{reconstruction loss}} + \underbrace{\lambda_{\text{ASC}}\mathcal{L}_{\text{ASC}}}_{\text{sum-to-one constrain loss}}\\ + \underbrace{\lambda_{\text{per}}\mathcal{L}_{\text{per}} + \lambda_{\text{deg}}\mathcal{L}_{\text{deg}}}_{\text{model-constrained loss}}.\end{aligned} \tag{14}$$

The target HrHSI often contains complicated objects of different sizes and contents, exhibiting abundant spatial information, such as texture and edge. However, the solutions of L_2 loss often lack high-frequency content with overly smooth textures, which leads to the loss of spatial details in the outcome and produces unsatisfactory images. In contrast, the L_1 loss tends to recover subtle and sharp details and hence is chosen as the loss criterion in our method.

Reconstruction Loss: Since two coupled autoencoders establish the basic framework of our EU2ADL, the first loss is the reconstruction loss stemming from inputs and their corresponding reconstructions, which can be computed as

$$\mathcal{L}_{\text{rec}} = \left\|\mathbf{Z} - \widehat{\mathbf{Z}}\right\|_1 + \left\|\mathbf{Y} - \widehat{\mathbf{Y}}\right\|_1. \tag{15}$$

Sum-to-One Loss: In order to satisfy the constrains given by (4), the ASC loss is introduced as follows:

$$\mathcal{L}_{\text{ASC}} = \|\mathbf{1}_{\text{WH}} - \mathbf{A}\mathbf{1}_P\|_1 + \left\|\mathbf{1}_{\text{wh}} - \widetilde{\mathbf{A}}\mathbf{1}_P\right\|_1. \tag{16}$$

Model-Constrained Loss: Apart from the two loss functions mentioned above, a hybrid model-constrained loss, including two terms, is proposed, with one perceptual loss for abundance maps and another degradation-guided loss for the target product.

We first propose a degradation-guided loss with explicit physical meanings. As the name suggests, the loss term aims to guide our network to generate the desired result that satisfies the degradation model shown in (1), which can be calculated as

$$\mathcal{L}_{\text{deg}} = \left\| f_{\text{PSF}}\left(\widehat{\mathbf{X}}; W_{\text{PSF}}\right) - \mathbf{Y}\right)\right\|_1 + \left\| f_{\text{SRF}}\left(\widehat{\mathbf{X}}; W_{\text{SRF}}\right) - \mathbf{Z}\right)\right\|_1 \tag{17}$$

where $f_{\text{PSF}}(\cdot)$ denotes the spatial degradation containing a spatial blurring and a downsampling operator and $f_{\text{SRF}}(\cdot)$ denotes the spectral degradation using the SRF. To better illustrate this idea, Fig. 3 shows the proposed degradation-guided loss. First, HrMSI $\mathbf{Z}$ and LrHSI $\mathbf{Y}$ are fed into our proposed EU2ADL to generate HrHSI $\widehat{\mathbf{X}}$. Then, the gray and blue arrows aim to spatially and spectrally degrade $\widehat{\mathbf{X}}$, which correspond to the first and second terms in (17), respectively. Finally, under the constraint of the proposed degradation-guided loss, two degraded versions of $\widehat{\mathbf{X}}$ should be identical to the inputs. However, the parameters of PSF and SRF, i.e., W_{PSF} and W_{SRF}, are often unavailable in most cases, and hence, their side information is usually given in advance for the following fusion process. For example, Qu et al. [57] and Wang et al. [58], [59], treated these sensor-related parameters as known priors and works [14], [15] need the spectral coverage between HrMSI and LrHSI. To overcome the limitations mentioned above, we design an independent ADLnet to automatically estimate these unknown parameters and hence realize a blind fusion task. The estimate process will be detailed in Section IV-D.

The second term is inspired by (3) that describes an important relationship between abundance maps $\mathbf{A}$ and $\widetilde{\mathbf{A}}$, i.e., $\widetilde{\mathbf{A}} = \mathbf{P}\mathbf{A}$. Some works [57], [59] intend to model this correlation to improve their fusion performance. However, they simply adopt a duplication operation to align their different spatial sizes and

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

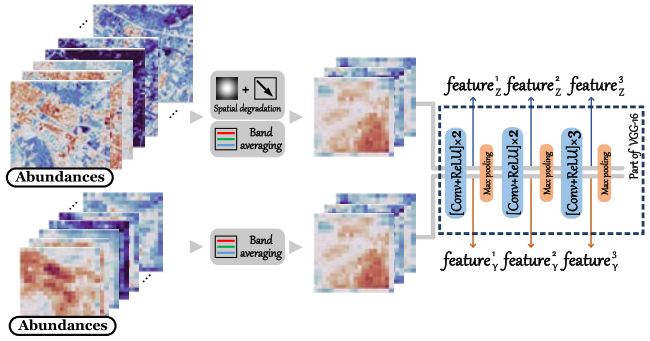

Fig. 4. Illustration for perceptual abundance loss.

hence inevitably introduce more uncertainty into the network training. To solve the problem, we substitute the PSF kernel estimated by ADLnet for the duplication operation, enabling a more precise constraint. However, due to the inevitable noise and the estimation error from **P**, this relationship cannot be perfectly satisfied. If strong pixel-level constraints, i.e., plain L_1 or L_2 loss, are imposed on the abundance relationship, uncertainty and negative influences may be introduced. In contrast, the perceptual loss is widely used to capture the differences in the feature space via well-pretrained networks, aiming to pursue the match at a semantic and perceptual level instead. Considering this important property of perceptual loss, we adopt it to realize the abundance constraint, allowing some degree of variation in the pixel level. Fig. 4 shows three main procedures to realize our proposed abundance loss. First, the spatial degradation is imposed on the abundance map **A** to generate desired **PA**. Since the VGG-16 network [67] for feature extraction is originally designed for natural images, we average adjacent bands into three before feeding them into the subsequent network. Eventually, three-band abundance maps are put into the VGG-16 network, and three features (i.e., $\text{feature}_Z^{1,2,3}$ and $\text{feature}_Y^{1,2,3}$) before the max-pooling layer are chosen as feature maps for comparison. The process of perceptual loss can be mathematically expressed as

$$\mathcal{L}_{\text{per}} = \left\| \phi_{\text{vgg}}^{1,2,3}\left(f_{\text{average}}(f_{\text{PSF}}(\mathbf{A};\, W_{\text{PSF}}))\right) - \phi_{\text{vgg}}^{1,2,3}\left(f_{\text{average}}\left(\widetilde{\mathbf{A}}\right)\right) \right\|_1. \tag{18}$$

Here, $f_{\text{PSF}}(\cdot)$ is used to degrade the abundance map **A** and obtain desired **PA**, $f_{\text{average}}(\cdot)$ represents the averaging operator, and $\phi_{\text{vgg}}^{1,2,3}(\cdot)$ are three chosen feature maps of VGG-16 network.

2) Bells and Whistles: To satisfy the nonnegativity constraint, the clamp function is deployed at the end of the encoder and decoder parts to shrink the abundance value and the reconstructed product into the range [0, 1]. Besides, the clamp operator is also imposed on endmembers after each parameter updating to guarantee their nonnegativity properties.

3) Implementation Details: Our networks are trained with the Adam optimizer [68] under the PyTorch environment [69]. We set the learning rate as 0.001 for ADLnet with 4000 training epochs. The linear decay schedule is adopted to EU2ADL from 7000 to 14 000 epochs, and the initial learning rate is also set as 0.001. Kaiming initialization [70] is applied to initialize the weights of our networks. Finally, ω and μ in (13) are

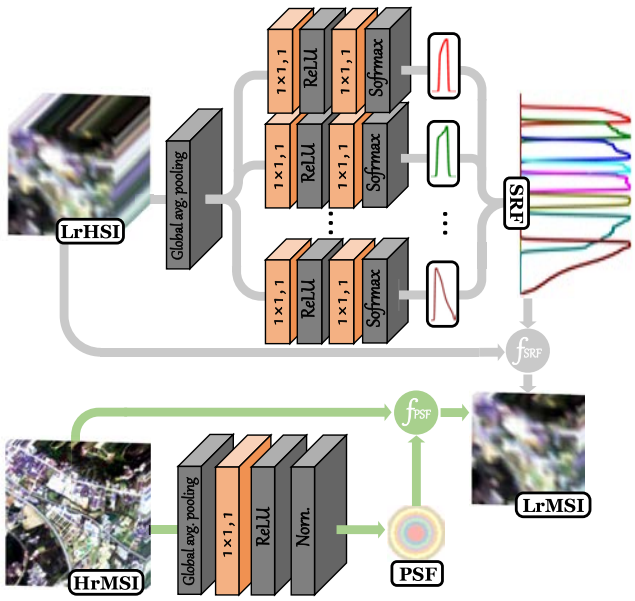

Fig. 5. Illustration for the proposed ADLnet.

empirically set as 0.0005; the weight of λ_{rec}, λ_{ASC}, λ_{per}, and λ_{deg} controlling the different loss terms in (14) are set as 100, 0.01, 1, and 10, respectively.

