Model Inspired Autoencoder for Unsupervised Hyperspectral Image Super-Resolution

Jianjun Liu\textsuperscript{(a), Member, IEEE, Zebin Wu\textsuperscript{(b), Senior Member, IEEE, Liang Xiao\textsuperscript{(c), Member, IEEE,} and Xiao-Jun Wu\textsuperscript{(d), Member, IEEE}}

Abstract—This article focuses on hyperspectral image (HSI) super-resolution that aims to fuse a low-spatial-resolution HSI and a high-spatial-resolution multispectral image to form a high-spatial-resolution HSI (HR-HSI). Existing deep learning-based approaches are mostly supervised that rely on a large number of labeled training samples, which is unrealistic. The commonly used model-based approaches are unsupervised and flexible but rely on handcrafted priors. Inspired by the specific properties of model, we make the first attempt to design a model-inspired deep network for HSI super-resolution in an unsupervised manner. This approach consists of an implicit autoencoder network built on the target HR-HSI that treats each pixel as an individual sample. The nonnegative matrix factorization (NMF) of the target HR-HSI is integrated into the autoencoder network, where the two NMF parts, spectral and spatial matrices, are treated as decoder parameters and hidden outputs, respectively. In the encoding stage, we present a pixelwise fusion model to estimate hidden outputs directly and then reformulate and unfold the model’s algorithm to form the encoder network. With the specific architecture, the proposed network is similar to a manifold prior-based model and can be trained patch by patch rather than the entire images. Moreover, we propose an additional unsupervised network to estimate the point spread function and spectral response function. Experimental results conducted on both synthetic and real datasets demonstrate the effectiveness of the proposed approach.

Index Terms—Autoencoder, hyperspectral image (HSI), nonnegative matrix factorization (NMF), super-resolution, unfolding.

I. INTRODUCTION

HYPERSPECTRAL image (HSI) is a kind of 3-D image taken at different spectral bands, with its spectral range covering hundreds of contiguous and narrow bands that span the visible to infrared spectrum. The high spectral resolution of HSIs enables accurate identification of materials and thereby promotes various applications, such as military, agriculture, and mineralogy. Due to the limited incident energy, there is always a tradeoff between spectral resolution, spatial resolution, and signal-to-noise ratio of images when designing the imaging sensors [1]–[6]. Thus, the spatial resolution of HSIs is usually sacrificed, which impedes the subsequent tasks. Conversely, conventional multispectral images (MSIs) at much lower spectral resolution can be acquired with higher spatial resolution. An economical HSI super-resolution solution is to instead record a low-spatial-resolution HSI (LR-HSI) and a high-spatial-resolution MSI (HR-MSI) and to fuse them into a target high-spatial-resolution HSI (HR-HSI) [2], [3], [5]. The HSI super-resolution solution breaks the limitations of the imaging sensors and has shown a good performance in practice.

HSI super-resolution that fuses an LR-HSI with an HR-MSI has attracted great attention [2], [3], [5]. This fusion problem arises from the pansharpening problem that fuses a low-spatial-resolution MSI or HSI with a high-spatial-resolution panchromatic image [1], [4], [6]. Generally, the conventional approaches proposed for the pansharpening problem can be extended to solve HSI super-resolution, but the fusion process of HSI super-resolution is relatively more complicated than that of pansharpening due to the rich spectral information. Related fusion approaches can be roughly divided into four categories: component substitution [7], multiresolution analysis [8], model-based approaches, and deep learning-based approaches. Among these categories, the research of model-based approaches is the most classic one, and deep learning-based approaches have been the most active one recently.

Model-based approaches consider building optimization models to obtain the target image. Given two observed images, they design fidelity terms and exploit spectral/spatial priors to enforce the desired result. Some approaches treat the target image as a variable and recover the target image entirely, such as group spectral embedding [9], clustering manifold structure [10], nonlocal patch tensor sparse representation [11], and structured sparse low-rank representation [12]. Most approaches consider separating the target image into parts and regenerating it via the recovered parts. There are many decomposition strategies by making assumptions about the target image. Examples are that it lives in a low-dimensional subspace and the subspace-based models are solved by exploiting prior knowledge, such as piecewise smooth [13], dictionary learning [14], tensor multirank [15], low tensor-train
rank [16], and truncated matrix decomposition [17]; or that
it can be represented linearly by pure spectral signatures
and the endmember and abundance matrices are recovered
simultaneously [18]–[21]; or that it can be sparsely represented
by an overcomplete spectral dictionary and different priors
are used to obtain the spectral dictionary and coefficients
[22]–[26] or by approaches that separate the target image
by tensor decomposition and update each component iteratively
[27]–[31]. Moreover, there are some approaches that
build models to estimate the point spread function (PSF) and
spectral response function (SRF) [13], [32]. The entire process
of model-based approaches is unsupervised. Although these
models are flexible and their theory is relatively complete,
they rely on handcrafted priors and there are many empirical
parameters to tune.

Deep learning-based approaches are data-driven. They build
deep neural networks to solve the related fusion problems and
produce the target image by feeding observed images into the
network. Some approaches enhance the ability to fuse images
in the network structures, such as 3-D convolutional neural net-
works (CNNs) [33], residual networks [34], multiscale struc-
tures [35], pyramid networks [36], attention networks [37],
[38], cross-mode information [39], dense networks [40], [41],
and adversarial network [42]–[44]. Some approaches use
detailed information from high-spatial-resolution conventional
images to improve the performance [45]–[48]. Inspired by the
specific properties of model, some form a hybrid of model-
and deep learning-based approaches [49]–[51], and some use
the deep unfolding technique to ease the construction of
networks [52]–[55]. These approaches have shown good per-
formance in exploiting the relationship between the observed
and target images. However, they are mostly supervised that
require plenty of labeled samples to train the networks, which
limits their applications in many scenarios.

There are some deep learning-based approaches developed
for HSI super-resolution that are performed in an unsupervised
manner. For instance, Dian et al. [56] introduced a CNN
denoiser to regularize the fusion model. Zhang et al. [57]
integrated the deep image prior into the fusion model and
thereby presented a unified unsupervised network for HSI
super-resolution. Qu et al. [58] exploited an unsupervised
approach composed of two autoencoder networks, which are
coupled through a shared decoder. Wang et al. [59] proposed a
variational probabilistic autoencoder framework implemented
by CNNs for HSI super-resolution. Yao et al. [60] proposed
a two-stream convolutional autoencoder framework inspired
by coupled spectral unmixing (CSU) and introduced a cross-
attention module to improve the performance. Uezato et al.
[61] designed a network composed of an encoder–decoder
network and a deep decoder network. Zheng et al. [62]
proposed a network consisting of three coupled autoencoder
networks, inspired by CSU, where the three autoencoder
networks are coupled through two convolutional layers. Most
approaches are built on the autoencoder architecture. Similar
to model-based approaches, the construction of networks relies
too much on human experience.

