Model-Guided Deep Unfolded Fusion Network With Nonlocal Spatial-Spectral Priors for Hyperspectral Image Super-Resolution

Abdolraheem Khader, *Student Member, IEEE*, Jingxiang Yang, *Member, IEEE*, and Liang Xiao, *Member, IEEE*

***Abstract*—Due to the physical boundaries, fusing low spatial resolution hyperspectral (LrHSI) with high spatial resolution multispectral (HrMSI) images is a hot and promising area for obtaining hyperspectral that have high spatial-spectral resolution images (HrHSI). Effectively formulating the fundamental features of hyperspectral images (HSI), such as global spectral correlation, nonlocal spatial correlation, as well spatial-spectral correlation, is complex in HSI-MSI fusion. Moreover, the fusion process is highly affected by the degradation systems, where these systems are not known in real scenarios. To this end, in this article, we proposed a model-guided deep unfolded fusion network with nonlocal spatial-spectral priors (MGDuNLSS-net) that can maintain the essential features of the HSIs and implicitly estimates the degradation process in an adequate running time. Specifically, the proposed method is designed based on subspace representation in an iterative manner and unrolling its steps toward a deep network as an end-to-end framework. This approach contains two submodules, fusion [nonlocal spatial-spectral block (NLSSB)] and imaging system submodules. The former submodule is proposed to exploit the images' intrinsic characteristics to improve the preservation of spectral and spatial details. NLSSB contains two nonlocal self-similarity (NLSS) layers embedded between two bidirectional simple recurrent unit (BSRU) layers. The recurrent calculation, as well as refined components to maintain the global spectral correlation, are the light recurrence operation and highway network, while 3-D convolutions in the BSRU can retain the spatial-spectral correlation. The NLSS layer can efficiently and effectively model long-range spatial contexts, which is designed based on criss-cross attention. The later submodule is used to refine the prediction of the degradation process at any iteration via backprojecting the estimated fused image to the observed pair, which can ensure the good performance of fusion. Compared with state-of-the-art fusion approaches, three remote sensing datasets are used to validate the proposed approach's performance.**

***Index Terms*—Deep learning, global spectral correlation, hyperspectral and multispectral images fusion, imaging systems, subspace representation.**

I. Introduction

DUE to the cost, physical limitation, and complexity constraints in remote sensing, obtained optic images with high spectral resolution usually maintain a more subordinate spatial resolution compared to images of lower spectral resolution [1], [2]. The high-spatial-spectral resolution image has great importance in various applications, such as environment monitoring, target tracking, and military investigation [3], [4], [5]. Therefore, in seeking to achieve a high-spatial-spectral resolution image using the available images, considerable strategies have been developed to tackle this problem by fusing the high spectral resolution image with a high spatial resolution image [6]. Unlike single-image superresolution methods that optimize the relationship between LrHSI and HrHSI [7], [8], image fusion methods using two complementary images (LrHSI and HrMSI) are more challenging and can lead to more satisfactory performance. The image fusion problem is an ill-posed inverse task, where an image that has a low spatial and high spectral resolution is further improved in spatial resolution by utilizing a supplementary image with low spectral and high spatial resolution, in which both these images depict the identical view and are appropriately coregistered [9], [10], [11]. The widespread image fusion scheme is pansharpening [12], [13], [14]; herein, a multispectral image (MSI) is improved by operating the high-resolution wide-band panchromatic (PAN) image. An equivalent task is improving a hyperspectral image (HSI) with either an MSI or PAN [15]. When using the MSI to enhance the low-resolution HSI, the problem is named hyperspectral and multispectral image fusion, which is more challenging than the pansharpening problem due to both MSI and HSI containing spectral and spatial information that should be preserved [16], [17].

Hyperspectral (HS) and multispectral (MS) image fusion can be either model based or learning based. The model-based strategies of HS-MS image fusion are carried out by formulating an optimization function that can be solved analytically or iteratively. In this regard, the subspace representation model is extensively utilized in the fusion problem due to the nature of HS and MS images properties that can be represented in low-rank matrices. However, based on the subspace model,

Manuscript received 13 March 2023; revised 25 April 2023; accepted 27 April 2023. Date of publication 2 May 2023; date of current version 24 May 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61571230, Grant 61871226, Grant 62001226, in part by Natural Science Foundation of Jiangsu Province under Grant BK20200465, in part by Jiangsu Provincial Social Developing Project under Grant BE2018727, in part by the Fundamental Research Funds for the Central Universities under Grant JSGP202204 and Grant 30920021134. *(Corresponding authors: Jingxiang Yang; Liang Xiao.)*

Abdolraheem Khader and Jingxiang Yang are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: abdolraheem@njust.edu.cn; yang123jx@njust.edu.cn).

Liang Xiao is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China, and also with the Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: xiaoliang@mail.njust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3272370

the two observed images are factorized into their spectral signatures and the coefficient (abundance) matrix, after then use the estimated subspace of HSI and the abundance matrix of MSI to reconstruct the high-resolution hyperspectral image (HrHSI). Matrix and tensor factorization are appropriate techniques to accomplish the fusion; thus, many studies are proposed based on these techniques [16], [18], [19], [20], [21]. While model-based approaches gained good performance in terms of quantitative and qualitative metrics, they have serious shortcomings that can limit their performance. Since these methods commonly work in an iterative manner, time computation is one of the limitations that model-based methods can face; furthermore, they have many parameters that need to be accurately tuned, which is another challenge. On the other hand, the model-based methods rely significantly on the observation model, which is unknown in real scenarios. Therefore, a few model-based approaches are tried to estimate the imaging model [22] in two steps that can increase the computation time; further, the imaging models are learned from the observed pair images with no information from the fused image.

Recently, deep learning attained impressive prosperity in computer vision [23] and image processing [24]. Therefore, many researchers proposed HSI-MSI fusion works based on deep learning either in supervised [25], [26], [27] or unsupervised methods [28]. The supervised deep HSI-MSI fusion methods utilize the convolutional neural network (CNN) to optimize the mapping function between the observed pairs and the ground truth to estimate the HrHSI in a black-box fashion, where the imaging models are enforced to estimate implicitly with no supervision. Although good performance is attained by these methods and reduces the computation time, the black-box manner can bind their further improvement; additionally, the simple convolution layer extracts the feature locally that limits it from exploiting the important essential features of the HSI like nonlocal spatial, global spectral, as well spatial-spectral correlation. However, to elevate the limitations of the black-box networks, various studies are proposed based on deep learning by unfolding the variational model's solution into CNN and introducing spatial and/or spectral attention modules, which makes the network interpretable, has a physical meaning, and more reliable to extract the intrinsic characteristics of the HSI [29], [30], [31], [32]. Even though these methods obtained good enhancement for deep learning networks, they either cannot fully utilize the fundamental properties of the HSI or the predicted fused image may not be projected back to estimate the original observed pair to enhance the prediction of the imaging models.

Inspired by the success of model-based and deep learning-based techniques, in this article, we proposed model-guided deep unfolded fusion network with nonlocal spatial-spectral priors (MGDuNLSS-Net) for HSI-MSI Fusion. We formulate the fusion method in this article based on the subspace representation model in an iterative fashion and unfold it toward the network, which can be trained in an end-to-end model that jointly estimates the observation systems and the target HrHSI. The proposed technique includes two submodules: the imaging system and the fusion sub-modules. In the proposed method, instead of learning the degradation systems implicitly without supervision, the imaging system module takes the fused image at the previous stage and projects it back to the original HrMSI and LrHSI. By this means, the imaging system module helps the proposed MGDuNLSS-Net to supervise and perfectly predict the degradation parameters, such as downsampling constraints and the blur matrix. Furthermore, the fusion submodule in the proposed method at any stage takes the errors between the original observed pair and the estimated pair by the former module and fuses them to enhance the fused image in the previous stage. To fully utilize the crucial features of HSI for fusion, the proposed fusion module is fashioned by a nonlocal spatial-spectral block (NLSSB) that contains two nonlocal self-similarity layers (NLSS) embedded between two bidirectional simple recurrent unit layers (BSRU), where the critical characteristics of the HSI such as the global spectral, nonlocal spatial, and spatial-spectral correlations can be significantly exploited. Specifically, the recurrent calculation and refined components to maintain the global spectral correlation are the light recurrence and highway network, while 3-D convolutions in the BSRU can retain the spatial-spectral correlation. The NLSS layer can efficiently and effectively model long-range spatial contexts. The contributions of this article are listed below:

1) We proposed an end-to-end unfolded network (termed MGDuNLSS-Net) with no hand-crafted parameters tuning that can iteratively learn the imaging systems and the targeted HrHSI. The proposed network is based on a subspace representation model that makes it interpretable, has a physical meaning, and is easy to train.
2) While the degradation model highly influences the fusion process, and it is unknown, the proposed method can supervise and correctly learn the degradation process by its imaging system submodule, which projects the fused image back to estimate the original LrHSI and HrMSI.
3) We proposed a nonlocal spatial-spectral block to enhance the fusion process by utilizing more robust features from the fusion images that can consider the global spectral correlation, which is calculated and refined by the light recurrence and highway network in the BSRU layer, where the spatial-spectral correlation is obtained by the NLSS layer which ensures maintaining the long-range spatial contexts. Moreover, the nonlocal spatial correlation can be maintained by 3-D convolutions that existed in BSRU.
4) Extensive experiments using three remote sensing databases (specifically, Pavia Center, Chikusei, and Cuprite Mine) are conducted to verify the performance of the proposed MGDuNLSS-Net.

The rest of this article is organized as the following. Section II investigates previous studies briefly. The proposed method and implementation details are stated in Section III. Section IV represents the experiments and discussions. Finally, Section V concludes the article.

II. Related Works

We briefly review the previous works in HSI-MSI fusion in this section, which can be roughly grouped into model-based and learning-based methods. Then, considering latent HSI statistics, several priors are used for fusion.

A. Model-Based HSI-MSI Fusion

HSI-MSI methods solve the fusion problem by formulating an appropriate variational model and carefully choosing proper prior to regularizing the optimization problem, where the degradation systems should be known or appropriately computed in advance. By utilizing the HSI's low-rank property, in [18], the authors incorporated sparse prior for approximate low-rank tensor to exploit the latent features of HIS, where the fusion optimization problem is dissolved by utilizing the alternating direction method of multipliers (ADMM). Guo et al. [33] enhanced the spatial details preservation by integrating the total variation (TV) with sparse coding priors and proposed 2D-CNBTD based on a coupled nonnegative block-term tensor model. The nonlocal structure tensor total variation regularization to improve the extracting the nonlocal image self-similarity and local structural image regularity of the images is used to enhance the fusion result in [10]. On the other hand, numerous HSI-MSI fusion techniques based on LMM are proposed that can use matrix or tensor factorization techniques to solve the fusion problem. In these approaches, the observed images are alternatively factorized into the spectral basis and abundance matrix, where the spectral signatures of LrHSI and the coefficient matrix of HrMSI can obtain the fused image. For example, Yokoya et al. [34] introduced coupled nonnegative matrix factorization (CNMF) technique, which combines the sparsity of the coefficient matrix and the nonnegativity constraints. To further enhance the spatial resolution of the desired fused image, a nonlocal self-similarity regularizer is combined with the sparse prior in a unified fusion model, which alternatively and iteratively approximates the dictionary and abundance matrix as stated in [35] and [36]. The subspace techniques for HSI-MSI fusion [22], [37], [38] are developed based on the subspace, which differs from the unmixing model by ignoring the nonnegativity of the dictionary and the coefficient matrix and the sum-to-one constraint of the coefficients matrix. Veganzones et al. [39] underlined that the HrHSI is low rank locally and that each local region's spectral basis and coefficients are computed individually.

Contrary to the matrix factorization methods, the tensor factorization techniques treat the HSI as a 3-D tensor. Borsoi et al. [40] proposed a coupled tensor decomposition approach using purely algebraic and optimization procedures to improve the fusion method that considers the modifications in an additive model confined in space and spectral range. The work in [41] engaged the coupled Tucker decomposition, where the fundamental global spatial-spectral information crosswise the different modes are extracted by capturing the global interdependencies. This method combines coupled Tucker factorization and the local submanifold structures in a joined model. Ma et al. [42] devised a fusion approach that joins sparse and smooth regularizations on low-rank tensor decomposition using the logarithmic sum function to eliminate superfluous information in both spectral and spatial realms. Zhang et al. [43] devised two graphs in spectral and spatial fields, the former derived from the LrHSI image for spectral smoothness and the latter drawn for the spatial consistency from the HrMSI image. Integrated all regularizers finally reconstruct the fusion model in this work. However, these model-based HSI-MSI methods are time consuming and rely heavily on the observation systems, which is not available in reality. Moreover, the hand-crafted parameters and choosing an appropriate prior are more challenges that these techniques can face.

B. Learning-Based HSI-MSI Fusion

Considering deep learning, many researchers utilize the CNN ability to extract features and learn the nonlinear maps between the input and output to tackle the fusion task. However, the deep learning-based techniques are constructed by staking many convolutional layers in one branch that take the concatenated LrHSI and HrMSI as input, and their output is the predicted fused image [26], [44]. To enrich the network capability of feature model, two-branches networks are investigated. For example, Yang et al. [25] proposed a deep network with two branches one to extract features from HrMSI, the second to take the LrHSI and capture its features, and finally, the HrHSI is constructed from the features captured from these inputs. In [45], an iterative refinement is proposed that takes the HrMSI and LrHSI in two branches to refine the fused image iteratively. Han et al. [46] considered that the LrHSI and HrMSI have very diverse spatial structures and proposed a multiscale CNN approach. This method has two branches, one to preserve the spectral features and the second for maintaining the spatial features, which progressively shrinks and expands the feature size of the LrHSI and LrMSI, respectively.

Considering the benefits of the model-based and learning-based strategies, many researchers tried to combine model based with deep learning to enhance the performance of the deep network in HSI-MSI fusion. Liu et al. [47] introduced a deep HSI-MSI fusion approach based on unrolling the solution of a subspace-based optimization model. The fusion problem in this work is first solved by ADMM and then unrolling the proposed algorithm's steps toward the network, which makes the proposed network more interpretable. The spatial-spectral dual-optimization model-driven deep network is proposed in [48], incorporating the spatial and spectral priors of the input LrHSI and HrMSI in the proposed network. Wang et al. [49] devised a deep HSI-MSI fusion method with interimage variability that treat flexible image priors and interimage changeability of images from various sensors. This method is formulated as a posteriori model and uses CNN to learn the image's priors. On the other hand, the researchers tried to enhance the capability of the deep methods by incorporating the attention mechanism. For example, to address the challenge of complicated fusion of three different data types, in [50], HyperNet is built based on a uniform fusion technique. Specifically, multiple specially constructed multiscale-attention-enhance blocks that use multiscale convolution to dynamically capture information from various reception fields as well as two attention techniques to improve the extraction ability of information lengthwise the spatial and spectral domains, respectively, are used to extract the spatial details of the PAN and MSI. Based on a variational probabilistic generative model, the RAFnet approach for recurrent attention fusion is presented in [51] that utilizes two variational autoencoders to ensure both spectral and spatial information preservation. Hu et al. [52] introduced the Fusformer method that increases the receptive field of the network by self-attention (SA) mechanism to improve the extracting of global relationships in features.

