A Multiresolution Details Enhanced Attentive Dual-UNet for Hyperspectral and Multispectral Image Fusion

Jian Fang, Jingxiang Yang, *Member, IEEE*, Abdolraheem Khader, *Member, IEEE*, and Liang Xiao, *Member, IEEE*

***Abstract*—The fusion-based super-resolution of hyperspectral images (HSIs) draws more and more attention in order to surpass the hardware constraints intrinsic to hyperspectral imaging systems in terms of spatial resolution. Low-resolution (LR)-HSI is combined with a high-resolution multispectral image (HR-MSI) to achieve HR-HSI. In this article, we propose multiresolution details enhanced attentive dual-UNet to improve the spatial resolution of HSI. The entire network contains two branches. The first branch is the wavelet detail extraction module, which performs discrete wavelet transform on MSI to extract spatial detail features and then passes through the encoding–decoding. Its main purpose is to extract the spatial features of MSI at different scales. The latter branch is the spatio-spectral fusion module, which aims to inject the detail features of the wavelet detail extraction network into the HSI to reconstruct the HSI better. Moreover, this network uses an asymmetric feature selective attention model to focus on important features at different scales. Extensive experimental results on both simulated and real data show that the proposed network architecture achieves the best performance compared with several leading HSI super-resolution methods in terms of qualitative and quantitative aspects.**

***Index Terms*—Attention mechanism, discrete wavelet transform, hyperspectral image (HSI), multiscale, UNet.**

I. Introduction

THE hyperspectral images (HSIs) are those that provide dense spectral sampling at each pixel [1]. Compared with natural images, HSIs contain a wider spectral range, where the channel division of the spectrum is muchly detailed, and the number of channels can reach tens to hundreds. HSIs can discriminate some similar materials and is suitable for remote sensing applications such as classification [2], [3], object recognition [4], change detection [5], [6], disaster [7], and biodiversity [8]. Nevertheless, due to the limitations of the imaging characteristics of the hyperspectral camera itself and the image acquisition environment in practical scenarios, the direct acquisition of HSIs with high spatial resolution is difficult [9]. As such, there is an increasing interest in fusing low-resolution (LR)-HSI with high-resolution multispectral images (HR-MSI) to achieve HR-HSI by enhancing the calculation method of LR image quality for hyperspectral imaging [10]. The HSI fusion methods can be categorized into component substitution (CS), multiresolution analysis (MRA), model-driven, and deep learning methods.

The CS methods [11], [12] and the MRA methods [13], [14] inherit from the traditional pan-sharpening methods. The CS methods decompose the LR-HSI image into spectral and spatial information, then replace the spatial information with HR-MSI, and finally invert this process to obtain HR-HSI. The MRA methods employ multiscale decomposition to obtain HR-MSI spatial detail information to be injected into the corresponding band of HSI. In spite of the fact that CS and MRA fusion methods are effective in injecting the spatial detail of MSI into HSI, they tend to cause more severe spectral distortion.

Model-driven methods are based on mathematical models for HSI–MSI fusion, and representative methods include Bayesian-based, matrix factorization, and tensor representations. Bayesian distribution-based HSI fusion methods [15], [16], [17] use a Bayesian dictionary and sparse coding to reconstruct HSI. Taking advantage of the presence of target images in low-dimensional subspaces, Wei et al. [15] proposed a variational-based method to fuse HSI–MSI. Based on matrix decomposition, the work in [18], [19], [20], [21], and [22] utilized the high correlation between spectral bands to decompose the HS image into a coefficient matrix and spectral basis, which turns the HSI fusion problem into a problem of estimating the coefficient matrix and spectral basis. The spectral basis was extracted for LR-HSI in [20]. Sparse coding was extracted for HR-MSI using G-SOMP+ [20], and finally, the HR-HSI was attained using sparse coding and spectral basis. Based on tensor representation, the work in [23], [24], and [25] treat the HSIs as tensors without destroying the spatial-spectral structure, so tensor decomposition may be a better solution to the image fusion problem. Dian et al. [23] proposed a nonlocal sparse tensor decomposition HSI super-resolution method. This method decomposes HSIs into estimates of sparse core tensors and dictionaries, where

Manuscript received 17 September 2022; revised 17 November 2022; accepted 10 December 2022. Date of publication 13 December 2022; date of current version 23 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61871226, Grant 61571230, and Grant 62001226; in part by the Jiangsu Provincial Social Developing Project under Grant BE2018727; and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20200465. *(Corresponding authors: Liang Xiao; Jingxiang Yang.)*

The authors are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: fangjian@njust.edu.cn; yang123jx@mail.nwpu.edu.cn; abdolraheem@njust.edu.cn; xiaoliang@mail.njust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3228941

dictionaries and core tensors are learned from LR-HSI and HR-MSI. The model-based approach makes full use of image a priori factors, such as sparsity, low rank, and global similarity. Regardless of the good interpretability of the method, the complex correlation and nonlinear features of HSIs are difficult to depict with these priori factors, and the fusion performance is limited.

With the vigorous development of deep learning, more and more researchers pay attention to the direction of deep HSI fusion. Some researchers [26], [27], [28] utilize deep learning networks to reconstruct HR-HSI to learn degradation models. Dian et al. [26] used convolutional neural networks (CNNs) to learn image priors, and then combined the image priors with traditional HSI fusion algorithms. In [29], the observation model and the estimated fusion process were optimized, and a deep learning blind algorithm for HSI fusion was proposed. The CNN denoising-based method (CNN-Fus) algorithm using subspace representation and CNN denoising was proposed in [30]. Some researchers have proposed deep HSI fusion algorithms using model-driven methods. An iterative HSI super-resolution algorithm with a deep HSI denoiser was suggested in [31], which is based on likelihood and deep image prior domain knowledge. Since deep learning requires a large number of training samples, but there are not many samples in real scenarios, some researchers presented an unsupervised deep HSI fusion method. Wei et al. [32] exploited deep neural networks to capture the statistics of images and proposed an unsupervised recursive HSI super-resolution method using pixel-aware refinement. While the deep learning-based fusion methods have achieved excellent performance, they have room for improvement in terms of spatial detail. Embedding a multiscale spatial feature extraction module in a deep network has been shown to be effective in alleviating problems such as the easy blurring of boundaries [33]. Second, the multiscale feature module fuses feature information from several different scales, which can suppress the noise passed by shallow features and recover the spatial structure detail information of the fused image more effectively in the decoding stage and improve the fusion effect of the model. However, most current deep learning methods use the connectivity of HSI and MSI along the spectral channel as the input to the network; this does not fully take into account the underlying multiscale spatial information.

Several researchers have proposed the fusion of LR-HSI and HR-MSI at different scales to obtain HR-HSI. Zhou et al. [34] proposed a pyramid fully convolutional network to solve the MSI and HSI fusion problem. This network comprises two subnetworks; the first is to extract LR-HSI's spectral information by the convolutional kernel and encode them as deep features; The second subnetwork is intended to combine HR-MSI pyramids with encoded deep features to acquire HR-HSI. The method proposed in [35] solved the HR-MSI and LR-HSI fusion problem. In this network, the deep features of LR-HSI are gradually enlarged by deconvolution, and then the deep features of LR-HSI and HR-MSI are fused at different scales. However, this structure ignores the basic and shallow features of MSI. Therefore, the work [36] introduced a dual UNet (DUNet) fusion method, which first used the encoding–decoding network to extract MSI spatial features at different scales and then used these scale features to inject them into the UNet network. Previous work has used pooling, convolution, and upsampling operations to extract multiscale information from HR-MSIs and LR-HSIs to fuse HR-HSIs. Such an approach requires a large number of parameters to learn the detailed information of the MSIs, and the spatial details learned are not necessarily those needed in the fused images. In contrast, discrete wavelets have also been shown to extract the spatial details of images well in experiments on single HSI super-resolution. Therefore, the method proposed in this article uses the multiscale wavelet details extracted by the discrete wavelet transform and combines them with convolution to extract multiscale information from MSIs.

Inspired by the above problems, this article designs a multiresolution details enhanced attentive dual-UNet (MDA-DUNet). As shown in Fig. 1, this network can be divided into four parts. The first part is the detail extraction network to extract the spatial detail information of MSI. The second part is the spatio-spectral encoding module, which integrates the details of the above network and the detail extraction encoding module. The third part is the asymmetric feature selective attention module (AFSAM), which selects the vital information from the multiscale information of the spatio-spectral encoding module. Finally, the spatio-spectral decoding module incorporates the obtained features from the AFSAM with the features from the detail extraction decoding module and the spatio-spectral encoding module to produce the ultimate fused image.

The main contributions of this article are listed below:

1) In this article, an MDA-DUNet network is proposed to fuse LR-HSI and HR-MSI to obtain HR-HSI. The proposed framework can fully exploit the multiscale information of MSI and HSI for better spatio-spectral fusion.
2) A wavelet detail extraction module has been designed to learn wavelet detail features using a deep network of encoder and decoder structures. The discrete wavelet transform is used in this module to extract multiscale detail features from multispectral images, and the decoder and decoder structures are combined to extract multiscale detail features. In this way, the extracted wavelet features are not only involved in the encoding process but also in the decoding process, thus maximizing the use of the spatial detail features of the MSI.
3) In addition, an attention module for asymmetric feature selection is being designed in this article. The asymmetric features refer to the deep features in the UNet network where the spatial size and channels output at each scale are inconsistent. Asymmetric features in UNet are selected by this module using a spatial-spectral attention mechanism, which provides a significant performance improvement compared to the simple use of splicing.

The rest of this article is organized as follows. The dual-branch network approach with asymmetric attention and wavelet subband injection is described in Section II, followed by simulation experimental results and analysis in Section III, then real-data experimental results and analysis in Section IV, and finally, Section V concludes this article.