Inspired by the specific properties of model, we consider
constructing an unsupervised network by referencing some
models and propose a model-inspired autoencoder (MIAE) for
unsupervised HSI super-resolution. Specifically, we perform
nonnegative matrix factorization (NMF) on the target HR-
HSI to maintain its intrinsic structure and thereby propose an
implicit autoencoder network for HR-HSI by integrating its
NMF model. In the autoencoder network, each hyperspectral
pixel is treated as an individual sample, and the two NMF parts
of the target HR-HSI, i.e., spectral and spatial matrices, are
Treated as decoder parameters and hidden outputs, respectively.
Since the inputs of the autoencoder network are unknown,
we take the two observed images as inputs and present a
pixelwise fusion model to estimate each hidden output vector
directly. The pixelwise fusion model is solved by the gradient
descent algorithm, and the algorithm is reformulated and
unfolded to form the encoder network. The loss function is just
built on the mechanism of spectral and spatial degradations,
and an additional blind estimation network is proposed to esti-
mate the PSF and SRF. Compared with the existing HSI super-
resolution approaches, some of the innovative characteristics
of MIAE are highlighted as follows.

1) MIAE is an unsupervised deep learning-based approach
that involves only one implicit autoencoder. The autoen-
coder network treats each pixel as an individual sample,
and thus, the proposed network can be treated as a kind
of manifold prior-based model and can be trained patch
by patch to accelerate the training process.

2) MIAE is constructed by referencing models, and thus,
the construction of the network is relatively concise.
The NMF of the target HR-HSI is integrated into the
autoencoder, and the encoder network is inspired by the
pixelwise fusion model.

3) An additional unsupervised network is proposed to esti-
mate the PSF and SRF from the two observed images
directly.

The remainder of this article is organized as follows.
Section II proposes the proposed MIAE and its relationship
to the model-based approaches, as well as the blind estimation
network. In Section III, the effectiveness of MIAE is demon-
strated through experiments on three synthetic datasets and
one real dataset. Section IV provides the concluding remarks.

II. PROPOSED APPROACH

Fig. 1 shows the overall architecture of MIAE. The details
of the proposed network are described as follows.

A. NMF-Inspired Autoencoder

NMF is a useful dimension reduction method [63]. It can
capture the intrinsic structure of the data and represent the

Fig. 1. Overall architecture of MIAE.
data in a sparse manner. The properties of NMF indicate that it can facilitate the inference process of super-resolution if we perform NMF on the target HR-HSI. Let us represent the target HR-HSI as a matrix \( \hat{X} \in \mathbb{R}^{N_B \times N_H N_W} \), where \( N_B \) denotes the spectral band and \( N_H \) and \( N_W \) denote the spatial height and width, respectively. NMF aims to factor \( \hat{X} \) into two rank-\( J \) (\( J < \min\{N_B, N_H N_W\} \)) nonnegative matrices, i.e.,

\[
\hat{X} \approx AS,
\]

(1)

where the spectral matrix \( A \in \mathbb{R}^{N_B \times J} \geq 0 \) and the spatial matrix \( S \in \mathbb{R}^{J \times N_H N_W} \geq 0 \) with \( \geq \) being a componentwise inequality.

NMF can be integrated into an autoencoder network [64]–[66]. Equation (1) can be rewritten as

\[
\hat{x}_i \approx A s_i \quad \forall i
\]

(2)

where \( i = 1, 2, \ldots, N_H N_W \) and \( \hat{x}_i \in \mathbb{R}^{N_B} \) and \( s_i \in \mathbb{R}^J \) are the column vectors of \( \hat{X} \) and \( S \), respectively. Let \( \hat{x}_i \) represent the reconstructed vector and \( s_i \) represent the hidden output vector, and we can construct the following autoencoder network:

\[
x_i \xrightarrow{f(\cdot)} s_i \xrightarrow{g(\cdot)} \hat{x}_i \quad \forall i
\]

(3)

where the input data \( x_i \in \mathbb{R}^{N_B} \) are the column vector of the observed HR-HSI \( X \in \mathbb{R}^{N_B \times N_H N_W} \). The network (3) consists of two networks \( f(\cdot) \) and \( g(\cdot) \). \( f(\cdot) \) (described in Section II-B) is the proposed model-inspired encoder network with \( s_i = f(x_i; \theta) \), where \( \theta \) denotes all trainable parameters involved in the network. In hyperspectral unmixing [64]–[66], sum-to-one constraint is added to enforce the hidden vector, i.e., \( \sum_j s_{ij} = 1 \) with \( j \in \mathbb{Z}^J \) being a vector of all 1s. We do not intend to finish the two tasks of fusion and unmixing at once and only use the nonnegative constraint. Specifically, \( s_i \) is enforced using \( C^L_0(s_i) \) when designing the network, where \( C^L_0(\cdot) \) is a clamp function that forces all elements of the input vector/matrix into the range [0, 1].

The decoder network \( g(\cdot) \) derived from (2) is the reverse of the NMF process. It can be expressed as \( \hat{x}_i = g(s_i; A) = A s_i \), where the spectral matrix \( A \) is treated as the trainable weight matrix. When designing the network, the weight matrix \( A \) is enforced using \( C^L_0(A) \) and implemented by applying the clamp function every iteration. To restrict the scale of outputs, the clamp function is applied to \( \hat{x}_i \), i.e., \( \hat{x}_i = C^L_0(A s_i) \).

If the input data \( X \) are given, one can build a loss function for the autoencoder network (3) as follows:

\[
\mathcal{L} = \|X - \hat{X}\|_{1,1}
\]

(4)

where \( \| \cdot \|_{1,1} \) represents the absolute sum of all the matrix elements. Then, one can train the network (3) by feeding the \( N_H N_W \) hyperspectral pixels.

