X-Shaped Interactive Autoencoders With Cross-Modality Mutual Learning for Unsupervised Hyperspectral Image Super-Resolution

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

***Abstract*— Hyperspectral image super-resolution (HSI-SR) can compensate for the incompleteness of single-sensor imaging and provide desirable products with both high spatial and spectral resolution. Among them, unmixing-inspired networks have drawn considerable attention due to their straightforward unsupervised paradigm. However, most do not fully capture and utilize the multimodal information due to their limited representation ability of constructed networks, hence leaving large room for further improvement. To this end, we propose an X-shaped interactive autoencoder network with cross-modality mutual learning between hyperspectral and multispectral data, XINet for short, to cope with this problem. Generally, it employs a coupled structure equipped with two autoencoders, aiming at deriving latent abundances and corresponding endmembers from input correspondence. Inside the network, a novel X-shaped interactive architecture is designed by coupling two disjointed U-Nets together via a parameter-shared strategy, which not only enables sufficient information flow between two modalities but also leads to informative spatial-spectral features. Considering the complementarity across each modality, a cross-modality mutual learning module (CMMLM) is constructed to further transfer knowledge from one modality to another, allowing for better utilization of multimodal features. Moreover, a joint self-supervised loss is proposed to effectively optimize our proposed XINet, enabling an unsupervised manner without external triplets supervision. Extensive experiments, including super-resolved results in four datasets, robustness analysis, and extension to other applications, are conducted, and the superiority of our method is demonstrated.**

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

Manuscript received 10 May 2023; revised 26 June 2023; accepted 25 July 2023. Date of publication 31 July 2023; date of current version 10 August 2023. This work was supported by the National Key Research and Development Program of China under Grant 2021YFB3900502. *(Corresponding author: Ke Zheng.)*

Jiaxin Li and Zhi Li are 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; lizhi21@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).

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).

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.2023.3300043

I. Introduction

HYPERSPECTRAL images (HSIs) are widely employed in remote sensing fields due to their powerful ability in object identification [1], [2], [3]. However, HSIs are usually accompanied by low-spatial-resolution due to the imaging principles, which greatly jeopardizes their subsequent application [4], [5]. By contrast, multispectral images (MSIs) with low-spectral-resolution can provide richer spatial details and possess great potential in fine-grained tasks. Therefore, it is possible to yield a high-resolution HSI (HrHSI) product by fusing low-resolution HSI (LrHSI) and corresponding high-resolution MSI (HrMSI) [6]. In this article, we focus on this burgeoning field and call it multispectral-aided HSI super-resolution (MSI-aided HSI-SR) or HSI super-resolution (HSI-SR) for short [7]. Fig. 1(a) shows an illustration of this technique.

In general, mainstream algorithms of HSI-SR are mainly divided into three groups, namely, detail injection-based, optimization-based, and deep-learning (DL)-based methods [8], [9]. Detail injection-based methods originate from the pansharpening for fusing low-resolution MSIs and high-resolution panchromatic (PAN) images. Despite their facilitation, spatial-spectral distortions are nonnegligible in the outcomes. Differently, optimization-based methods treat the HSI-SR task as an ill-posed inverse problem and establish a minimizing function by resorting to hand-crafted priors and sensor-related parameters, i.e., point spread function (PSF) and spectral response function (SRF).

The past decade has witnessed the great potential of DL in different domains, thanks to the availability of large datasets and the growth of computing power [10], [11]. Naturally, DL technology is also successfully applied to the HSI-SR task and achieves impressive progress. Generally, we classify existing DL-based methods into two groups according to their training paradigms, i.e., supervision and un-supervision, and give each of them an illustration in Fig. 1(b) and (c), respectively. Specifically, the former aims to explore latent relationships between inputs and ground-truth from large training triplets, i.e., observed HSI-MSI correspondence and target HrHSI. However, the unavailability of HrHSI in real scenarios compels researchers to synthesize reduced-scale training triplets, in which original data are spatially downsampled in order to regard the observed LrHSI as the ground-truth.

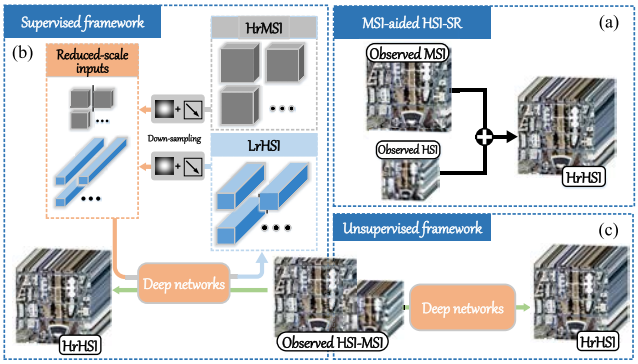

Fig. 1. (a) MSI-aided HSI-SR. (b) and (c) Paradigm under the supervised and unsupervised frameworks, respectively.

However, it is hard for reduced-scale training triplets to fully represent original spatial-spectral information, which may hinder networks from digging out representative features for performance improvement. Besides, the demand for large training triplets in the HSI-SR task makes it inherently impractical in real applications. Hence, the unsupervised paradigm has become a research hotspot in recent years, which depends purely on one full-scale HSI-MSI correspondence and hence mitigates the need for large training sets.

Along this line, unmixing-inspired networks become one of the mainstream approaches to realizing the unsupervised paradigm, where the HSI-SR task is reformulated as the estimation of abundances and endmembers of unknown HrHSI. Among them, uSDN [12], CUCaNet [13], and HyCoNet [14] are three representative works. However, these methods are prone to their limited representation ability due to poorly crafted structures and lack of reasonable constraints, and hence further lead to unnecessary spectral-spatial distortions under the unsupervised framework. Therefore, the above-mentioned issues motivate us to further improve existing approaches.

In this article, we propose an X-shaped interactive autoencoder network with cross-modality mutual learning (XINet), in which two coupled autoencoders serve as the backbone to derive unknown abundances and corresponding endmembers of target HrHSI. In order to enhance the representability of XINet, we construct a novel X-shaped interactive architecture equipped with three primary components. Specifically, the initial feature extraction module (IFEM) is first tailored for HrMSI and LrHSI to emphasize their spatial and spectral features, respectively. Following that, a deep feature interaction module (DFIM) is performed to further capture hierarchical spatial-spectral features while guaranteeing cross-modality interaction between two isolated branches. Finally, the abundance generation module (AGM) is deployed to yield the final abundances with physical constraints. Moreover, a cross-modality mutual learning module (CMMLM) is implemented at different stages of the architecture, aiming to transfer multiscale knowledge from one modality to another for better utilization of multimodal information. By resorting to our proposed joint self-supervised loss, XINet can produce high-fidelity outcomes without the supervision of HrHSI. In summary, the contributions of this article are as follows.

1) A novel X-shaped interactive architecture is designed by coupling two disjointed U-Nets together via a parameter-shared strategy. Inside the architecture, it is further divided into three components, with IFEM in the head, DFIM in the body, and AGM in the tail, which can not only enable adequate information propagation between two independent modalities but also lead to powerful feature representations with multiscales and multidepths.
2) A CMMLM is constructed. Specifically, it is capable of conveying knowledge from one modality to another via an elaborately designed five-stage scheme and hence allows for full utilization of multimodal information for better fusion performance.
3) We construct a joint self-supervised loss that purely depends on observed HSI-MSI correspondence. By considering the spatial property of the under-studying scene, a winsorized pixel-aware total variation (WPATV) is introduced to adaptively characterize the smoothness of abundances. Besides, spatial and spectral reconstruction losses (SPEs) are proposed to separately guarantee the quality in spatial and spectral domains. Finally, with the joint cooperation of three loss terms, spatial details, and spectral fidelity can be highly guaranteed.
4) The SR performance of XINet on four publicly available datasets is qualitatively and quantitatively demonstrated. Moreover, its robustness to the interference of Gaussian noise and variation of the scale factor is confirmed. Beyond that, its potential in other applications, i.e., hypersharpening and classification, is also verified.

The organization of this article is as follows. Section II introduces the development of the MSI-aided HSI-SR task. Sections III and IV elaborate on the idea and structure of XINet. Section V validates the effectiveness of the proposed network for HSI-SR. Finally, Section VI summarizes the article.

II. Related Work

In this section, we group mainstream algorithms into two categories, i.e., traditional and DL-based methods, and give a brief introduction to each of them.

A. Traditional Methods

By dividing the HSI-SR task into multiple parallel pansharpening problems, detail injection-based methods, including component substitution and multiresolution analysis, have been successfully applied in this domain. For example, Gomez et al. [15] utilized specific assignment rules to adapt wavelet-based tools for the HSI-SR task. Along this line, Zhang and He [16] employed 3-D wavelet transform to further enhance spatial information. Later, Chen et al. [17] proposed a unified framework and successfully extend most pansharpening algorithms to HSI-SR.

However, the aforementioned methods usually ignore the underlying relationship between observed HSI-MSI correspondence and unknown HrHSI, hence leading to limited fusion results. By contrast, the HSI-SR task is cast as a minimizing problem under the optimization-based framework,

in which matrix factorization and tensor representation are two prevailing strategies to establish the target model. For example, Kawakami et al. [18] attempted to estimate the basis and corresponding coefficients in a two-stage manner under the framework of matrix factorization. Similarly, Akhtar et al. [19] used the nonparametric Bayesian representation to construct the target image in a four-stage scheme. Considering the property of locally low rank, Veganzones et al. [20] partition the image into multiple patches and establish parallel fusion models for each patch. Fang et al. [21] adopted the super-pixel technique to cluster similar pixels and then estimate corresponding abundances by exploiting their local structures. Liu et al. [22] introduced spatial and spectral matrices to jointly establish a four-step fusion framework. Instead of stretching 3-D data cubes into 2-D matrices, tensor representation is becoming an alternative solution owning to its high-dimensional representation for HrHSI, such as Tucker decomposition, block-term decomposition, and Canonical decomposition. For example, Li et al. [23] estimated the core tensor and three dictionaries to generate the final image under a coupled tensor factorization. Considering the property of nonlocal self-similarities, Dian et al. [24] performed the tensor factorization within multiple clustered groups in a semiblind situation. Xu et al. [25] employed ℓ_1-norm and TV regularization terms to separately characterize the sparsity and smoothness of unknown tensors. Ding et al. [26] formulated the HSI-SR task into a coupled block-term decomposition, in which unmixing priors can be naturally incorporated into latent variables. Xu et al. [27] first constructed nonlocal similar patches and then perform coupled Canonical decomposition to explore their correlations. Latter, Xu et al. [28] introduced the tensor ring to fully characterize their spatial-spectral information and successfully achieve a robust fusion model. He et al. [29] drew on the strengths of both Canonical and Tucker decomposition to further exploit the low-rank property of HrHSI. Xu et al. [30] cluster nonlocal patches to model their spatial-spectral similarities based on the $t-product$.

B. DL-Based Methods

By resorting to powerful network architectures, DL-based approaches are capable of implicitly extracting prior information concealed in large datasets [31], [32]. Specifically, early works attempt to concatenate upsampled LrHSI with HrMSI to form the input of different networks, including 3-D-CNN [33], ResNet-based network [34], and DenseNet-like network [35]. However, these pioneer works not only ignore the diversity across different modalities but also lead to insufficient utilization of original features. Therefore, existing methods tend to employ a multistream structure to comprehensively explore their spatial-spectral information. To achieve this goal, Xu et al. [36] and Han et al. [37] designed a two-branch structure with one branch for multiscale feature extraction and another branch for target reconstruction. Similarly, Zhan et al. [38] and Wu et al. [39] exploited spatial information by resorting to the attention mechanism. Considering the poor generalization of existing networks, Hu et al. [40] constructed a lightweight network by jointly adopting spectral-spatial attention and finally yield competitive fusion outcomes. Later, Ran [41] et al. proposed a general fusion framework with less computational burden, which exhibits excellent performance in various resolution enhancement tasks. Gao et al. [42] proposed two novel loss terms, called self-supervised loss and separable loss, to promote feature discrimination and physical interpretation. To further enhance spatial details and preserve spectral information, Zhu et al. [43] proposed a two-stage fusion framework, in which band correlation is considered to further refine the coarse outcome from the first stage. Recently, Guo et al. [44] proposed a unified fusion framework to jointly realize registration, degradation estimation, and multisatellite fusion, showing superiority in different fusion situations. Considering the limitation of CNN, some works [45], [46] attempt to employ transformer-based architectures to capture the global relationship within features. In addition to these purely DL-based methods, some researchers start to combine optimization- and DL-based methods together by resorting to unfolding or embedding strategy [47], [48], [49].