However, the learning-based methods for HSI-MSI fusion inspired outstanding performance and significantly reduced computation time. Moreover, the observation models can be learned implicitly in the deep learning methods from the data with no hand-crafted tuning of the parameters, contrary to model-based methods. Although the advantages of the learning-based methods, there is still room for improvement that can be investigated. The network architecture may be in a black-box manner or not fully interpretable, and further, it possibly will not consider important information such as the spectral-spatial correlation due to the limited receptive field of the classic convolutional layer. Furthermore, deep learning methods learn the degradation models implicitly, which lacks supervision to guarantee good estimation of estimated imaging models, which also can be inspected more.

III. Model-Guided Deep Unfolded Fusion Network With Nonlocal Spatial-Spectral Priors

A. *Problem Formulation*

The aim of fusing the observed pair of low-spatial resolution hyperspectral image (LrHSI) and the high-spatial-resolution multispectral image (HrMSI) is the achievement of a high-spatial-resolution hyperspectral image (HrHSI). Let us assume $\mathbf{X} \in \mathbb{R}^{C \times MN}$ is target HrHSI, where C is the numeral of spectral bands while the spatial height and width are denoted by M and N, respectively. Based on the latent low-rank property of HSIs, it can be represented via the subspace representation model. The subspace model has advantages such as efficient computation time because the number of spectral bases is lesser than the number of channels of the HSIs ($K < C$), it can fully achieve significant correlations among the spectral channels, and its semiunitary property. Thus, based on the subspace representation model, HrHSI can be expressed as the following:

$$\mathbf{X} = \mathbf{DA} \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{C \times K}$ is the spectral signatures (spectral basis) with K atoms, and $\mathbf{A} \in \mathbb{R}^{K \times MN}$ is the abundance (fractional coefficients) matrix that holds the sparse vectors of entire HrHSI's pixels. The subspace representation model omitted the sparsity of the abundance matrix. By this means, each pixel in $\mathbf{X}$ is believed as a linear combination between the spectral signatures and the coefficients. Moreover, based on the subspace model, the nonnegative restriction of $\mathbf{D}$ is ignored.

The spatial downsampled version of $\mathbf{X}$ can be regarded as the LrHSI, denoted by $\mathbf{Y} \in \mathbb{R}^{C \times mn}$, where m $(m < M)$ and n $(n < N)$ are the spatial height and width of the LrHSI, respectively. Therefore, LrHSI can be carried out by the next imaging model:

$$\mathbf{Y} = \mathbf{DAB} = \mathbf{D}\tilde{\mathbf{A}} \tag{2}$$

where $\mathbf{B} \in \mathbb{R}^{MN \times mn}$ denotes the degradation blur matrix constructed by the convolution of the sensor's point spread function (PSF) and the HrHSI spectral channels pursued by a downsampling operator, and $\tilde{\mathbf{A}}(\tilde{\mathbf{A}} = \mathbf{AB})$ is a spatial downgraded abundance matrix. Meantime, the HrMSI is a version of $\mathbf{X}$ that spectrally downsampled; hence the HrMSI (denoted by $\mathbf{P} \in \mathbb{R}^{c \times MN}$) can be achieved by the following imaging system:

$$\mathbf{P} = \mathbf{RDA} = \widetilde{\mathbf{D}}\mathbf{A} \tag{3}$$

where $\mathbf{R} \in \mathbb{R}^{c \times C}$ represents the spectral downsample matrix, and $\widetilde{\mathbf{D}}(\widetilde{\mathbf{D}} = \mathbf{RD})$ is the spectral downsampled dictionary.

The HrHSI can be reconstructed from the HrMSI and HrHSI by jointly using imaging models (2) and (3) to iteratively and alternatively optimize the spectral basis and its interrelated abundance and, finally, use (1). However, the next minimization problem can be used to approximate the spectral signatures and abundance matrix:

$$\min_{\mathbf{D},\mathbf{A}} f_1(\mathbf{Y}, \mathbf{DAB}) + f_2(\mathbf{P}, \mathbf{RDA}) \tag{4}$$

where the first term ensures spectral preservation in the estimated spectral basis, and the second indicates spatial similarity to the coefficients matrix. Furthermore, the optimization problem (4) assumed that the spatial and spectral degradation models ($\mathbf{B}$ and $\mathbf{R}$) are available, which is not true in reality. Many researchers implicitly learned the degradation models from the training dataset to overcome the problem of unknown degradation models with no supervision [45]. Therefore, the degradation operators can be learned as follows:

$$\min_{\mathbf{B},\mathbf{R}} f\left(\widetilde{\mathbf{D}}\mathbf{AB}, \mathbf{RD}\tilde{\mathbf{A}}\right) \tag{5}$$

where f function impose data integrity. Based on the two optimization problems (4) and (5), the HrHSI is reconstructed by solving the fusion problem in two steps, learning the imaging model operators using (5) and then using the estimated imaging models to solve (4). However, there are limitations to this two-step method. The imaging model is exclusively estimated using data from $\mathbf{P}$ and $\mathbf{Y}$. In practicality, lacking knowledge from $\mathbf{X}$, the estimation of the imaging model could not be correctly resolved. Another simple factorization approach uses a prebuilt CNN like autoencoder to directly train a mapping function between the two inputs $\mathbf{P}$ and $\mathbf{Y}$ estimate the abundance matrix $\mathbf{A}$, which can be expressed as follows:

$$\min_{\Theta} \|\mathbf{X} - f_{\Theta}(\mathbf{P}, \mathbf{Y})\|_F^2 \tag{6}$$

where Θ denotes the learning parameters of the CNN network. Using autoencoder as an example CNN network, (6) assumes the decoder weights as the spectral basis, while the output of the encoder represents the abundance matrix. In this solution, the spectral signatures extracted from $\mathbf{Y}$ and the coefficient matrix extracted from $\mathbf{P}$ are used to produce the target HrHSI. However, like the optimization problem (5), the obtained HrHSI ($\mathbf{X}$) from optimizing problem (6) might not be back-projected to the observed pair $\mathbf{P}$ and $\mathbf{Y}$. The preservation of the spectral and spatial information is therefore not assured. Moreover, most networks are designed to omit unimportant data to determine high-level visual characteristics, resulting in information loss by default and making spectral and spatial preservation more challenging.

B. Architecture of the Proposed Network

Although solving optimization problems (4) and (5) iteratively and alternatively gained good performance [22], it has severe limitations, such as computation complexity and handcrafted tuning of the parameters. Therefore, deep learning can be a brilliant choice for accomplishing the fusion problem. With the revolution of deep learning and its success in image processing and computer vision, many fusion strategies on the basis of deep learning have been introduced lately. Even though these methods achieve superior performance, the black-box fashion of these methods makes their training process hard, and they have no physical interpretation of their internal process.

Lately, an interpretable deep convolutional network with an architecture based on the model-based algorithm steps can profit from the merits of model-based and deep learning techniques. Thereby, in this study, we design Model-Guided Deep Unfolded Networks with Non-local Spatial-Spectral Priors (MGDuNLSS-Net). The proposed network unified the two-step solution mentioned above in an end-to-end manner while the consistency between the output and observed input pair is also incorporated. In other words, the proposed method imposes data consistency. The fusion and imaging models are expressed as follows:

$$\begin{aligned}\mathbf{X}_{\text{out}} &= f_{\theta_1}(\mathbf{P}, \mathbf{Y}) \\ [\mathbf{P}_{\text{out}}, \mathbf{Y}_{\text{out}}] &= f_{\theta_2}(\mathbf{X}_{\text{out}}) \end{aligned} \tag{7}$$

where f_{θ_1} denotes the fusion model and f_{θ_2} is the learning imaging model. θ_1 and θ_2 are the learnable parameters. The proposed method is built on an iterative back-projection refining process. Many image processing tasks, including image reverse filtering [53], picture super-resolution [54], and denoising [55] have exhibited similar concepts. Thereby, the proposed model can be stated as follows:

$$\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)} + f_{\theta_1}\left(\Delta\mathbf{P}^{(t-1)}, \Delta\mathbf{Y}_{up}^{(t-1)}\right) \tag{8}$$

where $\mathbf{X}^{(t)}$ and $\mathbf{X}^{(t-1)}$ represent the estimated HrHSI in the current and previous iterations, respectively. $\Delta\mathbf{P}^{(t-1)}$ represents the map error between the observed $\mathbf{P}$ and the estimated $\mathbf{P}^{(t-1)}$. $\Delta\mathbf{Y}_{up}^{(t-1)}$ denotes the upsampled error between the back-projected $\mathbf{Y}^{(t-1)}$ in the preceding iteration and its corresponding observed LrHSI $\mathbf{Y}$. The following equation can obtain these error maps:

$$\left[\Delta\mathbf{Y}^{(t-1)}, \Delta\mathbf{P}^{(t-1)}\right] = [\mathbf{Y}, \mathbf{P}] - \left[\mathbf{Y}^{(t-1)}, \mathbf{P}^{(t-1)}\right] \tag{9}$$

TABLE I
DETAILS OF SIMULATED LAYERS OF THE PROPOSED METHOD

layer	Type	Size	N-of-filters	Stride
$\mathbf{D}$	Conv	3×3	C	1
$\mathbf{D}^{\text{T}}$	Conv	3×3	K	1
$\mathbf{B}$	Conv	$a \times a$	K	a
$\widetilde{\mathbf{D}}$	Conv	3×3	c	1

In this table, a represents the downsampling factor.

where $\mathbf{Y}^{(t-1)}$ and $\mathbf{P}^{(t-1)}$ can be obtained by f_{θ_2} in (7). By this means, the proposed method has two submodels, one for fusion tasks and the second for learning imaging systems by projecting the estimated fused HrHSI back to the observed pair LHSI and HrMSI, which allows truthful approximation of the imaging models, as can be seen in Fig. 1(a). These two submodels work iteratively and correct each other until convergence. First, to drive the proposed method's imaging model submodule, given the estimated $\mathbf{X}$ at any iteration t, we can factorize it into $\mathbf{A}$ and $\mathbf{D}$. In this regard, thank the subspace representation model's semiunitary advantage, where $\mathbf{D^T D} = \mathbf{I}$ [19], [56], therefore, the coefficients matrix can be obtained as follows:

$$\mathbf{A} = \mathbf{D}^{\text{T}}\mathbf{X}. \tag{10}$$

In our proposed network, we simulate the multiplication process by convolutional filters as stated in [57]; thereby, we use a convolutional layer with filter size $z \times z$ and the number of filters K that can estimate $\mathbf{D}^{\text{T}}$ automatically as a trainable operator through the learning process. Meanwhile, we can obtain the spatial degraded coefficients matrix by applying the degradation blur matrix $\mathbf{B}$ to the obtained $\mathbf{A}$ from (10), according to the imaging model of (2). Similar to a layer used to simulate the transposed subspace, the degradation blur matrix $\mathbf{B}$ is done by convolutional filters. This process can be expressed as the following:

$$\tilde{\mathbf{A}} = \text{conv}_{z\times z}(\mathbf{A}) \tag{11}$$

where $\text{conv}_{z\times z}$ denotes a convolutional layer with kernel size $z \times z$. In this context, we imitate $\mathbf{D}$ and $\widetilde{\mathbf{D}}$ by two convolutional operations of size $z \times z$ with C and c filters, respectively. Table I shows the details of these simulated layers. In a few words, the fused HrHSI at any iteration can be projected back to its corresponding observed pairs after obtaining its coefficients by (10), and the spatial downgraded abundance by (11), and finally, using (2) and (3) to predict LrHSI and HrMSI, respectively. In this way, the proposed model can supervise the degradation process. After obtaining $\mathbf{P}^{(t-1)}$ and $\mathbf{Y}^{(t-1)}$ by back-projection of $\mathbf{X}^{(t-1)}$ at iteration $(t-1)$th, the error between the predicted HrMSI, LrHSI, and its corresponding inputs can be calculated using (9). Subsequently, the error maps are concatenated along the channels dimension and feeding to the fusion model f_{θ_1}. Finally, the latest version of fused HrHSI $(\mathbf{X}^{(t)})$ at iteration t can be estimated by adding the output of f_{θ_1} and $\mathbf{X}^{(t-1)}$ as shown in (8). In the following section, we investigate the fusion submodule of the proposed method.