### B. Pixelwise Fusion Model-Inspired Encoder Network

In HSI super-resolution, the input data \( X \in \mathbb{R}^{N_B \times N_H N_W} \) are not given. One cannot train the network (3) directly. Instead, we have two observed (i.e., degenerated) images of \( X \), an LR-HSI \( Y \in \mathbb{R}^{N_B \times N_H N_W} \) and an HR-MSI \( Z \in \mathbb{R}^{N_B \times N_H N_W} \), where \( N_B < N_B \) is the multispectral band and \( N_B \) and \( N_H N_W \) are the spatial sizes. We assume that \( N_H = r N_B \) and \( N_W = r N_B \) with \( r > 1 \) being the resolution ratio. The observations \( Y \) and \( Z \) can be modeled as spatially degraded and spectrally degraded versions of \( X \). Specifically, these two degeneration processes can be written as

\[
Y \approx XBD \quad (5)
\]

\[
Z \approx RX \quad (6)
\]

where the PSF \( B \in \mathbb{R}^{N_B N_W \times N_B N_H} \) is the spatial blur, \( D \in \mathbb{R}^{N_B N_W \times N_B N_H} \) is the spatial downsampling, and \( R \in \mathbb{R}^{N_B \times N_B} \) is the SRF of multispectral sensor.

In (3), the network needs to be trained by feeding the input data pixel by pixel. It is the key to the success of autoencoder. For the degeneration processes, (6) can be rewritten as a pixelwise formulation, i.e., \( z_i \approx Rx_i \) with \( x_i \in \mathbb{R}^{N_B} \) being the column vector of \( Z \), whereas (5) cannot because of the coupling matrices \( B \) and \( D \).

We consider resizing \( Y \) to the same size as \( X \) using bilinear interpolation, in order to approximate \( X \) pixel by pixel at the spectral level. \( x_i \) can be obtained by solving

\[
\min_{x_i} \frac{1}{2} \|z_i - RX_i\|^2 + \frac{\lambda}{2} \|y_i - RX_i\|^2 \quad \forall i
\]

(7)

where \( \lambda > 0 \) is the regularization parameter and \( y_i \in \mathbb{R}^{N_B} \) is the column vector of \( Y_i \in \mathbb{R}^{N_B \times N_B} \) with \( Y \) being a bilinear interpolated version of \( Y \).

In order to construct the encoder network, it is unnecessary to solve \( x_i \) and then design \( f(\cdot) \), which will lead to error accumulation. We can treat \( x_i \) as an implicit variable and design \( f(\cdot) \) by using \( z_i \) and \( y_i \) directly, i.e.,

\[
(z_i, y_i) \xrightarrow{f(\cdot)} x_i \xrightarrow{g(\cdot)} \hat{x}_i \quad \forall i
\]

(8)

Specifically, we want to obtain the hidden layer output \( s_i \) by solving the following pixelwise fusion model:

\[
\min_{s_i} \frac{1}{2} \|z_i - RA s_i\|^2 + \frac{\lambda}{2} \|y_i - As_i\|^2 \quad \forall i
\]

(9)

and design \( f(\cdot) \) by unfolding all steps of its algorithm as network layers. Notably, in (9), both \( R \) and \( A \) are treated as new trainable parameters to facilitate the design of the encoder network. In the model-based HSI super-resolution, our previous works [17], [51] have shown the effectiveness of (9).

Although (9) has an analytic solution, it is not suitable as the encoder network and is difficult to implement by a network. Equation (9) can be solved by the gradient descent algorithm as

\[
s_i^k = s_i^{k-1} - \eta (\bar{A}^T \bar{A}s_i^{k-1} - \bar{A}^T z_i + \lambda A^T A s_i^{k-1} - \lambda A^T y_i)
\]

(10)

where \( \bar{A} = RA \in \mathbb{R}^{N_B \times J} \), \( \eta > 0 \) is the step, and \( k = 1, 2, \ldots, K \) represents the \( k \)th iteration. To better design the network, input data and intermediate variables are distinguished by rewriting (10) as

\[
s_i^k = (I - \eta \bar{A}^T \bar{A} - \eta \lambda \bar{A}^T A)s_i^{k-1} + \eta \bar{A}^T z_i + \eta \lambda \bar{A}^T y_i
\]

(11)

where \( I \) represents the identity matrix. According to the \( K \) iterations of (11), the proposed encoder network is mainly
a structure of \( K \) stages. Fig. 2 shows the details of \( f(\cdot) \) when \( K = 3 \). In (11), all variables \( s^k_i \), \( z_i \), and \( y^1_i \) are left multiplied by a matrix. This process is implemented using a fully connected layer followed by a leaky ReLU, and (11) can be rewritten as

\[
s^k_i = f^k\left(s^{k-1}_i; \theta^k_i\right) + \eta f_z(z_i; \theta_z) + \eta \lambda f_y(y^1_i; \theta_y)
\]

where \( \{f^k\}^K_{k=2}, f_z \) and \( f_y \), represent the modules designed for the multiplication of matrix and vector and \( \{\theta^k\}^K_{k=1}, \theta_z \), and \( \theta_y \), represent the trainable parameters involved in the corresponding networks. The red dotted boxes in Fig. 2 show the layers of \( \{f^k\}^K_{k=2}, f_z \) and \( f_y \), where an additional fully connected layer is used in \( f_y \) for the purpose of feature extraction. In (12), the three modules are combined linearly. To break the fixed format of optimization model and provide more flexibility, the linear combination is implemented by concatenating these modules and performing a fully connection and a leaky ReLU. Equation (12) can be rewritten as

\[
s^k_i = f^k\left(f^{k-1}_i; \theta^k_i\right), z_i; \theta_z, y^1_i; \theta_y
\]

where \( \{f^k\}^K_{k=1}, \theta_{i} \), and \( \theta_{i} \), represent the modules and trainable parameters for the linear combination, respectively. Finally, by performing \( f^k \) from 1 to \( K \), we can obtain the hidden output by \( s_i = C_0^k(f^K) \) and have the trainable parameters \( \theta = (\theta_z, \theta_y, \{\theta_{i} \}^K_{k=1}, \{\theta_{i} \}^K_{k=1}) \).