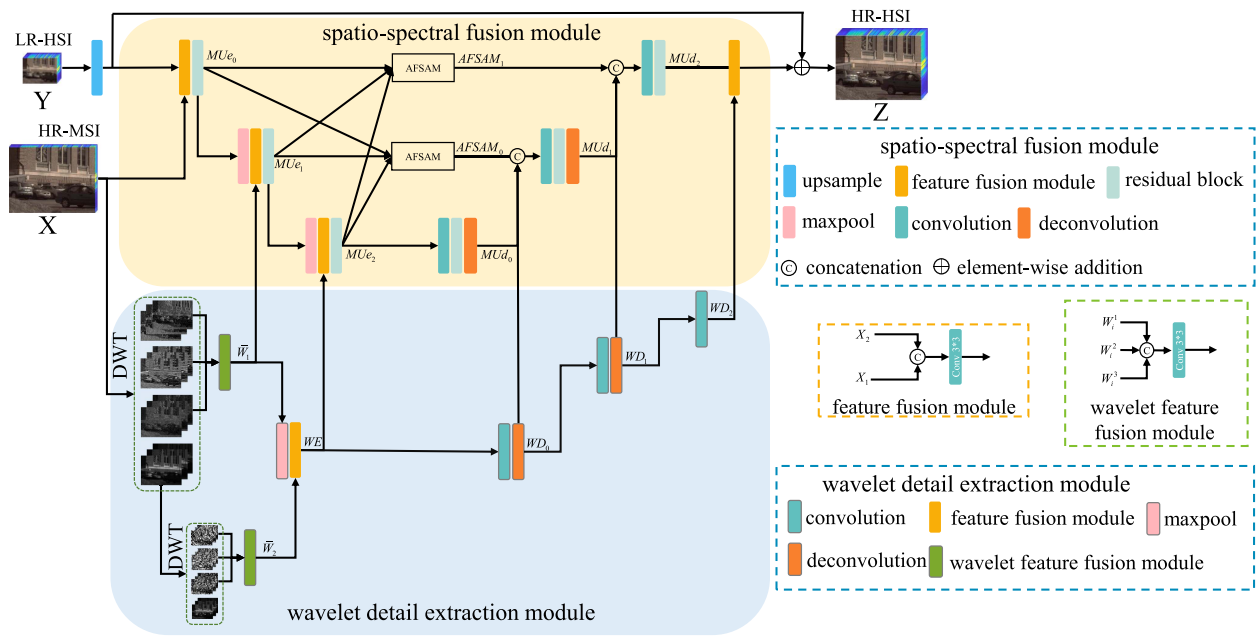

Fig. 1. Overall structure of the proposed network.

II. Methodology

In this section, our proposed method is described in detail. The structure of the proposed network is shown in Fig. 1. Let $\mathbf{X} \in \mathbb{R}^{H\times W\times b}$ denotes the HR-MSI, where H, W, and b represent the dimensions of the spatial height, width, and spectral band number, respectively. Let $\mathbf{Y} \in \mathbb{R}^{h\times w\times B}$ denotes the LR-HSI, where h, w, and B represent the dimensions of the spatial height, width, and spectral band number, respectively. The method proposed in this article consists of four parts, namely, wavelet detail extraction module, spatio-spectral encoding module, asymmetric feature selective module, and spatio-spectral decoding module. The high-frequency details are extracted from the wavelet detail extraction network by discrete wavelet transform, and then high-frequency spatial information of different deep and scales is extracted purely from MSI by an encoding–decoding. The spatio-spectral fusion module is accountable for incorporating spatial information from all phases of the high-frequency detail extraction network into the HSI for detailed enhancement. The asymmetric feature selective module extracts the asymmetric features of the spatio-spectral encoding module, then extracts the important features using the spatio-spectral attention mechanism, and finally integrates the extracted features into the decoding.

A. Wavelet Detail Extraction Module

Wavelet transforms [37] are effective methods for analyzing an image's message since they decompose the image into low-pass sub-band images and multiscale directed high-frequency sub-band images. According to [38], wavelet transform in a CNN was favorable for single-image super-resolution. A wavelet residual network was proposed in [39] for computed tomography image reconstruction, which uses wavelet detail to enhance image quality.

Discrete wavelet transform [40] extracts high-frequency detail features of HR-MSI. In this article, Haar discrete wavelet transform (filter bank is "DB1") is used to extract high-frequency details of multiresolution from MSI, in which low-pass filter and high-pass filter banks are represented. The image passes through a low-pass filter for the low-frequency sub-band image, whose ranks and rows are all at the d-scale, which can be obtained by the following equation:

$$C_d = \bar{\Phi}^{(d)}\bar{\Phi}^{(d)}(C_{d-1}) \tag{1}$$

where C_d represents the dth scale of low-frequency sub-band image.

Images of high-frequency sub-bands in the three directions are defined as

$$W_d^1 = \bar{\Phi}^{(d)}\bar{\Psi}^{(d)}(C_{d-1}) \tag{2}$$

$$W_d^2 = \bar{\Psi}^{(d)}\bar{\Phi}^{(d)}(C_{d-1}) \tag{3}$$

$$W_d^3 = \bar{\Psi}^{(d)}\bar{\Psi}^{(d)}(C_{d-1}) \tag{4}$$

where $\bar{\Phi}\bar{\Psi}(C)$ represents the convolution of C with the separable filter $\bar{\Phi}\bar{\Psi}$, C_{d-1} represents the low-pass sub-band image at the dth scale, and W_d^1, W_d^2, and W_d^3 represent high-frequency sub-band images in horizontal, vertical, and diagonal directions at the d-scale, respectively. We extract the features from the image using a wavelet feature fusion module, as shown in Fig. 1. We concatenate the three high-frequency detail features of wavelets W_i^1, W_i^2, and W_i^3 and further refine the concatenated features using 3×3 convolutional layer, which can be written as follows:

$$\bar{W}_i = \text{conv}_{3\times 3}\left(\text{cat}\left(W_i^1, W_i^2, W_i^3\right)\right) \tag{5}$$

where $\text{conv}_{3\times 3}$ represents the convolution operation, and its kernel size is 3×3, cat indicates a channel dimension concatenate operation. The low-frequency sub-band image C_1 and

high-frequency sub-band images W_1^1, W_1^2, and W_1^3 are obtained from the MSI through discrete wavelet transform. These three high-frequency sub-band images are concatenated together in the channel dimension, and the first feature of the encoding is obtained using the convolution of 3×3, denoted by $\bar{W}_1$. Since the low-frequency sub-band image C_1 also has high-frequency information, discrete wavelet transform is used for the MSI low-frequency sub-band image to obtain the low-frequency sub-band image C_2 and the high-frequency sub-band images W_2^1, W_2^2, and W_2^3, and concatenate them in the channel dimension. These three high-frequency sub-band images result in a high-frequency image, and then this image undergoes a 3×3 convolution, represented by $\bar{W}_2$.

The deep features $\bar{W}_1$ are first subjected to a maximum pooling operation and concatenated in the high-frequency deep feature $\bar{W}_2$ in channel dimension, then passed through a convolutional layer of size 3×3 to finally obtain the deep features of the second output of the encoding module. This process is represented as

$$WE = F\left(\bar{W}_2, \text{maxp}\left(\bar{W}_1\right)\right). \tag{6}$$

where $F(X_1, X_2) = \text{conv}_{3\times3}(\text{cat}(X_1, X_2))$, X_1 and X_2 are deep features, and maxp$(\cdot)$ is the pooling operation for the maximum channel dimension.

A wavelet detail extraction decoding module is intended to complement the information from the spatio-spectral encoding module. This decoding module consists of a deconvolution layer of stride size 2 and a convolution layer with a kernel size 3×3. The deep features of the three outputs of the decoding module are expressed as

$$WD_n = \begin{cases} \text{dec}\left(\text{conv}_{3\times3}\left(\bar{W}_3\right)\right), & n = 0 \\ \text{dec}\left(\text{conv}_{3\times3}\left(WD_{n-1}\right)\right), & n = 1 \\ \text{conv}_{3\times3}\left(WD_{n-1}\right), & n = 2 \end{cases} \tag{7}$$

where dec$(\cdot)$ is deconvolution to up-sample feature, and WD_n represents the outputted deep feature of the decoding module at the nth stage of detail extraction.

B. Spatio-Spectral Encoding Module

The LR-HSI is preprocessed; that is, the space size of the LR hyperspectral is sampled at most the same as that of the spectrum, which can be achieved as follows:

$$\overline{\mathbf{Y}} = \text{Up}\left(\mathbf{Y}\right) \tag{8}$$

where $\overline{\mathbf{Y}}$ represents the LR-HSI after upsampling, and Up$(\cdot)$ represents a spatial upsampling operation. The upsampling method is bilinear interpolation with a scale factor of 8.

To address the problem of the training error increasing rather than decreasing after adding too many layers. He et al. [41] proposed the residual network. Residual blocks are used in this step to extract deep features; hence, they will be briefly discussed below. The residual block consists of two convolution layers with a size of 3×3 and the ReLU activation function. This process is formulated as

$$\text{RB}\left(X_{\text{in}}\right) = \text{conv}_{3\times3}\left(\delta\left(\text{conv}_{3\times3}\left(X_{\text{in}}\right)\right)\right) + X_{\text{in}} \tag{9}$$

where X_{in} is the input feature, and δ is the ReLU function. The up-sampled HSI and MSI are first concatenated in the same channel dimension as the original input, followed by learning the concatenated image using a convolutional layer of 3×3 size and a residual block, and finally obtaining the first output feature of the deep feature encoding module. The deep detail features are concatenated with the maximum pooling features of the $(n$-1)th output of the coding module, and the concatenated deep features are then trained using a convolutional layer of size 3×3 and a residual block to obtain the nth output feature of the decoding module. This process is expressed as

$$\text{MU}e_n = \begin{cases} \text{RB}\left(F\left(\overline{\mathbf{Y}}, \mathbf{X}\right)\right), & n = 0 \\ \text{RB}\left(F\left(\text{maxp}\left(\text{MU}e_{n-1}\right), \bar{W}_1\right)\right), & n = 1 \\ \text{RB}\left(F\left(\text{maxp}\left(\text{MU}e_{n-1}\right), WE\right)\right), & n = 2 \end{cases} \tag{10}$$

where MUe_n represents the output feature of the encoding module at the nth stage.

C. AFSAM

In the encoding–decoding structure [42], the encoding part performs layer-by-layer downsampling using pooling layers, and the decoding part performs layer-by-layer upsampling using deconvolution. The spatial information in the original input image is gradually recovered along with the details in the image. The resulting LR image is eventually mapped to a pixel-level, HR image. To compensate for the information lost in the downsampling of the encoding stage, the UNet [43] uses a splicing operation between the encoding and decoding of the network to fuse the feature maps at the corresponding positions in the two processes. The decoding is able to retain more HR detail information contained in the high-level feature maps during upsampling, thus better recovering the spatial detail information of the original image.