Unfortunately, the superior performance of the above-mentioned methods is largely driven by reduced-scale training datasets, which renders them inherently impractical in real applications. Considering the straightforward and practical optimization paradigm of the unsupervised framework, researchers start to refocus their efforts on this burgeoning field. Among them, deep image prior is an effective strategy to recover the target image under the guidance of degradation models [50], [51], [52]. For example, Zhang et al. [53] proposed a two-stage fusion framework with one stage for coarse image generation and another stage for quality refinement. Later, Zhang et al. [54] introduced the image-specific statistics into the first stage in order to enrich its representation ability and improve the fusion performance. The approach proposed in this article falls within another direction, namely, associating spectral unmixing with the HSI-SR task, which relies on the assumption that each pixel is the linear combination of several distinct materials. To be specific, the first attempt is made by Qu et al. [12] and achieved a significant improvement compared with traditional methods. Following the same line, Yao [13] et al. and Zheng et al. [14] proposed their CNN-based architectures to execute more sufficient feature learning. Liu et al. [55] embed nonnegative matrix factorization into the autoencoder network to formulate a pixel-level fusion model. However, existing unmixing-inspired networks still have limited representation ability due to the poorly crafted architecture and lack of effective constraints, leaving large room for further development. Therefore, the above-mentioned issues stimulate us to develop a more powerful network, named XINnet, for breaking through the existing bottleneck.

III. Problem Formulation

HSI-SR aims to reconstruct desirable HrHSI $\mathcal{X} \in \mathbb{R}^{H\times W\times C}$ with both high spatial and spectral resolution from observed LrHSI $\mathcal{Y} \in \mathbb{R}^{h\times w\times C}$ and HrMSI $\mathcal{Z} \in \mathbb{R}^{H\times W\times c}$, where H/h, W/w, and C/c denote the height, width, and channel size, respectively. In general, $w \ll W, h \ll H$, and $c \ll C$

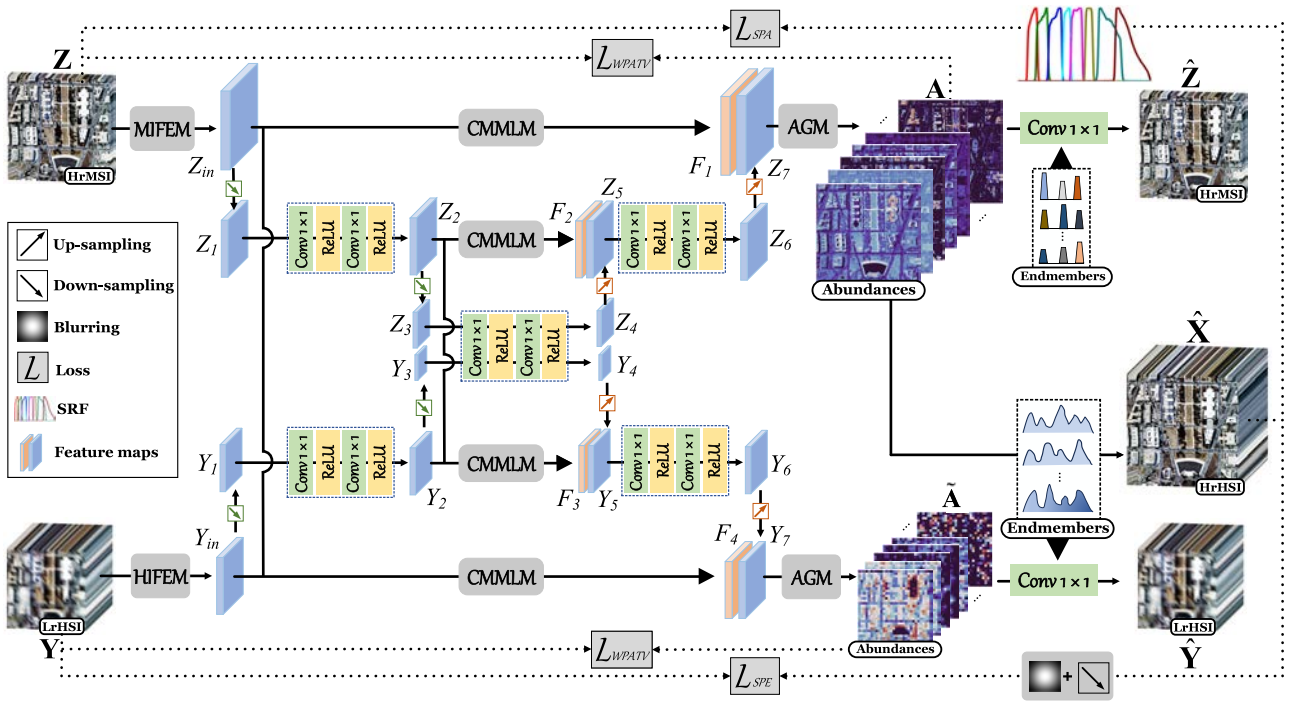

Fig. 2. Overall structures of proposed XINet, where the left depicts the corresponding legend.

are satisfied due to the unavoidable imaging constraint. For simplicity, 3-D cubes are unfolded into matrix forms, i.e, $\mathbf{X} \in \mathbb{R}^{HW\times C}$, $\mathbf{Y} \in \mathbb{R}^{hw\times C}$, and $\mathbf{Z} \in \mathbb{R}^{HW\times c}$, with each column representing one spectral band.

Given the input $\mathbf{Y}$ and $\mathbf{Z}$, their correlation with $\mathbf{X}$ can be established in terms of the degradation models, i.e.,

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

where $\mathbf{P} \in \mathbb{R}^{wh\times WH}$ denotes PSF blurring and spatial downsampling operator and $\mathbf{S} \in \mathbb{R}^{C\times c}$ denotes the spectral degradation operator. $\mathbf{N}_y$ and $\mathbf{N}_z$ include modeling errors and sensor noises.

By virtue of the linear spectral mixing model, each pixel in HrHSI can be linearly expressed by several distinct signatures (endmember) and their corresponding coefficients (abundances), which can be written as

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

Here, $\mathbf{E} \in \mathbb{R}^{P\times C}$ and $\mathbf{A} \in \mathbb{R}^{WH\times P}$ represent endmember matrix with a number of P spectral signatures and corresponding abundance matrix, respectively. $\mathbf{N}$ is the error.

Combining linear spectral mixing model (2) and degradation models (1) leads to

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

Here, $\widetilde{\mathbf{A}} \in \mathbb{R}^{wh\times P}$ and $\widetilde{\mathbf{E}} \in \mathbb{R}^{P\times c}$ can be treated as degraded $\mathbf{A}$ and $\mathbf{E}$ in the spatial and spectral domain, respectively. Therefore, recovering HrHSI is basically equivalent to inferring $(\mathbf{A}, \mathbf{E})$ from input $(\mathbf{Y}, \mathbf{Z})$ by solving the following problem:

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

where $\|\cdot\|_F$ and $\mathbf{1}$ symbolize the Frobenius norm and all-one vector, respectively. Based on this straightforward idea, we construct XINet architecture step-by-step to derive unknown endmember matrix $\mathbf{E}$ and corresponding abundance matrix $\mathbf{A}$ under the unsupervised framework.

IV. Methodology

A. Overall Network Architecture

As shown in Fig. 2, our proposed XINet can be divided into two phases, with one phase for abundance estimation and another phase for joint modality reconstruction and endmember inferring. Precisely, the first phase can be expressed as

$$\mathbf{A}, \widetilde{\mathbf{A}} = X_{\text{en}}(\mathbf{Z}, \mathbf{Y}; \mathbf{W}_{\text{en}}) \tag{5}$$

where $X_{\text{en}}(\cdot)$ denotes the X-shape interactive encoder aiming at generating unknown $(\mathbf{A}, \widetilde{\mathbf{A}})$ from the input $(\mathbf{Z}, \mathbf{Y})$ and $\mathbf{W}_{\text{en}}$ is corresponding parameters to be learn.

To recover input $(\mathbf{Z}, \mathbf{Y})$ from $(\mathbf{A}, \widetilde{\mathbf{A}})$ under the spectral unmixing framework, 1×1 linear convolution without any bias is employed in the second phase to serve as the decoder, which can be formulated as

$$\widehat{\mathbf{Z}} = f_{\text{de}}\left(\mathbf{A}; \mathbf{W}_{f,\text{de}}\right) \tag{6}$$

$$\widehat{\mathbf{Y}} = g_{\text{de}}\left(\widetilde{\mathbf{A}}; \mathbf{W}_{g,\text{de}}\right) \tag{7}$$

where $f_{\text{de}}(\cdot)$ and $g_{\text{de}}(\cdot)$ denote the decoders, where trainable parameters $\mathbf{W}_{f,\text{de}}$ and $\mathbf{W}_{g,\text{de}}$ in each decoder can be naturally interpreted as $\widetilde{\mathbf{E}}$ and $\mathbf{E}$, respectively.

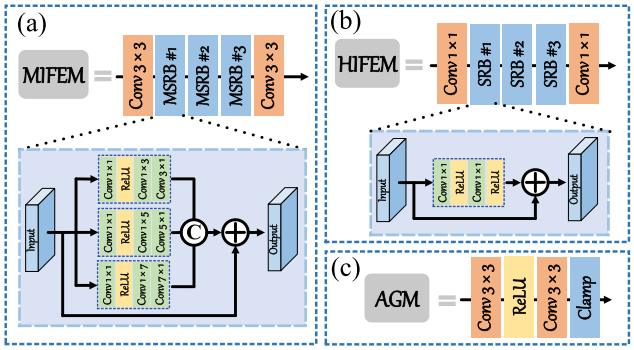

Fig. 3. Detailed structures of (a) MIFEM, (b) HIFEM, and (c) AGM.

Under the guidance of our proposed self-supervised loss, unknown $\mathbf{E}$ and $\mathbf{A}$ can be obtained by purely utilizing observed HSI-MSI correspondence. After that, HrHSI $\widehat{\mathbf{X}}$ is obtained according to (2)

$$\widehat{\mathbf{X}} = g_{\text{de}}\left(\mathbf{A}; \mathbf{W}_{g,\text{de}}\right). \tag{8}$$

B. X-Shaped Interactive Encoder

Considering the weak representation ability of previous networks, we designed an X-shaped interactive architecture as the backbone of the encoder part in XINet, in which two disjointed U-Nets are coupled together via a parameter-shared strategy. Despite the great success achieved by U-Net architectures, its potential in HSI-SR still remains to be explored. Therefore, three cascaded components, namely, IFEM, DFIM, and AGM, are elaborately constructed by fully considering the characteristic of input data and the issues faced by existing HSI-SR networks.

1) Initial Feature Extraction Module: Considering the discrepancy existing in HSI-MSI correspondence, two modality-oriented modules are designed to guarantee the sufficiency of initially extracted features, with one module named MIFEM extracting the spatial details of HrMSI and another module named HIFEM exploiting spectral information of LrHSI. The details are shown in Fig. 3(a) and (b), respectively.