**C. Loss Function and Training Strategy**

To train the autoencoder network (8), we cannot use the loss function (4) anymore since the input data \( x_i \) are just an implicit variable that does not exist. Instead, we have two observed images \( Y \) and \( Z \) to work with and can only build loss function by using the observations and the target \( \hat{X} \). Similar to (5) and (6), the two degenerated images of \( \hat{X} \) are considered, and the outputs LR-HSI \( \hat{Y} \) and HR-MSI \( \hat{Z} \) can be modeled as \( \hat{Y} \approx \hat{X}BD \) and \( \hat{Z} \approx R\hat{X} \). Then, the loss function (4) can be rewritten as

\[
L = ||Z - \hat{Z}||_{1,1} + \gamma \||Y - \hat{Y}||_{1,1}
\]

where \( \gamma > 0 (\gamma = 1 \) is used in this work) is used to balance the spatial and spectral terms.

When training the network (8) using the loss function (14), one has to combine all pixels \( X_i \) into an image \( \hat{X} \) since \( \hat{Y} \) and \( \hat{X} \) are coupled together by BD. In other words, one has to train (8) by feeding \( Z \) and \( Y^1 \) entirely. Despite this, we can still train (8) by using small patches to accelerate the training process. Specifically, we divide \( Z \) and \( Y^1 \) into overlapped patches so that the patches cover all pixels and then discard the pixels affected by spatial blur \( B \) at the boundaries of these patches when computing the loss function (14).

**D. Blind Estimation Network**

The PSF \( B \) and the SRF \( R \) are required to train the proposed autoencoder network. By combining (5) and (6), we have

\[
\text{ZBD} \approx \text{RY}.
\]

By imposing some physical constraints, one can obtain \( B \) and \( R \) by solving

\[
\begin{align*}
\min_{B, R} & \ ||ZBD - RY||_{1,1} \\
\text{s.t.} & \ B \geq 0, \ R \geq 0, \ R1_{NB} = 1_{NB}.
\end{align*}
\]

Problem (16) can be solved by some optimization algorithms. Instead, we solve (16) by training a network. Specifically, we treat \((Z, Y)\) as inputs and \((B, R)\) as trainable parameters. Then, the blind estimation network can be constructed by using the following loss function:

\[
L' = ||\hat{Z} - \hat{Y}||_{1,1}
\]
E. Relationship to Model-Based Approaches

The proposed MIAE can be regarded as a kind of specific fusion model. By combining (1), (8), and (14), MIAE can be rewritten as

$$\min_{A, \theta} \|Z - RAS\|_{1,1} + \gamma \|Y - ASBD\|_{1,1}$$

s.t. \( s_i = f(z_i, y_i^1; \theta) \) \( \forall i \)

\[1 \geq A \geq 0, \quad 1 \geq S \geq 0, \quad 1 \geq AS \geq 0. \tag{18} \]

In (18), \( f(\cdot) \) can be thought of as a nonlinear mapping function, and each \( s_i \) is only associated with the input \( z_i \) and \( y_i^1 \) that correspond to its spatial position. The constraint \( s_i = f(z_i, y_i^1; \theta) \) acts as a manifold regularization that embeds the combination of \( z_i \) and \( y_i^1 \) into a low-dimensional space \( \mathbb{R}^d \). \( s_i = f(z_i, y_i^1; \theta) \) also acts as a self-supervised deep prior regularization that only uses itself as training data.

III. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, experiments on both synthetic and real datasets are conducted to demonstrate the performance of the proposed MIAE. Before the following experiments, all datasets are conducted to demonstrate the performance of the proposed MIAE. Before the following experiments, all datasets are scaled to the range \([0, 1]\). The quality of the fused images in the synthetic datasets is assessed with root-mean-squared error (RMSE), peak signal-to-noise ratio (PSNR), spectral angle mapper (SAM), relative dimensionless global error (ERGAS), and universal image quality index (UIQI) [2], [3], [5].

A. Synthetic Datasets and Implementation Details

Three real-life HSI datasets, University of Pavia (PaviaU), Kennedy Space Center (KSC), and Washington DC Mall (DC), are manipulated to use as synthetic reference images for the simulation experiments.

1) The PaviaU dataset is acquired by the Reflective Optics System Imaging Spectrometer (ROSIS), with a spectral range of 0.43–0.86 \(\mu\)m. The ROSIS sensor is characterized by 115 spectral bands and 103 remaining after removal of noisy bands. This image, with a size of \(610 \times 340\) pixels, has a spatial resolution of 1.3 m per pixel. We select a 256 pixel part as the reference image.

2) The KSC dataset is acquired by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS), with a spectral range of 0.4–2.5 \(\mu\)m. The AVIRIS sensor is characterized by 224 spectral bands and the number of spectral bands is reduced to 176 by removing water absorption bands. The size of this image is \(512 \times 614\) with a spatial resolution of 18 m. We select the up-left 512 \(\times 512\) pixel part as the reference image.

3) The DC dataset is acquired by the hyperspectral digital imagery collection experiment (HYDICE) image, with a spectral range of 0.4–2.4 \(\mu\)m. The HYDICE sensor is characterized by 210 spectral bands, and bands in the region where the atmosphere is opaque have been removed, leaving 191 bands. This image, with a size of \(1208 \times 307\) pixels, has a spatial resolution of about 2.8 m. We select a 512 \(\times 256\) pixel part as the reference image.

For each reference image, we generate the two observed images, LR-HSI and HR-MSI, according to the Walds protocol [67]. To generate the LR-HSI, we spatially blur the reference image and then downsample it by a factor of 8 \((r = 8)\) in each direction. A Gaussian blur of 15 \(\times 15\) pixels, with a mean of 0 and a standard deviation of 3.40, is applied to each band of the reference image. To generate the HR-MSI, \(\mathbf{R}\) is derived from the spectral response of the IKONOS satellite. We generate a four-band image by averaging the bands of the reference image according to the spectral response profiles of the RGB and NIR bands. To account for ubiquitous noise or error, moderate Gaussian noise is added to the LR-HSI (SNR = 30 dB) and the HR-MSI (SNR = 40 dB).

We implement and train the proposed network and blind estimation network using the PyTorch framework. As discussed in Section II-C, we divide the observed images into patches to accelerate the training process. Taking the HR-MSI as a reference, \(40 \times 40\)-pixel overlapping patches with a stride of 24 are extracted for training. The batch sizes are 25 for the PaviaU and DC datasets and 50 for the KSC dataset. An Adam optimizer is used to train the network for 10,000 iterations. The learning rate is initialized as \(5 \times 10^{-3}\) and gradually decayed by multiplying \(1 - (1/9000)\) max(0, iteration – 1000), where ‘iteration’ represents the current number of iterations. As for the blind estimation network, it is trained by feeding the observed images entirely, the total number of iterations is 5000, and the learning rate is set as \(5 \times 10^{-5}\). The code of the proposed MIAE will be available on https://github.com/liuofficial.