An asymmetric feature fusion module (AFFM) was proposed in [44] to improve the deblurring image performance using multiscale features. This module turns the multiscale features of the encoding into the same spatial dimension, then concatenates these features again in the channel dimension, afterward does the convolution of the concatenated features, and finally concatenates the obtained convolution with the features of the decoding. Although this method can make good use of multiscale features, simple splicing cannot be exploited to greater advantage. Multikernel networks [45] were proposed to extract scale-important features. This article proposes an asymmetric selective attention mechanism by combining the multikernel networks and the asymmetric fusion module.

Fig. 2 shows an example of MUe_0 based on the AFSAM. First of all, deconvolution and convolution are used to change the space size; and channel number size of input deep features to the same size as MUe_0; and this process can be written as the following:

$$\text{Su}_n = \begin{cases} \text{MU}e_n, & n = 0 \\ \text{conv}_{3\times3}\left(\text{dec}\left(\text{MU}e_n\right)\right), & n = 1 \\ \text{conv}_{3\times3}\left(\text{dec}\left(e_n\right)\right), & n = 2. \end{cases} \tag{11}$$

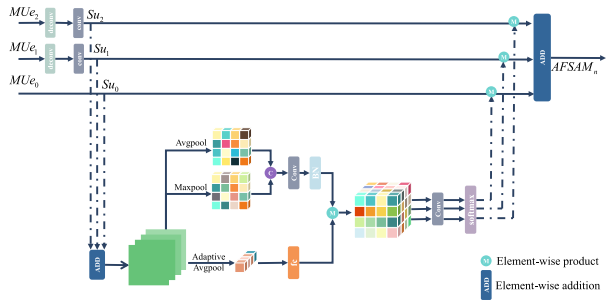

Fig. 2. AFSAM.

Then, adding the three obtained deep features element by element to obtain the deep feature, which is expressed as

$$\mathrm{Su} = \sum_{i=0}^{2} \mathrm{Su}_i \tag{12}$$

where Su denotes the element-by-element sum of the three deep features.

An extraction process of spatial and channel attention weights is introduced so as to select the important features using the spatial and channel attention mechanisms of multiresolution features.

The global averaging pooling operation is first used to extract the global perceptual field of Su so that the channel attention weights of the deep feature Su can be obtained, and each feature channel is abstracted as a feature point. This process is defined as

$$Cs = gp(Su) \tag{13}$$

where $gp(\cdot)$ represents the space dimension average pooling operation.

A two-layer multilayer perception network is used to carry out nonlinear feature transformation to construct the correlation between feature graphs. This process is formulated as

$$Cz = fc(Cs) = \delta\left(\mathcal{B}(Cs)\right) \tag{14}$$

where $\mathcal{B}$ represents the batch normalization [46].

In order to obtain the spatial attention weights of the deep features, two deep features with constant spatial dimension and one channel dimension are first obtained by average pooling and maximum pooling for the Su channel dimension. Then, the two deep features are concatenated together in the channel dimension, and this process is expressed as

$$Ss = \mathrm{cat}\left(\mathrm{avgp}(Su), \mathrm{maxp}(Su)\right) \tag{15}$$

where $\mathrm{avgp}(\cdot)$ is the pooling operation for the average channel dimension.

The spatial attention weight is obtained by convolution layer with a size of 7×7 calculation for Ss, and the process is determined as

$$Sz = \mathrm{conv}_{7\times7}(Ss) \tag{16}$$

where Sz is the spatial attention weight.

The obtained spatial and channel attention weights are multiplied to obtain the spatial-spectral attention weight, which is defined as

$$Sc = Cz \cdot Sz. \tag{17}$$

Three 1×1 convolutions obtain three spatio-spectral attention weights, which can be rewritten as

$$SC_i = \mathrm{conv}_{1\times1}(Sc), i = 0, 1, 2 \tag{18}$$

where $\mathrm{conv}_{1\times1}$ represents a convolution layer with size of 1×1.

The softmax function is used for the three attention weights to obtain $Sa + Sb + Sc = 1$, which ends up with the following equations:

$$Sa = \frac{e^{SC_0}}{e^{SC_0} + e^{SC_1} + e^{SC_2}} \tag{19}$$

$$Sb = \frac{e^{SC_1}}{e^{SC_0} + e^{SC_1} + e^{SC_2}} \tag{20}$$

$$Sc = \frac{e^{SC_2}}{e^{SC_0} + e^{SC_1} + e^{SC_2}} \tag{21}$$

where Sa, Sb, and Sc represent the spatial-spectral attention weights of Su_0, Su_1, and Su_2, respectively.

The attention module is obtained by multiplying the spatio-spectral attention weight and deep feature, and the output of the module is obtained by adding the three attention modules element-by-element. This process is expressed as

$$\mathrm{AFSAM}_1 = Sa \cdot \mathrm{Su}_0 + Sb \cdot \mathrm{Su}_1 + Sc \cdot \mathrm{Su}_2 \tag{22}$$

where AFSAM_1 is based on MUe_0, the space size and channel number of MUe_1 and MUe_2 are changed to be the same as MUe_0 by deconvolution and convolution operations, and then Su is obtained by adding. While AFSAM_0 is based on MUe_1, the space size and channel number of MUe_0 and MUe_2 are changed to be the same as MUe_1 by deconvolution or pooling and convolution operations, and then Su is added.

D. Spatio-Spectral Decoding Module

The AFSAM and detail extraction decoding module are used to construct the spatial-spectral decoding module, and the fusion results are obtained by the ReLU activation function. The first output of the decoding module is obtained by extracting the deep feature of MUe_2 using convolution, residual block, and deconvolution. Afterward, the output AFSAM_{n-1} of the AFSAM, the output MUd_{n-1} of the spatio-spectral encoding module, and the output WD_{n-1} of the detail extraction decoding module are spliced together. Then, the convolution layer with a size of 3×3 and residual blocks are purposed. Finally, the nth output of the spatio-spectral decoding module is obtained by deconvolution, which can be formulated as follows:

$$\mathrm{MUd}_n = \begin{cases} \mathrm{dec}\left(\mathrm{RB}\left(\mathrm{conv}_{3\times3}\left(\mathrm{MUe}_2\right)\right)\right), & n = 0 \\ \mathrm{RB}(\mathrm{conv}_{3\times3}(\mathrm{cat}(\mathrm{AFSAM}_{n-1}, \mathrm{MUd}_{n-1}, WD_{n-1})))), & n = 1, 2 \end{cases} \tag{23}$$

where MUd_n represents the decoding module of the nth stage. After that, the output MUd_2 of the decoding and the decoding WD_2 of detail extraction are spliced on the channel dimension, and features are extracted by convolution. Then, the up-sampled

HS image and extracted features are added element-by-element. Finally, the fusion image is obtained by using the ReLU activation function. This process is described as

$$\hat{\mathbf{Z}} = \delta\left(F\left(MUd_2, WD_2\right) + \overline{\mathbf{Y}}\right) \tag{24}$$

where $\hat{\mathbf{Z}}$ denotes the reconstructed image.

E. Loss Function

In our MDA-DUNet network, training is achieved by minimizing the following loss function:

$$\mathcal{L}(\Theta) = \arg\min_{\Theta} \sum_{i=1}^{N} \left\| \mathcal{F}\left(X_i, Y_i; \Theta\right) - Z_i \right\|_1 \tag{25}$$

where X_i, Y_i, and Z_i indicate the ith pair of LR-HSI, HR-MSI, and the original HR-HSI, respectively. $\mathcal{F}(\cdot, \Theta)$ denotes the reconstructed HSI patch by the network with parameters Θ. During network training, the ADAM [47] optimizer of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is used with the learning rate $\gamma = e^{-4}$ and the number of iterations is 1000 and the batch size is 32. All experiments are performed on PyTorch in Windows 10 using an Inter(R) Core(TM) i7-9700 CPU and NVIDIA 2080TI GPU.

III. Experimental Results

A. Comparison Methods

The proposed method is compared with seven current mainstream HSI image super-resolution methods, including three traditional methods, namely coupled nonnegative matrix factorization (CNMF) [18],[1] the subspace regularized method (HySure) [48],[2] and coupled spectral unmixing (CSU) [19],[3] and four deep learning methods, namely deep HSI sharpening method (DHSIS) [26],[4] deep blind iterative fusion network (DBIN) [29],[5] CNN-Fus [30],[6] a novel model-guided deep convolutional network (MoG-DCN) [31],[7] and a dual U-Net (DUNet). For a fair comparison, the same data preprocessing is used in all methods, and the deep learning-based methods among the methods compared are trained using the code provided by the authors with the proposed parameters on the same training data and the same protocol for evaluating the experimental results of all methods.

B. Experimental Dataset

Three publicly simulated hyperspectral imaging datasets are used to verify the performance of the proposed method, i.e., Columbia Computer Vision Laboratory (CAVE) dataset,[8] Harvard dataset,[9] the Interdisciplinary Computational VisionLab (ICVL) dataset,[10] and the Chikusei dataset.[11] With a spatial size of 512×512, a band range of 400–700 nm, a wavelength interval of 10 nm, and 31 spectral bands, the CAVE dataset comprises 32 indoor HSIs. The Harvard dataset consists of 50 indoor and outdoor HSIs, which have a spatial size of 1040×1392, a band range of 420–720 nm, a wavelength interval of 10 nm, and 31 spectral bands. The ICVL dataset comprises 201 HSIs with a spatial size of 1300×1392, a band range of 400–700 nm, a wavelength interval of 10 nm, and 31 spectral bands. For convenience, in the experiments, we crop the top-left 1024×1024 pixels from Harvard and ICVL datasets for training and testing the proposed method. The Chikusei dataset contains airborne HSI taken by visible and near-infrared imaging sensors in agricultural and urban areas of Chikusei, Ibaraki Prefecture, Japan. This hyperspectral dataset has 128 bands in the spectral range of 363–1018 nm, and the scene consists of 2517×2335 pixels. After removing black borders from the spatial domain, the centered 2048×2048 pixels were cropped and extracted for use in our experiments. Partial images of the test set for these four datasets are shown in Fig. 3.