The fusion submodule seeks to improve the fineness of the predicted HrHSI made by the previous layer by fusing the concatenated error between the observed pair and their corresponding backprojected pair. However, given the concatenated features from the imaging submodule $\mathbf{E} \in \mathbb{R}^{C+c\times M\times N}$, the fusion submodule f_{θ_1} refine the fused HrHSI of the previous layer without losing spatial and spectral fusion according to (8)

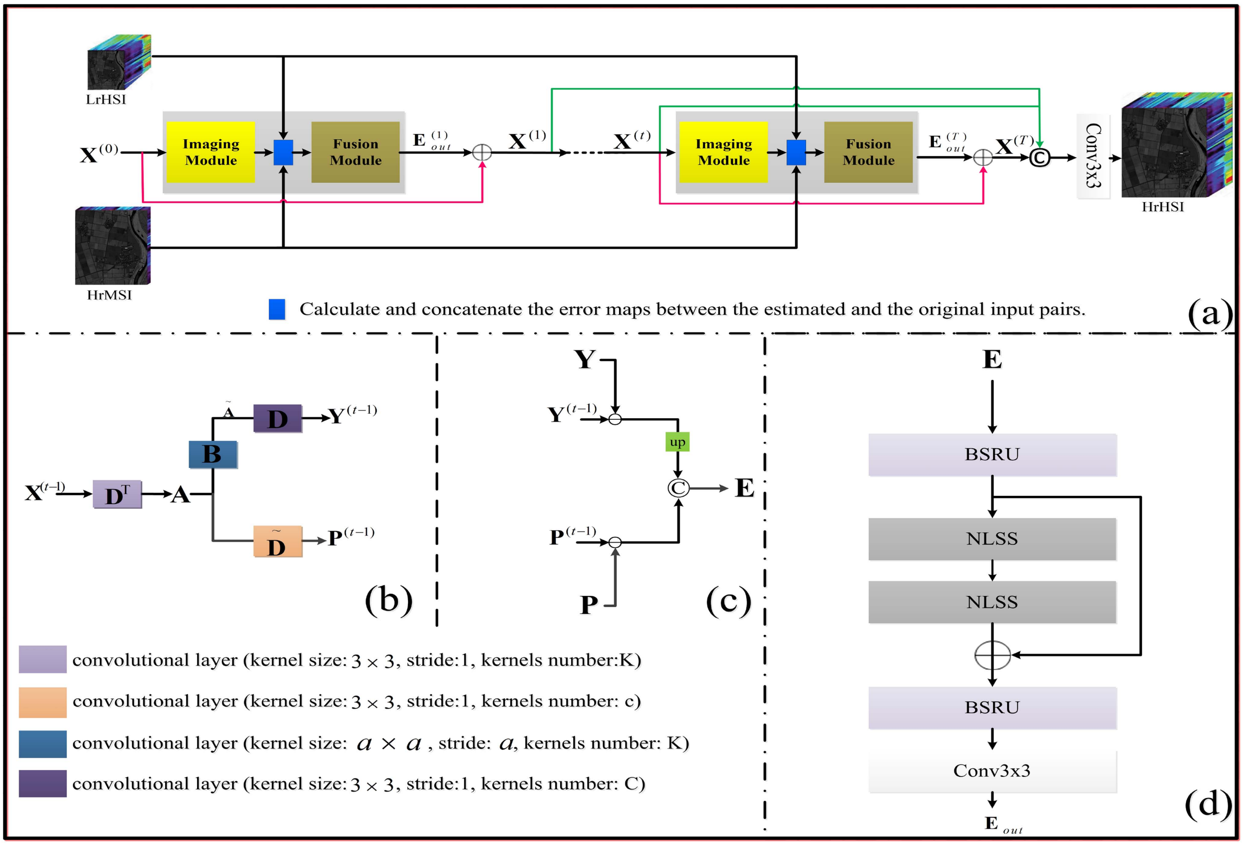

LrHSI
X^(0)
Imaging Module
Fusion Module
E^(1)_out
X^(1)
X^(t)
Imaging Module
Fusion Module
E^(T)_out
X^(T)
Conv3x3
HrHSI
HrMSI
Calculate and concatenate the error maps between the estimated and the original input pairs.
(a)
X^(t−1)
D^T
A
B
Ã
D
Y^(t−1)
D̃
P^(t−1)
(b)
Y
Y^(t−1)
up
E
P^(t−1)
P
(c)
convolutional layer (kernel size: 3 × 3, stride:1, kernels number:K)
convolutional layer (kernel size: 3 × 3, stride:1, kernels number: c)
convolutional layer (kernel size: a × a, stride: a, kernels number: K)
convolutional layer (kernel size: 3 × 3, stride:1, kernels number: C)
E
BSRU
NLSS
NLSS
BSRU
Conv3x3
E_out
(d)

Fig. 1. Illustration of (a) the overall structure of the proposed NLSS-DuNet method, (b) detailed processes in the imaging submodule, (c) calculate and concatenate the error maps of the observed pair and the corresponding estimated pair, and (d) the nonlocal spatial-spectral block (NLSSB) for the fusion submodule. © denotes the concatenation process and ⊕ is elementwise addition.

by using the output features $\mathbf{E}_{\text{out}}$. Three key aspects, global spectral correlation, nonlocal self-similarity, and spatial-spectral correlation, ensure good enhancement of the fused image $\mathbf{X}$. However, we proposed a nonlocal spatial-spectral block (NLSSB) that can fully exploit global spectral correlation, nonlocal self-similarity, and spatial-spectral correlation, as shown in Fig. 1(d). The proposed NLSSB is designed based on a bidirectional simple recurrent unit and a nonlocal self-similarity that can extract and reconstruct the features in different aspects. As can be seen in Fig. 1(d), the architecture of the proposed NLSSB comprises two bidirectional simple recurrent unit layers (BSRU) to guarantee global spectral and spatial-spectral correlation priors. Furthermore, we offer nonlocal self-similarity (NLSS) blocks in the proposed NLSSB to ensure long-range spatial dependencies. However, we embed two NLSS layers between the two BSRU layers. Finally, a convolutional layer ends the proposed NLSSB to make the output feature $\mathbf{E}_{\text{out}}$ has bands number as same as the target HrHSI. In the subsequent two sections, we introduce the specifics of the suggested bidirectional simple recurrent unit (BSRU) and nonlocal self-similarity (NLSS).

1) Bidirectional Simple Recurrent Unit (BSRU): HSI's can be considered a collection of gray images ordered sequentially, with highly spectrally correlated bands. With the success of recurrent neural networks (RNN), it can be instrumental in modeling the prior for sequential data. Therefore, we incorporate the bidirectional simple recurrent unit (BSRU) in the proposed fusion submodule block to take advantage of the global spectral correlation prior to the HSI. However the simple recurrent unit (SRU) is a simple recurrent unit with significant parallel processing and capability for sequence representation, which can be expressed as the following:

$$\begin{aligned} q_i &= \sigma(W_q E_i + u_q \odot s_{i-1} + b_q) \\ s_i &= q_i \odot s_{i-1} + (1 - q_i) \odot (W E_i) \\ r_i &= \sigma(W_r E_i + u_r \odot s_{i-1} + b_r) \\ h_i &= r_i \odot s_i + (1 - r_i) \odot E_i \end{aligned} \tag{12}$$

where q_i and r_i are vectors resulting in the update and reset gates, respectively. s_i and s_{i-1} are interior state vectors for the current and previous time. At the same time, the output is denoted by h_i and $\odot$ is elementwise multiplication. W_q, W, W_r are parameter matrices and u_q, u_r, b_q, and b_r are parameter vectors to be learned during training. σ is the logistic sigmoid function restraining the outputs for taking values ranging from 0 to 1.

In general, SRU contains two submodules [58]. The first submodule is light recurrence, which takes the input E_i and creates the current interior state s_i that symbolizes the sequential information by performing a linear interpolation between the preceding interior state s_{i-1} and the present observation WE_i, where the forget gate q_i controls the weightiness. The current internal state s_i weighted by the reset gate and the input E_i are averaged by the second submodule (highway network) using a skip connection. However, to estimate the spectral correlation of HSI by using SRU where its input is one band from the HSI, the spectral correlation statistics can be exploded from the first band to the ith band by s_i. The spectral correlation can be computed by averaging the correlation features in the former $i-1$ channels and present WE_i using the update gate q_i. At the same time, the way that controls the merge of the spectral statistics enclosed in the recent channel with the spectral correlation statistics in preceding channels is done by using r_i. Visibly, we can consider that the SRU is appropriate for capturing the global spectral correlation among the different channels of HSI's based on the investigation mentioned above.

Based on simple recurrent unit, the network can just process HSIs with a limited geographic dimension after it has been trained due to the elementwise multiplication between u_q and u_r with s_{i-1} that their spatial size and feature dimension are same. Therefore, to overcome this limitation and make the network has fewer training parameters, we can remove these vectors since pointwise multiplication terms have little effect on the capability for representation, as stated in [58]. Furthermore, to improve the capability of representation, we as well do an extra transformation on E_i using W_h rather than directly fusing with it as in SRU. However, after removing bias too, the SRU can be modified and reformulated as follows:

$$\begin{aligned}q_i &= \sigma(W_q E_i)\\ s_i &= q_i \odot s_{i-1} + (1-q_i) \odot (W_s E_i)\\ r_i &= \sigma(W_r E_i)\\ h_i &= r_i \odot s_i + (1-r_i) \odot (W_h E_i).\end{aligned} \tag{13}$$

In (13), the sequential information is calculated using an update gate in the light recurrence as in SRU, while the reset gate is for various purposes that can be regarded as the improvement of the light recurrence, creating the sequential representation more truthful. As stated in [59], we express the update and reset gates as follows:

$$\begin{aligned}Q &= \sigma(W_q * E)\\ R &= \sigma(W_r * E)\end{aligned} \tag{14}$$

where $E \in \mathbb{R}^{L_{in} \times C + c \times M_{in} \times N_{in}}$ and $*$ denote the input features and convolutional operator, respectively. W_q and W_r represent 3-D convolutions of size $3 \times 3 \times 3$. To extract local spatial-spectral correlation efficiently, we introduce 3-D convolution to produce E_s and E_h, this process is expressed as the following:

$$\begin{aligned}E_s &= \tanh(W_s * E)\\ E_h &= \tanh(W_h * E)\end{aligned} \tag{15}$$

where $\tanh$ is the nonlinear activation function. Moreover, in each iteration, from $i = 1$ to the band's number, we take the ith band form Q, R, E_s, and E_h then the feature map is computed as follows:

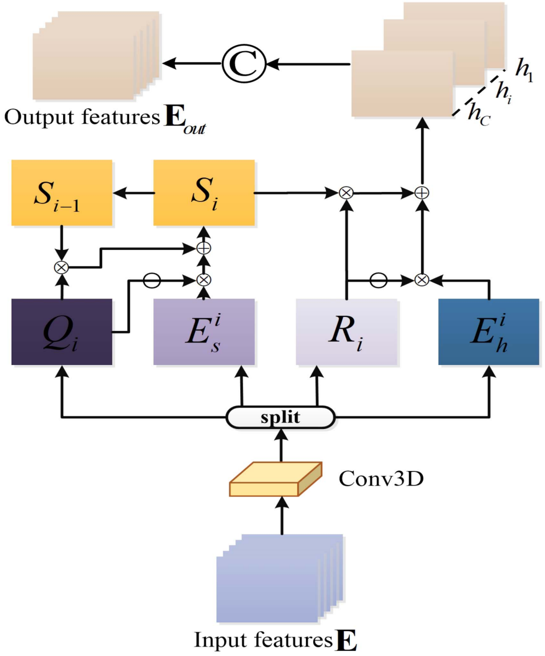


Fig. 2. Whole design of the SRU layer in the proposed network. © denotes concatenation, ⊖ represent subtraction, ⊕ is elementwise addition, and ⊗ is multiplication.

$$\begin{aligned}S_i &= Q_i \odot S_{i-1} + (1-Q_i) \odot E_s^i\\ h_i &= R_i \odot S_i + (1-R_i) \odot E_h^i.\end{aligned} \tag{16}$$

Eventually, E' features map can be achieved by concatenating all h_i, where $i = 1$ to C. Fig. 2 illustrates the detailed operations of the modified SRU, which we can in an arbitrary way, call it SRU.

However, it is visible that we can see (16) computes the global spectral information for an ith band only governed by the information from 1st to $(i-1)$th bands with no care of the information from the $(i+1)$th to the Cth bands. Therefore, to capture more accurate global spectral information that considers all HSI bands of the HSIs, we introduce the bidirectional SRU (BSRU) block in our proposed NLSSB. The BSRU calculates the global spectral information for the ith band in opposite directions from the 1st to $(i-1)$th bands and from Cth to $(i+1)$th bands, then we add them to each other. Moreover, we use two BSRUs as the first layer of the proposed NLSSB and the last layer, which can reconstruct the output features.

2) Nonlocal Self-Similarity (NLSS): We assume the HSIs are gray images linked together, with a global spectral correlation that can be good information for image recovery in the last section. From the spatial point of view, we can regard the HSI as a repeated small batch across all spatial dimensions. Therefore, nonlocal self-similarity prior is favorable for extracting contextual information, which enhances image recovery. That context information inside HSIs cannot be adequately utilized by the classic convolution network, which can only capture the

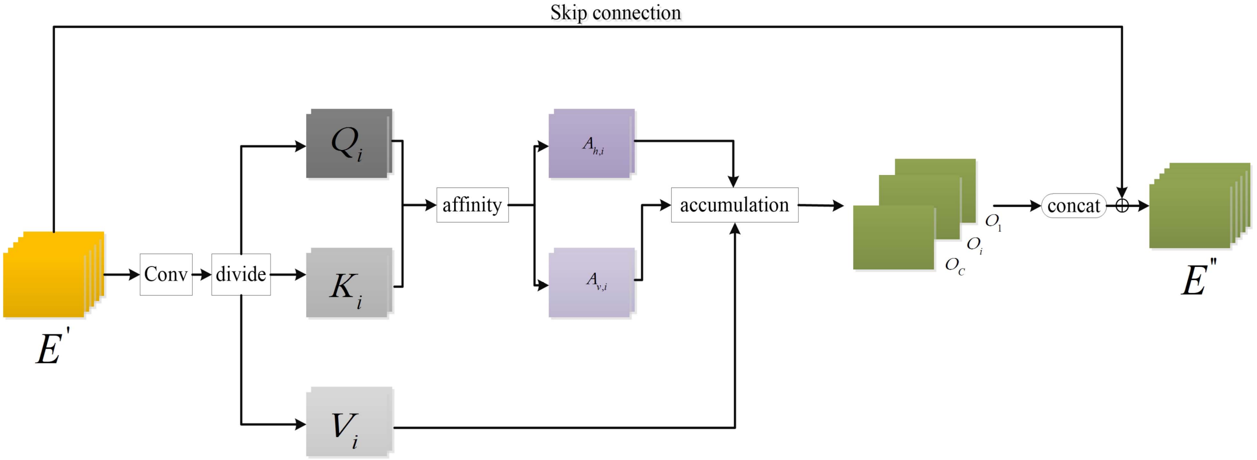


Fig. 3. Nonlocal self-similarity (NLSS).

local context information. However, we embed the nonlocal self-similarity (NLSS) layer in the proposed network, which can utilize the long-range relationships between pixels to their maximum potential for better image enhancement. Various nonlocal prior are proposed models in the literature to utilize the similarity between the pixels in different regions. The nonlocal module proposed in [60] extracts the spatial correlation by computing the distance between each pixel, which is computationally complicated and memory-intensive. Criss-cross attention [61] is less computationally and memory-intensive because it substitutes the typical densely linked graph with numerous subsequent sparsely connected networks. The criss-cross attention module aims to improve the pixelwise representative capacity by gathering context information in both vertical and horizontal directions. Significantly different features receive lower attention weights in the recurrent criss-cross attention module, whereas similar features receive higher attention weights. To this end, we use criss-cross attention to design the NLSS layer in the proposed method.

The proposed NLSS layer is designed based on the criss-cross attention module. However, the NLSS produces a new output feature map for any given input feature map, where the output combines contextual data for every pixel along its criss-cross pathway. In this way, only contextual data in the horizontal and vertical dimensions are included in the output feature map, with ignore the pixels not in the same line vertically or horizontally of the pixel, making it insufficient to capture the nonlocal contextual features through all image's pixels. Therefore, the output feature is passed into the NLSS layer once more and generates the feature map with more affluent and denser context information. Consequently, each pixel in the second output collects contextual information from all pixels. Briefly, the nonlocal self-similarity feature is generated by two NLSS layers, as depicted in Fig. 1(d).