D. Attention-Embedded Degradation Learning Network

In this section, we propose ADLnet to estimate unknown degradation parameters by transforming them into learnable attention maps. According to the degradation model shown in (1), LrHSI **Y** and HrMSI **Z** are the spatially and spectrally degraded version of HrHSI **X**, respectively. To obtain the unknown degradation parameters (i.e., **P** and **S**), we derive another relationship between input **Y** and **Z**, namely, the spectrally degraded version of LrHSI $\mathbf{Y}_{\downarrow}$ should be close to the spatially degraded version of HrMSI $\mathbf{Z}_{\downarrow}$, which can be expressed as

$$\begin{aligned} \mathbf{Y}_{\downarrow} &= \mathbf{YS} = (\mathbf{PX} + \mathbf{N}_{\mathbf{y}})\mathbf{S} \\ \mathbf{Z}_{\downarrow} &= \mathbf{PZ} = \mathbf{P}(\mathbf{XS} + \mathbf{N}_{\mathbf{z}}) \\ \mathbf{Y}_{\downarrow} &= \mathbf{Z}_{\downarrow}. \end{aligned} \tag{19}$$

Here, $\mathbf{Y}_{\downarrow}$ and $\mathbf{Z}_{\downarrow}$ can be treated as two latent images with both low spatial and spectral resolutions, and we call them low-resolution MSI (LrMSI). Based on (19), the loss function of our ADLnet (see Fig. 5) is given as follows:

$$\mathcal{L}_{\text{ADLnet}} = \|\mathbf{Y}_{\downarrow} - \mathbf{Z}_{\downarrow}\|_1. \tag{20}$$

To be specific, the core of SRF is basically a weighting process followed by a summation operator along the spectral axis, which aims to give each channel a specific weight to synthesize one corresponding band [14], [23], [66]. Similarly, channel attention is to adaptively derive a channelwise attention map to recalibrate the weight of each channel [71], whose idea is highly consistent with SRF. Hence, spectral attention is adopted here to derive the desired SRF, as shown in the upper half of Fig. 5. Specifically, global averaging pooling is first imposed on LrHSI **Y** to obtain a 1×1 descriptor, and then,

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

a network with c parallel branches (note that c is the spectral number of HrMSI) is followed up, each of which is designed to learn one spectral response. In each branch, a modified version of squeeze-and-excitation network (SE-net) [72] is chosen as the main component, in which the last sigmoid activation is replaced by the softmax activation to satisfy sum-to-one and nonnegative constraints. Once obtaining the estimated SRF, a process containing channelwise multiplication and a summation operator (represented as notation f_{SRF} in Fig. 5) is followed to simulate the spectral degradation and generate the first LrMSI $\mathbf{Y}_{\downarrow}$, which can be calculated as

$$\mathbf{Y}_{\downarrow} = f_{\text{SRF}}\left(\mathbf{Y}; \text{SE}_{\text{mod}}\left(\mathbf{Y}; \mathbf{W}_{\text{SE}_{\text{mod}}}\right)\right) \tag{21}$$

where the outputs of $\text{SE}_{\text{mod}}(\cdot)$ denote the learned SRF and $f_{\text{SRF}}(\cdot)$ represents the process to generate $\mathbf{Y}_{\downarrow}$ using extracted SRF.

On the other hand, the purpose of PSF is principally to mix the information of nearby pixels with given weights [73], such as Gaussian kernel and mean kernel. Correspondingly, spatial attention aims to focus on the most important regions by giving each pixel a learnable weight, whose learning process can be intuitively utilized to derive our needed PSF. As shown in the lower half of Fig. 5, we first adopt global averaging pooling to shrink the spatial size of input HrMSI $\mathbf{Z}$ into that of our desired PSF, and then, a combination of 1×1 convolution, ReLU activation, and normalization operator is used to derive unknown PSF. Finally, a band-by-band convolution process with learned PSF kernel and specific strides (represented as notation f_{PSF} in Fig. 5) is deployed to generate the second LrMSI $\mathbf{Z}_{\downarrow}$, which can be modeled as

$$\mathbf{Z}_{\downarrow} = f_{\text{PSF}}(\mathbf{Z}; \text{Network}(\mathbf{Z}; \mathbf{W}_{\text{Network}})) \tag{22}$$

where the output of $\text{Network}(\cdot)$ denotes the learned PSF kernel and $f_{\text{PSF}}(\cdot)$ represents the convolution process to generate $\mathbf{Z}_{\downarrow}$ with extracted PSF kernel and given strides. It is worth mentioning that some previous works also intend to estimate these unknown degradation parameters by simply treating them as trainable parameters of convolutional layers [14], [60], [74]. Since the parameter updating cannot guarantee their physical constraints, i.e., sum-to-one and nonnegativity, they have to truncate their estimated parameters into reasonable scopes every iteration. In contrast, advisable operators can be embedded into our ADLnet to naturally satisfy these restrictions and hence successfully get rid of these tedious steps.

V. Experiments and Results

In this section, comprehensive experiments are conducted to verify the performance of our proposed method. First, the influence of some key parameters is discussed, including the number of blocks $\mathbf{N}$ and endmembers $\mathbf{P}$. Second, an ablation study is followed to investigate the effectiveness of the proposed PSOS-Net, ADLnet, and loss functions. Finally, four public HSI datasets are used, i.e., PRISMA dataset, Houston dataset, TianGong-1 dataset, and Chikusei dataset, to compare our method with other ten state of the arts.

TABLE I
Dataset Specifications

	PRISMA	Houston	TianGong-1	Chikusei
Size of HrHSI	$400 \times 400 \times 69$	$400 \times 400 \times 46$	$240 \times 240 \times 54$	$400 \times 400 \times 110$
Spectral range of HrHSI (nm)	434 - 2442	380 – 1050	413 - 887	363 - 1018
Scale factor	8	10	12	16
Size of LrHSI	$50 \times 50 \times 69$	$40 \times 40 \times 46$	$20 \times 20 \times 54$	$25 \times 25 \times 110$
Size of HrMSI	$400 \times 400 \times 8$	$400 \times 400 \times 8$	$240 \times 240 \times 8$	$400 \times 400 \times 8$

A. Datasets and Setup

1) Datasets: The PRISMA datasets are provided by the Italian Space Agency for the 2022 Hyperspectral Pansharpening Challenge,[1] which consist of four co-registered hyperspectral and panchromatic images [75]. One HSI with 69 bands and 400×400 pixels is selected as the reference image. The Houston dataset is released by the Hyperspectral Image Analysis Laboratory[2] for the 2018 IEEE GRSS Data Fusion Competition. These hyperspectral data have 48 bands with 1-m spatial resolution covering a spectral range of 380–1050 nm [76]. A subimage of $400 \times 400 \times 46$ pixels is utilized as the reference image after discarding noise and water absorption bands. The TianGong-1 datasets[3] are originally produced by Tiangong-1 Hyperspectral Imager for scene classification [77]. A representative subimage titled "city-015-VNI-2013041514" with $240 \times 240 \times 54$ pixels in the visible near-infrared spectral range is used as the reference image. The Chikusei dataset[4] was taken by Headwall Hyperspec-VNIR-C imaging sensor in Chikusei with original 2517×2335 pixels and 128 bands from 363 to 1018 nm [78]. A subimage with size of $400 \times 400 \times 110$ is cropped as the reference data by removing noisy bands.

According to the widely acknowledged Wald' protocol [79], we aim to construct the corresponding LrHSI and HrMSI from the aforementioned reference images. Since the spatial degradation represents the process to blur and downsample the reference image, we first apply an isotropic Gaussian kernel on each band of reference data and then downsample every $r \times r$ pixels in the spatial domain, where r is the scaling factor, to generate the simulated LrHSI. In the setting of our experiments, r is set as 8, 10, 12, and 16 in the four datasets. The spectral response with eight bands in WorldView 2 is used to synthesize HrMSI [80]. Table I summarizes the main specifications of datasets used in this article.

2) Compared Methods: Ten representative methods are chosen for the performance comparison, including one detail injection-based method, i.e., SFIM [81], three matrix decomposition-based methods, i.e., G-SOMP+ [22], CSU [23], and CNMF [66], two tensor representation methods, i.e., STEREO [29] and CSTF [26], and four DL-based methods, i.e., UDALN [53], HyCoNet [14], CUCaNet [15], and MIAE [60].