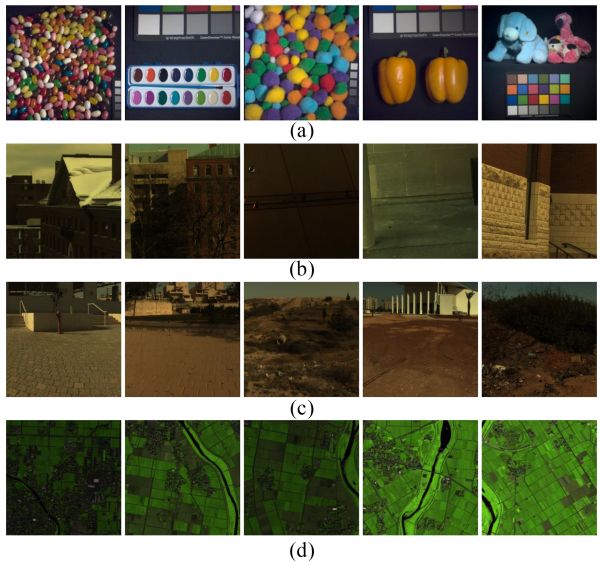

Fig. 3. Examples of the testing images selected from (a) CAVE, (b) Harvard, (c) ICVL datasets, and (d) Chikusei datasets.

The LR-HSI for the four datasets is acquired by a Gaussian filter of $r \times r$ (the mean value is 0, the standard deviation is 2) and down-sampling every r pixels in the vertical and horizontal directions of each band of the reference image, namely, the extraction factor is $r \times r$. The HR-MSI of the same scene is simulated by spectrally downsampling the HR-HSI using the subspectral sampling matrix $\mathbf{R}$, where $\mathbf{R}$ adopts the Nikon D700 camera response function.[12] For the Chikusei dataset, given the diversity of hyperspectral sensors, the spectral response function R of IKONOS satellite[13] was used to generate HR-MSI. At the same time, the observed images from these datasets are

[1] http://naotoyokoya.com/Download.html
[2] https://github.com/alfaiate/HySure
[3] https://github.com/lanha/SupResPALM
[4] https://github.com/renweidian/DHSIS
[5] https://github.com/wwhappylife/Deep-Blind-Hyperspectral-Image-Fusion
[6] https://github.com/renweidian/CNN-FUS
[7] https://github.com/chengerr/Model-Guided-Deep-Hyperspectral-Image-Super-resolution
[8] http://www.cs.columbia.edu/CAVE/databases/multispectral/
[9] http://vision.seas.harvard.edu/hyperspec/
[10] http://icvl.cs.bgu.ac.il/hyperspectral/
[11] http://naotoyokoya.com/Download.html
[12] https://www.maxmax.com/spectral_response.htm
[13] https://www.satimagingcorp.com/satellite-sensors/ikonos/

TABLE I
AVERAGE MPSNR, RMSE, ERGAS, SAM, UIQI, AND MSSIM RESULTS OF THE ABOVE METHODS ON THE CAVE DATASET WITH GAUSSIAN BLUR KERNEL AND SCALING FACTORS OF 8, 16, AND 32

Scale factor	Indices	CNMF [18]	HySure [48]	CSU [19]	DHSIS [26]	DBIN [29]	CNN-Fus [30]	MoG-DCN [31]	DUNet[36]	MDA-DUNet
$s = 8$	MPSNR↑	34.3027	34.7822	35.6888	46.2977	45.7831	44.5789	46.3167	46.2823	**47.1945**
	RMSE↓	5.4723	5.3232	4.7694	1.4653	1.5233	1.8994	1.4217	1.4759	**1.3292**
	ERGAS↓	2.6006	2.4181	2.2064	0.6641	0.6781	0.8689	0.645	0.6655	**0.5976**
	SAM↓	7.8920	11.5451	7.8787	3.8452	3.6035	5.4241	3.5967	4.3319	**3.4568**
	UIQI↑	0.771	0.8043	0.7986	0.9242	0.926	0.8735	0.9278	0.9128	**0.9343**
	MSSIM↑	0.9388	0.9107	0.9531	0.9904	0.9925	0.985	0.9924	0.9885	**0.9929**
$s = 16$	MPSNR↑	26.4938	28.5437	30.9482	41.4026	42.1841	40.6231	43.3771	41.5953	**44.2974**
	RMSE↓	13.2830	11.3402	7.9380	2.7170	2.4506	3.4303	2.1443	2.5925	**1.9877**
	ERGAS↓	3.0507	2.5826	1.8842	**0.5773**	1.0442	0.6637	0.9085	1.1360	0.8314
	SAM↓	11.5676	18.5054	11.4524	5.8874	4.6400	8.0049	4.6794	6.3316	**4.3369**
	UIQI↑	0.5741	0.6894	0.6904	0.8858	0.9110	0.8326	0.9130	0.8792	**0.9258**
	MSSIM↑	0.8326	0.7907	0.9042	0.9818	0.9889	0.9678	0.9893	0.9784	**0.9901**
$s = 32$	MPSNR↑	21.4637	21.0979	26.6490	38.6731	37.6474	33.9185	38.6788	36.7921	**41.0863**
	RMSE↓	24.7266	26.2164	12.8840	3.8456	4.7977	7.2042	3.7066	4.4249	**3.0119**
	ERGAS↓	2.5815	3.0316	1.4651	**0.3938**	1.8571	0.7463	1.5293	1.9562	1.2222
	SAM↓	18.5055	30.9459	14.3465	8.2859	7.1366	20.4831	7.8913	10.3624	**6.0760**
	UIQI↑	0.3847	0.5258	0.5719	0.8483	0.8692	0.7424	0.8394	0.8029	**0.8990**
	MSSIM↑	0.6785	0.6487	0.8485	0.9727	0.9769	0.8794	0.9747	0.9453	**0.9841**

Notes: The ↑ indicates that the larger the value, the better the performance, and the ↓ indicates that the smaller the value, the better the performance. The best and second-best results are in bold and underlined, respectively.

used as reference images. In experiments, we performed spatial enhancements of factors 8, 16, and 32.

The first 20 HSIs from the CAVE dataset are used for the training process, and the last 12 HSIs for testing. For Harvard dataset, the first 30 HSIs are used for the training process and the last 20 HSI are used as testing images. For the ICVL dataset, 50 datasets are selected from the 201 datasets for the experiments. The first 30 HSIs are used as the training dataset, and the next 20 HSIs are used as the testing dataset in the experiments. Since deep learning needs a large number of data as training sets, blocks of these training HSI are used as training sets for training the proposed network. In the case of upscaling factor 8, the size of the LR-HSI block is $4 \times 4 \times 31$, the size of the HR-MSI block is $32 \times 32 \times 3$, and the HR-HSI block is $32 \times 32 \times 31$, respectively; in the case of upscaling factor 16, the size of the LR-HSI block is $2 \times 2 \times 31$, the size of the HR-MSI block is $32 \times 32 \times 3$, and the HR-HSI block is $32 \times 32 \times 31$, respectively; in the case of upscaling factor 32, the size of the LR-HSI block is $1 \times 1 \times 31$, the size of the HR-MSI block is $32 \times 32 \times 3$, and the HR-HSI block is $32 \times 32 \times 31$, respectively.

In the Chikusei dataset, we selected images of 1024×2048 pixels in size from the top region of the images for training, while cropping the rest of the images into nine nonoverlapping $512 \times 512 \times 128$ as the test data. In the case of upscaling factor 8, the size of the LR-HSI block is $4 \times 4 \times 128$, the size of the HR-MSI block is $32 \times 32 \times 4$, and the HR-HSI block is $32 \times 32 \times 128$, respectively; in the case of upscaling factor 16, the size of the LR-HSI block is $2 \times 2 \times 128$, the size of the HR-MSI block is $32 \times 32 \times 4$, and the HR-HSI block is $32 \times 32 \times 128$, respectively; in the case of upscaling factor 32, the size of the LR-HSI block is $1 \times 1 \times 128$, the size of the HR-MSI block is $32 \times 32 \times 4$, and the HR-HSI block is $32 \times 32 \times 128$, respectively.

C. Quantitative Indicators

This article uses six evaluation metrics to quantitatively evaluate the difference between the fused images and the reference images. For example, mean peak signal-to-noise ratio (MPSNR), spectral angle mapping (SAM) [49], mean structural similarity indicator (MSSIM) [50], erreur relative global adimensionnelle synthese (ERGAS) [51], root mean square error (RMSE), and universal image quality index (UIQI) [52]. In contrast to MPSNR, MSSIM, and UIQI (larger is better), RMSE, ERGAS, and SAM are negatively correlated with image quality (smaller the better).

D. Experimental Results

Tables I–IV show the evaluation results of the different methods. For the CAVE, Harvard, ICVL, and Chikusei datasets, different fusion methods are first evaluated on 10, 20, 20, and eight test datasets, respectively, and then, the mean of the evaluation metrics is calculated. According to Tables I–IV, the proposed method achieve the higher MPSNR, UIQI, and MSSIM metrics and the lower RMSE, ERGAS, and SAM metrics. This means that the HSI reconstructed by the proposed method has a better spatial structure and less spectral distortion than the comparison methods.