Specifically, two 3×3 layers are separately deployed at the head and tail of MIFEM for adapting the channel dimension. Between two layers, three multiscale residual blocks (MSRBs) are cascaded to further dig out their multiscale spatial information. Precisely, MSRB is divided into three branches, each of which attempts to extract corresponding features with specific receptive-fields, i.e., 3×3, 5×5, and 7×7. However, large convolutional kernels are prone to increase the computational burden of our network. Thus, we factorize symmetric $n \times n$ convolution into asymmetric $1 \times n$ and $n \times 1$ convolution, e.g., 3×3 kernel into cascaded 1×3 and 3×1 kernels, which can not only reduce the network parameter but also lead to more representative features [56], [57]. After that, feature maps at three scales are aggregated to form a more compact representation and then added to the input through the residual learning strategy,

Beyond that, a similar layout is applied for HIFEM to exploit the spectral features of LrHSI. The major difference lies in the kernel size, in which 1×1 point-wise convolution replaces original multiscale convolutions in MIFEM. In each SRB, we directly adopt a one-stream residual structure, which enables HIFEM to emphasize the spectral domain without involving neighboring pixels. The formula of IFEM can be summarized as

$$\begin{aligned} Z_{\text{in}} &= f_{\text{MIFEM}}(\mathbf{Z}) \\ Y_{\text{in}} &= f_{\text{HIFEM}}(\mathbf{Y}) \end{aligned} \tag{9}$$

where Z_{in} and Y_{in} denote the outputs of $f_{\text{MIFEM}}(\cdot)$ and $f_{\text{HIFEM}}(\cdot)$, respectively.

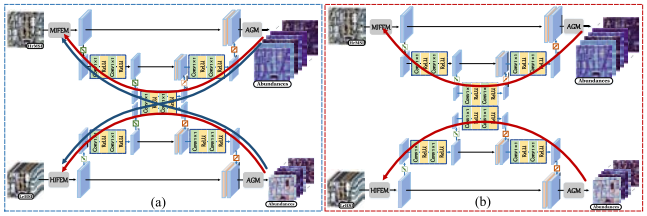

Fig. 4. Gradient flow (a) with and (b) without the bridging path.

2) Deep Feature Interaction Module With Mutual Learning: DFIM with mutual learning is proposed to enhance the interaction and utilization of multimodal information, which consists of four primary parts, namely, the downsampling path, bridging path, CMMLM, and upsampling path.

3) Downsampling Path: The first part attempts to gradually narrow down spatial dimensions of extracted initial features and expands their receptive-fields to capture global information. Between two downsampling operators, a feature extraction block with two 1×1 layers and two ReLU nonlinearity is added to increase their channel sizes while keeping their spatial resolution unchanged. The first part for two branches can be expressed as

$$\begin{cases} Z_1 = Z_{\text{in}} \downarrow \\ Z_2 = f_1(Z_1) \\ Z_3 = Z_2 \downarrow \end{cases} \text{and} \begin{cases} Y_1 = Y_{\text{in}} \downarrow \\ Y_2 = f_2(Y_1) \\ Y_3 = Y_2 \downarrow \end{cases} \tag{10}$$

respectively. Here, $f_1(\cdot)$ and $f_2(\cdot)$ denote the feature extraction block in two branches and $\downarrow$ represents the downsampling operator with a reduction factor of four by employing bilinear interpolation.

4) Bridging Path: Considering the importance of multimodal interaction, the bridging path is constructed to couple two disjointed U-Nets together, hence leading to a novel X-shaped architecture. The advantage of such a design is briefly shown in Fig. 4, with each curve representing the direction of back-propagation. As can be seen from the blue curves, gradient information from one branch can be naturally transmitted into another through this parameter-shared bridging path. Otherwise, information flow will be inevitably blocked, leading to insufficient interaction between two modalities. This process can be defined as

$$\begin{aligned} Z_4 &= f_b(Z_3) \\ Y_4 &= f_b(Y_3) \end{aligned} \tag{11}$$

where $f_b(\cdot)$ denotes the parameter-shared bridging path which contains two 1×1 layers and two ReLU activations.

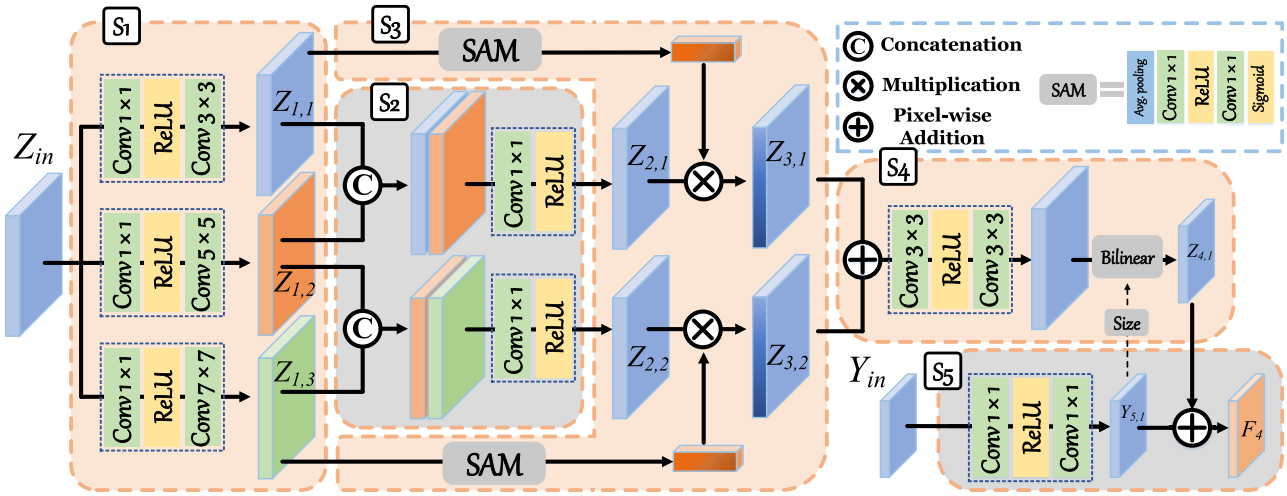

Fig. 5. Overall structures of proposed CMMLM.

5) Cross-Modality Mutual Learning Module: The most significant part of U-Nets is the skip connection which is in charge of concatenating the feature maps of each stage in the downsampling path to the corresponding stages in the upsampling path, enabling the network to combine the low- and high-level features effectively. To further utilize the cross-modality complementary information, CMMLM is designed to replace the original skip connection, which encourages input modalities to learn from each other for the full exploration of cross-modal features. The detailed structure is shown in Fig. 5.

The whole process of CMMLM can be divided into five stages, namely, multiscale feature extraction, feature fusion, feature self-recalibration, feature adjustment, and feature injection. We take features F_4 generated from Z_{in} and Y_{in} as an example to demonstrate the working mechanism of this tailor-designed module. Specifically, considering the potential of multiscale representations in tackling complicated scenes with diverse structures, we intentionally employ a three-stream layout with three receptive-fields, i.e., 3×3, 5×5, and 7×7, in the first stage, which can be described as

$$Z_{1,1}, Z_{1,2}, Z_{1,3} = f_{S1}(Z_{\text{in}}) \tag{12}$$

where $f_{S1}(\cdot)$ denotes the first stage to derive three multiscale features $Z_{1,1}$, $Z_{1,2}$, and $Z_{1,3}$, with each branch containing two convolutional layers and a ReLU activation unit. However, the representative ability of outcome features is still limited without considering their correlation. Therefore, feature maps from stage one are concatenated in pairs and fed into the second stage for a deep feature fusion, which is given by

$$Z_{2,1}, Z_{2,2} = f_{S2}\big(\text{cat}\big(Z_{1,1}, Z_{1,2}\big), \text{cat}\big(Z_{1,2}, Z_{1,3}\big)\big) \tag{13}$$

where $f_{S2}(\cdot)$ represents the feature fusion stage consisting of two branches, each of which has a 1×1 layer and a ReLU unit, and cat is the concatenation operator. Nevertheless, simply fusing multiscale features is prone to produce redundant information and resulting in the loss of important representations. Considering this deficiency, attention-based weight coefficients derived from the outputs of stage one is utilized to recalibrate fused features, enabling the network to suppress trivial information and highlight salient ones, which can be expressed as

$$\begin{cases} Z_{3,1} = Z_{2,1} * \text{SAM}\big(Z_{1,1}\big) \\ Z_{3,2} = Z_{2,2} * \text{SAM}\big(Z_{1,3}\big) \end{cases} \tag{14}$$

where SAM stands for the spectral attention module which aims to calculate weight coefficients for self-recalibration and $*$ is the channel-wise multiplication. Concretely, each SAM first employs an average-pooling to shrink input features into a vector, and then two 1×1 layers are utilized to compress and restore its channel size, respectively. Finally, One Sigmoid unit produces the final weight coefficients. Through the above-mentioned three stages, abundant multiscale features can be achieved to a great extent. In order to successfully inject them into another modality, scale adjustment is required to eliminate their spatial-spectral discrepancy, i.e.,

$$Z_{4,1} = \big(f_{S4}\big(Z_{3,1} + Z_{3,2}\big)\big) \Updownarrow \tag{15}$$

where $f_{S4}(\cdot)$ is to resize the spectral dimension using two 3×3 convolutional layers and a ReLU activation unit, and $\Updownarrow$ denotes the spatial alignment resorting to bilinear interpolation. The last stage is to transmit $Z_{4,1}$ from the HrMSI branch to complement the feature learning of the LrHSI branch for better utilization of multimodal information, which is calculated as

$$F_4 = f_{S5}(Y_{\text{in}}) + Z_{4,1} \tag{16}$$

where $f_{S5}(\cdot)$ aims to extract features of Y_{in} before receiving complementary information from $Z_{4,1}$. The whole process of CMMLM for two modalities can be summarized as

$$\begin{cases} F_4 = \textbf{CMMLM}(Z_{\text{in}}, Y_{\text{in}}) \\ F_3 = \textbf{CMMLM}(Z_2, Y_2) \end{cases} \tag{17}$$

and

$$\begin{cases} F_1 = \textbf{CMMLM}(Y_{\text{in}}, Z_{\text{in}}) \\ F_2 = \textbf{CMMLM}(Y_2, Z_2) \end{cases} \tag{18}$$

respectively.

6) Upsampling Path: The last part of DFIM is the upsampling path that gradually restores their spatial resolution while reducing the spectral dimensions under the guidance of CMMLM. Between two upsampling operators, the feature extraction block with the same structure as that in the downsampling path is employed to realize the contraction of spectral dimension. Thus, we formulate this process as

$$\begin{cases} Z_5 = Z_4 \uparrow \\ Z_6 = f_3(\text{cat}(F_2, Z_5)) \\ Z_7 = Z_6 \uparrow \end{cases} \text{and} \begin{cases} Y_5 = Y_4 \uparrow \\ Y_6 = f_4(\text{cat}(F_3, Y_5)) \\ Y_7 = Y_6 \uparrow \end{cases} \quad (19)$$

where $f_3(\cdot)$ and $f_4(\cdot)$ are the feature extraction blocks and $\uparrow$ represents the upsampling operator with a factor of four by employing bilinear interpolation.

7) Abundance Generation Module: Considering the fact that multimodal information has been fully exploited and utilized through the first two components, a simple structure is preferable to construct AGM. As shown in Fig. 3(c), a 3×3 layer is first employed to compress the channel size of input features, followed by a ReLU unit. After that, the second 3×3 layer aims to map the channel size into P. Considering the physical properties of abundances, a Clamp function is deployed at the bottom of AGM to guarantee their nonnegativity constraint. In this way, each channel of feature outputs can be interpreted as a fractional abundance map of one spectral signature. This process can be expressed as

$$\begin{cases} \mathbf{A} = f_{\text{AGM}}(\text{cat}(F_1, Z_7)) \\ \widetilde{\mathbf{A}} = f_{\text{AGM}}(\text{cat}(F_4, Y_7)) \end{cases} \quad (20)$$

where $f_{\text{AGM}}(\cdot)$ denotes the AGM. $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ are corresponding abundances of LrHSI and HrMSI, respectively.