3) Evaluation Metrics: We will evaluate these methods from quantitative and visual perspectives. First, six

[1] https://www.ieee-whispers.com/hyperspectral-pansharpening-challenge/
[2] https://hyperspectral.ee.uh.edu/
[3] http://www.msadc.cn/main/setsubDetail?id=1369487569196158978
[4] http://naotoyokoya.com/Download.html

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

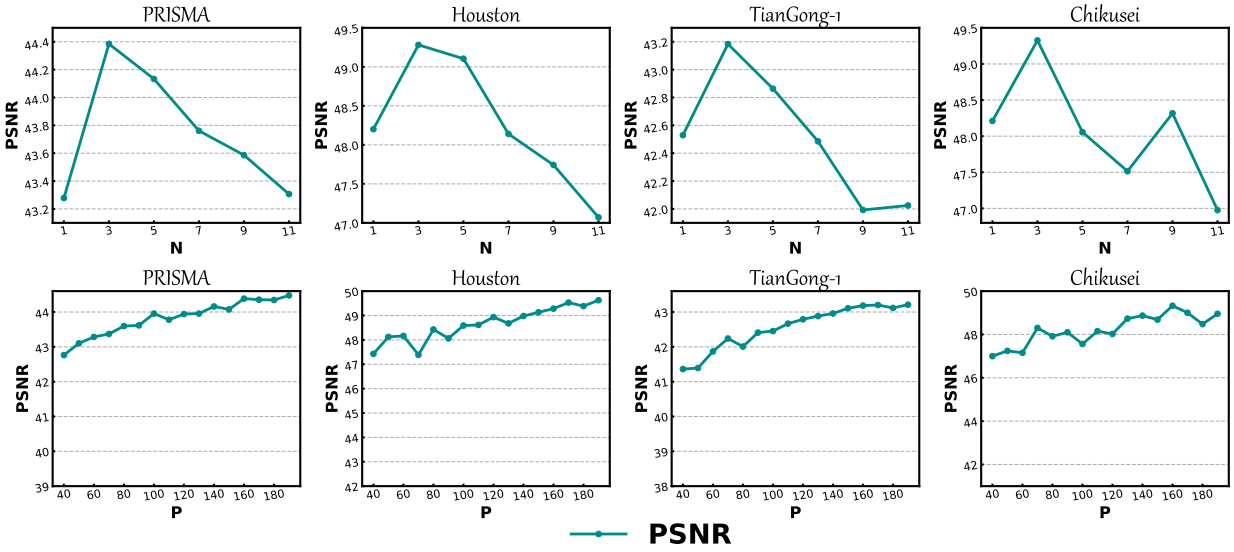

Fig. 6. PSNR curves under different setting of N and P in four datasets. The first and second lines exhibit the relationship between PSNR and the number of blocks N and endmembers P, respectively.

TABLE II

ABLATION STUDY OF PSOS-NET IN FOUR DATASETS. THE BEST ONE IS SHOWN IN BOLD. THE ↑ AND ↓ MEAN THAT HIGHER AND LOWER VALUES CORRESPOND TO BETTER RESULTS, RESPECTIVELY

Datasets	Methods	SAM↓	PSNR↑	ERGAS↓	RMSE↓	SSIM↑	UIQI ↑
PRISMA	w/o PSOS-Net	3.0546	42.7698	1.3080	0.0130	0.9610	0.9942
	with PSOS-Net	**2.6649**	**44.3848**	**1.1599**	**0.0112**	**0.9652**	**0.9953**
Houston	w/o PSOS-Net	1.2415	47.0358	0.2692	0.0041	0.9951	0.9995
	with PSOS-Net	**1.0594**	**49.2815**	**0.2431**	**0.0034**	**0.9957**	**0.9996**
TianGong-1	w/o PSOS-Net	1.2273	41.3873	0.2455	0.0080	0.9855	0.9995
	with PSOS-Net	**1.0714**	**43.1837**	**0.1865**	**0.0062**	**0.9875**	**0.9996**
Chikusei	w/o PSOS-Net	1.0764	48.0803	0.6208	0.0045	0.9948	0.9985
	with PSOS-Net	**0.9986**	**49.3260**	**0.6012**	**0.0043**	**0.9953**	**0.9986**

well-established metrics are selected for quantitative evaluation, including root-mean-square error (RMSE) and peak signal-to-noise ratio (PSNR) for the spatial fidelity [82], spectral angle mapper (SAM) for the spectral recovery [83], relative dimensionless global error in synthesis (ERGAS) for global measures [84], structure similarity (SSIM), and universal image quality index (UIQI) based on image statistics [85], [86]. Second, we choose three forms of visualization, i.e., SAM heatmap, mean relative absolute error (MRAE) heatmap, and the residual heatmap at one specific band, to assist the performance comparison. Specifically, SAM heatmap aims to calculate the spectral similarity in each spectral vector, MRAE heatmap is based on the relative absolute difference between different bands, and the residual heatmap gives a representative band to show the spatial reconstruction quality.

B. Parameters Discussion

Two parameters, i.e., the number of blocks **N** and endmembers **P**, determine the structure of our designed network, and hence, we will evaluate their influence on the fusion performance of EU2ADL. First, our encoder part consists of a two-stream SSTS-Net, which mainly contains **N** cascaded blocks for modality-salient representation learning, and hence, the parameter **N** plays a key role in our network. As shown in the first row of Fig. 6, the PSNR curves for the four datasets reach the peak value when N arrives at 3 and then drop rapidly with N continuing to increase. Hence, we set N as 3 in four datasets. Second, the number of endmember **P** controls the kernel number of EU2ADL and hence affects the representation ability of encoder part [87], [88], [89]. Consequently, we will determine the number of **P** in terms of fusion performance. From the second row, it can be clearly seen that the PSNR curves show an upward trend as **P** increases from 40 to 160, and then, the values only have a marginal rise, even a sudden decrease when **P** keeps on rising. Since the increase of **P** comes with extra network parameters and computational stress, we prefer to choose $\mathbf{P} = 160$ for the four datasets in the following experiments. Besides, we have to argue that the number of **P** is actually larger than that of distinct materials in the scene in order to further consider the nonlinear mixture phenomenon and incorporate the underlying spectral variability, enabling our network to better represent scene spectra under complex situations and result in a satisfactory separation of the spectral signal sources.

C. Ablation Study

In this section, we verify the effectiveness of the proposed PSOS-Net for enhanced representation learning, ADLnet for

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

TABLE III
ABLATION STUDY OF ADLNET IN FOUR DATASETS. THE BEST ONE IS SHOWN IN BOLD

Datasets	Methods	SAM	PSNR	ERGAS	RMSE	SSIM	UIQI
PRISMA	w/o ADLnet	**2.6067**	**44.4293**	**1.1483**	**0.0110**	**0.9657**	**0.9958**
	with ADLnet	2.6649	44.3848	1.1599	0.0112	0.9652	0.9953
Houston	w/o ADLnet	**1.0514**	**49.6382**	**0.2349**	**0.0033**	**0.9958**	**0.9996**
	with ADLnet	1.0594	49.2815	0.2431	0.0034	0.9957	**0.9996**
TianGong-1	w/o ADLnet	**1.0601**	**43.2897**	**0.1854**	**0.0061**	**0.9877**	**0.9996**
	with ADLnet	1.0714	43.1837	0.1865	0.0062	0.9875	**0.9996**
Chikusei	w/o ADLnet	**0.9908**	**49.5654**	**0.5985**	**0.0043**	**0.9954**	**0.9986**
	with ADLnet	0.9986	49.3260	0.6012	**0.0043**	0.9953	**0.9986**

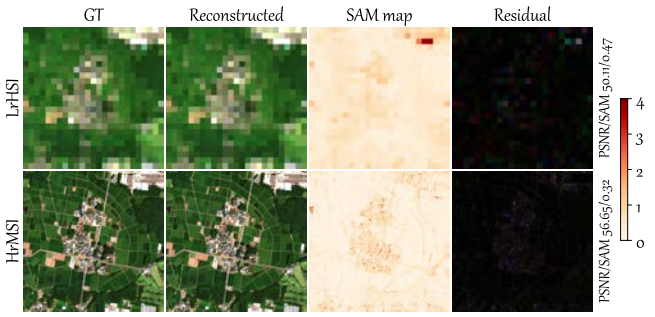

Fig. 7. Reconstructed LrHSI and HrMSI from generated HrHSI Chikusei using estimated PSF and SRF, along with their quality measures.

TABLE IV
ABLATION STUDY IN THE CHIKUSEI DATASET WITH A COMBINATION OF DIFFERENT LOSS FUNCTIONS

No.	Loss function	SAM	PSNR	ERGAS
(a)	$\mathcal{L}_{rec} + \mathcal{L}_{ASC}$	27.4905	16.8450	16.2701
(b)	$\mathcal{L}_{rec} + \mathcal{L}_{ASC} + \mathcal{L}_{dup}$	25.3793	17.9379	12.0629
(c)	$\mathcal{L}_{rec} + \mathcal{L}_{ASC} + \mathcal{L}_{per}$	21.2323	19.1878	11.2074
(d)	$\mathcal{L}_{rec} + \mathcal{L}_{ASC} + \mathcal{L}_{deg}$	1.4978	48.1336	0.6645
(e)	$\mathcal{L}_{rec} + \mathcal{L}_{ASC} + \mathcal{L}_{deg} + \mathcal{L}_{per}$	0.9986	49.3260	0.6012

the degradation model learning, and deliberately designed loss functions for improving the fusion performance.