Since the fused HR-HSIs shown in Figs. 4, 6, 8, and 10 are close to each other, the visual heat maps of mean squared error

TABLE II
AVERAGE MPSNR, RMSE, ERGAS, SAM, UIQI, AND MSSIM RESULTS OF THE ABOVE METHODS ON THE HARVARD DATASET WITH GAUSSIAN BLUR KERNEL AND SCALING FACTORS OF 8, 16, AND 32

Scale factor	Indices	CNMF [18]	HySure [48]	CSU [19]	DHSIS [26]	DBIN [29]	CNN-Fus [30]	MoG-DCN [31]	DUNet[36]	MDA-DUNet
$s = 8$	MPSNR↑	42.1368	40.7493	40.3709	45.5197	45.6731	44.8322	45.9283	43.9534	**45.9495**
	RMSE↓	2.5026	2.9602	3.1537	1.6187	1.5959	1.7804	**1.5441**	1.9841	1.5541
	ERGAS↓	1.0874	1.3166	1.3741	0.9167	0.8321	1.0152	0.8166	0.9481	**0.8071**
	SAM↓	2.812	3.4675	3.2612	2.4422	2.3599	2.825	**2.3161**	2.8572	2.3336
	UIQI↑	0.8612	0.8652	0.858	0.8905	0.8938	0.8697	0.8947	0.8791	**0.8953**
	MSSIM↑	0.9715	0.9694	0.9719	0.9809	0.9823	0.9768	**0.9825**	0.9725	0.9820
$s = 16$	MPSNR↑	36.9782	35.3255	36.5105	42.8090	45.3744	42.6716	45.0437	44.4802	**45.4405**
	RMSE↓	4.9101	6.1251	4.9642	2.3485	**1.6668**	2.7132	1.7728	1.9087	1.7173
	ERGAS↓	0.9225	1.1836	0.9875	0.7011	0.9292	**0.6539**	0.9238	0.9568	0.8676
	SAM↓	3.4848	5.9667	3.7168	3.2989	2.6072	4.0998	2.5327	2.7993	**2.5184**
	UIQI↑	0.7629	0.8244	0.8053	0.8754	0.8819	0.8574	0.8890	0.8829	**0.8904**
	MSSIM↑	0.9352	0.9261	0.9549	0.9755	**0.9846**	0.9702	0.9821	0.9782	0.9814
$s = 32$	MPSNR↑	32.3801	30.7656	33.1200	41.8741	43.8482	35.4913	44.4170	43.9202	**44.5370**
	RMSE↓	8.6199	11.1601	7.4266	2.6509	2.4367	6.9825	2.3293	2.2582	**2.1369**
	ERGAS↓	0.7405	1.0187	0.7251	**0.4070**	1.0626	1.0330	0.9590	0.9987	0.9348
	SAM↓	4.4584	10.2215	4.4245	3.7737	3.3934	12.9450	3.1407	3.1968	**2.9889**
	UIQI↑	0.6407	0.7496	0.7224	0.8696	0.8770	0.7932	**0.8877**	0.8809	0.8875
	MSSIM↑	0.8781	0.8544	0.9247	0.9739	**0.9832**	0.9135	0.9814	0.9783	0.9808

TABLE III
AVERAGE MPSNR, RMSE, ERGAS, SAM, UIQI, AND MSSIM RESULTS OF THE ABOVE METHODS ON THE ICVL DATASET WITH GAUSSIAN BLUR KERNEL AND SCALING FACTORS OF 8, 16, AND 32

Scale factor	Indices	CNMF [18]	HySure [48]	CSU [19]	DHSIS [26]	DBIN [29]	CNN-Fus [30]	MoG-DCN [31]	DUNet[36]	MDA-DUNet
$s = 8$	MPSNR↑	42.7429	41.0795	41.5649	49.7071	50.3108	48.1709	51.9058	52.0298	**52.4231**
	RMSE↓	2.0934	2.5572	2.3981	0.9145	0.8548	1.1159	0.6858	0.6855	**0.6522**
	ERGAS↓	0.7413	1.0101	0.8611	0.4968	0.4412	0.6827	0.3505	0.3486	**0.3381**
	SAM↓	1.7214	2.7046	2.0627	1.3958	1.2835	1.9537	1.1225	1.1609	**1.0877**
	UIQI↑	0.9125	0.9243	0.9239	0.9485	0.9536	0.9214	0.9589	0.9593	**0.9616**
	MSSIM↑	0.9837	0.9841	0.9872	0.9941	0.9957	0.9883	0.9961	0.9959	**0.9964**
$s = 16$	MPSNR↑	36.5197	35.4180	36.6287	47.9687	48.9072	45.8316	50.9774	50.3195	**51.6987**
	RMSE↓	4.5278	5.1881	4.3334	1.1802	1.0498	1.5045	0.8039	0.8799	**0.7419**
	ERGAS↓	0.7000	0.7966	0.7030	**0.3120**	0.5424	0.4794	0.3936	0.4356	0.3738
	SAM↓	2.5573	5.3196	2.8482	1.6989	1.5081	2.6787	1.2457	1.3957	**1.1618**
	UIQI↑	0.8295	0.8888	0.8715	0.9400	0.9461	0.9197	0.9573	0.9527	**0.9598**
	MSSIM↑	0.9546	0.9536	0.9723	0.9931	0.9945	0.9885	0.9960	0.9953	**0.9962**
$s = 32$	MPSNR↑	30.9816	30.7356	33.0675	43.6325	38.2204	36.0483	48.6688	48.6118	**50.3153**
	RMSE↓	8.7894	9.1533	6.6887	1.9399	5.6025	5.9965	1.0929	1.1271	**0.9087**
	ERGAS↓	0.6550	0.6892	0.5261	**0.2862**	2.1582	1.0829	0.4821	0.5252	0.4358
	SAM↓	3.7144	9.5154	3.5467	2.8897	6.9125	10.2727	1.5149	1.6788	**1.3516**
	UIQI↑	0.7040	0.8234	0.7951	0.9180	0.8561	0.8536	0.9513	0.9475	**0.9560**
	MSSIM↑	0.9044	0.8980	0.9517	0.9872	0.9713	0.9395	0.9953	0.9948	**0.9959**

images are depicted in Figs. 5, 7, 9, and 11 to visually highlight their differences, where the blue color indicates small errors and the red color indicates significant errors.

1) Results on CAVE Dataset: As can be seen from Fig. 5, the fusion result of CNMF and HySure has a large area of the reconstruction error. The fused images of the CSU method have significant reconstruction errors on apples. DHSIS, DBIN, CNN-Fus, MoG-DCN, and DUNet methods have partial reconstruction errors on reconstructed apples. The proposed method has the least reconstruction error. The per-band MPSNR and MSSIM of the reference and fused images are shown in Figs. 12(a) and 13(a). As can be seen from these figures, the proposed has the highest index on each band. Each band SAM of the reference image and the fused image is shown in Fig. 14(a). As can be seen from the figure, our method has the lowest index on each band.

2) Results on Harvard Dataset: As can be seen from Fig. 7, the fusion results of CNMF, HySure, CSU, MoG-DCN, and

TABLE IV
AVERAGE MPSNR, RMSE, ERGAS, SAM, UIQI, AND MSSIM RESULTS OF THE ABOVE METHODS ON THE CHIKUSEI DATASET WITH GAUSSIAN BLUR KERNEL AND SCALING FACTORS OF 8, 16, AND 32

Scale factor	Indices	CNMF [18]	HySure [48]	CSU [19]	DHSIS [26]	DBIN [29]	CNN-Fus [30]	MoG-DCN [31]	DUNet[36]	MDA-DUNet
$s = 8$	MPSNR↑	23.2415	21.5400	23.2267	27.1123	29.8938	26.1037	29.3254	29.9550	**30.3253**
	RMSE↓	27.5074	30.0995	26.8957	21.2084	12.2755	21.7984	14.3779	**11.7224**	12.0186
	ERGAS↓	11.4413	14.0526	11.5262	8.1020	4.8705	8.3993	6.2087	4.9811	**4.2696**
	SAM↓	4.2121	6.5111	3.3637	16.7390	3.5885	16.5571	**2.8050**	4.6340	3.8437
	UIQI↑	0.6871	0.5923	0.7042	0.6566	0.8680	0.5887	0.8231	0.8598	**0.8711**
	MSSIM↑	0.7792	0.6993	0.7831	0.6818	0.9167	0.6121	0.8923	0.9122	**0.9318**
$s = 16$	MPSNR↑	19.8532	17.6746	21.6445	25.5592	26.9923	24.8118	27.6016	28.0212	**28.6987**
	RMSE↓	36.2786	44.7360	31.9213	21.7916	17.6094	29.4535	15.5616	14.9517	**14.4477**
	ERGAS↓	16.0684	10.6288	13.5389	**4.2079**	6.6532	4.6009	6.2994	5.9951	6.0467
	SAM↓	7.6024	12.4582	4.8723	18.8792	**2.6261**	24.7458	4.1391	6.3858	3.0856
	UIQI↑	0.4261	0.4361	0.6118	0.5055	0.8101	0.4670	0.7721	0.7614	**0.8294**
	MSSIM↑	0.6104	0.5546	0.6965	0.6632	0.8929	0.6018	0.8543	0.8279	**0.9035**
$s = 32$	MPSNR↑	17.4211	14.3532	20.3353	25.2770	26.2270	24.5158	27.3570	27.5868	**27.9183**
	RMSE↓	46.3714	64.5680	37.5688	22.8627	17.6542	34.2943	**15.4334**	15.5390	15.6251
	ERGAS↓	5.3359	7.7948	3.9384	**2.2812**	7.6213	2.4418	6.1329	6.3360	6.9211
	SAM↓	10.7349	18.6708	5.9179	21.9690	7.0000	35.9602	7.0012	6.8453	**3.5564**
	UIQI↑	0.2755	0.3313	0.5402	0.5140	0.6768	0.4371	0.7020	0.7382	**0.8025**
	MSSIM↑	0.4981	0.4366	0.6342	0.6884	0.7742	0.5750	0.8019	0.8104	**0.8777**

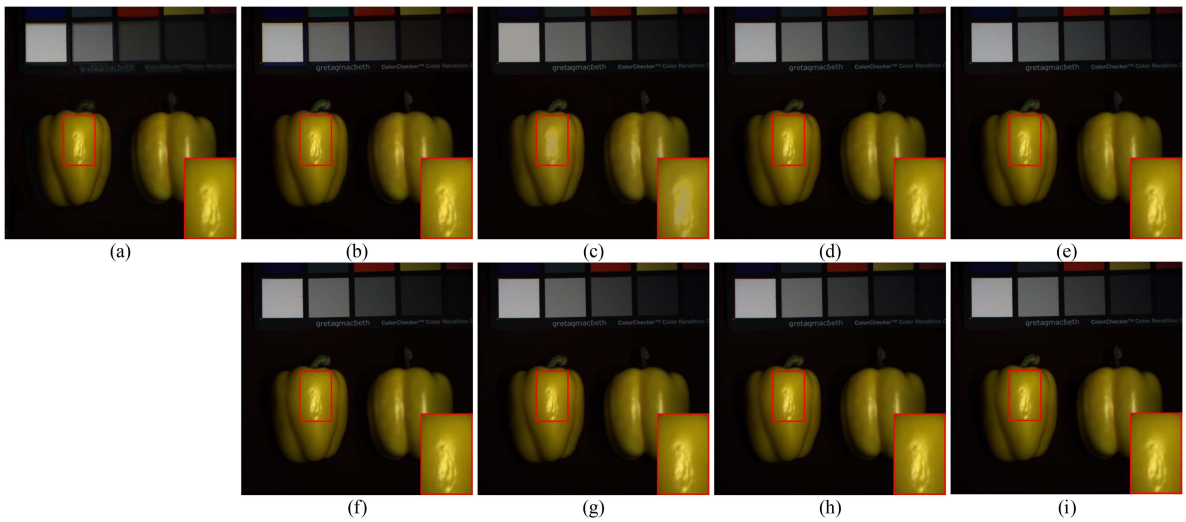

Fig. 4. Fused image of real and fake apples with pseudocolor composite map (bands 30, 20, 10). (a) CNMF [18]. (b) HySure [48]. (c) CSU [19]. (d) DHSIS [26]. (e) DBIN [29]. (f) CNN-Fus [30]. (g) MoG-DCN [31]. (h) DUNet [36]. (i) MDA-DUNet.