C. Loss Functions

The loss function of XINet is divided into two parts, which can be defined as follows:

$$\mathcal{L}_{\text{overall}} = \underbrace{\mathcal{L}_{\text{basic}}}_{\text{basic loss}} + \underbrace{\lambda_{\text{joint}}\mathcal{L}_{\text{joint}}}_{\text{joint self-supervised loss}} \quad (21)$$

where $\mathcal{L}_{\text{basic}}$ and $\mathcal{L}_{\text{joint}}$ represent a basic loss and a joint self-supervised loss, respectively, and λ_{joint} is the weight coefficient. Considering the fact that $L2$ loss tends to overly smooth the reconstruction image, $L1$ loss is chosen as our criterion for a better recovery of high-frequency details.

1) Basic Loss: The basic loss includes two terms, namely, reconstruction loss $\mathcal{L}_{\text{rec}}$ and abundance sum-to-one loss (ASC) $\mathcal{L}_{\text{ASC}}$, which is calculated as

$$\mathcal{L}_{\text{basic}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{ASC}}\mathcal{L}_{\text{ASC}} \quad (22)$$

where λ_{rec} and λ_{ASC} are the weight factors of two loss terms whose detailed forms are shown as follows:

$$\begin{cases} \mathcal{L}_{\text{rec}} = \left\|\mathbf{Z} - \widehat{\mathbf{Z}}\right\|_1 + \left\|\mathbf{Y} - \widehat{\mathbf{Y}}\right\|_1 \\ \mathcal{L}_{\text{ASC}} = \|\mathbf{1}_{WH} - \mathbf{A}\mathbf{1}_P\|_1 + \left\|\mathbf{1}_{wh} - \widetilde{\mathbf{A}}\mathbf{1}_P\right\|_1. \end{cases} \quad (23)$$

2) Joint Self-Supervised Loss: Apart from the basic loss, we further construct a joint self-supervised loss that purely utilizes observed HSI-MSI correspondence as signals to optimize our proposed XINet. Overall, it contains three terms, including a WPATV, spatial reconstruction loss (SPA), and SPE, which is defined as

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{WPATV}} + \mathcal{L}_{\text{SPA}} + \mathcal{L}_{\text{SPE}}. \quad (24)$$

3) Winsorized Pixel-Aware Total Variation: Considering that adjacent pixels in hyperspectral scenes tend to represent similar materials, we employ a total variation (TV) regularizer to guarantee this important feature and encourage piecewise smooth of unknown abundances, which is modeled as

$$\begin{cases} \text{HTV}\left(\widetilde{\mathbf{A}}\right) = \sum_{j=1}^{P} \text{STV}\left(\widetilde{\mathbf{A}}_j\right) \\ \text{HTV}(\mathbf{A}) = \sum_{j=1}^{P} \text{STV}\left(\mathbf{A}_j\right) \end{cases} \quad (25)$$

where $\text{HTV}(\cdot)$ is to calculate TV regularization of multiband images by adding their gradient in each band together, where the single-band version $\text{STV}(\cdot)$ is defined as

$$\text{STV}(X) = \|\nabla_h X\|_1 + \|\nabla_v X\|_1 \quad (26)$$

where ∇_h and ∇_v are horizontal and vertical first-order differences, respectively. However, HSIs usually possess complicated structures with different spatial properties, such as homogeneous and edge areas. If the regularization strength is identically and equally imposed on different regions without considering their diversity, the regularization constraint will be inevitably degraded, which may lead to the over-smoothing of edge areas or weak suppression in homogeneous regions. Thus, it is necessary to adaptively adjust the smooth strength according to spatial structure in each pixel location.

How to determine pixel-aware weights is the key to our proposed WPATV. Fortunately, gradient information is an advisable indicator to represent the spatial structure of each pixel. Considering the availability of observed images, the weight coefficients can be directly obtained from input HSI-MSI correspondence, and the process can be formulated as

$$\begin{cases} G_{\widetilde{\mathbf{A}}} = \sum_{j=1}^{C} \text{STV}\left(\mathbf{Y}_j\right) \\ G_{\mathbf{A}} = \sum_{j=1}^{c} \text{STV}\left(\mathbf{Z}_j\right). \end{cases} \quad (27)$$

Fig. 6(a) shows the gradient distribution of $G_{\mathbf{A}}$ in the Chikusei dataset. We can find that most pixels are distributed between 0 and 1. However, there also exist some outliers with large values ranging from 2 to 5, if the pixel-aware weight is determined by directly normalizing $G_{\widetilde{\mathbf{A}}}$ and $G_{\mathbf{A}}$ without considering these outliers, the discrimination between edges and homogeneous areas will be severely weakened. Therefore, winsorization is introduced to mitigate the negative effect, which can be expressed as

$$G_{\widetilde{\mathbf{A}}(m,n)} = \begin{cases} G_{\widetilde{\mathbf{A}}(m,n)}, & G_{\widetilde{\mathbf{A}}(m,n)} \le T_{\widetilde{A}} \\ T_{\widetilde{A}}, & G_{\widetilde{\mathbf{A}}(m,n)} > T_{\widetilde{A}} \end{cases} \quad (28)$$

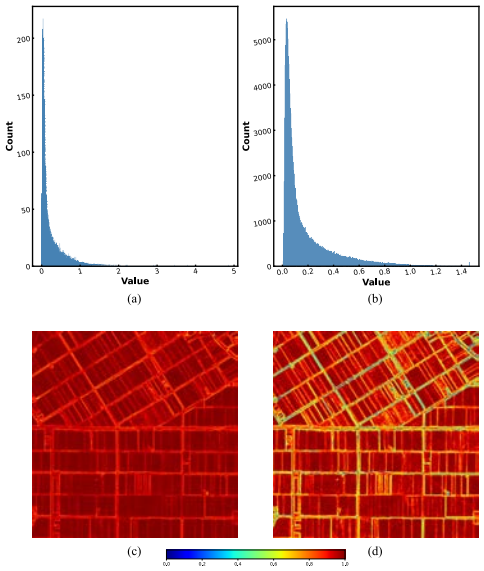

Fig. 6. Statistical results and corresponding pixel-aware weight maps before and after the winsorization operator. (a) and (b) Statistical results before and after the winsorization operator, respectively. (c) and (d) Pixel-aware weight maps before and after the winsorization operator, respectively.

and

$$G_{\mathbf{A}(m,n)} = \begin{cases} G_{\mathbf{A}(m,n)}, & G_{\mathbf{A}(m,n)} \leq T_A \\ T_A, & G_{\mathbf{A}(m,n)} > T_A \end{cases} \tag{29}$$

where $G_{\mathbf{A}(m,n)}$ and $G_{\mathbf{A}(m,n)}$ are the gradient value at the location of (m, n). $T_{\widetilde{A}}$ and T_A represent the top 99.95% value in $G_{\widetilde{\mathbf{A}}}$ and $G_{\mathbf{A}}$, respectively. Fig. 6(b) shows the statistical result after winsorization. Based on the winsorized gradient, the pixel-aware weight is constructed as follows:

$$\begin{cases} W_{\widetilde{\mathbf{A}}} = 1 - \dfrac{G_{\widetilde{\mathbf{A}}} - \mathrm{Min}\left(G_{\widetilde{\mathbf{A}}}\right)}{\mathrm{Max}\left(G_{\widetilde{\mathbf{A}}}\right) - \mathrm{Min}\left(G_{\widetilde{\mathbf{A}}}\right)} \\ W_{\mathbf{A}} = 1 - \dfrac{G_{\mathbf{A}} - \mathrm{Min}(G_{\mathbf{A}})}{\mathrm{Max}(G_{\mathbf{A}}) - \mathrm{Min}(G_{\mathbf{A}})} \end{cases} \tag{30}$$

where $\mathrm{Max}(\cdot)$ and $\mathrm{Min}(\cdot)$ denote the maximum and minimum value, respectively. Fig. 6(c) and (d) exhibits the weight maps before and after winsorization. It can be observed in Fig. 6(c) that most weights are close to 1, with less contrast between fields and their boundaries. However, the discrimination between the two of them can be noticeably magnified in Fig. 6(d) once eliminating these outliers. Finally, introducing (30) into (25) leads to our proposed WPATV

$$\mathcal{L}_{\mathrm{WPATV}} = \mathrm{HTV}\left(\widetilde{\mathbf{A}}\right) \otimes W_{\widetilde{\mathbf{A}}} + \mathrm{HTV}(\mathbf{A}) \otimes W_{\mathbf{A}} \tag{31}$$

where $\otimes$ is the pixel-wise multiplication.

4) Spatial-SPE: The objective of spatial-SPE is to guarantee the spatial details and spectral fidelity of the reconstructed image, respectively. According to (1), the spatially and spectrally degraded version of output should be close to input HSI-MSI correspondence, respectively. Thus, the loss function can be defined as follows:

$$\begin{cases} \mathcal{L}_{\mathrm{SPE}} = \left\| \left(f_{\mathrm{spa}}\left(\widehat{\mathbf{X}}; W_{\mathrm{PSF}}\right) - \mathbf{Y} \right) \right\|_1 \\ \mathcal{L}_{\mathrm{SPA}} = \left\| \left(f_{\mathrm{spe}}\left(\widehat{\mathbf{X}}; W_{\mathrm{SRF}}\right) - \mathbf{Z} \right) \right\|_1 \end{cases} \tag{32}$$

TABLE I
SIMULATION RESULTS OF FOUR DATASETS

	PRISMA	Washington DC Mall	TianGong-1	Chikusei
Size of HrHSI	$400 \times 400 \times 63$	$300 \times 300 \times 191$	$240 \times 240 \times 54$	$400 \times 400 \times 110$
Spectral range of HrHSI (nm)	434 - 2456	401 – 2473	413 - 887	363 - 1018
Scale factor	8	10	12	16
Size of LrHSI	$50 \times 50 \times 63$	$30 \times 30 \times 191$	$20 \times 20 \times 54$	$25 \times 25 \times 110$
Size of HrMSI	$400 \times 400 \times 8$	$300 \times 300 \times 8$	$240 \times 240 \times 8$	$400 \times 400 \times 8$

where $f_{\mathrm{spa}}(\cdot)$ attempts to realize spatial degradation through a group-wise convolution operation with PSF blurring kernel W_{PSF} and $f_{\mathrm{spe}}(\cdot)$ represents the spectral pooling by utilizing SRF W_{SRF}. It is worth mentioning that sensor-related parameters W_{PSF} and W_{SRF} can be given as prior information or estimated from input HSI-MSI correspondence in advance.

V. EXPERIMENTS AND RESULTS

In this section, we first introduce datasets and corresponding setups in our experiments. Then, parameters discussion and ablation study are conducted to analyze the influence of different components. Besides, we compare our method with 11 state-of-the-art algorithms in four widely-used datasets. Following that, robust analysis is established to verify the superiority of XINet in different situations. Beyond that, its potential in other applications, i.e., hypersharpening and classification, is also assessed.

A. Datasets and Setups

1) Datasets: Four publicly available datasets are chosen to verify the performance of different methods, including the PRISMA dataset, Washington DC Mall dataset, TianGong-1 dataset, and Chikusei dataset. Specifically, the PRISMA dataset acquired by the PRISMA mission totally contains four co-registered hyperspectral and PAN images, in which the second HSI is chosen for comparison. Washington DC Mall dataset is obtained by the Hydice sensor over the Mall in Washington, with a size of $1280 \times 307 \times 210$. The TianGong-1 dataset is collected for hyperspectral scene classification and includes a number of 204 images ranging from visible to short-wave infrared. One image titled "city-014-VNI-2013041514" is selected for comparison. The Chikusei dataset is captured by the Headwall Hyperspec-VNIR-C imaging sensor over the Chikusei area, which includes 2517×2335 pixels and 128 spectral bands.