As shown in Fig. 3, three convolutions layers are used to products query, key, and value for the input feature $E' \in \mathbb{R}^{L\times C+c\times M\times N}$ as the following:

$$Q = \text{conv}(E')$$
$$K = \text{conv}(E')$$
$$V = \text{conv}(E') \tag{17}$$

where $Q \in \mathbb{R}^{L'\times C+c\times M\times N}$, $K \in \mathbb{R}^{L'\times C+c\times M\times N}$, and $V \in \mathbb{R}^{L\times C+c\times M\times N}$ represent the query, key, and value, respectively, $L' < L$ to reduce the computation complexity. After the convolutions operation, the query, key, and value are divided into bands to get $\{Q_i, K_i, V_i\}$, where i depicts the number of bands. Furthermore, the horizontal attention map $A_{h,i}$ can be achieved by the affinity process based on Q_i and V_i as follows:

$$\{d_{\text{e},i}\}_j = Q_{\text{e},i}\{\Omega_{\text{e},i}^{\text{T}}\}_j$$
$$A_{h,i} = \text{Softmax}(D_{h,i}) \tag{18}$$

where e represents the location in feature maps, while the jth element extracted from K through the same column with location e is denoted by $\{\Omega_{\text{e},i}^{\text{T}}\}_j$. $d_{\text{e},i}$ is the relationship between $Q_{\text{e},i}$ and Ω_{e}, where $D_{h,i} = \{d_{\text{e},i1}\ldots d_{\text{e},iM}\}$. The vertical correlation map $A_{v,i}$ can be achieved with steps as in (18). Finally, the accumulation process on $\{A_{h,i}, A_{v,i}\}$ and V_i, this process can be formulated as the following:

$$O_{h,i_{\text{e}}} = A_{h,i}\Phi_{i_e}$$
$$O_{v,i_{\text{e}}} = A_{v,i}\Phi_{i_e}$$
$$O_i = O_{h,i} + O_{v,i} \tag{19}$$

where O_i represents the output, Φ_i denotes the collections obtained from V_i, which is in the exact row or column as the location e. After that, we concatenate all $\{O_i\}, i = 1\ldots C$ to obtain the output O with the same number of bands as the input. Eventually, the output feature E'' is obtained by utilizing the skip connections between the input and the output as the following:

$$E'' = O + E'. \tag{20}$$

Fig. 3. shows the detailed operations of the NLSS layer. However, we utilize NLSS in the proposed nonlocal spatial-spectral block by embedding two NLSS layers. In this regard, the two NLSS layers are embedded between the two bidirectional simple recurrent unit layers, as seen in Fig. 1(d).

Finally, as shown in Fig. 1(a), the introduced network generates fused HrHSI at each iteration and enhances it in the next iteration. However, these different versions of the output fused image have a piece of complementary information that can enhance the final fused image. Therefore, the final fused HrHSI can be achieved by the following process:

$$\mathbf{X}_{\text{out}} = \text{conv}\left(\text{cat}\left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(T)}\right)\right) \tag{21}$$

where cat represents the concatenation process and conv is a convolutional layer with filter size 3×3 and the number of features C.

C. Loss Function

The proposed model-guided deep unfolded networks with non-local spatial-spectral priors (MGDuNLSS-Net) need to train in an end-to-end fashion. Therefore, given the learnable parameters of the proposed that denoted Θ and the training samples $\mathbf{X}^{(i)}$, $\mathbf{Y}^{(i)}$, and $\mathbf{P}^{(i)}$; where $i = 1, 2, \ldots, I$ represents the number of training batches, the proposed network's loss function may indeed be expressed as follows:

$$L(\Theta) = \frac{1}{I}\sum_{i}\left\|\text{MGDuNLSS-Net}_{\Theta}(\mathbf{P}^{(i)}, \mathbf{Y}^{(i)}) - \mathbf{X}^{(i)}\right\|_1 \tag{22}$$

where $\|.\|_1$ denotes L_1 loss function, and Θ is the trainable parameters of MGDuNLSS-Net.

IV. Experimental Results

In this part, the effectiveness of our proposed MGDuNLSS-Net is verified by comprehensive experiments on three benchmarks remote sensing HSI databases and compares its performance with different last state-of-the-art HrMSI-LrHSI fusion methods. Moreover, we examine and demonstrate the computation time of the proposed network compared to previous comparison works.

A. Experimental Datasets

Three benchmarks of remote sensing datasets, namely, Pavia Center [62], Chikusei [63], and Cuprite Mine [64] datasets, are utilized to validate the proposed technique. A brief description and the setting of these databases are given below.

1) Pavia Center Dataset: This dataset is taken by ROSIS sensor for the period of a flight campaign over the urban area of Pavia, Northern Italy. 1096×715 pixels are the spatial size of this HSI, with 102 spectral channels. The spectral band domain is 430 to 860 nm, while 1.3 m is its geometric resolution. After removing the bands with low SNR, 93 spectral bands are preserved for our experiment. We cropped 250×250 pixels for the testing dataset, and nonoverlapping residual pixels are employed for training and validation. The original HSI is used as reference; however, to simulate the LrHSI from the reference dataset, we apply the Gaussian filter of size 7×7 using standard deviation 2 and downsampled both vertical and horizontal dimensions by 5 pixels. Meanwhile, the HrMSI is simulated by an IKONOS-like reflectance spectral response. Both LrHSI and HrMSI are corrupted by 30 dB and 35 dB i.i.d Gaussian noise, respectively. For training, we divided the training image into batches with the size of $40 \times 40 \times 3$, $8 \times 8 \times 93$, and $40 \times 40 \times 93$ for LrHSI, HrMSI, and HrHSI, respectively.

2) Chikusei Dataset: Headwall Hyperspec-VNIR-C imaging sensor observed the Chikusei HSI over urban and agricultural zones in Chikusei with spatial size 2517×2332 pixels and 128 spectral bands. The spectral bands span the wavelength range from 363 to 1018 nm. For convenience, we clipped 2048×2048 pixels for the experiment. Then we take 512×512 pixels for testing and the remaining pixels for training and validation. While the original HSI is saved as a reference HSI for training and testing, Gaussian filter of size 7×7 with a standard deviation of 2 is used to simulate the LrHSI and subsampling every 8 pixels in the spatial dimensions. Moreover, the IKONOS satellite spectral response function is applied to generate the HrMSI. After then, we added an i.i.d Gaussian noise (35 and 30 dB) to HrMSI and LrHSI, respectively. As well we train the network in small blocks of size $64 \times 64 \times 3$, $8 \times 8 \times 128$, and $64 \times 64 \times 128$, for HrMSI, LrHSI, and HrHSI, respectively.

3) Cuprite Mine: This dataset is observed by the AVIRIS over Nevada, United States. The spatial resolution of Cuprite Mine HSI is 512×512 with 224 spectral bands. With an interval of 10, the spectral bands covered wavelengths ranging from 400 to 2500 nm. After taking out the channels with low SNR and water absorptions (1,2, 105–115, 150–170, 223–224), 188 spectral channels remain for our experiment. 256×256 pixels are selected for testing, and then we train the proposed network on the remainder pixels. The LrHSI is generated by applying a 7×7 Gaussian kernel with std 2 and downsampled by factor 4 in both spatial directions. The six bands that matched the visible and mid-infrared range spectral bands of the USGS/NASA Landsat7 satellite (480, 560, 660, 830, 1650, and 2220 nm) are used directly to simulate the HrMSI. An i.i.d Gaussian noise (30 and 35 dB) is used to corrupt the simulated LrHSI and HrMSI, respectively. The training blocks size of $32 \times 32 \times 6$, $8 \times 8 \times 188$, and $32 \times 32 \times 188$ for HrMSI, LrHSI, and HrHSI, respectively.

B. Comparison Methods

Seven state-of-the-art fusion techniques are used in the comparison with our proposed method to examine the performance of the proposed technique further. Hyperspectral subspace-based regularized fusion (HySure) method [22], coupled sparse tensor factorization (CSTF) method [65], coupled nonnegative matrix factorization (CNMF) method [34], coupled spectral unmixing (CSU) method [66], regularizing HSI and MSI fusion by CNN denoiser (CNN-Fus) method [56], deep spatiospectral attention CNNs (HSRNet) method [67], and multihierarchical cross transformer for hyperspectral and multispectral image fusion (MCT-NET) method [68]. The former four methods are model-driven with no training, and the later three methods are based on deep learning techniques, which need to train them using the training data. The implementation codes of all these methods are provided by the authors and are publicly available. For a fair comparison, we use the testing dataset for these methods as the same as in the proposed method, and they are evaluated with the same metrics.

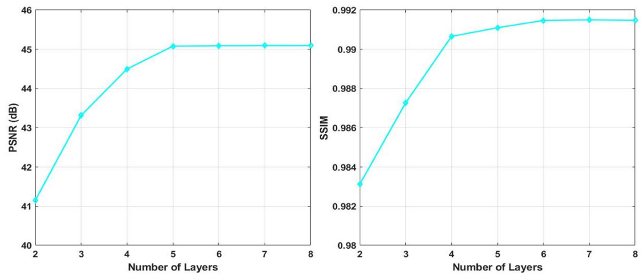


Fig. 4. PSNR and SSIM curves are functions of the number of layers for Chikusei dataset.

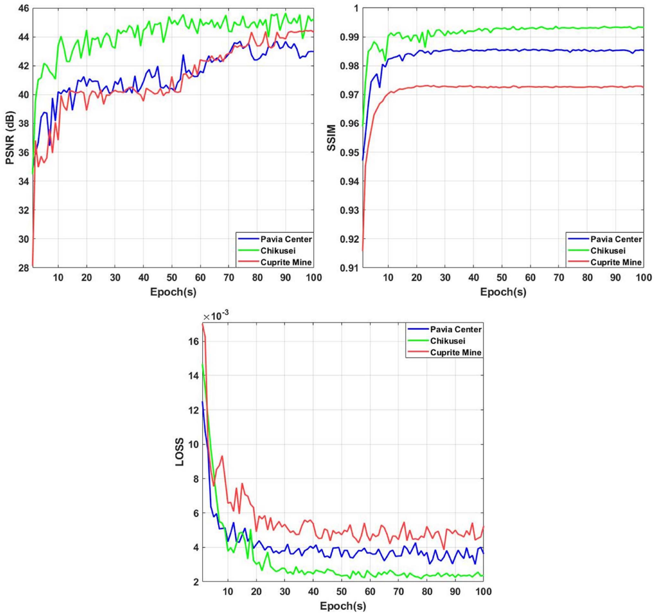


Fig. 5. Average PSNR, SSIM, and loss function curves are functions of the proposed method's epochs for the training datasets.

TABLE II
INFLUENCE OF DIFFERENT INITIALIZATION OF MGDUNLSS-NET TESTED ON PAVIA CENTER DATASET [62]

Initialization	Pavia Center dataset [62]				
	PSNR (dB)	SAM	ERGAS	UIQI	SSIM
Best value	∞	0	0	1	1
Random	43.53	2.46	2.1061	0.9886	**0.9884**
Zero	**43.74**	**2.37**	**2.0850**	**0.9896**	0.9882

The best results are typed in bold.

TABLE III
QUANTITATIVE PERFORMANCE MEASUREMENTS (PSNR (DB), SAM, ERGAS, UIQI, AND SSIM) OF THE PROPOSED MGDUNLSS-NET IN COMPARISON WITH STATE-OF-THE-METHODS (HYSURE, CNMF, CSTF, CSU, HSRNET, MCT-NET, CNN-FUS) ON THE PAVIA CENTER DATASET

Mehtod (s)	Pavia Center dataset [62]				
	PSNR (dB)	SAM	ERGAS	UIQI	SSIM
Best value	∞	0	0	1	1
HySure [22]	39.69	5.05	3.0483	0.9789	0.9673
CNMF [34]	39.95	4.69	3.2649	0.9715	0.9746
CSTF [65]	40.22	3.29	3.5048	0.9806	0.9810
CSU [66]	41.01	3.32	2.9373	0.9828	0.9776
HSRNet [67]	43.26	2.49	2.2900	0.9837	0.9846
MCT-NET [68]	41.53	2.92	2.8935	0.9718	0.9791
CNN-Fus [56]	43.39	2.61	2.6559	0.9889	0.9859
MGDuNLSS-Net	**43.74**	**2.37**	**2.0850**	**0.9896**	**0.9882**

The optimum result is written in bold.

TABLE IV
QUANTITATIVE PERFORMANCE MEASUREMENTS (PSNR (DB), SAM, ERGAS, UIQI, AND SSIM) OF THE PROPOSED MGDUNLSS-NET IN COMPARISON WITH STATE-OF-THE-METHODS (HYSURE, CNMF, CSTF, CSU, HSRNET, MCT-NET, CNN-FUS) ON THE CHIKUSEI DATASET

Mehtod (s)	Chikusei dataset [63]				
	PSNR (dB)	SAM	ERGAS	UIQI	SSIM
Best value	∞	0	0	1	1
HySure [22]	41.32	3.13	2.4966	0.9782	0.9703
CNMF [34]	42.50	3.48	2.6179	0.9791	0.9761
CSTF [65]	43.09	2.96	2.3182	0.9834	0.9846
CSU [66]	42.75	2.91	2.9373	0.9808	0.9792
HSRNet [67]	43.96	2.53	1.6942	0.9851	0.9839
MCT-NET [68]	43.28	2.39	1.7787	0.9812	0.9827
CNN-Fus [56]	44.73	2.27	1.8634	0.9844	0.9870
MGDuNLSS-Net	**45.09**	**2.10**	**1.5970**	**0.9903**	**0.9916**

The optimum result is written in bold.

C. Quality Measurement Metrics

Given the predicted fused ($\mathbf{X}_{\text{out}}$) and target ground-truth ($\mathbf{X}$) images, the five quantitative criteria used to validate the effectiveness of the performance of the MGDuNLSS-Net approach are listed below:

1) Peak Signal-to-Noise Ratio (PSNR): The PSNR is utilized to measure how similar the predicted fused image and the reference image are, and this metric can be formulated as follows:

$$\text{PSNR}(\mathbf{X}, \mathbf{X}_{\text{out}}) = \frac{1}{C} \sum_{t=i}^{C} \text{PSNR}\left(\mathbf{X}^{(i)}, \mathbf{X}_{\text{out}}^{(i)}\right). \tag{23}$$

With the increasing of PSNR value, the quality of the reconstructed image is better, and ∞ is the optimum value of PSNR.

2) Universal Image Quality Index (UIQI): The best value of UIQI is one, which is calculated on a sliding window and then takes the mean across whole windows and all spectral channels. The UIQI of two windows x and x_{out} can be calculated as the following:

$$\text{UIQI}(x, x_{\text{out}}) = \frac{4\mu_x \mu_{x_{\text{out}}}}{\mu_x^2 + \mu_{x_{\text{out}}}^2} \frac{\sigma_{x,x_{\text{out}}}^2}{\sigma_x^2 + \sigma_{x_{\text{out}}}^2} \tag{24}$$

where μ and σ represent the mean and standard variance, respectively. $\sigma_{x,x_{\text{out}}}^2$ is the covariances between x and x_{out}.