1) Influence of PSOS-Net: Different from the previous work whose encoder part only depends on two independent networks, a parameter-shared PSOS-Net is proposed and embedded into our EU2ADL for the modality-interacted representation enhancement. In this section, the performance gain provided by PSOS-Net is evaluated by removing or retaining this subnetwork. In Table II, it is clearly shown that all indicators obtained with PSOS-Net are greatly improved than that obtained w/o PSOS-Net. Precisely, EU2ADL w/o PSOS-Net extracts the abundance maps only depending on SSTS-Net and hence lacks interaction between two modalities. In contrast, PSOS-Net enables our EU2ADL to have sufficient information exchange between different branches and provides great potential for better fusion results. From the above comparisons, the positive impact gained from PSOS-Net is convincingly exhibited.

2) Influence of ADLnet: We present ADLnet to adaptively estimate the unknown degradation parameters, i.e., PSF and SRF. To evaluate its performance, we conduct an extra experiment by providing EU2ADL with exact PSF and SRF and compared their results with blind ones. It can be observed in Table III that methods w/o ADLnet achieve better results in four datasets when compared with blind ones. However, their performance gaps are very small, which can confirm the strength of ADLnet. Besides, we further perform another experiment in the Chikusei dataset to check the reconstructed LrHSI and HrMSI using estimated PSF and SRF. As shown in Fig. 7, the first and second rows depict the ground-truth (GT) image, reconstructed images, map of SAM error between the two images, and the residual images of three chosen bands in LrHSI and HrMSI. It can be clearly seen that the reconstructed images have high spatial fidelity with less spectral distortions, indicating that our ADLnet can effectively learn these unknown parameters with high precision.

3) Influence of Loss Functions: We conduct an experiment in the Chikusei dataset to study the effectiveness of different loss functions proposed in (14). At the same time, we also investigate the performance of two abundance constraints, including a normal L_1 loss based on the duplication operator, i.e., $\mathcal{L}_{\text{dup}}$, and our proposed perceptual abundance loss $\mathcal{L}_{\text{per}}$.

In Table IV, the loss combination of (a), which contains a basic reconstruction loss and a sum-to-one loss, is treated as the baseline. We can see that it is difficult for our network to converge in the situation of (a). Inspired by previous works [57], [59], we first adopt a duplication operator, i.e., nearest neighbor interpolation, to upsample the abundance maps of LrHSI to the same resolution as HrMSI and then calculate their similarity using a simple L_1 loss, as shown in (b). Compared with (a), there are some improvements by adding the $\mathcal{L}_{\text{dup}}$ loss. When we replace $\mathcal{L}_{\text{dup}}$ with our proposed perceptual abundance loss $\mathcal{L}_{\text{per}}$ in (c), more significant promotions can be achieved. However, it is still hard for our method to achieve acceptable results, and hence, we study the impact of another proposed degradation-guided loss $\mathcal{L}_{\text{deg}}$ in (d). It is demonstrable that $\mathcal{L}_{\text{deg}}$ contributes significantly to the performance improvement in our network learning since all quality metrics surge dramatically compared with (b) and (c). More importantly, the performance is further boosted under the joint cooperation of the hybrid loss, as shown in (e).

D. Comparison With State of the Arts

In this section, EU2ADL is compared with ten state of the art methods from quantitative and visual perspectives in Tables V–VIII and Figs. 8–12, respectively.

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

TABLE V
QUALITY EVALUATION IN THE PRISMA DATASET. THE BEST ONE IS SHOWN IN BOLD

Method	SAM	PSNR	ERGAS	RMSE	SSIM	UIQI
SFIM	3.4257	37.0899	4.1488	0.0183	0.9645	0.9913
G-SOMP+	3.6360	39.0766	1.6567	0.0152	0.9429	0.9893
CSU	3.3003	38.7994	1.4587	0.0140	0.9057	0.9925
CNMF	3.0621	40.3642	1.2817	0.0128	0.9496	0.9938
STEREO	3.9409	40.4490	1.5863	0.0159	0.9202	0.9900
CSTF	3.2606	40.3678	1.4153	0.0136	0.9243	0.9932
UDALN	3.0205	40.6779	1.3336	0.0133	0.9452	0.9929
HyCoNet	2.9755	40.6387	1.3683	0.0121	0.9562	0.9924
CUCaNet	3.1637	40.3472	1.4326	0.0129	0.9614	0.9920
MIAE	2.8518	42.7023	1.2274	0.0115	0.9591	0.9934
Ours	**2.6649**	**44.3848**	**1.1599**	**0.0112**	**0.9652**	**0.9953**

1) PRISMA: The visualization results of all methods in the PRISMA dataset are shown in Fig. 8. Overall, most methods can generate products without noticeable distortions. Specifically, CNMF obtains good performance in traditional methods on account of its coupled strategy and special update rules. However, traditional methods are generally inferior to DL-based methods because of their limited representation ability. Clearly, our proposed EU2ADL outperforms other competitors in both spectral and spatial domains, followed by MIAE and UDALN, as shown in the second and third lines. Also, another example at band 10 is provided to convincingly exhibit the strength of EU2ADL since our method can generate most of the detailed information in the whole area. Moreover, we give the quantitative result in Table V to straightforwardly evaluate all methods. From Table V, we can draw the same conclusion that our method achieves the best performance in all metrics and obtain a large improvement in the PSNR value. In contrast, MIAE exhibits a competitive ability in preserving spectral information, which can be confirmed in Table V. However, G-SOMP+ and STEREO fail to deal with spectral distortion and lose many spatial details. More importantly, the PSNR curve as a function of band number is drawn in Fig. 12. It further demonstrates the superior reconstruction ability of our proposed method because the highest value in most bands is achieved by EU2ADL.

2) Houston: Fig. 9 shows the visualization results of all methods. Visually, UDALN and MIAE yield acceptable products with high spatial and spectral fidelity, while traditional methods show unsatisfying results in the building areas, especially in the roof regions. However, EU2ADL performs even better than UDALN and MIAE on the whole image, showing a strong ability in recovering roofs and shadow regions which other methods fail to deal with. Besides, Table VI summarizes the quality measures in the Houston dataset. From Table VI, a large gap can be observed between our method and other methods, especially in PSNR value, which forcefully confirms the strong performance of EU2ADL. Also, MIAE and UDALN exhibit competitive results compared with the other two DL-based unmixing methods, partly because HyCoNet and CUCaNet lack sufficient constraints in network learning. Tensor representation-based methods achieve better results than matrix decomposition-based methods since the former can perfectly preserve the original structure information of HSI. The PSNR curve shown in Fig. 12 also consistently verifies the superiority of our method, followed by MIAE.

3) TianGong-1: The visual results for TianGong-1 dataset are shown in Fig. 10. The product of matrix-based methods, such as G-SOMP+ and CSU, are accompanied by noticeable distortions, when compared with tensor representation-based methods, due partly to its destruction of spatial–spectral structures. Again, EU2ADL can provide better visual results with grained details and preserved spectral features when compared with HyCoNet and CUCaNet because reasonable constraints are introduced to guide our network toward an optimal value. Besides, MIAE and UDALN exhibit the closest results to our method, but they fail to handle the areas covered by green trees and blue roofs. However, our method can still give stable results in these areas with high fidelity. Table VII reports the quality measures for the TianGong-1 dataset, which gives basically identical conclusions as the visual evaluation. EU2ADL yields the best result in terms of all metrics, followed by MIAE and UDALN. In contrast, HyCoNet and CUCaNet behave poorly in spectral preservation, which can be further observed from the visual perspective. The PSNR curve shown in Fig. 12 exhibits the competitive result of our method. Though MIAE performs better in some bands, our method has advantages in the overall trend.

4) Chikusei: Fig. 11 shows the visual results of all methods in the Chikusei dataset. It is hard for traditional methods to tackle this complex scene where farmland surrounds residential areas. Hence, large spectral distortion inevitably occurs, especially in the building areas located in the middle and up-right of the image. UDALN, CUCaNet, and MIAE behave better in the building areas compared with traditional methods but still suffer from spectral distortions. HyCoNet and EU2ADL obtain competitive results in the middle buildings, but our method behaves better than HyCoNet in recovering the surrounding farmlands. Besides, EU2ADL also shows superior performance in spatial fidelity, as shown in the third and fourth rows. Table VIII summarizes the quantitative indexes in the Chikusei dataset, which further demonstrates our conclusions drawn from visual results. More concretely, EU2ADL achieves