For simplicity, a subimage is cropped from original datasets as the reference after discarding noisy bands. In order to evaluate their performance, simulated experiments are conducted according to the Wald protocol [58]. Precisely, Gaussian blurring and the downsampling operator with a specific scale factor are jointly utilized to generate LrHSI. Next, HrMSI is obtained through spectral degradation using the multispectral response in WorldView 2. The simulation results of four datasets are summarized in Table I.

2) Compared Methods and Evaluation Metrics: Our proposed XINet is compared with 11 representative methods, including three matrix-based methods, i.e., G-SOMP+ [59],

TABLE II
QUANTITATIVE METRICS WITH DIFFERENT NUMBER OF ENDMEMBERS P IN TIANGONG-1 DATASET

P	SAM	PSNR	ERGAS	SSIM	UIQI
10	1.3030	41.5836	0.2607	0.9834	0.9992
30	1.1909	43.5763	0.2217	0.9869	0.9993
50	1.1282	43.7325	0.2127	0.9887	0.9995
70	1.0703	44.6287	0.1915	0.9895	0.9996
90	1.0433	44.7334	0.1932	0.9896	0.9996
110	1.0683	44.5102	0.1936	0.9898	0.9996
130	1.0084	45.3571	0.1749	0.9900	0.9996
150	1.0390	44.8977	0.1879	0.9893	0.9996
170	1.0095	45.3975	0.1778	0.9903	0.9996
190	1.0305	45.0398	0.1856	0.9902	0.9996

CSU [60], and CNMF [61], three tensor-based methods, i.e., STEREO [62], CSTF [23], and SCOTT [63], and five DL-based methods, i.e., DBSR [54], HyCoNet [14], CUCaNet [13], MIAE [55], and SURE [64]. To quantitatively verify their performance, we select five evaluation metrics, namely, peak signal-to-noise ratio (PSNR) [65], spectral angle mapper (SAM) [66], relative dimensionless global error in synthesis (ERGAS) [67], structure similarity (SSIM), and universal image quality index (UIQI) [68], [69]. Beyond that, three kinds of visual forms are provided to aid the qualitative assessment, i.e., SAM heatmap, mean relative absolute error (MRAE) heatmap and the residual heatmap.

3) Implementation Details: Our network is implemented in PyTorch framework with the Adam optimizer [70]. The learning rate starts at 0.003 for the first 9000 epochs and then linearly decreases to zero in the second 9000 epochs. As for the weight factors in (21) and (22), λ_{joint}, λ_{rec}, and λ_{ASC} are set as 10, 100, and 0.01, respectively.

B. Parameters Discussion

The number of endmember P plays an important role in our proposed XINet, influencing the ability to represent complex scenes with subpixel phenomenon. In order to evaluate its impact on the fusion performance, we set P from 10 to 190 with an interval of 20. Table II shows the quantitative results in the TianGong-1 dataset. When P is small, the performance is not satisfactory. With the continuous rise of P, all indicators display a trend of improvement with little fluctuation. Arriving at 130, all metrics remain relatively stable when P keeps increasing. Considering the fact that the increase of P leads to extra parameter stress, we set P to 130 for the rest of the experiments. Actually, complex imaging environments and the intrinsic property of objects are prone to cause nonlinear mixing and introduce the so-called spectral variability phenomenon [71], [72]. In order to fully account for these spectral signatures and better represent under-studying scenes, P is set larger than the number of distinct materials.

C. Ablation Study

In this section, the influence of different components and loss functions is evaluated in terms of fusion performance.

1) Influence of Components: Several important modules or parts are designed in our proposed XINet for enhancing the reconstruction results. To assess their individual performance, we conduct an ablation study in the TianGong-1 dataset. Before beginning our discussion, the detailed structure of different variants in Table III needs introducing. Specifically, (a) is the first variant by replacing IFEM with plain 1×1 convolution, and (b) splits the X-shaped architecture into two isolated U-Nets, as shown in Fig. 4(b). CMMLM⌊ and CMMLM⌈ represent (17) and (18), respectively, which jointly constitute our proposed CMMLM.

From (a) to (e), all variants with incomplete configurations yield relatively poor performance compared with (f), which convincingly confirms the effectiveness of each component. Specifically, when IFEM is replaced by 1×1 layer, all metrics are degraded, which proves that the initially extracted features from IFEM can provide greater potential for subsequent reconstruction. Similarly, the absence of the bridging path also results in poor fusion quality, which indicates the effectiveness of the parameter-shared strategy and the importance of multimodal interaction. Besides, we further conduct an experiment to verify the strength of mutual learning. Comparing (c)/(d) with (e), it can be demonstrated that cross-modality information is helpful to enhance feature learning of another modality and further boosts fusion performance. As expected, when two modalities can simultaneously receive information from each other, the performance is further improved in (f).

2) Influence of Loss Functions: A joint self-supervised loss function is proposed, which enables our method to purely utilize the input HSI-MSI correspondence for optimization. In order to evaluate their influence, we conduct an ablation study in TianGong-1 dataset by gradually adding each loss term, and the results are shown in Table IV.

Specifically, $\mathcal{L}_{\text{basic}}$ in the experiment of (a) serves as the baseline for comparison. When adding $\mathcal{L}_{\text{SPA}}$ and $\mathcal{L}_{\text{SPE}}$, it is clearly shown that both spatial and spectral quality is significantly improved compared with (a), in which $\mathcal{L}_{\text{SPE}}$ achieves greater performance gain. As expected, the joint cooperation of $\mathcal{L}_{\text{SPA}}$ and $\mathcal{L}_{\text{SPE}}$ in (d) achieve more significant improvement. Besides, we also evaluate the proposed $\mathcal{L}_{\text{WPATV}}$ by comparing it with normal TV regularization. As can be seen from (d) to (f), the TV constraint is helpful to enhance fusion performance, but the improvement yielded by (e) is limited. Differently, considering the spatial structure of each pixel, $\mathcal{L}_{\text{WPATV}}$ achieves more significant enhancement in all metrics. To give a more straightforward illustration, the results of (d) and (f) in the Chikusei dataset are shown in Fig. 7. It can be clearly seen that the error in the field and their boundaries are effectively suppressed under the guidance of the adaptive regularization strength from $\mathcal{L}_{\text{WPATV}}$.

D. Comparison With State-of-the-Arts

To fully compare XINet with the other 11 state-of-the-arts, the visual and quantitative results of four datasets are shown

TABLE III
ABLATION STUDY OF DIFFERENT COMPONENTS IN TIANGONG-1 DATASET

No.	Configuration				Metric		
	IFEM	Bridging path	CMMLM↓	CMMLM↑	SAM	PSNR	ERGAS
(a)	☒	✓	✓	✓	1.0297	45.0433	0.1783
(b)	✓	☒	✓	✓	1.0373	44.9415	0.1798
(c)	✓	✓	☒	✓	1.0430	44.5213	0.1852
(d)	✓	✓	✓	☒	1.0473	44.6693	0.1836
(e)	✓	✓	☒	☒	1.0487	44.0731	0.1913
(f)	✓	✓	✓	✓	1.0084	45.3571	0.1749

TABLE IV
ABLATION STUDY OF DIFFERENT LOSS FUNCTIONS IN TIANGONG-1

No.	Loss function	SAM	PSNR	ERGAS
(a)	$\mathcal{L}_{basic}$	16.0276	20.4344	3.0562
(b)	$\mathcal{L}_{basic} + \mathcal{L}_{SPA}$	5.8044	32.9278	1.0363
(c)	$\mathcal{L}_{basic} + \mathcal{L}_{SPE}$	1.2207	41.0950	0.2535
(d)	$\mathcal{L}_{basic} + \mathcal{L}_{SPA} + \mathcal{L}_{SPE}$	1.0415	44.5085	0.1849
(e)	$\mathcal{L}_{basic} + \mathcal{L}_{SPA} + \mathcal{L}_{SPE} + \mathcal{L}_{TV}$	1.0327	44.6745	0.1831
(f)	$\mathcal{L}_{basic} + \mathcal{L}_{joint}$	1.0084	45.3571	0.1749

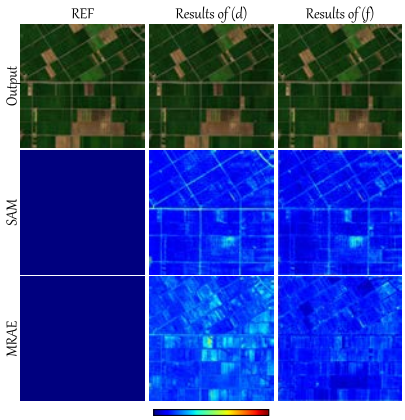

Fig. 7. Fusion results under the configuration of (d) and (f) in the Chikusei dataset.

TABLE V
QUANTITATIVE PERFORMANCE IN PRISMA DATASET

Method	SAM	PSNR	ERGAS	SSIM	UIQI
G-SOMP+	2.9564	36.3497	1.4414	0.9497	0.9908
CSU	2.8600	35.2396	1.5030	0.9042	0.9919
CNMF	2.7293	36.7590	1.3150	0.9462	0.9933
STEREO	3.3299	38.2371	1.4668	0.9335	0.9909
CSTF	2.6449	37.7894	1.2380	0.9461	0.9944
SCOTT	3.4780	38.3559	1.6023	0.9456	0.9871
DBSR	2.7557	39.1256	1.4824	0.9457	0.9887
HyCoNet	2.7243	38.2825	1.2500	0.9628	0.9936
CUCaNet	2.8682	36.2731	1.5242	0.9420	0.9924
MIAE	2.5901	40.1573	1.4129	0.9596	0.9893
SURE	2.3041	40.2827	1.1153	0.9574	0.9958
Ours	**2.1866**	**42.4385**	**1.1113**	**0.9631**	**0.9962**

in Figs. 8–12 and Tables V–VIII, respectively, in which the best one is shown in bold.

1) PRISMA: Fig. 8 shows the fusion results of 12 methods in the PRISMA dataset. Generally, DL-based methods obtain more reliable results than traditional ones and tensor-based approaches achieve more satisfactory products in the traditional category. Specifically, CNMF repairs more spatial details and leads to less spectral distortion compared with G-SOMP+ and CSU, partly because of its reasonable physical constraints and alternate unmixing process. The blind version of STEREO and SCOTT achieve better spatial reconstruction quality but exhibit noticeable spectral distortion compared with CSTF, which reflects the challenge of blind fusion tasks. As for the DL-based methods, XINet and SURE achieve satisfactory fusion quality in the spatial and spectral domains on account of their strong representation ability. Besides, MIAE also acquires competitive results in contrast to DBSR and HyCoNet, but CUCaNet suffers from noticeable spatial artifacts. It is worth mentioning that all methods obtain relatively poor results in the lake region partly due to its different reflection features with surrounding objects.

Moreover, the quantitative metrics listed in Table V further demonstrate the above-mentioned conclusions. Precisely, CNMF obtains relatively better results compared with G-SOMP+ and CSU. As for tensor-based methods, STEREO and SCOTT acquire higher PSNR with lower SAM, and CSTF achieves better spectral fidelity but with inferior spatial enhancement. DL-based methods outperform matrix- and tensor-based methods in most indicators, in which XINet achieves the best performance in all indicators. SURE obtains the second-best performance in the indicator of SAM and PSNR, and CUCaNet performs unsatisfactorily in PSNR, partly because of its insufficient utilization of multimodal information. More importantly, Fig. 12 shows the PSNR value of each band in four datasets. It can be seen that our method has obvious advantages in the overlapped bands and obtains an overall good performance in the nonoverlapped bands.