3) Relative Dimension Global Error in Synthesis (ERGAS): An overall measure of the predicted image's quality is calculated via a global index, and its best value is zero. The formula of ERGAS is written as the following:

$$\text{ERGAS}(\mathbf{X}, \mathbf{X}_{\text{out}}) = \frac{100}{d} \sqrt{\frac{1}{C} \sum_{i=1}^{C} \frac{\text{MSE}(\mathbf{X}^{(i)}, \mathbf{X}_{\text{out}}^{(i)})}{\mu(\mathbf{X}_{\text{out}}^{(i)})^2}} \tag{25}$$

where μ denotes the mean value of $\mathbf{X}_{\text{out}}^{(i)}$.

4) Structural Similarity Index (SSIM): The SSIM is a perceptual metric that determines how much the fusion process has degraded the quality of the estimated fused image. The SSIM of

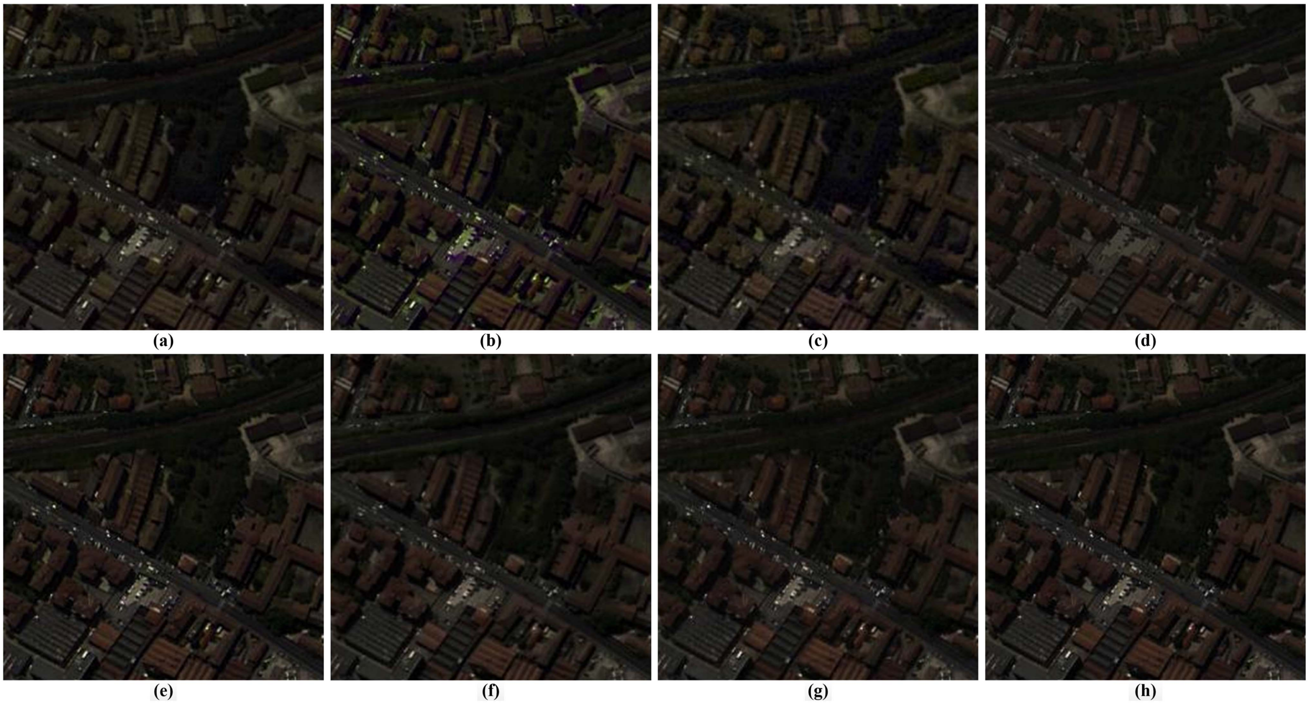


Fig. 6. Pseudocolor of the outcomes of the different experiments on the Pavia Center dataset. The obtained fused image has a spatial size of 250 × 250 (a) Fused image of HySure [22]. (b) Fused image of CNMF [34]. (c) Fused image of CSTF [65]. (d) Fused image of CSU [66]. (e) Fused image of HSRNet [67]. (f) Fused image of MCT-NET [68]. (g) Fused image of CNN-Fus [56]. (h) Fused image of MGDuNLSS-Net.

the overall bands can be expressed as

$$\mathrm{SSIM}(\mathbf{X}, \mathbf{X}_{\text{out}}) = \frac{1}{C}\sum_{i=1}^{C} \mathrm{SSIM}\left(\mathbf{X}^{(i)}, \mathbf{X}_{\text{out}}^{(i)}\right). \quad (26)$$

A higher value means better spatial features maintenance of the reconstructed image. One is the optimal value of SSIM.

5) Spectral Angle Mapper (SAM): SAM is used to calculate the spectral angle mapper among two vectors, which is the absolute value of the spectral angle (degree) between them. The optimum assessment amount of SAM is zero; therefore, there will not be any spectral distortion. The following equation formulates this process:

$$\mathrm{SAM}(\mathbf{X}, \mathbf{X}_{\text{out}}) = \frac{1}{N}\sum_{i=1}^{N} \operatorname{arcos}\frac{{\mathbf{X}_{\text{out}}^{(i)}}^{\mathrm{T}}\mathbf{X}^{(i)}}{\left\|\mathbf{X}_{\text{out}}^{(i)}\right\|_2 \left\|\mathbf{X}^{(i)}\right\|_2}. \quad (27)$$

D. Implementation Details

For the implementation of the proposed MGDuNLSS-Net method, which is an end-to-end network with learnable parameters that need to be initialized first, we used xavier_normal to initialize all learnable parameters. Although the random (from a uniform distribution on the interval [0,1]) and zero initializations of $\mathbf{X}^{(0)}$, $\mathbf{P}^{(0)}$, and $\mathbf{Y}^{(0)}$ have no big difference, zero initiation still work better for the proposed method, therefore, $\mathbf{X}^{(0)}$, $\mathbf{P}^{(0)}$, and $\mathbf{Y}^{(0)}$ are simply initialized by zero (please refer to Table II). The loss function is optimized by the ADAM optimizer with a momentum equal to 0.999, where a learning rate is initialized by 10^{-3} and halved every five epochs. The number of layers of the proposed MGDuNLSS-Net is set to five layers according to Fig. 4, which can balance performance and computational cost, while 16 is the number of mini-batch during the training process. According to Fig. 5, which displays the PSNR and SSIM as a function of training epochs, we set the number of training epochs equal to 100. Pytorch 1.12 is used to implement all experiments with NVIDIA GeForce RTX 3090.

E. Experimental Results

The comparison of performance using the Pavia Center dataset: Table III demonstrates the quantitative outcomes of the testing images in terms of PSNR (dB), SAM, ERGAS, UIQI, and SSIM, where the optimum score are typed in bold among the compression methods for clarity. This table demonstrates how, across all assessment metrics, our proposed MGDuNLSS-Net approach may significantly outperform other competing approaches with a minor variance between the reference image and the fused outcome. It is appealed that our techniques can better retain spectral and spatial data. According to this table, the proposed method obtained a minimum ERGAS value, meaning it has the slightest dynamic variation and shift of the fused image compared to its corresponding reference image. Furthermore, the fused method of the proposed method has the optimal spectral distribution of intensities compared to the other methods according to the obtained SAM values of different approaches. The proposed MGDuNLSS-Net method has better details preservation of the spatial structure as stated by the SSIM.

Moreover, the visual comparison of the outcome images and the corresponding error maps of the competing approaches for the testing images are shown in Figs. 6 and 7. Fig. 6 describes the pseudocolor (bands 30 for red, 20 for green, and 5 for

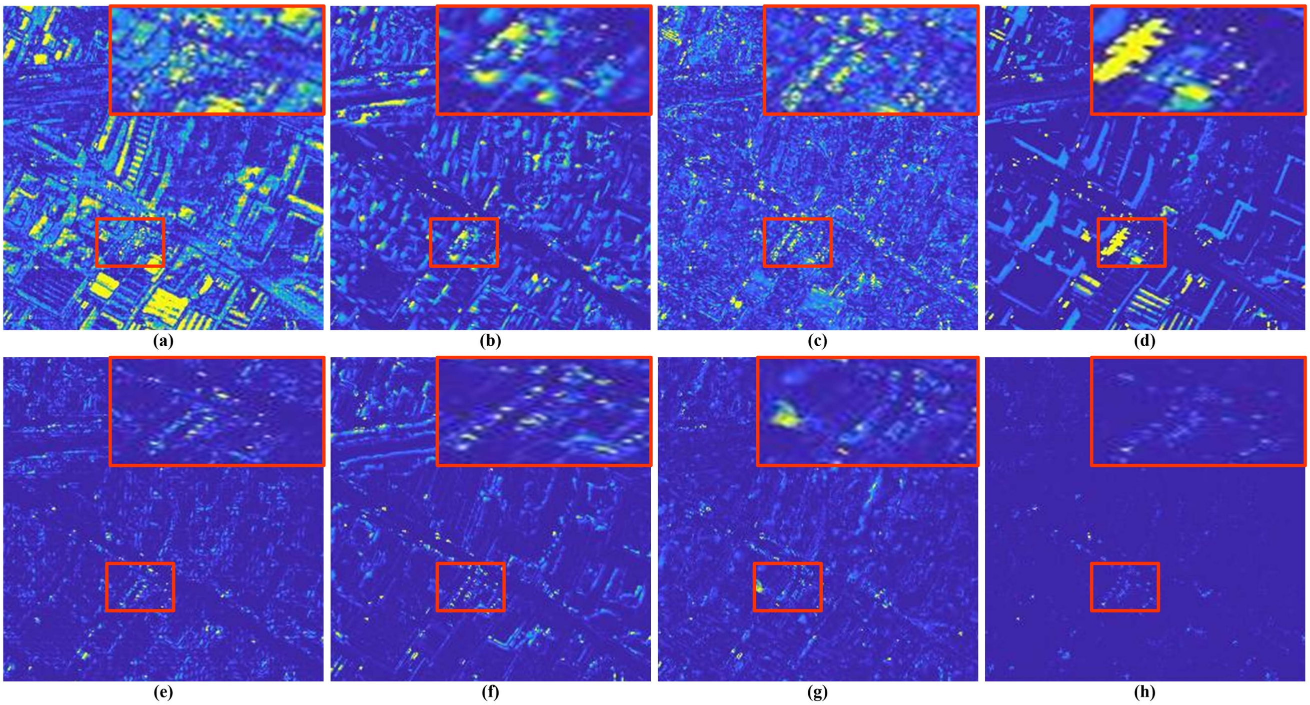


Fig. 7. Absolute errors map of the fusion results from Pavia Center database shown in Fig. 6. (a) HySure [22]. (b) CNMF [34]. (c) CSTF [65]. (d) CSU [66]. (e) HSRNet [67]. (f) MCT-NET [68]. (g) CNN-Fus [56]. (h) MGDuNLSS-Net.

blue) of the obtained fused images of different methods and the corresponding reference of the testing image from this dataset. To better show the difference between these fused images from the different approaches, the absolute error maps between the outcomes fused images and the reference are depicted in Fig. 7. To clarify the variance of the different methods more, the meaningful region of error images for the various techniques are zoomed and marked in a red box. It is obvious from the red boxes of the error maps that the proposed MGDuNLSS-Net approach achieves better in terms of less distortion and reconstructing the detailed structures than other comparison approaches, which confirms the value of the SSIM obtained by various comparison methods.

The comparison of performance using the Chikusei dataset: The objective outcomes of the testing image of this database in terms of PSNR (dB), SAM, ERGAS, UIQI, and SSIM are displayed in Table IV, with the optimum value written in bold. In this table, the proposed MGDuNLSS-Net has the maximum results in terms of PSNR, UIQI, and SSIM and minimum results in terms of SAM and ERGAS. According to these quantitative metrics, the proposed MGDuNLSS-Net attained better spatial structure preservation between the output and the reference image, lesser spectral distortion, least change, and dynamic shift compared to the previous works tested in this article. Fig. 8 depicts the visual result of the obtained fused image for the different methods and the reference of the testing data, displayed in false color by bands 90, 70, and 40 for red, green, and blue, respectively. This figure demonstrates that the proposed approach's result among the other approaches is closest to the reference image with better reconstructing spatial structures details. In order to reconnoiter the difference more visually, we showed the absolute error maps between the obtained outcome images from all approaches and the reference image in Fig. 9. It is evident that the error map between the fused image obtained by the proposed and the reference is the least fused error at both the smooth areas and edges of the image, which means is impeccably consolidates the quantitative measurements shown in Table IV.

TABLE V
QUANTITATIVE PERFORMANCE MEASUREMENTS (PSNR (DB), SAM, ERGAS, UIQI, AND SSIM) OF THE PROPOSED MGDUNLSS-NET IN COMPARISON WITH STATE-OF-THE-METHODS (HYSURE, CNMF, CSTF, CSU, HSRNET, MCT-NET, CNN-FUS) ON THE CUPRITE MINE DATASET

Mehtod (s)	Cuprite Mine dataset [64]				
	PSNR (dB)	SAM	ERGAS	UIQI	SSIM
Best value	∞	0	0	1	1
HySure [22]	38.36	1.46	3.6323	0.9249	0.9557
CNMF [34]	43.07	1.59	2.8013	0.9183	0.9677
CSTF [65]	41.83	1.62	4.1927	0.9207	0.9496
CSU [66]	41.92	1.43	3.7339	0.9363	0.9621
HSRNet [67]	43.96	1.29	2.7088	0.9435	0.9708
MCT-NET [68]	43.39	1.26	2.7903	0.9416	0.9681
CNN-Fus [56]	43.84	1.37	2.9519	0.9439	0.9713
MGDuNLSS-Net	**44.17**	**1.12**	**2.4131**	**0.9486**	**0.9764**

The optimum result is written in bold.