DUNet have large area reconstruction errors. DHSIS, DBIN, and CNN-Fus methods have partial reconstruction errors. The method in this article has the least reconstruction error. The per-band MPSNR and MSSIM of the reference and fused images are shown in Figs. 12(b) and 13(b). As can be seen from these figures, the proposed method has the highest index on each band. Each band SAM of the reference image and the fused image is shown in Fig. 14(b). According to this figure, the proposed method has the lowest index on each band.

3) Results on ICVL Dataset: As shown in Fig. 9, the fusion results of CNMF, HySure, and CSU have significant area reconstruction errors. The fused images of DHSIS and CNN-Fus methods have considerable reconstruction errors on the trees. DBIN, MoG-DCN, and DUNet methods have partial reconstruction errors. The proposed MDA-DUNet has the least reconstruction error. The per-band MPSNR and MSSIM of the reference and fused images are shown in Figs. 12(c) and 13(c). As can be seen, our proposed method has the highest index on each band. Each band SAM of the reference image and the fused image is shown in Fig. 14(c). Again, the method proposed in this article has the lowest index on each band.

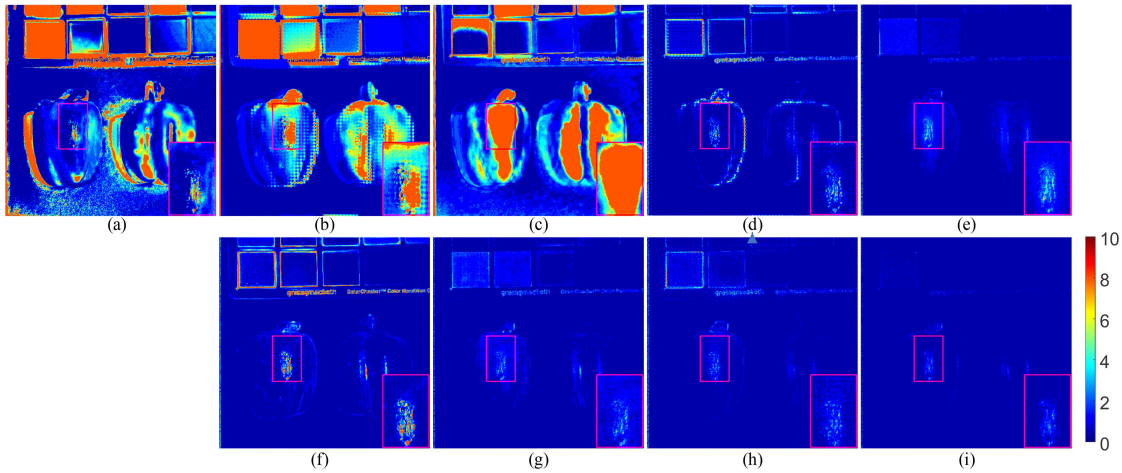

Fig. 5. Comparison method of fused images of real and fake apples with mean square error images. (a) CNMF [18]. (b) HySure [48]. (c) CSU [19]. (d) DHSIS [26]. (e) DBIN [29]. (f) CNN-Fus [30]. (g) MoG-DCN [31]. (h) DUNet [36]. (i) MDA-DUNet.

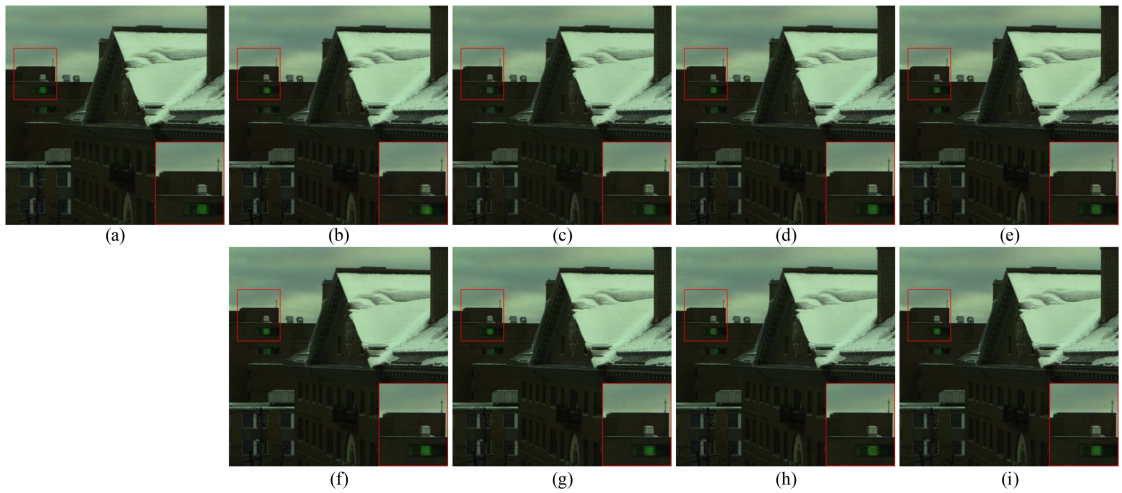

Fig. 6. Fused image of Img1 with pseudocolor composite map (bands 30, 20, 10). (a) CNMF [18]. (b) HySure [48]. (c) CSU [19]. (d) DHSIS [26]. (e) DBIN [29]. (f) CNN-Fus [30]. (g) MoG-DCN [31]. (h) DUNet [36]. (i) MDA-DUNet.

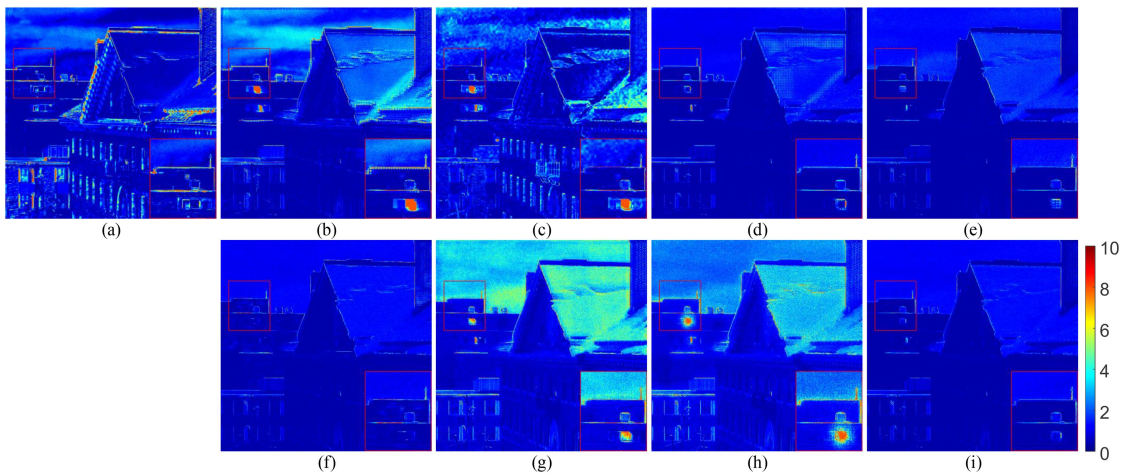

Fig. 7. Corresponding mean error images of the fused image comparison method for Img1. (a) CNMF [18]. (b) HySure [48]. (c) CSU [19]. (d) DHSIS [26]. (e) DBIN [29]. (f) CNN-Fus [30]. (g) MoG-DCN [31]. (h) DUNet [36]. (i) MDA-DUNet.

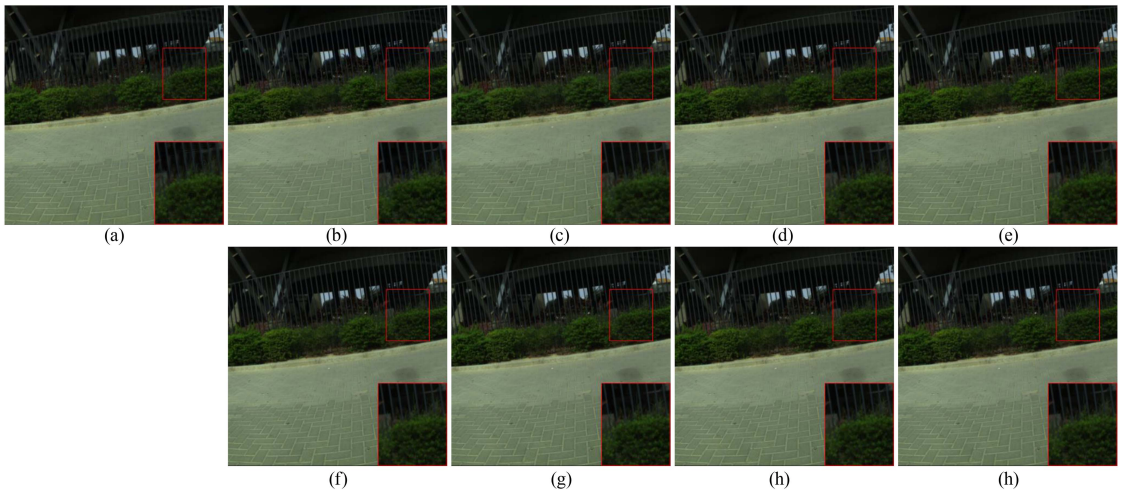

Fig. 8. Composite pseudocolor map of the fused image of Sami_0331-1019 (bands 30,20,10). (a) CNMF [18]. (b) HySure [48]. (c) CSU [19]. (d) DHSIS [26]. (e) DBIN [29]. (f) CNN-Fus [30]. (g) MoG-DCN [31]. (h) DUNet [36]. (i) MDA-DUNet.