2) Washington DC Mall: The visual results of the Washington DC Mall are shown in Fig. 9. Overall, matrix-based methods, i.e., G-SOMP+, CSU, and CNMF, achieve

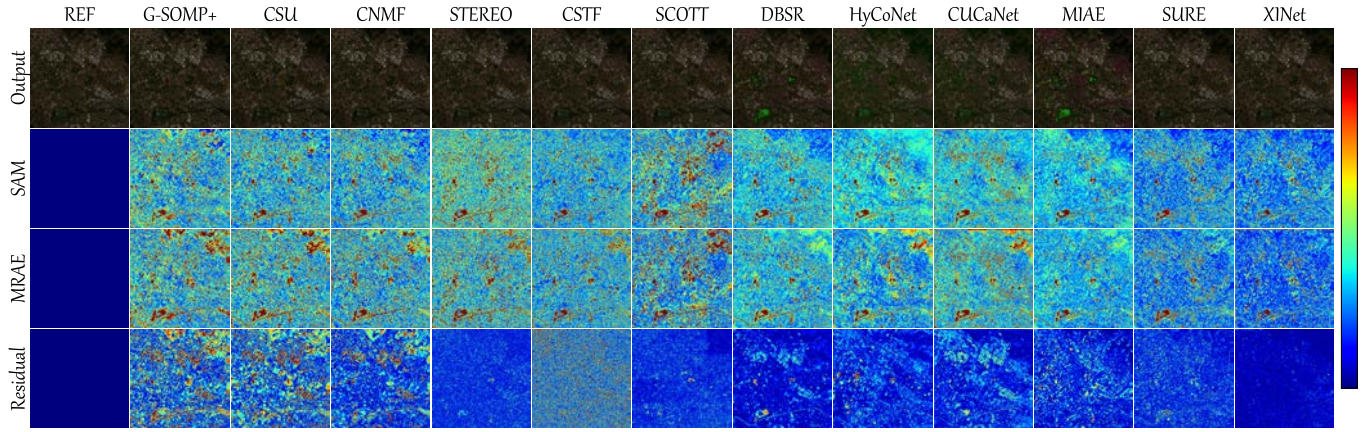

Fig. 8. Fusion results of 12 methods in PRISMA dataset. (First row) Pseudocolor R-G-B images of reconstructed outputs (R:14, G:8, and B:3). (Second row) Heatmap of SAM error. (Third row) Heatmap of MRAE. (Fourth row) Residual heatmap at band 11. The error range of three kinds of heat maps are [0, 6.5], [0, 0.15], and [0, 0.009], respectively.

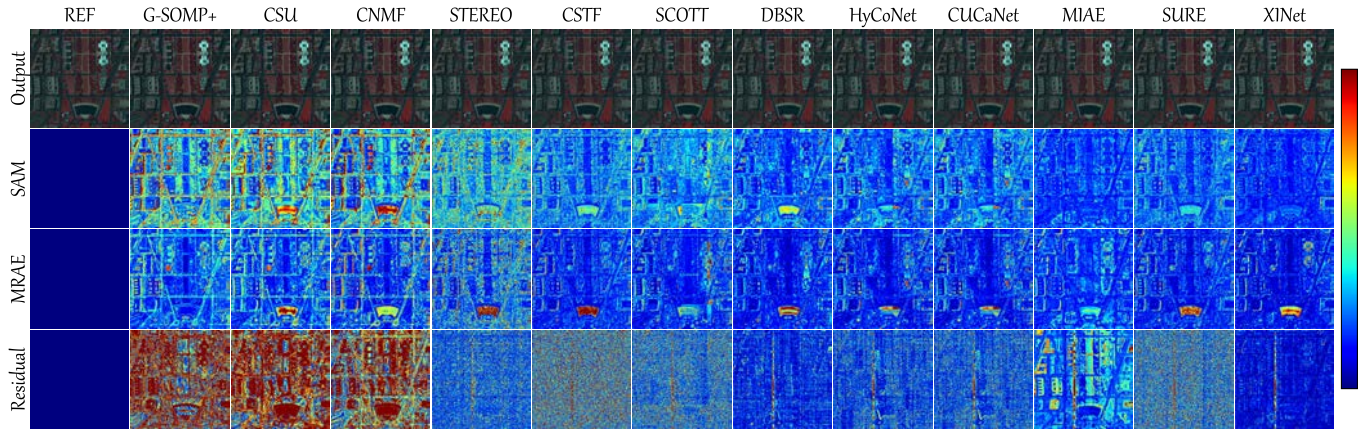

Fig. 9. Fusion results of 12 methods in Washington DC Mall dataset. (First row) Pseudocolor R-G-B images of reconstructed outputs (R:60, G:27, and B:17). (Second row) Heatmap of SAM error. (Third row) Heatmap of MRAE. (Fourth row) Residual heatmap at band 48. The error range of three kinds of heat maps are [0, 6.5], [0, 0.40], and [0, 0.008], respectively.

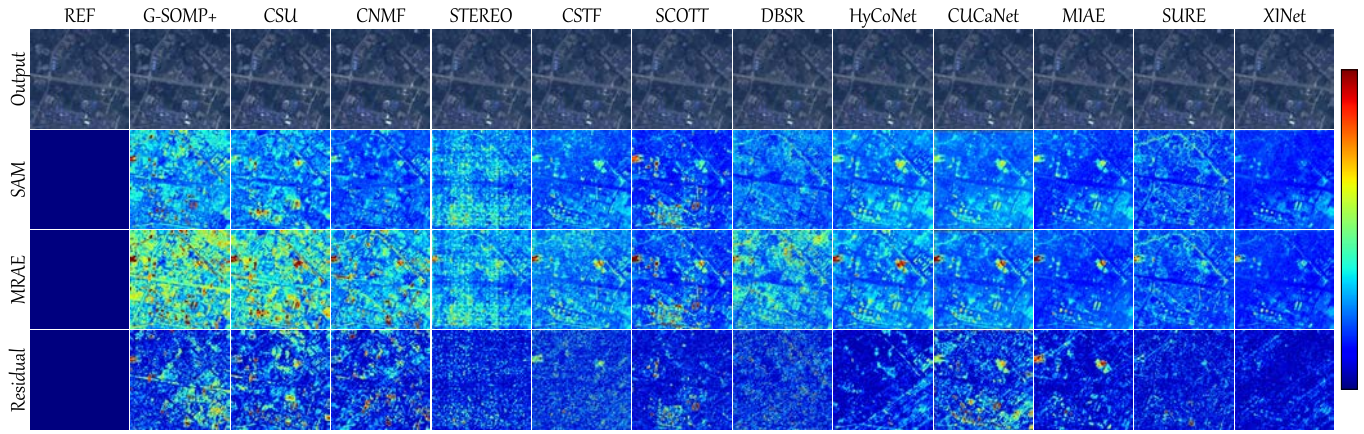

Fig. 10. Fusion results of 12 methods in TianGong-1 dataset. (First row) Pseudocolor R-G-B images of reconstructed outputs (R:29, G:19, and B:6). (Second row) Heatmap of SAM error. (Third row) Heatmap of MRAE. (Fourth row) Residual heatmap at band 48. The error range of three kinds of heat maps are [0, 5.5], [0.08], and [0, 0.026], respectively.

relatively poor results with noticeable spatial and spectral degradation. In terms of tensor-based methods, CSTF acquires better spectral preservation and spatial recovery compared with STEREO and SCOTT. All DL-based methods exhibit superior performance than traditional ones, in which XINet shows the best spatial-spectral reconstruction with blue. MIAE also exhibits competitive results, even better than ours in the middle-lower lake area. Besides, we also find that it

TABLE VI
QUANTITATIVE PERFORMANCE IN WASHINGTON DC MALL DATASET

Method	SAM	PSNR	ERGAS	SSIM	UIQI
G-SOMP+	2.9933	34.2878	1.6395	0.9476	0.9840
CSU	2.9542	34.3353	1.6211	0.9308	0.9813
CNMF	2.8402	34.8903	1.6233	0.9392	0.9797
STEREO	2.5790	36.3611	2.2822	0.9191	0.9628
CSTF	1.8249	37.7773	1.5025	0.9795	0.9838
SCOTT	1.9705	36.6655	1.8180	0.9737	0.9809
DBSR	1.6278	38.1332	1.6910	0.9798	0.9820
HyCoNet	1.6737	38.0249	1.6205	0.9796	0.9841
CUCaNet	1.7168	37.9092	1.6218	0.9794	0.9840
MIAE	1.3081	38.8602	2.3872	0.9724	0.9738
SURE	1.5861	38.6988	1.4810	0.9749	0.9843
Ours	**1.2301**	**39.8268**	**1.4605**	**0.9805**	**0.9867**

TABLE VII
QUANTITATIVE PERFORMANCE IN TIANGONG-1 DATASET

Method	SAM	PSNR	ERGAS	SSIM	UIQI
G-SOMP+	1.7714	36.1735	0.4124	0.9715	0.9980
CSU	1.5471	36.8678	0.3873	0.9722	0.9983
CNMF	1.2508	37.0527	0.3904	0.9834	0.9988
STEREO	1.6001	41.5475	0.2811	0.9706	0.9991
CSTF	1.3019	41.3339	0.2488	0.9813	0.9993
SCOTT	1.2381	41.1616	0.2461	0.9828	0.9993
DBSR	1.3416	40.5637	0.2658	0.9800	0.9990
HyCoNet	1.4548	41.7587	0.2524	0.9863	0.9991
CUCaNet	1.5125	41.2222	0.2902	0.9852	0.9992
MIAE	1.1705	43.9053	0.2062	0.9895	0.9994
SURE	1.2233	43.6018	0.2192	0.9820	0.9995
Ours	**1.0084**	**45.3571**	**0.1749**	**0.9900**	**0.9996**

TABLE VIII
QUANTITATIVE PERFORMANCE IN CHIKUSEI DATASET

Method	SAM	PSNR	ERGAS	SSIM	UIQI
G-SOMP+	1.2286	38.2567	0.7021	0.9710	0.9948
CSU	1.3406	37.8617	0.5647	0.9476	0.9930
CNMF	1.0593	40.5064	0.3524	0.9867	0.9954
STEREO	1.0095	42.0077	0.3717	0.9731	0.9970
CSTF	0.8152	43.1051	0.2975	0.9772	0.9980
SCOTT	0.9489	44.0998	0.3516	0.9856	0.9980
DBSR	0.8983	45.3765	0.2416	0.9869	0.9985
HyCoNet	0.9101	46.0336	0.2502	0.9908	0.9985
CUCaNet	0.9165	44.9185	0.2836	0.9908	0.9987
MIAE	0.8081	45.9483	0.2565	0.9871	0.9983
SURE	0.8121	44.2385	0.2686	0.9846	0.9983
Ours	**0.7685**	**47.4187**	**0.2400**	**0.9913**	**0.9989**

is challenging for most methods to recover reliable results in water areas. Surprisingly, G-SOMP+ and SCOTT obtain satisfactory outcomes in this region, since they deal with the whole image in a patch-based manner, which can better consider the local structures.

The quantitative assessment is reported in Table VI, providing consistent conclusions with visual results. Specifically, three matrix-based methods get lower values of PSNR and higher values of SAM, which reflects their limited reconstruction ability. STEREO has relatively poor performance compared with SCOTT, and CSTF achieves more acceptable results in both spatial and spectral domains. Importantly, our proposed XINet beats other competitors at all metrics, indicating the superior ability in enhancing spatial details and preserving spectral information. MIAE also gets a great performance in SAM and PSNR values but with higher ERGAS compared with the other DL-based methods. Besides, the PSNR curve shown in Fig. 12 demonstrates the overall good performance across all spectral bands achieved by our method.