The comparison of performance using the Cuprite Mine dataset: In order to validate the robustness of the proposed MGDuNLSS-Net's performance, we tested and investigated its performance on the Cuprite Mine database, which has the lowest number of training samples compared to the Pavia Center and Chikusei databases. Table V shows the numerical metrics in terms of PSNR (dB), SAM, ERAS, UIQI, and SSIM. Still, the

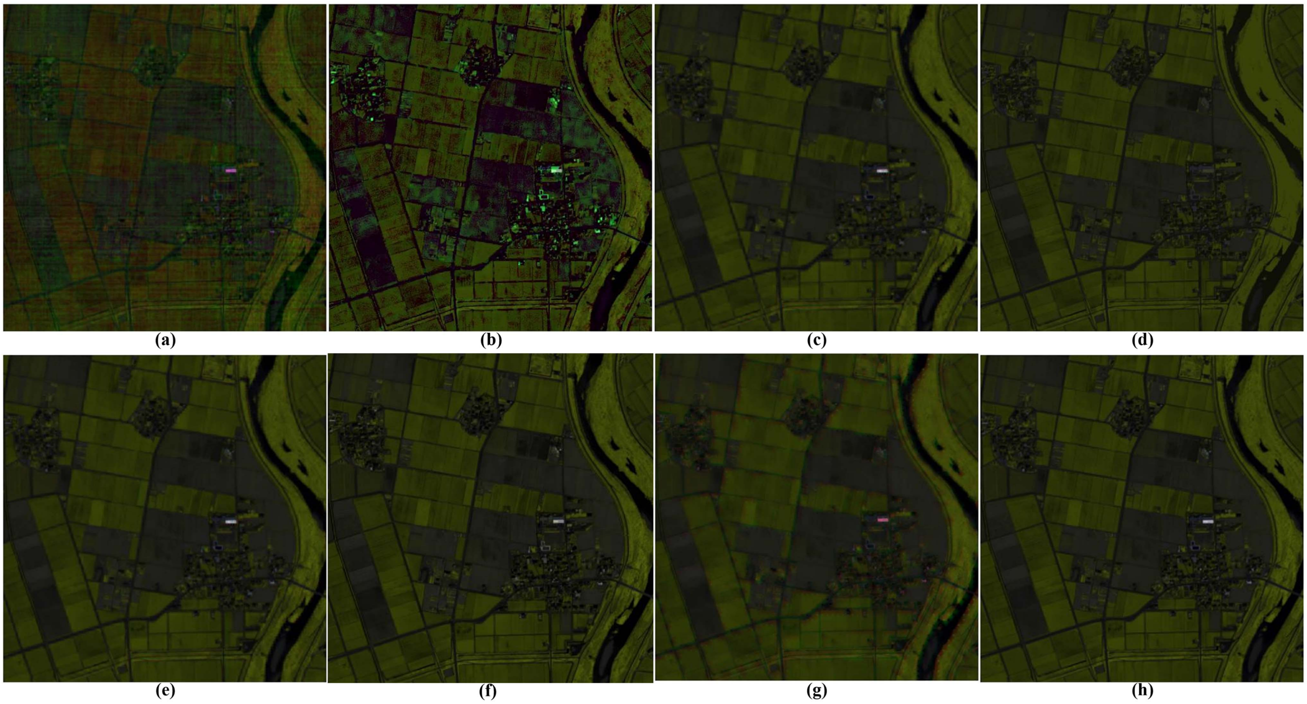


Fig. 8. Pseudocolor of the outcomes of the different experiments on the Chikusei dataset. The obtained fused image has a spatial size of 512 × 512. (a) Fused image of HySure [22]. (b) Fused image of CNMF [34]. (c) Fused image of CSTF [65]. (d) Fused image of CSU [66]. (e) Fused image of HSRNet [67]. (f) Fused image of MCT-NET [68]. (g) Fused image of CNN-Fus [56]. (h) Fused image of MGDuNLSS-Net.

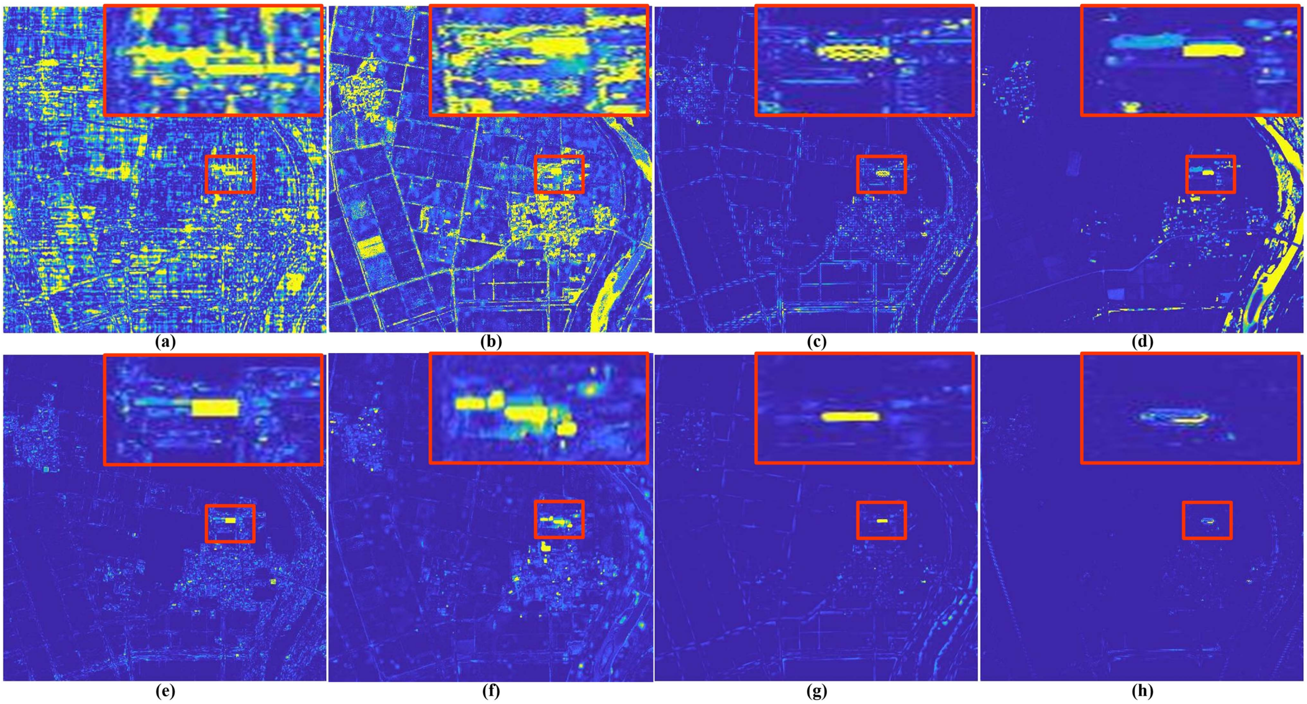


Fig. 9. Absolute errors map of the fusion results from Chikusei database shown in Fig. 8. (a) HySure [22]. (b) CNMF [34]. (c) CSTF [65]. (d) CSU [66]. (e) HSRNet [67]. (f) MCT-NET [68]. (g) CNN-Fus [56]. (h) MGDuNLSS-Net.

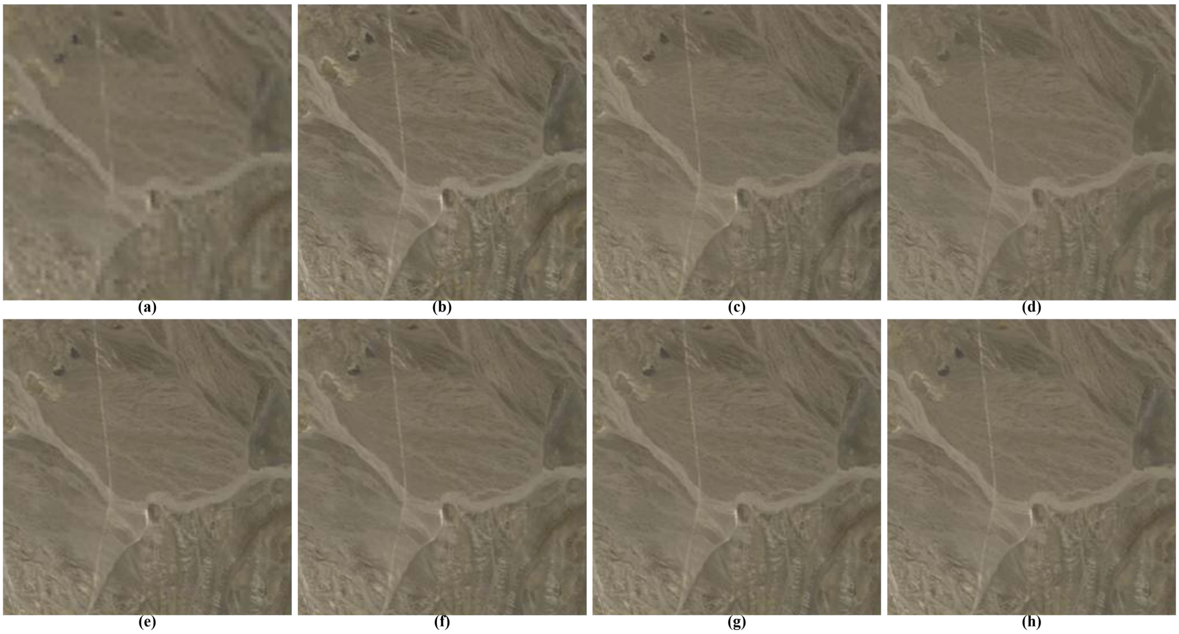


Fig. 10. Pseudocolor of the outcomes of the different experiments on the Cuprite Mine dataset. The obtained fused image has a spatial size of 256 × 256. (a) Fused image of HySure [22]. (b) Fused image of CNMF [34]. (c) Fused image of CSTF [65]. (d) Fused image of CSU [66]. (e) Fused image of HSRNet [67]. (f) Fused image of MCT-NET [68]. (g) Fused image of CNN-Fus [56]. (h) Fused image of MGDuNLSS-Net.

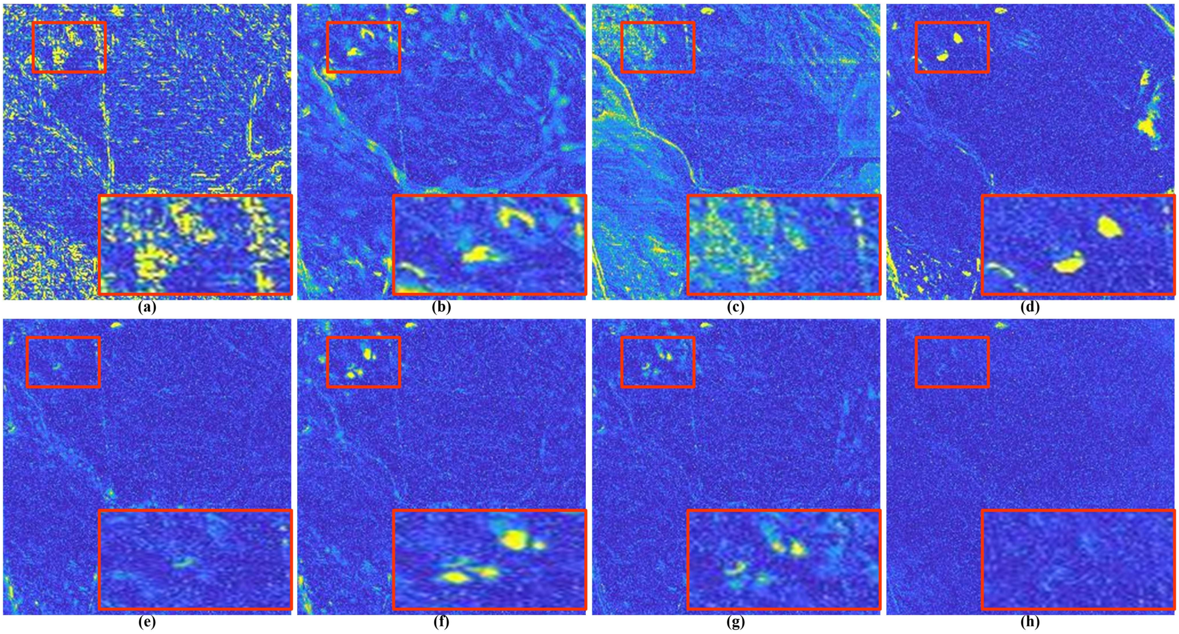
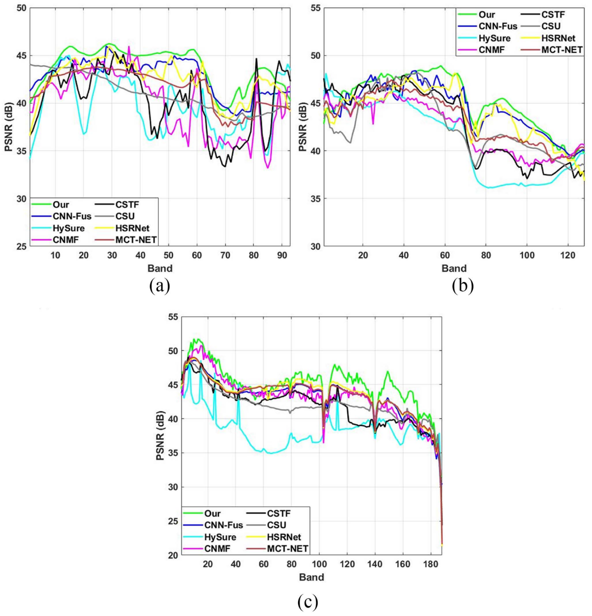


Fig. 11. Absolute errors map of the fusion results from Cuprite Mine database shown in Fig. 10. (a) HySure [22]. (b) CNMF [34]. (c) CSTF [65]. (d) CSU [66]. (e) HSRNet [67]. (f) MCT-NET [68]. (g) CNN-Fus [56]. (h) MGDuNLSS-Net.

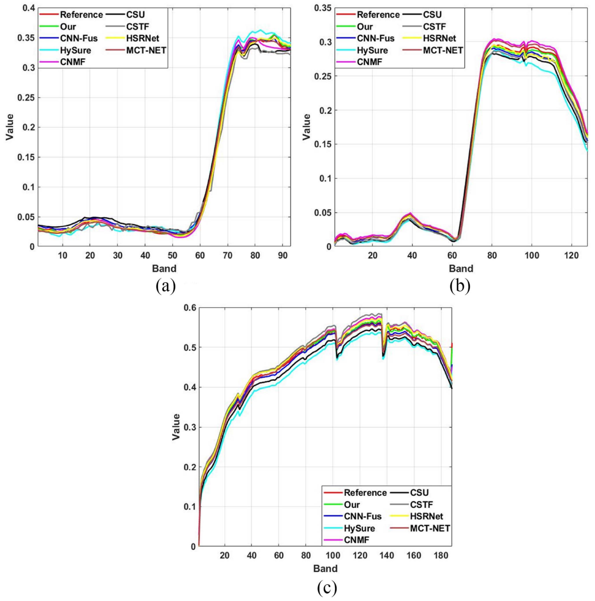
proposed method in this article has the best evaluation values among the quantitative measurements, with the lowest PSNR, SAM, and ERGAS and higher values in terms of UIQI and SSIM compared to other comparison techniques. To explore the performance of the proposed MGDuNLSS-Net and comparison approaches, we showed the fused images gained by these methods in Fig. 10. The RGB images of the fused images and the ground truth image shown in Fig. 10 are contained bands 70, 60, and 30 for red, green, and blue, respectively. To demonstrate the difference in the visual results, the errors between these achieved images and the reference shown in Fig. 11. Obviously, it can be seen from the marked error images that the obtained image of the proposed MGDuNLSS-Net approach is closest to the reference image than the other testing approaches, and it attains



(a) (b)

(c)

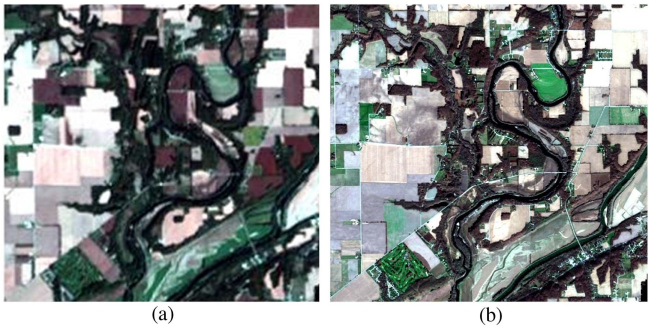
Fig. 12. Average PSNR curves as functions of all bands for the HSI reconstructed by the test methods. (a) Pavia Center dataset. (b) Chikusei dataset. (c) Cuprite Mine dataset.