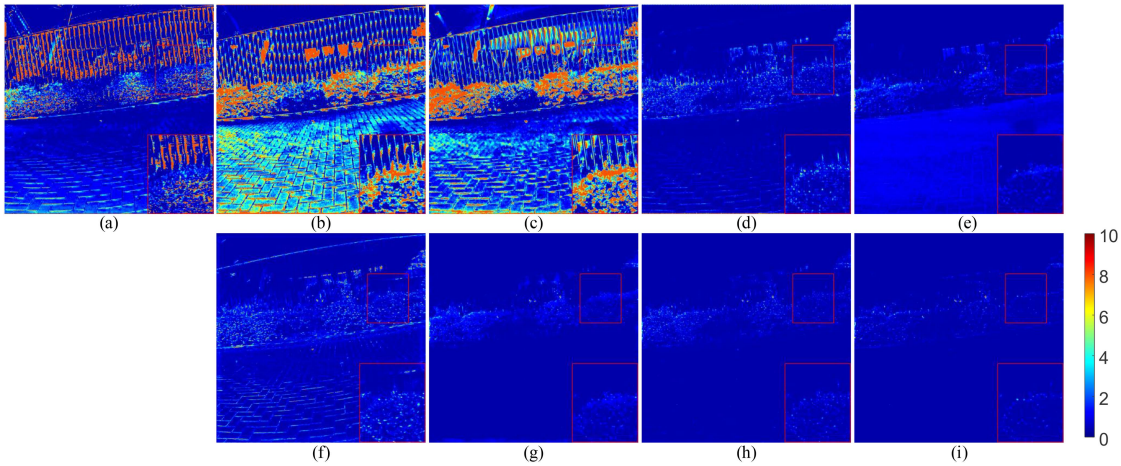

Fig. 9. Average error image corresponding to the fusion image comparison method of Sami_0331-1019. (a) CNMF [18]. (b) HySure [48]. (c) CSU [19]. (d) DHSIS [26]. (e) DBIN [29]. (f) CNN-Fus [30]. (g) MoG-DCN [31]. (h) DUNet [36]. (i) MDA-DUNet.

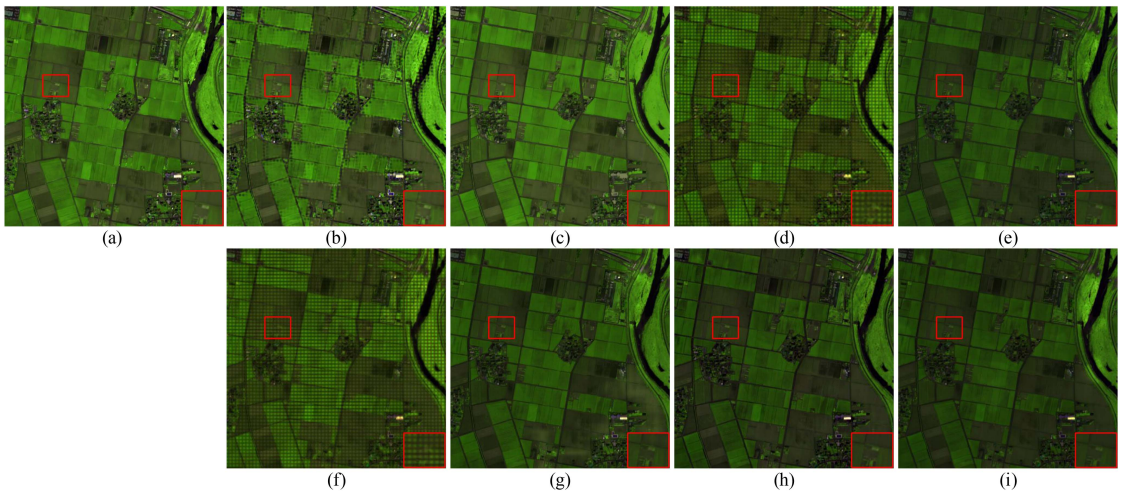

Fig. 10. Composite pseudocolor map of the fused image of Chikusei (bands 70,100,36). (a) CNMF [18]. (b) HySure [48]. (c) CSU [19]. (d) DHSIS [26]. (e) DBIN [29]. (f) CNN-Fus [30]. (g) MoG-DCN [31]. (h) DUNet [36]. (i) MDA-DUNet.

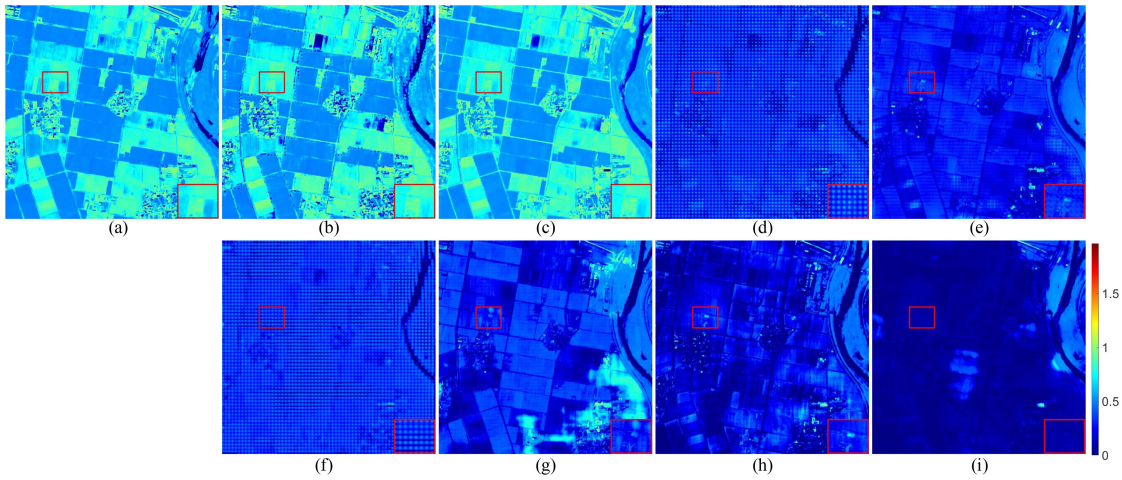

Fig. 11. Average error image corresponding to the fusion image comparison method of Chikusei. (a) CNMF [18]. (b) HySure [48]. (c) CSU [19]. (d) DHSIS [26]. (e) DBIN [29]. (f) CNN-Fus [30]. (g) MoG-DCN [31]. (h) DUNet [36]. (i) MDA-DUNet.

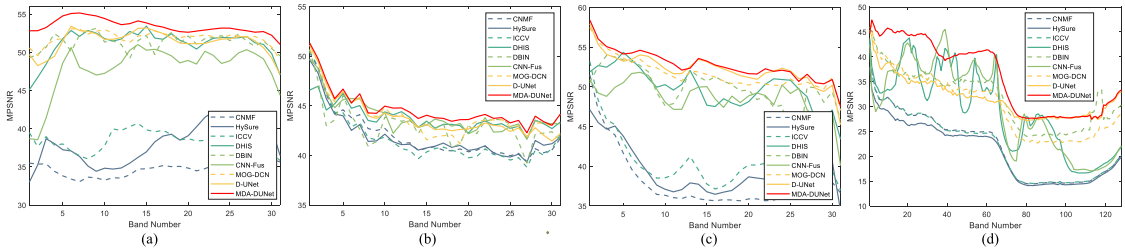

Fig. 12. Comparison of MPSNR curves in different bands. (a) Real and fake apples of CAVE datasets. (b) Imgc1 of Harvard datasets. (c) Sami_0331-1019 of ICVL datasets. (d) Chikusei datasets.

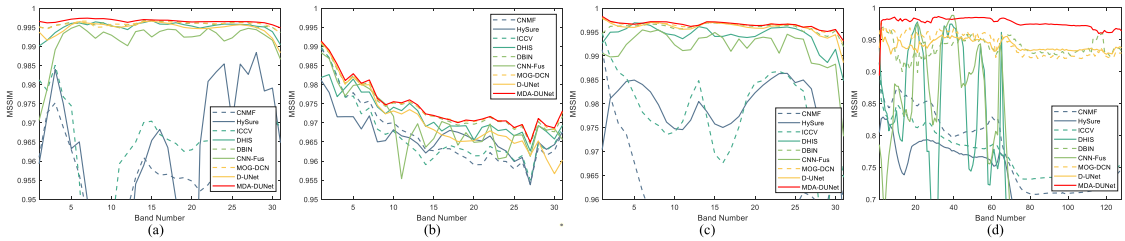

Fig. 13. Comparison of MSSIM curves for each band. (a) Real and fake apples of CAVE datasets. (b) Imgc1 of Harvard datasets. (c) Sami_0331-1019 of ICVL datasets. (d) Chikusei datasets.

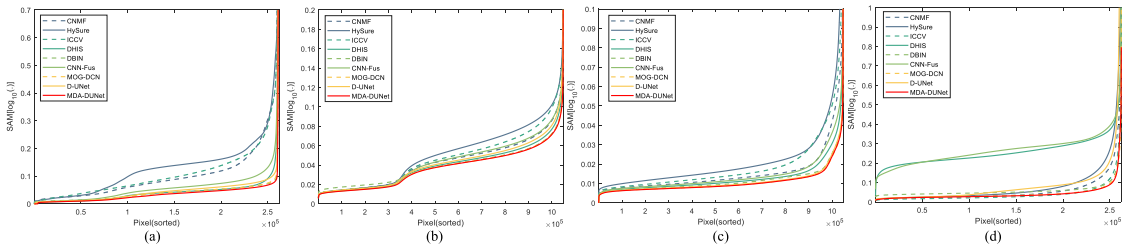

Fig. 14. Corresponding SAM curves of different methods are compared. (a) Real and fake apples of CAVE datasets. (b) Imgc1 of Harvard datasets. (c) Sami_0331-1019 of ICVL datasets. (d) Chikusei datasets.