B. Influence of Parameters

Two parameters, rank \(J\) and stage \(K\), need to be given when constructing the proposed network. In this set of experiments, we investigate them and show how they impact the quality measures of MIAE. Fig. 4 shows the PSNR results of MIAE as a function of \(J\) when \(K = 1, 2, \ldots, 5\). It can be seen that, for all datasets, the PSNR performance improves as \(J\) increases, but a large \(J\) will cause overfitting or performance degradation. Compared with small values of \(K\), a moderate \(K\) is better and a too-large \(K\) is prone to overfitting. Thus, \(K\) is eventually set as 80 for the PaviaU and KSC datasets and 30 for the DC dataset.

C. Experiment Results on Synthetic Datasets

1) Blind and Nonblind: Section II-D presents a blind estimation network for estimating the PSF and SRF.
This experiment is used to evaluate the estimated $\mathbf{B}$ and $\mathbf{R}$. Figs. 5 and 6 show the estimated $\mathbf{B}$ and $\mathbf{R}$, respectively, where the exact ones are also included for reference. The estimated $\mathbf{B}$ exhibits a sparse property and the curves of the estimated $\mathbf{R}$ are relatively coarse. Table I shows the quality measures of the proposed MIAE using the exact and estimated $\mathbf{B}$ and $\mathbf{R}$, that is, nonblind and blind cases. It can be seen that, for the PaviaU and KSC datasets, the performance degradation caused by blind estimation is very small when compared with the nonblind estimation, and for the DC dataset, the performance degradation is also not significant.

2) Influence of LR-HSI Interpolation: For the proposed MIAE, bilinear interpolation is used to upsample the LR-HSI to the same size of the target HR-HSI. This experiment shows how the interpolation method affects the performance of MIAE. Four interpolation methods are considered, i.e., bilinear interpolation, nearest interpolation, bicubic interpolation, and cubic spline interpolation. The quality measures to assess the different interpolation methods are given in Table II. In most cases, there is no obvious difference between these interpolation methods. The nearest interpolation performs slightly worse on the KSC dataset and the bicubic interpolation on the DC dataset.

3) Comparison With the State of the Art: Nine unsupervised methods, which can be divided into model- and deep learning-based approaches, are compared to evaluate the performance of MIAE. The model-based approaches consist of six methods. The first method is the baseline one (denoted by SYLV) that solves a Sylvester equation [68], and the next five methods are coupled NMF (CNMF) [18], CSU [19], NSSR [22], HySure [13], and NPTSR [11]. The deep learning-based approaches are CNNFUS [56] and three autoencoder-based methods, i.e., uSDN [58], HyCoNet [62], and also MIAE. The free parameters of the compared methods are tuned to be optimal with the test datasets, and the default training strategies are used for the deep learning-based approaches. All of the compared methods are blind, where the estimated $\mathbf{B}$ and $\mathbf{R}$ are used. For those methods that do not involve blind estimation procedures, $\mathbf{B}$ and $\mathbf{R}$ are estimated by the proposed blind estimation network.

The five quantitative results of the compared methods for the PaviaU dataset are shown in Table III with the best values marked in bold. It can be seen that all the methods outperform the baseline method SYLV, and the proposed MIAE method gives the best quantitative results followed by NPTSR and HyCoNet. Both the model- and deep learning-based approaches can yield good results. Fig. 7 shows the reference image and the fusion results of the compared methods in form of the 30th band gray and error images. Visually, it can be observed that the baseline method SYLV has severe spatial

---

1The code is derived from the toolbox on the website: https://openremotesensing.net/wp-content/uploads/2017/11/HSMSFusion_Toolbox.zip
2https://naotoyokoya.com/assets/zip/CNMF_MATLAB.zip
3https://github.com/lanha/SupResPALM
4http://see.xidian.edu.cn/faculty/wsdong/wsdong_Publication.htm
5https://github.com/alfaiate/HySure
6The code is provided by Dr. Xu.
7https://github.com/renweidian/CNN-FUS
8https://github.com/aicip/uSDN
9https://github.com/saber-zero/HyperFusion
TABLE II
QUALITY MEASURES OF MIAE USING DIFFERENT INTERPOLATION METHODS

| Method | best | bilinear | nearest | bicubic | spline | best | bilinear | nearest | bicubic | spline |
|--------|------|----------|---------|---------|--------|------|----------|---------|---------|--------|
| RMSE   | 0.0172 | 0.0172 | 0.0171 | 0.0172 | 0.0172 | 0.0427 | 0.0431 | 0.0428 | 0.0426 | 0.0145 |
| PSNR   | +∞    | 37.33   | 37.38   | 37.30   | 37.27   | 34.05  | 33.88   | 34.02   | 34.04   | 35.90   |
| SAM    | 0     | 2.43    | 2.43    | 2.41    | 2.43    | 7.00   | 7.11    | 7.01    | 7.01    | 1.84    |
| ERGAS  | 0     | 0.656   | 0.656   | 0.652   | 0.653   | 3.129  | 3.160   | 3.135   | 3.130   | 14.224  |
| UIQI   | 1     | 0.988   | 0.988   | 0.988   | 0.988   | 0.882  | 0.879   | 0.882   | 0.881   | 0.979   |

TABLE III
QUALITY MEASURES FOR THE PAVIAU DATASET USING DIFFERENT METHODS (THE BEST VALUES ARE HIGHLIGHTED)

| Method | SYLV | CNMF | CSU | NSSR | HySure | NPTSR | CNNFUS | uSDN | HyCoNet | MIAE |
|--------|------|------|-----|------|--------|-------|--------|------|---------|------|
| RMSE   | 0.1072 | 0.0196 | 0.0231 | 0.0236 | 0.0236 | 0.0194 | 0.0186 | 0.0237 | 0.0258 | 0.0188 | 0.0172 |
| PSNR   | 23.79 | 35.75 | 33.87 | 33.93 | 36.09 | 36.71 | 35.24 | 32.84 | 36.67 | 37.33 |
| SAM    | 12.62 | 2.62 | 2.89 | 3.21 | 2.70 | 2.64 | 3.16 | 3.49 | 2.66 | 2.43 |
| ERGAS  | 3.646 | 0.741 | 0.849 | 0.871 | 0.728 | 0.699 | 0.825 | 0.905 | 0.720 | 0.656 |
| UIQI   | 0.853 | 0.987 | 0.983 | 0.983 | 0.986 | 0.987 | 0.984 | 0.982 | 0.987 | 0.988 |

Fig. 7. Images (with a meaningful region marked and zoomed in three times for easy observation) and error maps at band 30 of HSI super-resolution results when applied to the PaviaU dataset. (a) Reference image. (b) SYLV. (c) CNMF. (d) CSU. (e) NSSR. (f) HySure. (g) NPTSR. (h) CNNFUS. (i) uSDN. (j) HyCoNet. (k) MIAE.