3) TianGong-1: The results for TianGong-1 are visually shown in Fig. VII. It can be clearly observed that CNMF recovers more spatial details compared with G-SOMP+ and CSU, even better than tensor-based methods in preserving spectral information. SCOTT obtains acceptable results in forest lands but fails to handle the building areas lying at the bottom of the image. Besides, the stitching line in SCOTT is easy to observe due to its block-based process. Compared with the other five DL-based methods, XINet can well reconstruct different ground objects without significant distortions. However, it is worth mentioning that reconstructing regularly arranged objects with small sizes is of great challenge for existing methods, such as blue roofs.

Table VII lists the numerical assessment of different methods in the TianGong-1 dataset, which reflects a similar pattern shown in the PRISMA dataset. Concretely, CNMF achieves higher PSNR and lower SAM values in matrix-based methods. STEREO is good at recovering spatial structures and details, while CSTF and SCOTT perform better in spectral fidelity. Notably, our method yields the best performance in terms of all metrics, especially in the indicator of PSNR. Though SURE and MIAE obtain the closest outcomes to XINet, there also exists an obvious gap in PSNR and ERGAS. Moreover, it can be seen from Fig. 12 that our method achieves the best score in almost all bands, which directly verifies the effectiveness of XINet.

4) Chikusei: Fig. 11 shows the outcomes and corresponding error maps of all methods. Clearly, three matrix-based perform poorly in most regions, especially the boundaries between each field. Though tensor-based methods obtain better results in the boundaries, they fail to guarantee the integrity and continuity of each field. Especially, stitching lines among different blocks can be clearly observed in the outcome of SCOTT. As for DL-based methods, DBSR, HyCoNet, and CUCaNet obtain relatively good performance in boundaries,

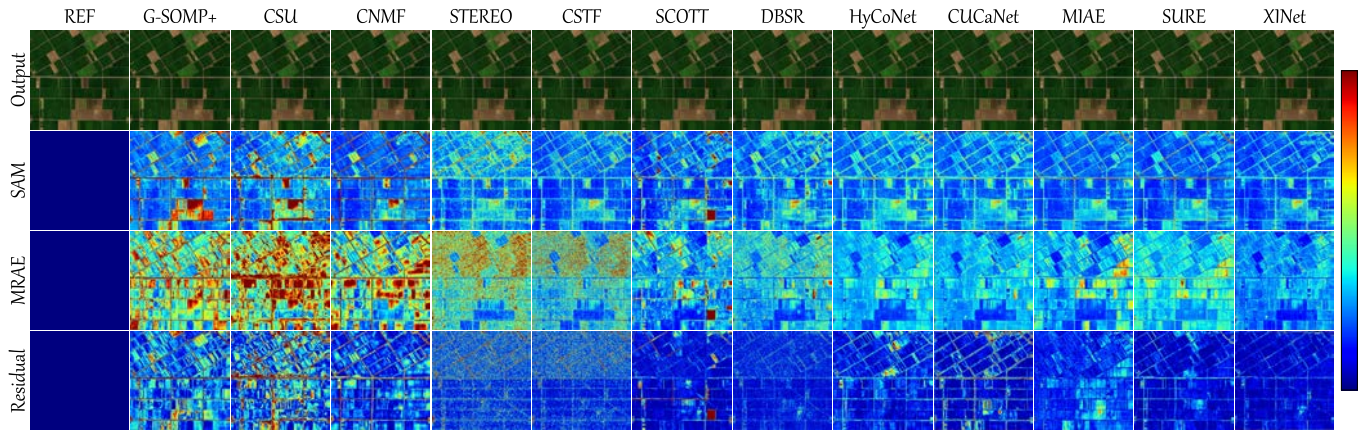

Fig. 11. Fusion results of 12 methods in Chikusei dataset. (First row) Pseudocolor R-G-B images of reconstructed outputs (R:56, G:36, and B:16). (Second row) Heatmap of SAM error. (Third row) Heatmap of MRAE. (Fourth row) Residual heatmap at band 26. The error range of three kinds of heat maps are [0, 3.0], [0, 0.08], and [0, 0.010], respectively.

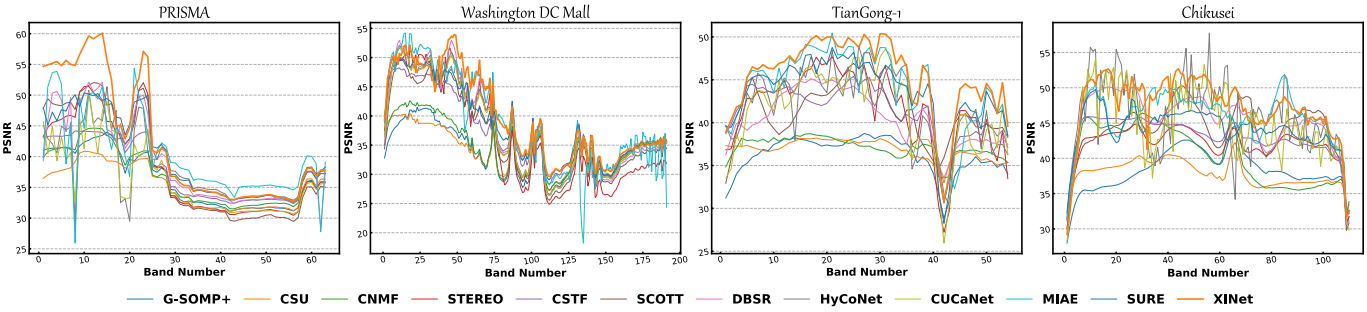

Fig. 12. PSNR value in each band of four datasets.

while MIAE achieves better results in recovering the field regions compared with SURE. Generally, our method achieves both superior outcomes in boundaries and field regions, partly thanks to the adaptive constraint imposed by $\mathcal{L}_{\text{WPATV}}$.

Table VIII summarizes the quantitative results in the Chikusei dataset, which further verifies the conclusion drawn from visual inspection. Specifically, our method obtains overall good results with the highest scores in all indicators, which convincingly demonstrates its ability in spatial recovery and spectral preservation. Besides, the PSNR curve of XINet shown in Fig. 12 also achieves competitive results compared with others.

E. Robustness Analysis

In this section, a robustness analysis is conducted to assess the performance of XINet under different scale factors and noise strengths, in which three competitors, i.e., CNMF, CSTF, and MIAE, along with the ~~Washington DC~~ Mall dataset are chosen as the benchmark.

1) Scale Factor: The fusion results under four different scale factors are shown in Fig. 13. It can be observed that the performance of all methods degrades with the increase of the scale factor, in which MIAE changes more dramatically than CNMF and CSTF. Despite the influence caused by the variation, XINet still outperforms other competitors in all cases, which demonstrates its superiority.

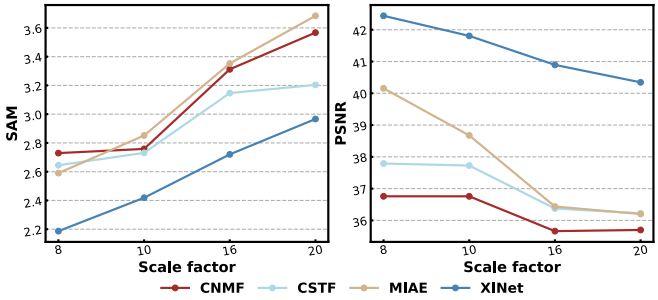

Fig. 13. Quantitative evaluation under four different scale factors in Washington DC Mall dataset.

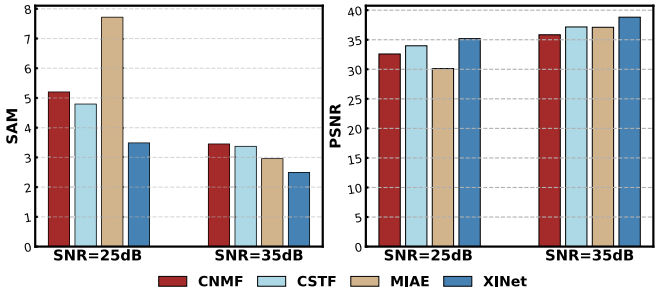

Fig. 14. Quantitative evaluation under different noise cases in Washington DC Mall dataset.

2) Noise: To evaluate their robustness against the Gaussian noise, two different strength is simulated and then added to the input HSI-MSI correspondence. As shown in Fig. 14,

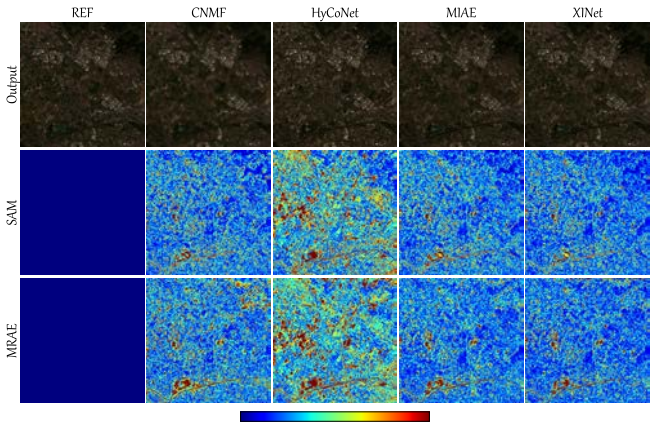

Fig. 15. Fusion results of hyperpansharpening in PRISMA dataset.

TABLE IX
QUANTITATIVE PERFORMANCE OF HYPERPANSHARPENING IN PRISMA DATASET

Method	SAM	PSNR	ERGAS	SSIM	UIQI
CNMF	5.1018	30.7775	2.1169	0.6835	0.9817
HyCoNet	7.1859	28.7147	2.6472	0.7235	0.9685
MIAE	4.9986	31.9832	1.8451	0.7995	0.9864
Ours	**4.6826**	**32.4961**	**1.7516**	**0.8013**	**0.9886**

Gaussian noise causes varying degrees of destruction to the fusion performance. Precisely, the quality of MIAE is severely damaged under the strength of 25 dB, while CNMF and CSTF are more robust than MIAE partly because of their reasonable constraints. By contrast, XINet still acquires the highest scores in all cases, because the proposed $\mathcal{L}_{\text{WPATV}}$ is capable of eliminating the latent noise contained in abundance maps.

F. Extension to Other Applications

Though our proposed XINet is designed for the HSI-SR task, we also investigate its potential on other two applications, that is, hyperpansharpening and classification.

1) Hyperpansharpening: Hyperpansharpening is a special case of the HSI-SR task, which aims to fuse one PAN image with high spatial resolution and one HSI with low spatial resolution to yield a product with both high spatial and spectral resolution. Considering the discrepancy between PAN images and MSIs, it is interesting and challenging for existing HSI-SR methods to tackle this problem. To assess the performance of XINet and three selected competitors, the simulation experiment is conducted in the PRISMA dataset. Precisely, the Gaussian kernel with a scale factor of 8 is employed to generate LrHSI, and the SRF of the PAN instrument in WorldView 2 is utilized to obtain the corresponding PAN image.

The outcomes of four approaches are reported in Table IX and Fig. 15. Generally, the performance of all methods is not as good as that obtained in MSI-aided HSI-SR, mainly due to the limited information carried by PAN images. Specifically, HyCoNet shows unsatisfactory results in this challenging task partly due to its limited ability for feature extraction. By contrast, XINet still achieves the best performance not only in quantitative metrics but also in visual results, largely thanks to its full utilization of multimodal information. However, there still exists a large room for our method to make progress in spectral preservation, which reflects the challenge of unsupervised hyperpansharpening.