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

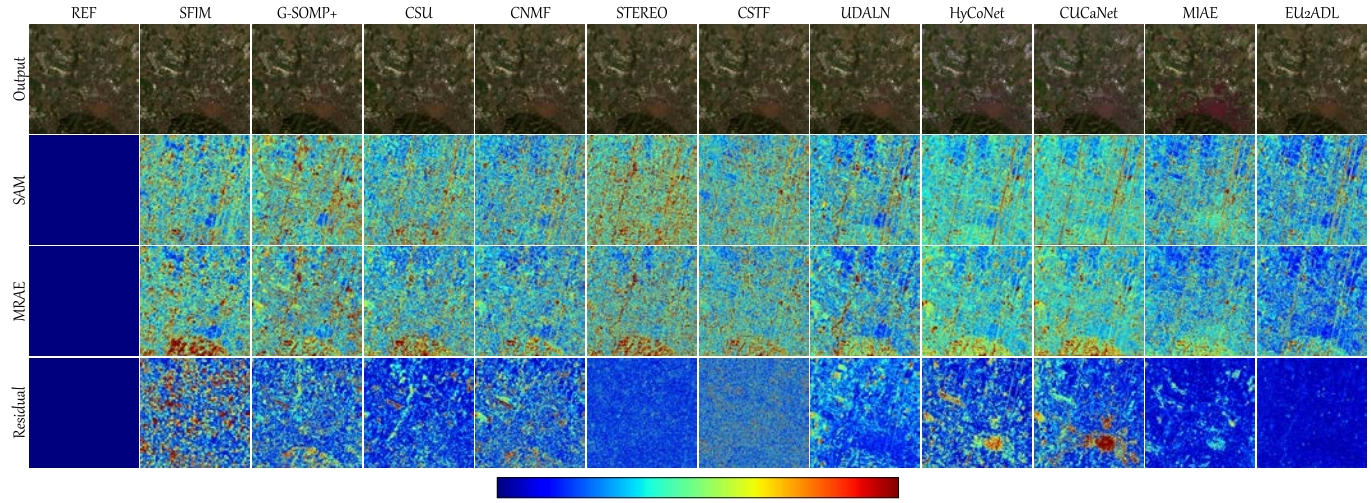

Fig. 8. Visualization of super-resolved PRISMA dataset. First row: color composite images of reconstructed outputs from different methods (R:14, G:8, and B:4). Second row: heatmap of SAM error. Third row: heatmap of MRAE. Fourth row: residual heatmap at band 10. The error range of three kinds of heat maps is [0, 6.5], [0, 0.15], and [0, 0.009].

TABLE VI
QUALITY EVALUATION IN THE HOUSTON DATASET. THE BEST ONE IS SHOWN IN BOLD

Method	SAM	PSNR	ERGAS	RMSE	SSIM	UIQI
SFIM	2.4014	34.2800	0.9395	0.0158	0.9732	0.9917
G-SOMP+	1.8888	39.0345	0.5428	0.0089	0.9897	0.9970
CSU	2.0575	39.3028	0.5233	0.0086	0.9842	0.9961
CNMF	1.6926	40.0283	0.4837	0.0081	0.9904	0.9968
STEREO	1.9761	41.7109	0.4130	0.0068	0.9782	0.9986
CSTF	1.6295	41.8930	0.3965	0.0067	0.9778	0.9988
UDALN	1.2060	43.3353	0.3414	0.0056	0.9951	0.9991
HyCoNet	1.9733	43.0014	0.4694	0.0071	0.9929	0.9979
CUCaNet	1.8013	42.9907	0.3845	0.0067	0.9938	0.9986
MIAE	1.1928	47.1067	0.2593	0.0042	0.9956	0.9994
Ours	**1.0594**	**49.2810**	**0.2431**	**0.0034**	**0.9957**	**0.9996**

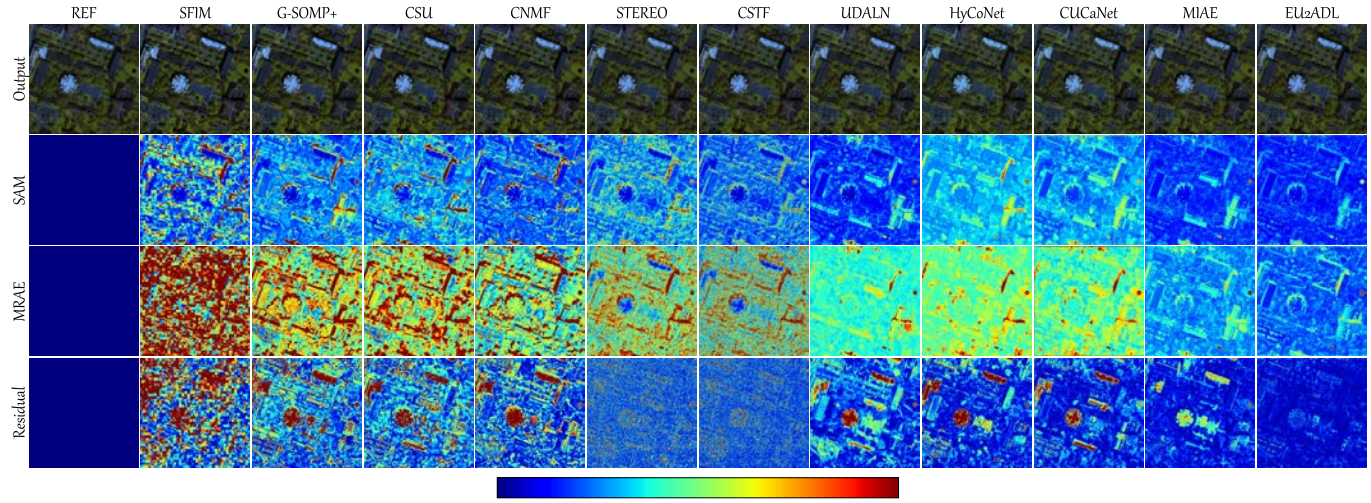

Fig. 9. Visualization of super-resolved Houston dataset. First row: color composite images of reconstructed outputs from different methods (R:46, G:30, and B:14). Second row: heatmap of SAM error. Third row: heatmap of MRAE. Fourth row: residual heatmap at band 10. The error range of three kinds of heat maps is [0, 6], [0, 0.06], and [0, 0.015].

the best performance in all cases, indicating our competitive ability in the HS-SR task, followed by HyCoNet. Also, the PSNR curve of our method in Fig. 12 shows competitive results in most bands.

E. Robustness Analysis

In this section, we will analyze the influence of scale factor and noise on the fusion performance. Four competitive

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

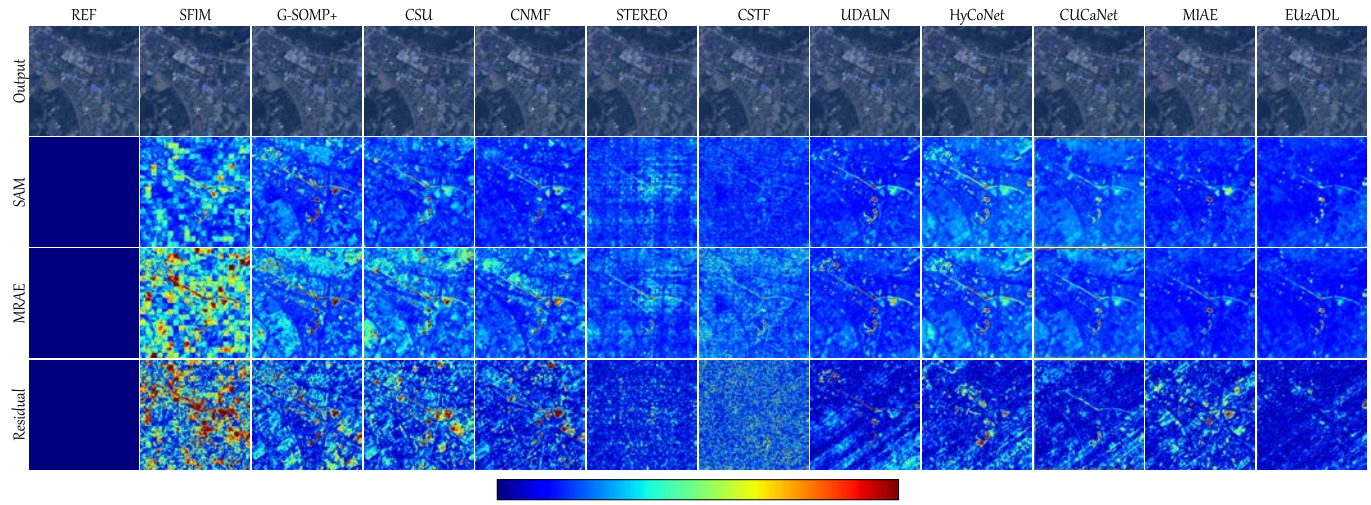

Fig. 10. Visualization of super-resolved TianGong-1 dataset. First row: color composite images of reconstructed outputs from different methods (R:29, G:19, and B:6). Second row: heatmap of SAM error. Third row: heatmap of MRAE. Fourth row: residual heatmap at band 48. The error range of three kinds of heat maps is [0, 6.5], [0, 0.1], and [0, 0.026].