TABLE IV
QUALITY MEASURES FOR THE KSC DATASET USING DIFFERENT METHODS (THE BEST VALUES ARE HIGHLIGHTED)

| Method | SYLV | CNMF | CSU | NSSR | HySure | NPTSR | CNNFUS | uSDN | HyCoNet | MIAE |
|--------|------|------|-----|------|--------|-------|--------|------|---------|------|
| RMSE   | 0.1574 | 0.0454 | 0.0465 | 0.0513 | 0.0453 | 0.0490 | 0.0534 | 0.0504 | 0.0441 | 0.0427 |
| PSNR   | 19.33 | 32.70 | 31.97 | 30.77 | 32.95 | 33.30 | 30.95 | 30.06 | 33.49 | 34.05 |
| SAM    | 23.23 | 7.78 | 8.02 | 8.65 | 7.64 | 7.29 | 9.01 | 9.14 | 7.22 | 7.00 |
| ERGAS  | 8.753 | 3.497 | 3.405 | 3.738 | 3.336 | 3.328 | 3.954 | 3.717 | 3.262 | 3.129 |
| UIQI   | 0.506 | 0.870 | 0.855 | 0.836 | 0.836 | 0.881 | 0.887 | 0.843 | 0.857 | 0.878 | 0.882 |

distortion and all other methods outperform it. MIAE and HyCoNet perform better than the other methods in terms of both zoomed region and error map. Fig. 10(a) shows the PSNR as a function of spectral band for the compared methods. It can be seen that the proposed MIAE method performs best in almost all bands followed by HyCoNet and NPTSR. Fig. 11(a) shows the SAM between the reference image and the fusion results for each pixel using the compared methods, with the pixels sorted by ascending error. As illustrated in this figure, MIAE consistently outperforms the others at the pixel level.

Table IV reports the five quality measures of the compared methods for the KSC dataset. From this table, we can see that the baseline method SYLV performs the worst, NPTSR gives the best UIQI result, and the proposed MIAE method performs best for the remaining four quality measures. NPTSR and HyCoNet are only inferior to MIAE. In Fig. 8, we show the reference image and the fusion results of the compared methods in form of the 30th band gray and error images. Visually, it can be observed that the reconstructed results of HyCoNet and MIAE are better than the others, and the baseline method SYLV gives the worst images. Fig. 10(b) shows PSNR as a function of the spectral band for the compared methods. MIAE, HyCoNet, and NPTSR achieve high results in most bands. Fig. 11(b) shows the SAMs for
Fig. 8. Images (with a meaningful region marked and zoomed in three times for easy observation) and error maps at band 30 of HSI super-resolution results when applied to the KSC dataset. (a) Reference image. (b) SYLV. (c) CNMF. (d) CSU. (e) NSSR. (f) HySure. (g) NPTSR. (h) CNNFUS. (i) uSDN. (j) HyCoNet. (k) MIAE.

Table V
QUALITY MEASURES FOR THE DC DATASET USING DIFFERENT METHODS (THE BEST VALUES ARE HIGHLIGHTED)

| Method | SYLV | CNMF | CSU | NSSR | HySure | NPTSR | CNNFUS | uSDN | HyCoNet | MIAE |
|--------|------|------|-----|------|--------|-------|--------|------|---------|------|
| RMSRE  | 0.1685 | 0.0283 | 0.0220 | 0.0366 | 0.0275 | 0.0261 | 0.0332 | 0.0278 | 0.0197  | 0.0145 |
| PSNR   | 39.53 | 32.61 | 32.48 | 29.90 | 32.13 | 32.77 | 32.39 | 29.24 | 31.63  | 35.90 |
| SAM    | 26.56 | 3.82 | 2.82 | 5.33 | 3.48 | 3.54 | 3.70 | 3.53 | 1.83  | 1.84 |
| RMAS   | 37.085 | 15.182 | 14.406 | 14.447 | 13.797 | 13.886 | 14.300 | 17.345 | 10.943  | 14.224 |
| UIQI   | 0.469 | 0.930 | 0.929 | 0.905 | 0.930 | 0.944 | 0.954 | 0.856 | 0.914  | 0.975 |

Fig. 9. Images (with a meaningful region marked and zoomed in three times for easy observation) and error maps at band 30 of HSI super-resolution results when applied to the DC dataset. (a) Reference image. (b) SYLV. (c) CNMF. (d) CSU. (e) NSSR. (f) HySure. (g) NPTSR. (h) CNNFUS. (i) uSDN. (j) HyCoNet. (k) MIAE.

D. Experiment Results on Real Data

The University of Houston (UH) dataset released by the 2018 IEEE GRSS Data Fusion Contest [69] is used to evaluate MIAE in practical applications. The original data are acquired by the National Center for Airborne Laser Mapping (NCALM), covering the UH campus and its surrounding urban areas. This experiment selects an LR-HSI and a high-resolution RGB (HR-RGB) image from this multimodal optical remote sensing datasets. The LR-HSI collected by the ITRES CASI-1500 sensor contains 4172 × 1202 pixels with a spatial resolution of 1 m and 48 spectral bands with a spectral range of 0.38–1.05 μm. The HR-RGB image collected by DiMAC ULTRALIGHT+ sensor contains 4172 × 1202 pixels with a spatial resolution of 1 m and 48 spectral bands with a spectral range of 0.38–1.05 μm. The HR-RGB image collected by DiMAC ULTRALIGHT+ sensor contains 83 440 × 24 040 pixels. Taking the LR-HSI as a reference, we select an area of 64 × 64 × 48 as our observation data and downsample the corresponding area of the HR-RGB to be a 512 × 512 × 3 size...
image, that is, the resolution ratio is $r = 8$. RGB images of the real dataset and the fusion results of the compared methods mentioned in Section III-C are given in Fig. 12. Visually, it can be seen that MIAE, NPTSR, and HySure give the good color and brightness results, and the result of the proposed MIAE is much closer to the HR-RGB image. Since the reference HR-HSI is not available, three quality measures without reference, spectral distortion index $D_s$, spatial distortion index $D_s$, and quality with no reference (QNR) [6], [70], are adopted to quantitatively measure the quality of the fusion results. For $D_s$, it is the average of the evaluation results of the three RGB channels. Table VI shows the three quality measures of the compared methods. It can be seen that the proposed MIAE gives the highest QNR.