TABLE X
CLASSIFICATION ACCURACY OVER INDIAN PINES DATASET

Method	OA	AA	Kappa
HrMSI	0.7456	0.7115	0.7092
CSTF	0.7189	0.7170	0.6783
MIAE	0.7320	0.7130	0.6929
Ours	**0.8178**	**0.7849**	**0.7918**

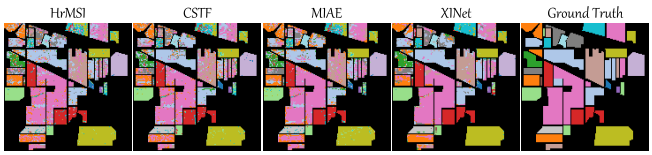

Fig. 16. Classification maps generated by SVM over Indian Pines dataset.

2) Classification: HSI-SR is a postprocessing technique to enhance the spatial resolution of HSIs and further promote their performance in downstream tasks. Therefore, the classification results of different super-resolved outcomes are evaluated in the Indian Pine dataset, in which original HrMSI, HrHSI reconstructed by CSTF, MIAE, and our proposed XINet are selected for comparison.

Also, the simulation experiment is implemented to generate input HSI-MSI correspondence. Specifically, a Gaussian kernel with a scale factor of 5 is utilized to generate LrHSI, and then the SRF of the multispectral instrument in WorldView 2 is employed to obtain HrMSI. Besides, we select a support vector machine (SVM) with RBF kernel as the classifier and choose 10% of the examples as the training data. Table X summarizes the classification accuracy in terms of overall accuracy (OA), average accuracy (AA), and Kappa coefficient (Kappa). It can be seen that our method achieves the highest scores in all metrics, which reflects our ability in spatial enhancement and spectral preservation from the side. Moreover, one interesting point worth mentioning is that the OA and Kappa values of CSTF and MIAE are even poor than those of HrMSI, which means simply introducing extra features into the classifier does not mean performance improvement. Fig. 16 further confirms the superiority of our method, especially in the lower part of the classification map.

VI. CONCLUSION

Considering the insufficient extraction and utilization of multimodal information in existing unsupervised HSI-SR networks, we propose an XINet to solve this problem. First, a novel X-shaped interactive architecture is designed as the backbone by coupling two isolated U-Nets together, which

not only enables sufficient information flow between two modalities but also leads to informative spatial-spectral features. Besides, CMMLM is embedded into multistages of the X-shaped encoder for better utilization of multimodal information. Moreover, a joint self-supervised loss is proposed to optimize our proposed XINet without the guidance of HrHSI. Extensive experiments are conducted to verify the effectiveness of proposed components, loss terms, and fusion performance. Furthermore, we evaluate the potential of XINet in the task of hyperpansharpening and classification, whose results convincingly demonstrate the ability of our method.

In the future, developing unsupervised networks for HSI-SR is still a research hotspot. Without the supervision of high-resolution images, it is extremely important to design powerful architectures and effective loss functions for feature extraction and learning guidance.

REFERENCES

[1] Y. Ding et al., "Unsupervised self-correlated learning smoothy enhanced locality preserving graph convolution embedding clustering for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536716.

[2] 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.

[3] D. Wang, L. Zhuang, L. Gao, X. Sun, M. Huang, and A. J. Plaza, "PDBSNet: Pixel-shuffle downsampling blind-spot reconstruction network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5511914.

[4] M. Wang, D. Hong, B. Zhang, L. Ren, J. Yao, and J. Chanussot, "Learning double subspace representation for joint hyperspectral anomaly detection and noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5507517.

[5] J. Zhou, W. Sun, X. Meng, G. Yang, K. Ren, and J. Peng, "Generalized linear spectral mixing model for spatial–temporal–spectral fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5533216.

[6] 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.

[7] G. Vivone, "Multispectral and hyperspectral image fusion in remote sensing: A survey," *Inf. Fusion*, vol. 89, pp. 405–417, Jan. 2023.

[8] N. Aburaed, M. Q. Alkhatib, S. Marshall, J. Zabalza, and H. Al Ahmad, "A review of spatial enhancement of hyperspectral remote sensing imaging techniques," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2275–2300, 2023.

[9] K. Ren, W. Sun, X. Meng, G. Yang, J. Peng, and J. Huang, "A locally optimized model for hyperspectral and multispectral images fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5519015.

[10] M. Wang et al., "Tensor decompositions for hyperspectral data processing in remote sensing: A comprehensive review," *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 1, pp. 26–72, Mar. 2023.

[11] W. Sun et al., "MLR-DBPFN: A multi-scale low rank deep back projection fusion network for anti-noise hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522914.

[12] 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.

[13] 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.

[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. 2021.

[15] R. B. Gomez, A. Jazaeri, and M. Kafatos, "Wavelet-based hyperspectral and multispectral image fusion," *Proc. SPIE*, vol. 4383, pp. 36–42, Jun. 2001.

[16] Y. Zhang and M. He, "Multi-spectral and hyperspectral image fusion using 3-D wavelet transform," *J. Electron. China*, vol. 24, no. 2, pp. 218–224, Mar. 2007.

[17] Z. Chen, H. Pu, B. Wang, and G.-M. Jiang, "Fusion of hyperspectral and multispectral images: A novel framework based on generalization of pan-sharpening methods," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 8, pp. 1418–1422, Aug. 2014.

[18] 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.

[19] N. Akhtar, F. Shafait, and A. Mian, "Bayesian sparse representation for hyperspectral image super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3631–3640.

[20] M. A. Veganzones, M. Simoes, G. Licciardi, N. Yokoya, J. M. Bioucas-Dias, and J. Chanussot, "Hyperspectral super-resolution of locally low rank images from complementary multisource data," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 274–288, Jan. 2016.

[21] L. Fang, H. Zhuo, and S. Li, "Super-resolution of hyperspectral image via superpixel-based sparse representation," *Neurocomputing*, vol. 273, pp. 171–177, Jan. 2018.

[22] J. Liu, Z. Wu, L. Xiao, J. Sun, and H. Yan, "A truncated matrix decomposition for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 8028–8042, 2020.

[23] 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.

[24] 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.

[25] T. Xu, T.-Z. Huang, L.-J. Deng, X.-L. Zhao, and J. Huang, "Hyperspectral image superresolution using unidirectional total variation with tucker decomposition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4381–4398, 2020.

[26] 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.

[27] Y. Xu, Z. Wu, J. Chanussot, P. Comon, and Z. Wei, "Nonlocal coupled tensor CP decomposition for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 348–362, Jan. 2020.

[28] 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.

[29] W. He, Y. Chen, N. Yokoya, C. Li, and Q. Zhao, "Hyperspectral super-resolution via coupled tensor ring factorization," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108280.

[30] 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.

[31] 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.

[32] L. Gao, J. Li, K. Zheng, and X. Jia, "Enhanced autoencoders with attention-embedded degradation learning for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5509417.

[33] 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.

[34] X.-H. Han and Y.-W. Chen, "Deep residual network of spectral and spatial fusion for hyperspectral image super-resolution," in *Proc. IEEE 5th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2019, pp. 266–270.

[35] 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.

[36] 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.

[37] X.-H. Han, Y. Zheng, and Y.-W. Chen, "Multi-level and multi-scale spatial and spectral fusion CNN for hyperspectral image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 4330–4339.

[38] T. Zhan et al., "A novel cross-scale octave network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5545816.
[39] H. Wu, J. Gui, Y. Xu, Z. Wu, Y. Y. Tang, and Z. Wei, "An efficient cross-modality self-calibrated network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5545312.
[40] 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.
[41] R. Ran, L.-J. Deng, T.-X. Jiang, J.-F. Hu, J. Chanussot, and G. Vivone, "GuidedNet: A general CNN fusion framework via high-resolution guidance for hyperspectral image super-resolution," *IEEE Trans. Cybern.*, vol. 53, no. 7, pp. 4148–4161, Jul. 2023, doi: 10.1109/TCYB.2023.3238200.
[42] 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.
[43] Z. Zhu, J. Hou, J. Chen, H. Zeng, and J. Zhou, "Hyperspectral image super-resolution via deep progressive zero-centric residual learning," *IEEE Trans. Image Process.*, vol. 30, pp. 1423–1438, 2021.
[44] A. Guo, R. Dian, and S. Li, "A deep framework for hyperspectral image fusion between different satellites," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 7939–7954, Jul. 2023, doi: 10.1109/TPAMI.2022.3229433.
[45] 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.
[46] S.-Q. Deng, L.-J. Deng, X. Wu, R. Ran, D. Hong, and G. Vivone, "PSRT: Pyramid Shuffle-and-Reshuffle transformer for multispectral and hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5503715.
[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] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5345–5355, Nov. 2018.
[49] J. Liu, D. Shen, Z. Wu, L. Xiao, J. Sun, and H. Yan, "Patch-aware deep hyperspectral and multispectral image fusion by unfolding subspace-based optimization model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1024–1038, 2022.
[50] 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.
[51] 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.
[52] 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.
[53] L. Zhang, J. Nie, W. Wei, Y. Zhang, S. Liao, and L. Shao, "Unsupervised adaptation learning for hyperspectral imagery super-resolution," in *Proc. IEEE Conf Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3073–3082.
[54] 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. 2021.
[55] 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.
[56] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[57] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1911–1920.
[58] 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.
[59] 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.
[60] 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.
[61] 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. 2012.
[62] 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.
[63] C. Prévost, K. Usevich, P. Comon, and D. Brie, "Hyperspectral super-resolution with coupled Tucker approximation: Recoverability and SVD-based algorithms," *IEEE Trans. Signal Process.*, vol. 68, pp. 931–946, 2020.
[64] H. V. Nguyen, M. O. Ulfarsson, J. R. Sveinsson, and M. D. Mura, "Deep SURE for unsupervised remote sensing image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5412613.
[65] 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.
[66] 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, May 1993.
[67] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?" in *Proc. Int. Conf. Fusion Earth Data*. Nice, France: SEE/URISCA, Jan. 2000, pp. 99–103.
[68] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.
[69] 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.
[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[71] L. Ren, D. Hong, L. Gao, X. Sun, M. Huang, and J. Chanussot, "Hyperspectral sparse unmixing via nonconvex shrinkage penalties," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5500415.
[72] L. Ren, D. Hong, L. Gao, X. Sun, M. Huang, and J. Chanussot, "Orthogonal subspace unmixing to address spectral variability for hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501713.

Jiaxin Li (Student Member, IEEE) received the B.E. degree from Chongqing University, Chongqing, China, in 2020. He is currently pursuing the Ph.D. degree in cartography and geography information systems 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. For more information GitHub: https://github.com/JiaxinLiCAS.

Ke Zheng received the B.S. degree in geographic information systems from Shandong Agricultural University, Taian, 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 with the College of Geography and Environment, Liaocheng University, Liaocheng, Shandong, China. His research interests include image processing, machine learning, deep learning, and their application in Earth vision.

Zhi Li received the B.Eng. degree in remote sensing science and technology from the School of Geography and Information Engineering, China University of Geosciences, Wuhan, China, in 2021. He is currently pursuing the Ph.D. degree in cartography and geography information systems with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include hyperspectral image processing, remote sensing classification, computer vision, and artificial intelligence.

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 systems from the Institute of Remote Sensing Applications, Chinese Academy of Sciences (CAS), Beijing, China, in 2007.

He is currently a Professor with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, CAS. He also 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 PI of ten scientific research projects at the national and ministerial levels, including projects by the National Natural Science Foundation of China (2016–2019, 2018–2020, and 2022–2025), and the National Key Research and Development Program of China (2021–2025). He has published more than 200 peer-reviewed papers, and there are more than 130 journal articles included in Science Citation Index (SCI). He was the coauthor of 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, and won the Second Prize of the State Scientific and Technological Progress Award in 2018. He received 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.

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 (via the part-time study) from The University of New South Wales, Canberra, ACT, Australia, in 1996.

She has 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. Her research interests include remote sensing, hyperspectral image processing, and spatial data analysis. 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, 3rd edition (1999) and 4th 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).

Dr. Jia received the Graduate Certificate in higher education from The University of New South Wales in 2005. She is the Editor-in-Chief of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.