TABLE VII
QUALITY EVALUATION IN THE TIANGONG-1 DATASET. THE BEST ONE IS SHOWN IN BOLD

Method	SAM	PSNR	ERGAS	RMSE	SSIM	UIQI
SFIM	1.6937	34.3666	0.4706	0.0143	0.9712	0.9974
G-SOMP+	1.4568	37.5235	0.3134	0.0104	0.9748	0.9988
CSU	1.3395	38.0308	0.2927	0.0094	0.9721	0.9989
CNMF	1.1640	38.7733	0.2746	0.0087	0.9826	0.9992
STEREO	1.2939	41.3143	0.2255	0.0072	0.9751	0.9994
CSTF	1.1384	40.0151	0.2418	0.0075	0.9702	0.9994
UDALN	1.1762	40.8782	0.2285	0.0074	0.9847	0.9993
HyCoNet	1.5245	40.1675	0.2677	0.0087	0.9827	0.9990
CUCaNet	1.5013	39.8469	0.2979	0.0094	0.9839	0.9992
MIAE	1.1444	42.4696	0.2129	0.0071	0.9869	0.9995
Ours	**1.0714**	**43.1837**	**0.1865**	**0.0062**	**0.9875**	**0.9996**

TABLE VIII
QUALITY EVALUATION IN THE CHIKUSEI DATASET. THE BEST ONE IS SHOWN IN BOLD

Method	SAM	PSNR	ERGAS	RMSE	SSIM	UIQI
SFIM	1.5159	36.2492	1.1064	0.0129	0.9805	0.9848
G-SOMP+	1.5879	39.9848	0.9302	0.0089	0.9888	0.9949
CSU	1.8886	37.0912	1.1093	0.0113	0.9711	0.9848
CNMF	1.3095	37.8300	0.9571	0.0107	0.9911	0.9954
STEREO	1.8479	44.3364	0.7067	0.0070	0.9783	0.9910
CSTF	1.2220	43.3144	0.6211	0.0061	0.9819	0.9942
UDALN	1.2145	45.5635	0.6774	0.0049	0.9882	0.9968
HyCoNet	1.1181	47.8634	0.6280	0.0048	0.9948	0.9979
CUCaNet	1.2785	46.1209	0.6726	0.0056	0.9943	0.9974
MIAE	1.1767	42.2955	0.7200	0.0072	0.9910	0.9980
Ours	**0.9986**	**49.3260**	**0.6012**	**0.0043**	**0.9953**	**0.9986**

methods from three categories, i.e., matrix decomposition, tensor representation, and DL, are chosen as the competitors according to their performance in previous experiments, and Houston dataset is given as an example.

1) Scale Factor: We examine the robustness of five methods against different scale factors, including 8, 10, 16, and 20, as shown in Fig. 13. As the scale factor increases, performance degradation inevitably occurs in general. Precisely, the SAM value of MIAE decreases from about 1.1 to 1.6, and the PSNR value drops from around 48.1 to 45.4. CNMF and CSTF exhibit poor performance, and their SAM value changes from about 1.4 to 2.0. Though the SAM and PSNR values of the proposed EU2ADL drop from about 1.0 and 49.7 to about 1.4 and 47.0, respectively, our method still outperforms other

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

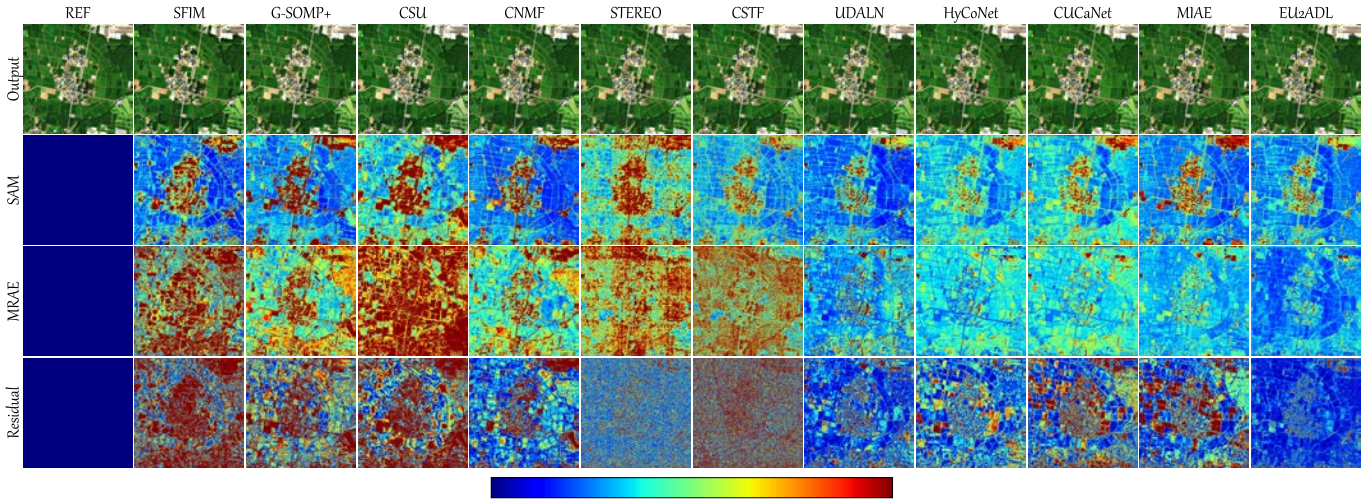

Fig. 11. Visualization of super-resolved Chikusei dataset. First row: color composite images of reconstructed outputs from different methods (R:56, G:36, and B:16). Second row: heatmap of SAM error. Third row: heatmap of MRAE. Fourth row: residual heatmap at band 45. The error range of three kinds of heat maps is [0, 2.7], [0, 0.08], and [0, 0.004].

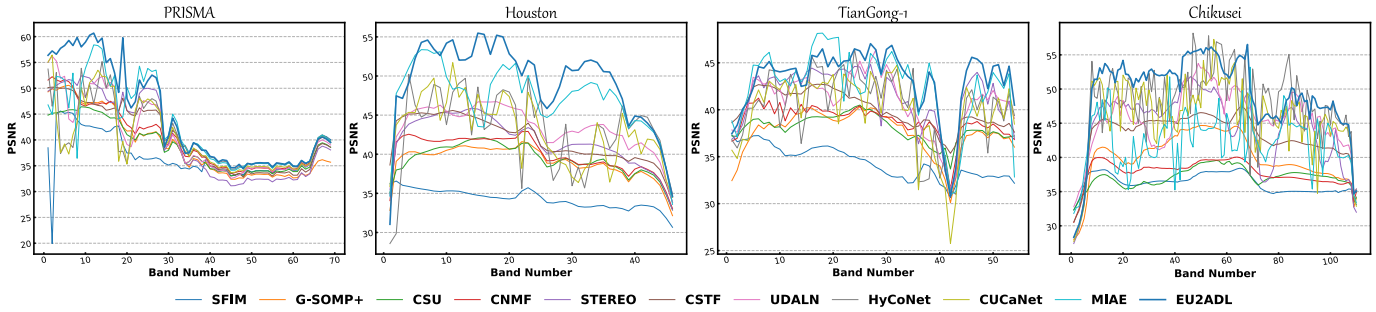

Fig. 12. PSNR value as a function of band number in four datasets.

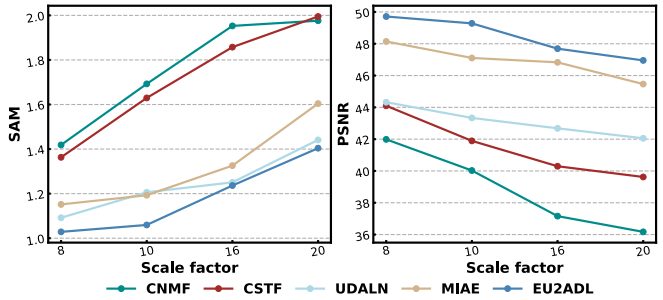

Fig. 13. Quantitative evaluation of five methods under four different scale factors in the Houston dataset.

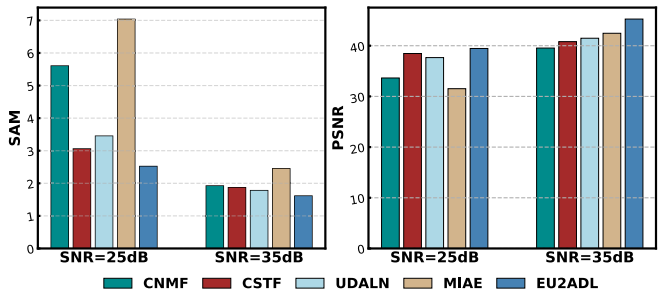

Fig. 14. Quantitative evaluation of five methods under different noise cases in the Houston dataset.

competitors in all cases, which confirms the superiority of EU2ADL.

2) Noise: After obtaining the LrHSI and HrHSI based on the simulation process described in Section V-A, the Gaussian noise is simultaneously added into the LrHSI and HrHSI. Fig. 14 presents the fusion results in the Houston dataset under two different noise signal-to-noise ratios (SNRs). Overall, it is hard for CNMF and MIAE to produce desirable results, which indicates their sensitivity to noise. Precisely, the SAM values of MIAE are about 7.0 and 2.4 under two different situations, and the corresponding results of CNMF are about 5.6 and 1.9. In contrast, our method still obtains the highest results in all noise cases, largely due to the proposed hybrid loss constraints. In the noise condition of 25 dB, our method achieves about 2.5 SAM values and 39.5 PSNR values. In the second situation, we obtain about 1.6 SAM values and 45.3 PSNR values.