(a) (b)

(c)

Fig. 13. Spectral signature of single-pixel for the test methods compared with the reference. (a) Pixel (125, 175) of testing HSI image from the Pavia Center dataset. (b) Pixel (83, 423) of the testing HSI image from the Chikusei dataset. (c) Pixel (142, 102) of test HSI image from the Cuprite Mine dataset.



(a) (b)

Fig. 14. Pseudocolor of the testing data of the real dataset Hyperion-Sentinel. (a) 210 × 210 pixels of LrHSI (bands 20, 5, 3). (b) 630 × 630 pixels of HrMSI (bands 3, 2, 1).

TABLE VI
Ablation Study of the Effectiveness of Different Architectures in MGDuNLSS-Net

BSRU	NLSS	Pavia Center dataset [62]				
		PSNR (dB)	SAM	ERGAS	UIQI	SSIM
Best value		∞	0	0	1	1
×	✓	42.27	2.51	2.5019	0.9790	0.9877
✓	×	42.91	2.39	2.2135	0.9884	0.9879
✓	✓	**43.74**	**2.37**	**2.0850**	**0.9896**	**0.9882**

The best results are typed in bold.

minimum restoration error at both the smooth areas edges of the testing image. This means that the proposed MGDuNLSS-Net method can more satisfactorily reconstruct the spatial details with better spectral maintenance than the other comparison methods.

Moreover, to further compare the fusion quality, we validate the performance of the proposed MGDuNLSS-Net method in terms of PSNR over all bands of the three testing datasets. Fig. 12 shows the PSNR (dB) value of the introduced technique with compare to the other testing techniques. It can be seen clearly that the proposed MGDuNLSS-Net outperformed the other comparison in most of the spectral bands across three databases. Furthermore, the spectral response attained by the different methods is validated in terms of pixel values of the fused images compared to the values of pixels from the ground truth. To this end, we select the pixels (125, 175), (83, 423), and (142, 102) from the testing part of the Pavia Center, Chikusei, and Cuprite Mine datasets, respectively, and displayed their values in Fig. 13. The curves of these pixel values show that the proposed method has similar and closed pixel values to the reference pixels compared to the other methods, which proves the MGDuNLSS-Net method's performance in this aspect.

F. Ablation Analysis

Through an ablation study, this section investigates the effect and importance of the bidirectional simple recurrent unit (BSRU) and the nonlocal self-similarity (NLSS) layers of the proposed MGDuNLSS-Net on Pavia center dataset with scale = 5. In this regard, we trained the proposed network in three different architectures. The first architecture includes only two nonlocal self-similarity (NLSS) layers in the fusion submodule, while in the second version, we built the fusion submodule by two bidirectional simple recurrent unit layers (BSRU) and removed the NLSS layers. Finally, the proposed method is trained in which the fusion submodule contains two NLSS blocks inserted between two BSRU layers. The quantitative results of these three different architectures are shown in Table VI, demonstrating the role of the proposed components of the fusion submodule (NLSS and BSRU).

TABLE VII
TIME EFFICIENCY OF PAVIA CENTER, CHIKUSEI, AND CUPRITE MINE DATASETS, RESPECTIVELY, AND GENERAL COMPARISON OF THE TESTING METHODS, **BOLD** TYPING MEANS BETTER

Method (s)	Advantage from model-based	Advantage from DL	Iterations	Supervise the degradation process	Running Time (S)
HySure [22]	**Yes**	No	Yes	No	208 / 362 / 271
CNMF [34]	**Yes**	No	Yes	No	96 / 131 / 118
CSTF [65]	**Yes**	No	yes	No	64 / 107 / 82
CSU [66]	**Yes**	No	yes	No	192 / 395 / 229
HSRNet [67]	No	**Yes**	**No**	No	2.48 / 3.16 / 2.733
MCT-NET [68]	No	**Yes**	**No**	No	8.04 / 13.68 / 10.82
CNN-Fus [56]	**Yes**	**Yes**	Yes	No	49 / 72 / 61
MGDuNLSS-Net	**Yes**	**Yes**	**No**	**Yes**	**0.901 / 1.337 / 1.156**

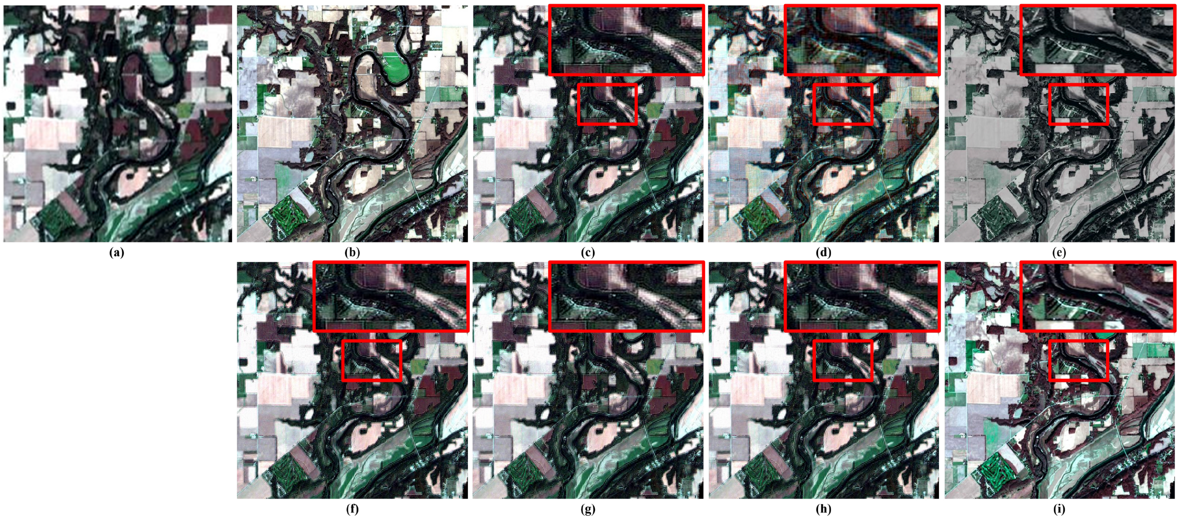


Fig. 15. Pseudocolor contains bands (R: 20, G: 5, and B: 3) of the outcomes of the different experiments on real datasets Hyperion-Sentinel. The obtained fused image has a spatial size of 630 × 630, where the Geometric resolution is 10 m. (a) Original LrHSI. (b) Original HrMSI. (c) Fused image of CNMF [34]. (d) Fused image of CSTF [65]. (e) Fused image of CSU [66]. (f) Fused image of HSRNet [67]. (g) Fused image of MCT-NET [68]. (h) Fused image of CNN-Fus [56]. (i) Fused image of MGDuNLSS-Net.

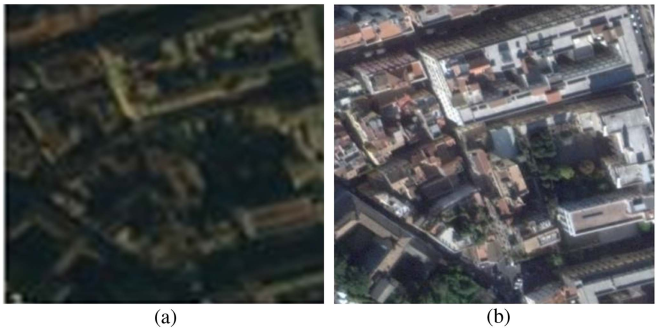


Fig. 16. Pseudocolor of the testing data of the real dataset WV2. (a) 90 × 90 pixels of LrMSI (bands 5, 3, 2). (b) 360 × 360 pixels of HrRGB.

G. Time Complexity of the Different Tested Methods

To compare the effectiveness of the testing approaches in terms of computational complexity, the running time of the different testing methods on the Pavia Center, Chikusei, and Cuprite Mine datasets is stated in Table VII. As shown in this table, the proposed MGDuNLSS-Net has efficient computation time. Specifically, HySure, CNMF, CSTF, and CSU are model-based methods that they need many iterations to converge. Therefore, the computation complexity mainly comes from the iteration need for these approaches. For CNN-Fus, while it benefits from model-based and deep learning but still has the model-based part needs iterations to complete the fusion process. In a few words, the CNN-Fus approach is not an end-to-end deep-learning framework. Although the proposed method is formulated in model-based problems, the model is fully unfolded toward a deep learning model that can be trained and tested in an end-to-end network. The speed benefit of the proposed MGDuNLSS-Net method fundamentally comes from deep learning and the subspace representation. While the comparison methods HSRNet and MCT-NET are end-to-end deep networks and with no iteration, still have a higher computation time in comparison to our proposed MGDuNLSS-Net. On the other hand, the observation model can highly influence the efficiency of the fusion process. While these models are not available in real scenarios, the testing methods CSTF, CSU, and CNN-Fus are used predefined degradation matrices, while the HySure, CNMF, HSRNet, and MCT-NET estimate the observation models with no supervision. To enhance the fusion outcomes, the proposed MGDuNLSS-Net supervises the prediction of the degradation process by projecting the estimated HrHSI at any iteration back to the original LrHSI and HrMSI and considering their error in the fusion process. Table VII reported a general and running time comparison of testing methods, including the proposed method.

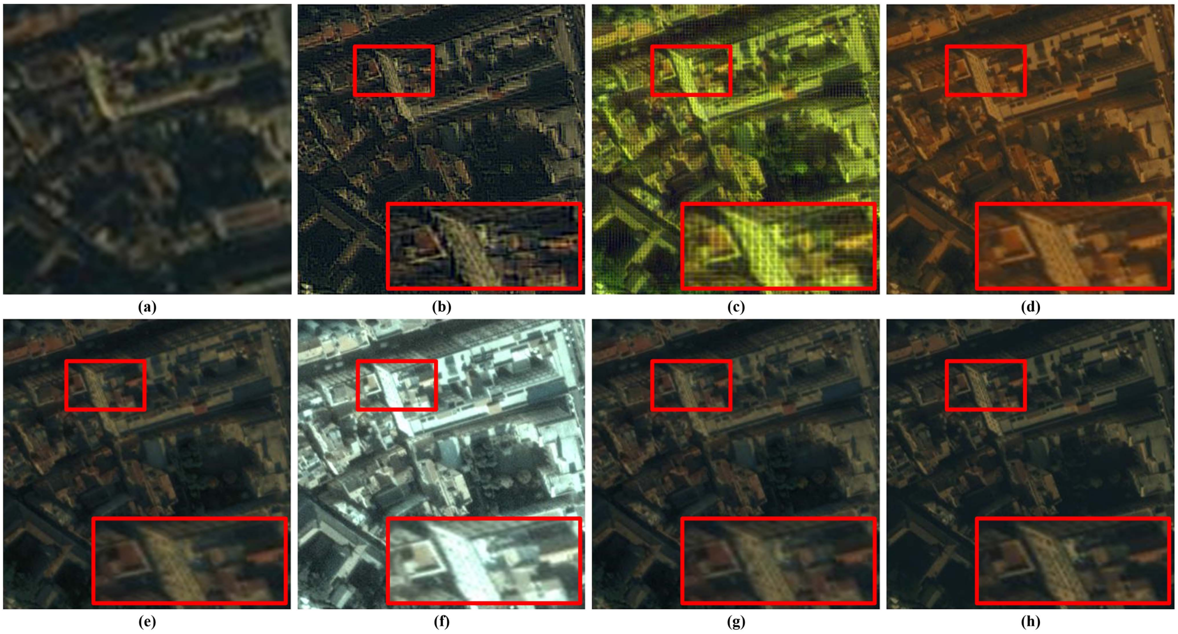


Fig. 17. Pseudocolor contains bands (R: 5, G: 3, and B: 2) of the outcomes of the different experiments on real datasets WV2. The obtained fused image has a spatial size of 360 × 360. (a) Original LrMSI. (b) Fused image of CNMF [34]. (c) Fused image of CSTF [65]. (d) Fused image of CSU [66]. (e) Fused image of HSRNet [67]. (f) Fused image of MCT-NET [68]. (g) Fused image of CNN-Fus [56]. (h) Fused image of MGDuNLSS-Net.

H. *Experiment With Real-Life Datasets*

In this section, we studied the performance of the proposed approach and the various comparison approaches by validating their effectiveness on real HSI-MSI datasets. In this regard, the hyperspectral and multispectral images taken by the Hyperion sensor toted by the Earth Observing-1 satellite (EO-1) and Sentinel-2 satellite, respectively, are utilized in our experiment. The spectral bands of the HSI span the wavelength domain from 400 to 2500 nm, comprising 242 channels, where 30 m is the Geometric resolution of this HSI with a spatial size of 2350 × 990. Nevertheless, 89 bands of the HSI are employed for the experiment after discarding the bands with low SNR and water absorption bands. On the other hand, the MSI contains 13 channels with 7050 × 2970 pixels as the spatial size. Therefore, the spatial resolution of this MSI is 10 m. However, four channels, including 490, 560, 665, and 842 nm, are selected as HrMSI in this experiment. Seeking convenience, we selected 630 × 630 and 210 × 210 pixels from the MSI and HSI, respectively, as the test set, which is shown in Fig. 14.