### E. Computational Efficiency

All of the above experiments are carried out using a desktop computer with an Intel Core i9-7900X CPU, a GeForce GTX 2080Ti GPU, and 64-GB memory. The computational efficiency of the compared methods mentioned in Section III-C3 is investigated. The first seven methods, SYLV, CNMF, CSU, NSSR, HySure, NPTSR, and CNNFUS, are performed using MATLAB, and the remaining three autoencoder-based methods, uSDN, HyCoNet, and MIAE, are implemented by the PyTorch framework. Table VII summarizes the running times of the first seven methods and the training times of the autoencoder-based methods, and the number of trainable parameters for each autoencoder network is reported in Table VIII.
Fig. 12. RGB images (with a meaningful region marked and zoomed in three times for easy observation) of HSI super-resolution results when applied to real dataset. (a) HR-RGB image. (b) LR-HSI. (c) SYLV. (d) CNMF. (e) CSU. (f) NSSR. (g) HySure. (h) NPTSR. (i) CNNFUS. (j) uSDN. (k) HyCoNet. (l) MIAE.

### TABLE VIII

| Method | uSDN | HyCoNet | MIAE |
|--------|------|---------|------|
| PaviaU | 37.9K| 377.7K  | 87.8K|
| KSC    | 48.9K| 389.5K  | 99.5K|
| DC     | 51.1K| 391.9K  | 21.7K|
| UH     | 2.3K | 368.6K  | 13.1K|

For the proposed MIAE, the training times of feeding the observed images entirely are also included in Table VII for reference, and the corresponding method is denoted by MIAE-ONE. It can be seen that MIAE takes less time to train the network than the other two autoencoder-based methods, and its trainable parameters are much less than HyCoNet. Ignoring the platform, MIAE is comparable to the model-based approaches.

### IV. CONCLUSION

This article has proposed an unsupervised MIAE network for HSI super-resolution. The proposed MIAE involves an implicit autoencoder network and the structures are concise. First, inspired by that performing NMF on the target HR-HSI can facilitate the inference process of super-resolution, the implicit autoencoder network is built on the target HR-HSI by integrating its NMF model, where the two NMF parts, spectral and spatial matrices, are treated as decoder parameters and hidden outputs respectively. The autoencoder network treats each hyperspectral pixel of the target HR-HSI as an individual sample, that is, the network is trained pixel by pixel. Second, the “implicit” indicates that the input pixel of the autoencoder network is unknown, and thus, a pixelwise fusion model taken the two observed images as inputs is presented to estimate the hidden layer vector directly. The pixelwise fusion model is simple and effective. Specifically, the LR-HSI is resized to the same size of the target HR-HSI using bilinear interpolation, in order to feed the network pixel by pixel. To break the fixed format of model and provide more flexibility, the gradient descent algorithm is used to solve the pixelwise fusion model, and the algorithm is reformulated and unfolded to form the encoder network. Finally, the loss function is built on the relationship between the target HR-HSI and the two observed images. With the specific pixelwise architecture, MIAE can be treated as a kind of manifold prior-based model and can be trained patch by patch to accelerate the training process. Moreover, a blind estimation network is proposed to estimate the PSF and SRF in an unsupervised manner. MIAE has been experimentally tested using three synthetic datasets and one real dataset, and the experimental results demonstrate its effectiveness. Although the results obtained by MIAE are very encouraging, further improvements, such as the application of convolutional autoencoder, should be pursued in future.

### ACKNOWLEDGMENT

The authors would like to thank Xu et al. [11], Simões et al. [13], Yokoya et al. [18], Lanaras et al. [19], Dong et al. [22], Dian et al. [56], Qu et al. [58], Zheng et al. [62], and Wei et al. [68] for providing their codes. They would like to thank NCALM and the Hyperspectral Image Analysis Laboratory, University of Houston (UH), for providing the UH datasets and the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society for supporting the annual Data Fusion Contest. They would also like to thank the anonymous reviewers for their constructive comments on this article.

### REFERENCES

[1] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce, “Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S data-fusion contest,” IEEE Trans. Geosci. Remote Sens., vol. 45, no. 10, pp. 3012–3021, Oct. 2007.
[2] L. Loncan et al., “Hyperspectral pansharpening: A review,” IEEE Trans. Geosci. Remote Sens., vol. 3, no. 3, pp. 27–46, Sep. 2015.
[3] N. Yokoya, C. Grohmfeld, and J. Chanussot, “Hyperspectral and multispectral data fusion: A comparative review of the recent literature,” IEEE Geosci. Remote Sens. Mag., vol. 5, no. 2, pp. 29–56, Jun. 2017.
[4] X. Meng, H. Shen, H. Li, L. Zhang, and R. Fu, “Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges,” Inf. Fusion, vol. 46, pp. 102–113, Jun. 2019.
Recent advances and new guidelines on hyperspectral and multispectral image fusion, "Inf. Fusion", vol. 69, pp. 40–51, May 2021.

A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods, "IEEE Geosci. Remote Sens. Mag.", vol. 9, no. 1, pp. 53–81, Mar. 2021.

A new look at IHS-like image fusion methods, "Inf. Fusion", vol. 2, pp. 177–186, May 2001.

Remote sensing image fusion using the curvelet transform, "Inf. Fusion", vol. 8, no. 2, pp. 143–156, 2007.

Multispectral and hyperspectral image fusion based on group spectral embedding and low-rank factorization, "IEEE Trans. Geosci. Remote Sens.", vol. 55, no. 3, pp. 1363–1371, Mar. 2017.

Exploiting clustering manifold structure for hyperspectral imagery super-resolution, "IEEE Trans. Image Process.", vol. 27, no. 12, pp. 5969–5982, Dec. 2018.

Nonlocal patch tensor sparse representation for hyperspectral image super-resolution, "IEEE Trans. Image Process.", vol. 28, no. 6, pp. 3034–3047, Jun. 2019.

Spatial–spectral structured sparse low-rank representation for hyperspectral image super-resolution, "IEEE Trans. Image Process.", vol. 30, pp. 3084–3097, 2021.