VI. Conclusion

In this article, we propose an EU2ADL to solve the MS-aid HS-SR task. Inspired by the unmixing mechanism, we design a novel encoder part for comprehensive feature extraction and information interaction and hence derive enhanced abundance maps for the subsequent reconstruction. Meanwhile, a hybrid model-constrained loss is proposed to guarantee the desirable product. On this basis, in order to get rid of the dependence on sensor-related parameters, we deliberately devise ADLnet to estimate unknown PSF and SRF by virtue

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

of the attention mechanism. Extensive experiments, including parameter discussion, ablation study in PSOS-Net, ADLnet, and loss functions, comparison with state of the arts, and robustness analysis, are conducted from quantitative and visual perspectives to show that our proposed EU2ADL and its components are effective and superior to the existing methods.

In future works, the proposed method will be further improved in two directions: one is to automatically determine the optimal hyperparameters, such as the number of endmembers, and the other is to design effective modules to realize feature extraction and interaction.

ACKNOWLEDGMENT

The authors would like to thank the IEEE GRSS IADF and the Hyperspectral Image Analysis Lab for providing the Houston dataset, Dr. Yokoya for providing the Chikusei dataset, the Italian Space Agency for providing the PRISMA dataset, and Kang Liu for providing the TianGong-1 dataset. They would also like to thank the authors who kindly provided their codes for comparison.

REFERENCES

[1] K. Shen, X. Yang, S. Lolli, and G. Vivone, "A continual learning-guided training framework for pansharpening," *ISPRS J. Photogramm. Remote Sens.*, vol. 196, pp. 45–57, Feb. 2023.

[2] Z. Zhang et al., "Multireceptive field: An adaptive path aggregation graph neural framework for hyperspectral image classification," *Expert Syst. Appl.*, vol. 217, May 2023, Art. no. 119508.

[3] Y. Ding, X. Zhao, Z. Zhang, W. Cai, N. Yang, and Y. Zhan, "Semi-supervised locality preserving dense graph neural network with ARMA filters and context-aware learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2021.

[4] L. Gao, D. Wang, L. Zhuang, X. Sun, M. Huang, and A. Plaza, "BS^{3}LNet: A new blind-spot self-supervised learning network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5504218.

[5] D. Wang, L. Gao, Y. Qu, X. Sun, and W. Liao, "Frequency-to-spectrum mapping GAN for semisupervised hyperspectral anomaly detection," *CAAI Trans. Intell. Technol.*, vol. 2023, pp. 1–16, Jan. 2023.

[6] M. Wang, Q. Wang, D. Hong, S. Roy, and J. Chanussot, "Learning tensor low-rank representation for hyperspectral anomaly detection," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 1–13, Jan. 2022.

[7] L. Liu, D. Hong, L. Ni, and L. Gao, "Multilayer cascade screening strategy for semi-supervised change detection in hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1926–1940, 2022.

[8] Y. Wang et al., "Mask DeepLab: End-to-end image segmentation for change detection in high-resolution remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 104, Dec. 2021, Art. no. 102582.

[9] R. Song, Y. Feng, W. Cheng, Z. Mu, and X. Wang, "BS2T: Bottleneck spatial–spectral transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532117.

[10] 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.

[11] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: A comparative review of the recent literature," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 2, pp. 29–56, Jun. 2017.

[12] R. Dian, S. Li, B. Sun, and A. Guo, "Recent advances and new guidelines on hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 69, pp. 40–51, May 2021.

[13] N. Liu, W. Li, Y. Wang, R. Tao, Q. Du, and J. Chanussot, "A survey on hyperspectral image restoration: From the view of low-rank tensor approximation," 2022, *arXiv:2205.08839*.

[14] K. Zheng et al., "Coupled convolutional neural network with adaptive response function learning for unsupervised hyperspectral super resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2487–2502, Mar. 2020.

[15] 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. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 208–224.

[16] J. Xiao, J. Li, Q. Yuan, M. Jiang, and L. Zhang, "Physics-based GAN with iterative refinement unit for hyperspectral and multispectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6827–6841, 2021.

[17] L. Zhang, J. Nie, W. Wei, Y. Li, and Y. Zhang, "Deep blind hyperspectral image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2388–2400, Jun. 2020.

[18] M. Selva, B. Aiazzi, F. Butera, L. Chiarantini, and S. Baronti, "Hyper-sharpening: A first approach on SIM-GA data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 3008–3024, Jun. 2015.

[19] D. Sylla, A. Minghelli-Roman, P. Blanc, A. Mangin, and O. H. F. d'Andon, "Fusion of multispectral images by extension of the pan-sharpening ARSIS method," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 5, pp. 1781–1791, May 2014.

[20] R. Kawakami, Y. Matsushita, J. Wright, M. Ben-Ezra, Y.-W. Tai, and K. Ikeuchi, "High-resolution hyperspectral imaging via matrix factorization," in *Proc. IEEE Conf Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 2329–2336.

[21] B. Huang, H. Song, H. Cui, J. Peng, and Z. Xu, "Spatial and spectral image fusion using sparse matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 3, pp. 1693–1704, Mar. 2014.

[22] N. Akhtar, F. Shafait, and A. Mian, "Sparse spatio-spectral representation for hyperspectral image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)* Cham, Switzerland: Springer, 2014, pp. 63–78.

[23] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3586–3594.

[24] X. Han, J. Yu, J.-H. Xue, and W. Sun, "Hyperspectral and multispectral image fusion using optimized twin dictionaries," *IEEE Trans. Image Process.*, vol. 29, pp. 4709–4720, 2020.

[25] X. Fu, S. Jia, M. Xu, J. Zhou, and Q. Li, "Fusion of hyperspectral and multispectral images accounting for localized inter-image changes," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517218.

[26] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4118–4130, Aug. 2018.

[27] Y. Bu et al., "Hyperspectral and multispectral image fusion via graph Laplacian-guided coupled tensor decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 648–662, Jan. 2021.

[28] R. Dian, S. Li, L. Fang, T. Lu, and J. M. Bioucas-Dias, "Nonlocal sparse tensor factorization for semiblind hyperspectral and multispectral image fusion," *IEEE Trans. Cybern.*, vol. 50, no. 10, pp. 4469–4480, Oct. 2020.

[29] C. I. Kanatsoulis, X. Fu, N. D. Sidiropoulos, and W.-K. Ma, "Hyperspectral super-resolution: A coupled tensor factorization approach," *IEEE Trans. Signal Process.*, vol. 66, no. 24, pp. 6503–6517, Dec. 2018.

[30] M. Ding, X. Fu, T. Z. Huang, J. Wang, and X. L. Zhao, "Hyperspectral super-resolution via interpretable block-term tensor modeling," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 3, pp. 641–656, Apr. 2021.

[31] Y. Xu, Z. Wu, J. Chanussot, and Z. Wei, "Nonlocal patch tensor sparse representation for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3034–3047, Jun. 2019.

[32] Y. Xu, Z. Wu, J. Chanussot, and Z. Wei, "Hyperspectral images super-resolution via learning high-order coupled tensor ring representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4747–4760, Nov. 2020.

[33] R. Dian, S. Li, and L. Fang, "Learning a low tensor-train rank representation for hyperspectral image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2672–2683, Sep. 2019.

[34] D. Jin, J. Liu, J. Yang, and Z. Wu, "High-order coupled fully connected tensor network decomposition for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[35] J. Li et al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102926.

[36] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.

[37] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-D-convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 639–643, May 2017.

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

[38] X.-H. Han, B. Shi, and Y. Zheng, "SSF-CNN: Spatial and spectral fusion with CNN for hyperspectral image super-resolution," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2506–2510.

[39] X. Zhang, W. Huang, Q. Wang, and X. Li, "SSR-NET: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5953–5965, Jul. 2021.

[40] S. Xu, O. Amira, J. Liu, C.-X. Zhang, J. Zhang, and G. Li, "HAM-MFN: Hyperspectral and multispectral image multiscale fusion network with RAP loss," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4618–4628, Jul. 2020.

[41] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7251–7265, Dec. 2022.

[42] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network," *Remote Sens.*, vol. 10, no. 5, p. 800, 2018.

[43] H. Gao, S. Li, and R. Dian, "Hyperspectral and multispectral image fusion via self-supervised loss and separable loss," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5537712.

[44] X. Wang, X. Wang, K. Zhao, X. Zhao, and C. Song, "FSL-Unet: Full-scale linked Unet with spatial–spectral joint perceptual attention for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539114.

[45] S. Liu, S. Liu, S. Zhang, B. Li, W. Hu, and Y.-D. Zhang, "SSAU-Net: A spectral–spatial attention-based U-Net for hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5542116.

[46] D. Shen, J. Liu, Z. Wu, J. Yang, and L. Xiao, "ADMM-HFNet: A matrix decomposition-based deep approach for hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2021.

[47] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "MHF-Net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1457–1473, Mar. 2022.

[48] J. Yang, L. Xiao, Y.-Q. Zhao, and J. C.-W. Chan, "Variational regularization network with attentive deep prior for hyperspectral–multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5508817.