While the proposed approach is a supervised learning approach that requires ground truth of the desired HSI at the training stage, which does not exist in reality, we simulated the training dataset from the remainder of pixels after the testing image was taken. The training dataset is obtained using Wald's protocol, where the spectral response function (SRF) and point spread function (PSF) matrices are created by following the strategy employed in [22]. In this context, the training samples of LrHSI, HrMSI, and HrHSI are partitioned into small blocks with a size of $5 \times 5 \times 89$, $15 \times 15 \times 4$, and $15 \times 15 \times 89$, respectively. Fig. 15 demonstrates the false-color RGB images obtained by the proposed approach, and the comparison approaches CNMF, CSTF, CSU, HSRNet, (MCT-NET), and CNN-Fus; herein, we ignore the outcome of HySure, which has the worst performance compared to the testing techniques on the synthesized datasets. The obtained results of the different methods tended to spectral distortion due to the HSI of Hyperion and MSI of S2 being collected at different times (around a month), where their endmembers are changed through this month. However, according to Fig. 15, the proposed MGDuNLSS-Net approach achieved the best outcome, which is nearest to the HrMSI compared to the other testing approaches.

Moreover, the real MSI dataset WV2[1] is used to further verify the proposed method's performance. The LrMSI of this database has eight spectral channels. In this regard, the spatial resolution of LrMSI is improved by fusing the LrMSI and HrRGB images that are contained in this real dataset to acquire HrMSI. Experimentally speaking, we cropped 90 × 90 and 360 × 360 pixels from the LrMSI and HrRGB as testing data, as can be seen in Fig. 16, and the remaining pixels are preserved for training purposes. The training pixels are allocated into small batches with the size of $16 \times 16 \times 3$, $4 \times 4 \times 8$, and $16 \times 16 \times 8$ for HrRGB, LrMSI, and HrMSI, respectively.

Fig. 17 depicts the results of the testing methods except for the HySure result, as we did in the first real dataset, including the proposed method. As can be seen from Fig. 17, the obtained results of CSTF, CSU, and MCT-NET are prone to spectral distortion, where CSU and MCT-NET are excessively bright, whereas CSTF has much texture artifact. As a result, the achieved image of CNMF is blurred with low spatial details. However, the best results are achieved by HSRNet, CNN-Fus,

[1] https://www.harrisgeospatial.com/Data-Imagery/Satellite-Imagery/HighResolution/WorldView-2

and the proposed method, although the results of HSRNet and CNN-Fus are susceptible to being blurred and bright.

V. Conclusion

In this article, we propose an efficient and effective model-guided deep unfolded fusion network with non-local spatial-spectral priors for hyperspectral and multispectral image fusion. The proposed method's architecture, which is an end-to-end network comprises two submodule. The first one is the fusion submodule, designed by the nonlocal spatial-spectral block (NLSSB) to ensure the full exploitation of the crucial features of the HSIs. NLSSB can successfully model the intrinsic characteristics of HSIs, such as global spectral, nonlocal spatial, and spatial-spectral correlation. Moreover, to further improve the performance of the fusion process that impressively be affected by an unknown degradation system, the proposed method estimates and keeps refining the imaging system at all iterations by backprojection of the obtained fused image at any iteration to the inputs pairs LrHSI and HrMSI that ensure good prediction of the degradation. For future work, the proposed approach can be enhanced by integrating and studying some other priors, such as sparse prior and total variation, or by further investigating the imaging system to ensure the estimation of the degradation model better.

References

[1] Y. Shang, J. Liu, J. Yang, and Z. Wu, "A model-inspired approach with transformers for hyperspectral pansharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7187–7202, 2022.

[2] R. Kawakami, Y. Matsushita, J. Wright, M. Ben-Ezra, Y.-W. Tai, and K. Ikeuchi, "High-resolution hyperspectral imaging via matrix factorization," in *Proc. CVPR*, 2011, pp. 2329–2336.

[3] J. Jiang, J. Ma, Z. Wang, C. Chen, and X. Liu, "Hyperspectral image classification in the presence of noisy labels," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 851–865, Feb. 2018.

[4] X. Mei, Y. Ma, C. Li, F. Fan, J. Huang, and J. Ma, "Robust GBM hyperspectral image unmixing with superpixel segmentation based low rank and sparse representation," *Neurocomputing*, vol. 275, pp. 2783–2797, 2018.

[5] X. Gong et al., "Feature matching for remote-sensing image registration via neighborhood topological and affine consistency," *Remote Sens.*, vol. 14, no. 11, 2022, Art. no. 2606.

[6] D. Sara, A. K. Mandava, A. Kumar, S. Duela, and A. Jude, "Hyperspectral and multispectral image fusion techniques for high resolution applications: A review," *Earth Sci. Inform.*, vol. 14, no. 4, pp. 1685–1705, 2021.

[7] J. Hou, Z. Zhu, J. Hou, H. Zeng, J. Wu, and J. Zhou, "Deep posterior distribution-based embedding for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 31, pp. 5720–5732, 2022.

[8] D. Liu et al., "An efficient unfolding network with disentangled spatial-spectral representation for hyperspectral image super-resolution," *Inf. Fusion*, vol. 94, pp. 94–111, 2023.

[9] R. A. Borsoi, T. Imbiriba, and J. C. M. Bermudez, "Super-resolution for hyperspectral and multispectral image fusion accounting for seasonal spectral variability," *IEEE Trans. Image Process.*, vol. 29, pp. 116–127, 2019.

[10] M. Xu, H. Pan, X. Wu, and Z. Jing, "Hyperspectral and multispectral image fusion via regularization on non-local structure tensor total variation," in *Proc. Int. Conf. Aerosp. Syst. Sci. Eng.*, 2023, pp. 225–238.

[11] W. Xie, J. Lei, Y. Cui, Y. Li, and Q. Du, "Hyperspectral pansharpening with deep priors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1529–1543, May 2019.

[12] L. Loncan et al., "Hyperspectral pansharpening: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 27–46, Sep. 2015.

[13] P. Zhuang, Q. Liu, and X. Ding, "Pan-GGF: A probabilistic method for pan-sharpening with gradient domain guided image filtering," *Signal Process.*, vol. 156, pp. 177–190, 2019.

[14] P. Guo, P. Zhuang, and Y. Guo, "Bayesian pan-sharpening with multiorder gradient-based deep network constraints," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 950–962, 2020.

[15] Y. Zheng, J. Li, Y. Li, K. Cao, and K. Wang, "Deep residual learning for boosting the accuracy of hyperspectral pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1435–1439, Aug. 2020.

[16] M. Zare, M. S. Helfroush, K. Kazemi, and P. Scheunders, "Hyperspectral and multispectral image fusion using coupled non-negative tucker tensor decomposition," *Remote Sens.*, vol. 13, no. 15, 2021, Art. no. 2930.

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

[18] X. Li, Y. Yuan, and Q. Wang, "Hyperspectral and multispectral image fusion via nonlocal low-rank tensor approximation and sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 550–562, Jan. 2021.

[19] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J.-Y. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3658–3668, Jul. 2015.

[20] Y. Zhou, L. Feng, C. Hou, and S.-Y. Kung, "Hyperspectral and multispectral image fusion based on local low rank and coupled spectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5997–6009, Oct. 2017.

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

[22] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3373–3388, Jun. 2015.

[23] J. Chai, H. Zeng, A. Li, and E. W. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," *Mach. Learn. Appl.*, vol. 6, 2021, Art. no. 100134.

[24] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," in *Classification in BioApps: Automation of Decision Making*, Berlin, Germany: Springer, 2018, pp. 323–350.

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

[26] J. Gao, J. Li, and M. Jiang, "Hyperspectral and multispectral image fusion by deep neural network in a self-supervised manner," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3226.

[27] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, 2021.

[28] 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, pp. 1–12, 2022.

[29] J. Yang, L. Xiao, Y.-Q. Zhao, and J. C.-W. Chan, "Variational regularization network with attentive deep prior for hyperspectral–multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5508817.

[30] Y. Sun, J. Liu, J. Yang, Z. Xiao, and Z. Wu, "A deep image prior-based interpretable network for hyperspectral image fusion," *Remote Sens. Lett.*, vol. 12, no. 12, pp. 1250–1259, 2021.

[31] D. Lei, X. Luo, L. Zhang, X. Li, Q. Liu, and W. Li, "An interpretable deep neural network for panchromatic and multispectral image fusion," in *Proc. 7th Int. Conf. Big Data Inf. Analytics*, 2021, pp. 71–78.

[32] T. You, C. Wu, Y. Bai, D. Wang, H. Ge, and Y. Li, "HMF-Former: Spatio-spectral transformer for hyperspectral and multispectral image fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5500505.

[33] H. Guo, W. Bao, K. Qu, X. Ma, and M. Cao, "Multispectral and hyperspectral image fusion based on regularized coupled non-negative block-term tensor decomposition," *Remote Sens.*, vol. 14, no. 21, 2022, Art. no. 5306.

[34] 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.

[35] K. Arias, E. Vargas, and H. Arguello, "Hyperspectral and multispectral image fusion based on a non-locally centralized sparse model and adaptive spatial-spectral dictionaries," in *Proc. 27th Eur. Signal Process. Conf.*, 2019, pp. 1–5.

[36] W. Dong et al., “Hyperspectral image super-resolution via non-negative structured sparse representation,” *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2337–2352, May 2016.

[37] A. Camacho, E. Vargas, and H. Arguello, “Hyperspectral and multispectral image fusion addressing spectral variability by an augmented linear mixing model,” *Int. J. Remote Sens.*, vol. 43, no. 5, pp. 1577–1608, 2022.

[38] Q. Wei, N. Dobigeon, and J.-Y. 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.

[39] 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.

[40] R. A. Borsoi, C. Prévost, K. Usevich, D. Brie, J. C. Bermudez, and C. Richard, “Coupled tensor decomposition for hyperspectral and multispectral image fusion with inter-image variability,” *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 3, pp. 702–717, Apr. 2021.

[41] Y. Bu et al., “Hyperspectral and multispectral image fusion via graph Laplacian-guided coupled tensor decomposition,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 648–662, Jan. 2020.

[42] F. Ma, F. Yang, and Y. Wang, “Low-rank tensor decomposition with smooth and sparse regularization for hyperspectral and multispectral data fusion,” *IEEE Access*, vol. 8, pp. 129842–129856, 2020.

[43] K. Zhang, M. Wang, S. Yang, and L. Jiao, “Spatial–spectral-graph-regularized low-rank tensor decomposition for multispectral and hyperspectral image fusion,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp. 1030–1040, Apr. 2018.

[44] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, “Pannet: A deep network architecture for pan-sharpening,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5449–5457.

[45] W. Wang, W. Zeng, Y. Huang, X. Ding, and J. Paisley, “Deep blind hyperspectral image fusion,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4150–4159.

[46] 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. Workshops*, 2019, pp. 4330–4339.

[47] 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.

[48] W. Dong, T. Zhang, J. Qu, Y. Li, and H. Xia, “A spatial–spectral dual-optimization model-driven deep network for hyperspectral and multispectral image fusion,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5542016.

[49] X. Wang, R. A. Borsoi, C. Richard, and J. Chen, “Deep hyperspectral and multispectral image fusion with inter-image variability,” *IEEE Trans. Geosci. Remote Sens.*, 2023.

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

[51] R. Lu, B. Chen, Z. Cheng, and P. Wang, “RAFnet: Recurrent attention fusion network of hyperspectral and multispectral images,” *Signal Process.*, vol. 177, 2020, Art. no. 107737.

[52] 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, 2022, Art. no. 6012305.

[53] X. Tao, C. Zhou, X. Shen, J. Wang, and J. Jia, “Zero-order reverse filtering,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 222–230.

[54] M. Irani and S. Peleg, “Improving resolution by image registration,” *Graphical Models Image Process.*, vol. 53, no. 3, pp. 231–239, 1991.

[55] Y. Romano and M. Elad, “Boosting of image denoising algorithms,” *SIAM J. Imag. Sci.*, vol. 8, no. 2, pp. 1187–1219, 2015.

[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] A. Khader, J. Yang, and L. Xiao, “NMF-DuNet: Nonnegative matrix factorization inspired deep unrolling networks for hyperspectral and multispectral image fusion,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5704–5720, 2022.

[58] T. Lei, Y. Zhang, S. I. Wang, H. Dai, and Y. Artzi, “Simple recurrent units for highly parallelizable recurrence,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 4470–4481.

[59] K. Wei, Y. Fu, and H. Huang, “3-D quasi-recurrent neural network for hyperspectral image denoising,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 363–375, Jan. 2020.

[60] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[61] Z. Huang et al., “CCNet: Criss-cross attention for semantic segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6896–6908, Jun. 2023, doi: 10.1109/TPAMI.2020.3007032.

[62] F. Dell’Acqua, P. Gamba, A. Ferrari, J. A. Palmason, J. A. Benediktsson, and K. Árnason, “Exploiting spectral and spatial information in hyperspectral urban data with high resolution,” *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 322–326, Oct. 2004.

[63] N. Yokoya and A. Iwasaki, “Airborne hyperspectral data over chikusei,” Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-0 5-27, 2016.

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

[65] 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.

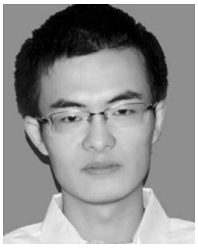
[66] C. Lanaras, E. Baltsavias, and K. Schindler, “Hyperspectral super-resolution by coupled spectral unmixing,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3586–3594.

[67] 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.

[68] X. Wang, X. Wang, R. Song, X. Zhao, and K. Zhao, “MCT-Net: Multi-hierarchical cross transformer for hyperspectral and multispectral image fusion,” *Knowl.-Based Syst.*, vol. 264, 2023, Art. no. 110362.

Abdolraheem Khader (Student Member, IEEE) received the B.S. and M.Sc. degrees in computer science from Karary University, Omdurman, Sudan, and Sudan University of Science and Technology, Khartoum, Sudan, in 2011 and 2014, respectively. He is currently working toward the Ph.D. degree in computer science from Nanjing University of Science and Technology Nanjing, China.

His research interests include the areas of deep learning and hyperspectral image superresolution.



Jingxiang Yang (Member, IEEE) received dual Ph.D. degrees in control theory and control engineering and engineering science from Northwestern Polytechnical University, Xi’an, China, and Vrije Universiteit Brussel, Brussels, Belgium, in 2019.

He is currently a Lecturer with the Nanjing University of Science and Technology, Nanjing, China. His research interests include deep learning and its applications in hyperspectral image processing.

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 (NJUST), Nanjing, China, in 1999 and 2004, respectively.

From 2009 to 2010, he was a Postdoctoral Fellow with the Rensselaer Polytechnic Institute, Troy, NY, USA. Since 2014, he has been the Deputy Director of the Jiangsu Key Laboratory of Spectral Imaging Intelligent Perception, Nanjing. He has served as the Second Director of the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, NJUST, where he is currently a Professor with the School of Computer Science. He has published more than 100 international journal articles including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING. His main research interests include inverse problems in image processing, computer vision and image understanding, pattern recognition, and remote sensing.