A convex formulation for hyperspectral image super-resolution via subspace-based regularization, "IEEE Trans. Geosci. Remote Sens.", vol. 53, no. 6, pp. 3373–3388, Jun. 2015.

Multispectral and multispectral image fusion based on a sparse representation, "IEEE Trans. Geosci. Remote Sens.", vol. 53, no. 7, pp. 3658–3668, Jul. 2015.

"Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization," "IEEE Trans. Image Process.", vol. 28, no. 10, pp. 5135–5146, Oct. 2019.

"Learning a low tensor-train rank representation for hyperspectral image super-resolution," "IEEE Trans. Neural Netw. Learn. Syst.", vol. 30, no. 9, pp. 2672–2683, Sep. 2019.

"A truncated matrix decomposition for hyperspectral image super-resolution," "IEEE Trans. Image Process.", vol. 29, pp. 8028–8042, 2020.

"SSR-NET: Spatial–spectral reconstruction network for hyperspectral and multispectral data fusion," IEEE Trans. Geosci. Remote Sens., vol. 50, no. 2, pp. 528–537, Feb. 2012.

"Hyper spectral super-resolution by coupled spectral unmixing," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 3586–3594.

"A spectral grouping and attention-driven bidirectional pyramid network," "IEEE Trans. Geosci. Remote Sens.", vol. 57, no. 8, pp. 5549–5563, Aug. 2019.

"Pan-sharpening using an efficient bidirectional pyramid network," "IEEE Trans. Geosci. Remote Sens.", vol. 60, no. 1, pp. 513–530, Jan. 2021.
[51] D. Shen, J. Liu, Z. Xiao, J. Yang, and L. Xiao, “A twice optimizing net with matrix decomposition for hyperspectral and multispectral image fusion,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 13, pp. 4095–4110, 2020.

[52] Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, and Z. Xu, “Multispectral and hyperspectral image fusion by MS/HS fusion net,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1585–1594.

[53] W. Wei, J. Nie, Y. Li, L. Zhang, and Y. Zhang, “Deep recursive network for hyperspectral image super-resolution,” IEEE Trans. Comput. Imag., vol. 6, pp. 1233–1244, 2020.

[54] W. Dong, C. Zhou, F. Wu, J. Wu, G. Shi, and X. Li, “Model-guided deep hyperspectral image super-resolution,” IEEE Trans. Image Process., vol. 30, pp. 5754–5768, 2021.

[55] D. Shen, J. Liu, Z. Wu, J. Yang, and L. Xiao, “ADMM-HFNet: A matrix decomposition-based deep approach for hyperspectral image fusion,” IEEE Trans. Geosci. Remote Sens., early access, Sep. 24, 2021, doi: 10.1109/TGRS.2021.312181.

[56] R. Dian, S. Li, and X. Kang, “Regularizing hyperspectral and multispectral image fusion by CNN denoiser,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 3, pp. 1124–1135, Mar. 2021.

[57] 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.

[58] 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.

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

[60] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, “Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 208–224.

[61] T. Uezato, D. Hong, N. Yokoya, and W. He, “Guided deep decoder: Unsupervised image pair fusion,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 87–102.

[62] 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.

[63] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

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

[65] Y. Qian, F. Xiong, Q. Qian, and J. Zhou, “Spectral mixture model inspired network architectures for hyperspectral unmixing,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 10, pp. 7418–7434, Oct. 2020.

[66] B. Palmsson, M. O. Ulfarsson, and J. R. Sveinsson, “Convolutional autoencoder for spectral-spatial hyperspectral unmixing,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 1, pp. 535–549, Jan. 2021.

[67] T. Ranchin and L. Wald, “Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation,” Photogramm. Eng. Remote Sens., vol. 66, no. 1, pp. 49–61, Jan. 2000.

[68] Q. Wei, N. Dobigeon, and J. Tourneret, “Fast fusion of multi-band images based on solving a Sylvester equation,” IEEE Trans. Image Process., vol. 24, no. 11, pp. 4109–4121, Nov. 2015.

[69] Y. Xu et al., “Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 12, no. 6, pp. 1709–1724, Jun. 2019.

[70] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, “Multispectral and panchromatic data fusion assessment without reference,” Photogramm. Eng. Remote Sens., vol. 74, no. 2, pp. 193–200, Feb. 2008.

Jianjun Liu (Member, IEEE) received the B.S. degree in applied mathematics and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, China, in 2009 and 2014, respectively. From 2018 to 2020, he was a Post-Doctoral Researcher with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong. He is currently an Associate Professor with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China. His research interests are in the areas of hyperspectral image classification, super-resolution, spectral unmixing, sparse representation, computer vision, and pattern recognition.

Zebin Wu (Senior Member, IEEE) received the B.S. and Ph.D. degrees in computer science from the Nanjing University of Science and Technology, Nanjing, China, in 2003 and 2008, respectively. He is currently a Professor with the School of Computer Science, Nanjing University of Science and Technology. His research interests include hyperspectral image processing, high-performance computing, and computer simulation.

Liang Xiao (Member, IEEE) received the B.S. degree in applied mathematics and the Ph.D. degree in computer science from the Nanjing University of Science and Technology, Nanjing, China, in 1999 and 2004, respectively. From 2009 to 2010, he was a Post-Doctoral Fellow at the Rensselaer Polytechnic Institute, Troy, NY, USA. He is currently a Professor with the School of Computer Science, Nanjing University of Science and Technology. His main research areas include inverse problems in image processing, scientific computing, data mining, and pattern recognition.

Xiao-Jun Wu (Member, IEEE) received the B.Sc. degree in mathematics from Nanjing Normal University, Nanjing, China, in 1991, and the M.S. and Ph.D. degrees in pattern recognition and intelligent system from the Nanjing University of Science and Technology, Nanjing, China, in 1996 and 2002, respectively. From 1996 to 2006, he taught at the School of Electronics and Information, Jiangsu University of Science and Technology, Zhenjiang, China, where he was promoted to a Professor. From 1999 to 2000, he was a fellow of the International Institute for Software Technology, United Nations University, Tokyo, Japan. From 2003 to 2004, he was a Visiting Researcher with the Centre for Vision, Speech, and Signal Processing (CVSSP), University of Surrey, Guildford, U.K. He is currently a Professor with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China. His research interests include pattern recognition, computer vision, and computational intelligence. He has over 300 publications in these areas.