[49] J.-F. Hu, T.-Z. Huang, L.-J. Deng, H.-X. Dou, D. Hong, and G. Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[50] K. Li, W. Zhang, D. Yu, and X. Tian, "HyperNet: A deep network for hyperspectral, multispectral, and panchromatic image fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 188, pp. 30–44, Jun. 2022.

[51] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang, "Hyperspectral image super-resolution with optimized RGB guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11653–11662.

[52] X. Han, J. Yu, J. Luo, and W. Sun, "Hyperspectral and multispectral image fusion using cluster-based multi-branch BP neural networks," *Remote Sens.*, vol. 11, no. 10, p. 1173, 2019.

[53] J. Li, K. Zheng, J. Yao, L. Gao, and D. Hong, "Deep unsupervised blind hyperspectral and multispectral data fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[54] T. Uezato, D. Hong, N. Yokoya, and W. He, "Guided deep decoder: Unsupervised image pair fusion," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 87–102.

[55] S. Liu, S. Miao, J. Su, B. Li, W. Hu, and Y.-D. Zhang, "UMAG-Net: A new unsupervised multiattention-guided network for hyperspectral and multispectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7373–7385, 2021.

[56] L. Zhang, J. Nie, W. Wei, Y. Zhang, S. Liao, and L. Shao, "Unsupervised adaptation learning for hyperspectral imagery super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3073–3082.

[57] Y. Qu, H. Qi, and C. Kwan, "Unsupervised sparse Dirichlet-Net for hyperspectral image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2511–2520.

[58] Z. Wang, B. Chen, H. Zhang, and H. Liu, "Unsupervised hyperspectral and multispectral images fusion based on nonlinear variational probabilistic generative model," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 721–735, Feb. 2022.

[59] Z. Wang, B. Chen, R. Lu, H. Zhang, H. Liu, and P. K. Varshney, "FusionNet: An unsupervised convolutional variational network for hyperspectral and multispectral image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 7565–7577, 2020.

[60] J. Liu, Z. Wu, L. Xiao, and X.-J. Wu, "Model inspired autoencoder for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522412.

[61] J. R. Patel, M. V. Joshi, and J. S. Bhatt, "A novel approach for hyperspectral image superresolution using spectral unmixing and transfer learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 1512–1515.

[62] J. S. Bhatt and M. V. Joshi, "Deep learning in hyperspectral unmixing: A review," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 2189–2192.

[63] J. R. Patel, M. V. Joshi, and J. S. Bhatt, "Spectral unmixing using autoencoder with spatial and spectral regularizations," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2021, pp. 3321–3324.

[64] L. Ren, Z. Ma, F. Bovolo, and L. Bruzzone, "A nonconvex framework for sparse unmixing incorporating the group structure of the spectral library," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5506719.

[65] R. Dian, S. Li, L. Fang, and Q. Wei, "Multispectral and hyperspectral image fusion with spatial–spectral sparse representation," *Inf. Fusion*, vol. 49, pp. 262–270, Sep. 2019.

[66] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2011.

[67] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[69] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Neural Inf. Process. Syst. (NIPS)*, vol. 32, 2019, pp. 1–15.

[70] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2015, pp. 1026–1034.

[71] M.-H. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 2022, pp. 1–38, Jan. 2022.

[72] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[73] Q. Wang and P. M. Atkinson, "The effect of the point spread function on sub-pixel mapping," *Remote Sens. Environ.*, vol. 193, pp. 127–137, May 2017.

[74] K. Zheng, L. Gao, D. Hong, B. Zhang, and J. Chanussot, "NonRegSRNet: A nonrigid registration hyperspectral super-resolution network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5520216.

[75] G. Vivone, A. Garzelli, Y. Xu, W. Liao, and J. Chanussot, "Panchromatic and hyperspectral image fusion: Outcome of the 2022 WHISPERS hyperspectral pansharpening challenge," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 166–179, 2023.

[76] B. Le Saux, N. Yokoya, R. Hansch, and S. Prasad, "2018 IEEE GRSS data fusion contest: Multimodal land use classification [technical committees]," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 1, pp. 52–54, Mar. 2018.

[77] K. Liu et al., "Scene classification dataset using the Tiangong-1 hyperspectral remote sensing imagery and its applications," *J. Remote Sens.*, vol. 24, pp. 1077–1087, Aug. 2020.

[78] N. Yokoya and A. Iwasaki, "Airborne hyperspectral data over Chikusei," Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep., SAL-2016-05-27, May 2016.

[79] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.

[80] C. Padwick, M. Deskevich, F. Pacifici, and S. Smallwood, "Worldview-2 pan-sharpening," in *Proc. ASPRS Annu. Conf.*, vol. 2630, 2010, pp. 1–14.

[81] J. G. Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *Int. J. Remote Sens.*, vol. 21, no. 18, pp. 3461–3472, Nov. 2010.

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.

[82] F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, and J. A. Benediktsson, "Quantitative quality evaluation of pansharpened imagery: Consistency versus synthesis," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1247–1259, Mar. 2016.
[83] F. A. Kruse et al., "The spectral image processing system (SIPS) interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, nos. 2–3, pp. 145–163, 1993.
[84] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?" in *Proc. Int. Conf. Fusion Earth Data*, 2000, pp. 99–103.
[85] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.
[86] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[87] C.-I. Chang, "A review of virtual dimensionality for hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp. 1285–1305, Apr. 2018.
[88] J. S. Bhatt and B. Chattopadhyay, "Virtual dimensionality of hyperspectral data: Use of multiple hypothesis testing for controlling type-I error," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2974–2985, 2020.
[89] V. S. Deshpande and J. S. Bhatt, "A practical approach for hyperspectral unmixing using deep learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

Lianru Gao (Senior Member, IEEE) received the B.S. degree in civil engineering from Tsinghua University, Beijing, China, in 2002, and the Ph.D. degree in cartography and geographic information system from the Institute of Remote Sensing Applications, Chinese Academy of Sciences (CAS), Beijing, in 2007.

He is currently a Professor with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, CAS. He has been a Visiting Scholar at the University of Extremadura, Cáceres, Spain, in 2014, and the Mississippi State University (MSU), Starkville, MS, USA, in 2016. In the last ten years, he was the Principal Investigator of ten scientific research projects at national and ministerial levels, including projects by the National Natural Science Foundation of China for the terms of 2016–2019, 2018–2020, and 2022–2025 and the National Key Research and Development Program of China for the term of 2021–2025. He has published more than 200 peer-reviewed articles, and there are more than 130 journal articles included in the Science Citation Index (SCI). He has coauthored three academic books, including *Hyperspectral Image Information Extraction*. He obtained 29 National Invention Patents in China. His research focuses on hyperspectral image processing and information extraction.

Dr. Gao was awarded the Outstanding Science and Technology Achievement Prize of the CAS in 2016 and was supported by the China National Science Fund for Excellent Young Scholars in 2017. He received the Second Prize of The State Scientific and Technological Progress Award in 2018 and the recognition of the Best Reviewers of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING in 2015 and the Best Reviewers of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING in 2017.

Jiaxin Li received the B.E. degree from Chongqing University, Chongqing, China, in 2020. He is currently pursuing the Ph.D. degree in cartography and geographic information system with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include multimodal remote sensing data fusion, hyperspectral image processing, and deep learning.

Ke Zheng received the B.S. degree in geographic information system from Shandong Agricultural University, Tai'an, China, in 2012, and the M.S. and Ph.D. degrees in remote sensing from the College of Geosciences and Surveying Engineering, China University of Mining and Technology (Beijing), Beijing, China, in 2016 and 2020, respectively.

He spent two years as a Post-Doctoral Associate with the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Science, Beijing. He is currently an Instructor at the College of Geography and Environment, Liaocheng University, Liaocheng, Shandong, China. His research interests include image processing, machine learning, deep learning, and their applications in Earth vision.

Xiuping Jia (Fellow, IEEE) received the B.Eng. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in January 1982, and the Ph.D. degree in electrical engineering and a Graduate Certificate in Higher Education from The University of New South Wales, Canberra, ACT, Australia, in 1996 and 2005, respectively, via part-time study.

She has had a lifelong academic career in higher education, for which she has continued passion. She is currently an Associate Professor at the School of Engineering and Information Technology, The University of New South Wales. She has published widely addressing various topics, including data correction, feature reduction, and image classification using machine learning techniques. She has coauthored the remote sensing textbook *Remote Sensing Digital Image Analysis* [Springer-Verlag, Third Edition (1999) and Fourth Edition (2006)]. She is the author of *Field Guide to Hyperspectral/Multispectral Image Processing* (SPIE, 2022). These publications are highly cited in the remote sensing and image processing communities with an H-index of 54 and an i-10-index of 189 (Google Scholar). Her research interests include remote sensing, hyperspectral image processing, and spatial data analysis.

Dr. Jia is the Editor-in-Chief of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on May 05,2023 at 02:50:35 UTC from IEEE Xplore. Restrictions apply.