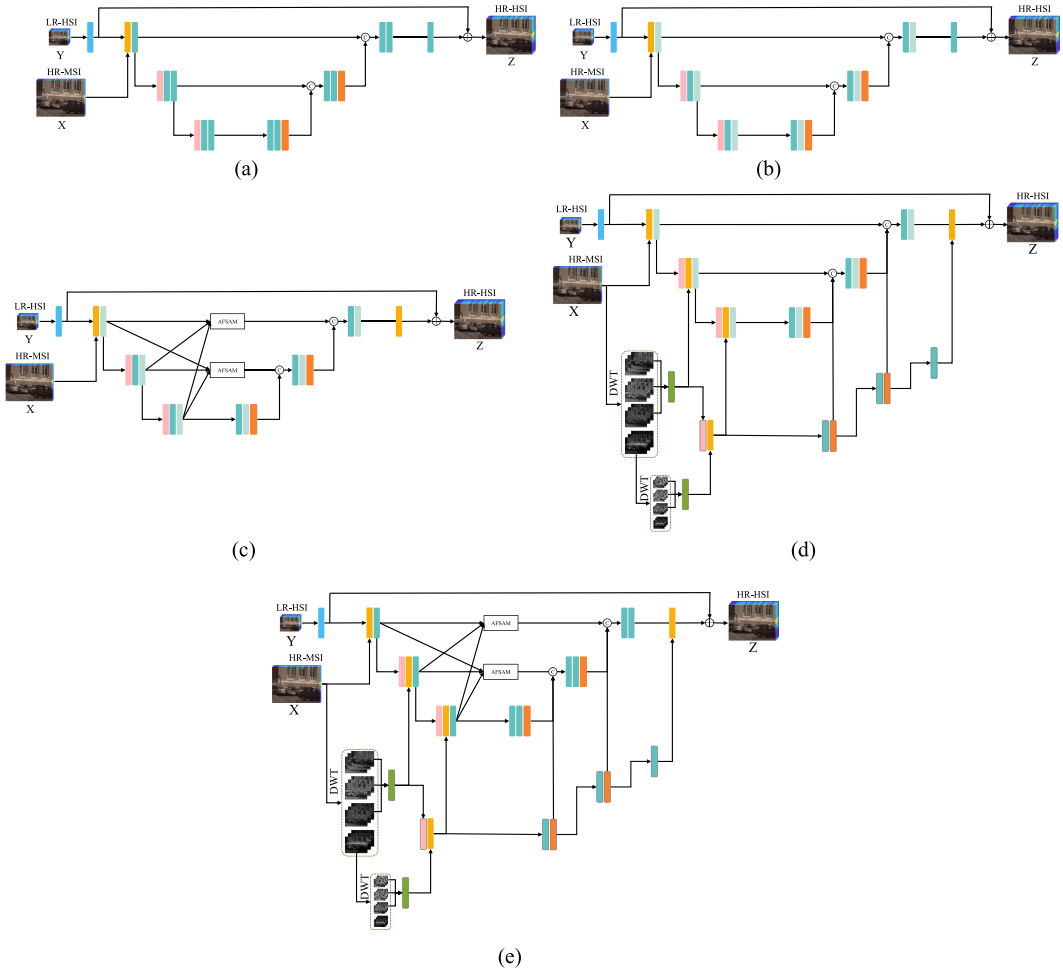

Fig. 15. Ablation study of the wavelet detail extraction module and the AFSAM. (a) UNet-1. (b) UNet-2. (c) UNet-3. (d) UNet-4. (e) UNet-5.

Fig. 16. Pseudocolor images of (a) MSIs (band 3, 2, 1) and (b) HSIs (band 20, 5, 3).

4) Results on Chikusei Dataset: As can be seen from Fig. 11, the fusion results of CNMF, HySure, and CSU have a large area of the reconstruction error. The fused images of the CSU method have a significant reconstruction error on the edge of the field dam. The DHSIS, DBIN, CNN-Fus, MoG-DCN, and DUNet methods have a partial reconstruction error on the reconstructed field dam. The method in this article has the smallest reconstruction error. The MPSNR and MSSIM for each band of the reference and fused images are shown in Figs. 12(d) and 13(d). It can be seen from these plots that the presented ones are mostly the highest values on each band. The SAM for each band for the reference and fused images is shown in Fig. 14(d). It can be seen from the plots that our method has the lowest index on each band.

5) Results on Different Noise Levels: In fusion tasks, MSIs and HSIs are often affected by noise [53], so noise is added to the images in this article. When we simulate HSI and MSI from HR-HSI, Gaussian noise is added to the HSI and MSI and the signal-to-noise ratio (SNR) varies from 10, 20, and 30 dB. For each noise level, we calculated the evaluation metrics for the CAVE dataset and then averaged them, as shown in Table V. As can be seen from the table, the proposed MDA-DUNet and DBIN methods show more advantages when the SNR is 10 and 20 dB. The MoG-DCN method achieves relatively good accuracy in the low-noise case.

6) Running Time Analysis and FLOPs: Table VI depicts the results of a quantitative comparison of the average running times and floating point operations per second (FLOPs) of the different methods on the CAVE dataset at 8× super-resolution. The table shows that the method in this article and the DUNet method, achieve better fusion performance in terms of running time and

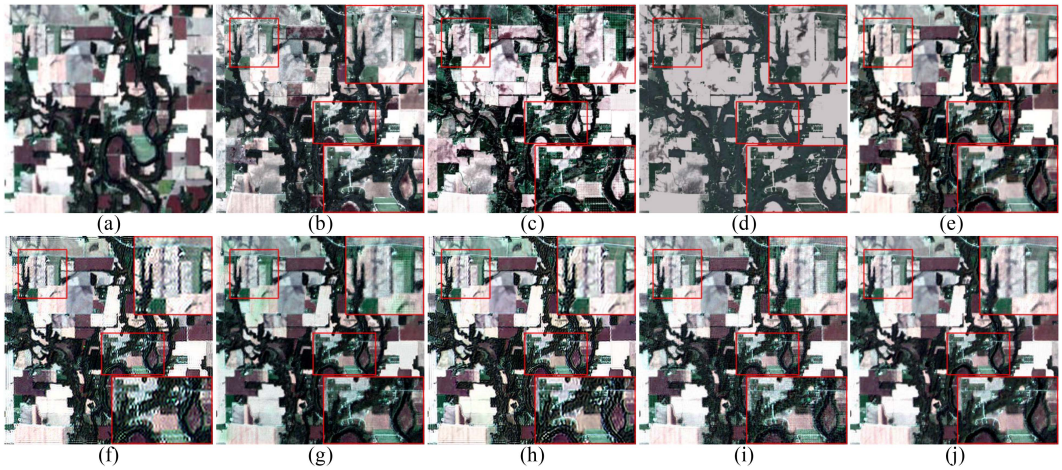

Fig. 17. Fused pseudocolor composite image of Hyperion-sentinel results (bands 20,5,3), size of the image is 600×600 with 10 m resolution. (a) Original LR-HSI. (b) CNMF [18]. (c) HySure [48]. (d) CSU [19]. (e) DHSIS [26]. (f) DBIN [29]. (g) CNN-Fus [30]. (h) MoG-DCN [31]. (i) DUNet [36]. (j) MDA-DUNet.

TABLE V
QUANTITATIVE COMPARISON OF THE DIFFERENT ALGORITHMS WAS CARRIED OUT ON THE CAVE DATASET BY ADDING DIFFERENT NOISES

Noise level	Indices	CNMF [18]	HySure [48]	CSU [19]	DHSIS [26]	DBIN [29]	CNN-Fus [30]	MoG-DCN [31]	DUNet[36]	MDA-DUNet
SNR=10	MPSNR↑	22.8397	22.3578	26.4966	21.9532	35.0170	23.0511	34.1785	28.6991	**35.8493**
	RMSE↓	21.3987	23.4086	13.5162	22.8121	5.2609	20.6863	5.6920	10.0710	**4.7064**
	ERGAS↓	8.2492	8.8265	5.3823	9.3475	2.1135	8.2105	2.3188	4.2946	**1.8811**
	SAM↓	34.3244	35.3193	21.5722	37.5199	9.9285	34.3705	9.7792	21.5142	**9.5671**
	UIQI↑	0.3590	0.3500	0.4542	0.3407	0.6739	0.3669	0.6732	0.4967	**0.6808**
	MSSIM↑	0.3249	0.3084	0.4746	0.2977	0.9150	0.2961	0.9141	0.6173	**0.9204**
SNR=20	MPSNR↑	31.9185	31.0712	32.9274	30.9609	38.1445	32.7003	38.0651	37.1938	**39.9881**
	RMSE↓	7.4887	8.0058	6.1999	8.0319	3.5779	6.8711	3.7158	3.8873	**2.9205**
	ERGAS↓	3.0593	3.2952	2.7268	3.3119	1.4923	2.7301	1.5339	1.6181	**1.2040**
	SAM↓	18.5690	19.4008	13.1142	23.0604	7.8114	19.8966	7.3319	10.4399	**6.3566**
	UIQI↑	0.5777	0.5856	0.6228	0.5516	0.7547	0.5810	0.7561	0.6929	**0.7719**
	MSSIM↑	0.7340	0.7313	0.8187	0.6774	0.9557	0.7115	0.9571	0.9150	**0.9642**
SNR=30	MPSNR↑	36.7218	34.3774	35.2222	35.4432	39.8444	41.2337	41.8825	41.0836	**43.5726**
	RMSE↓	4.1746	5.4637	4.7429	4.8047	3.2586	2.6061	2.3349	2.4893	**1.9206**
	ERGAS↓	1.9032	2.4352	2.2738	2.2500	1.4018	1.0935	1.0116	1.0720	**0.8258**
	SAM↓	11.0044	12.2398	10.0157	15.8815	6.5077	8.7933	5.6781	7.5541	**4.8732**
	UIQI↑	0.7157	0.7326	0.7134	0.6575	0.8217	0.7599	0.8337	0.7848	**0.8415**
	MSSIM↑	0.9155	0.8972	0.9259	0.8653	0.9745	0.9484	0.9807	0.9613	**0.9828**

TABLE VI
COMPARISON OF TEST RUN TIMES AND FLOPS FOR THE DIFFERENT ALGORITHMS WAS CARRIED OUT ON THE CAVE DATASET

Indices	CNMF [18]	HySure [48]	CSU [19]	DHSIS [26]	DBIN [29]	CNN-Fus [30]	MoG-DCN [31]	DUNet[36]	MDA-DUNet
Times(s)	19.5312	389.7733	164.9055	6.1281	36.0005	84.8657	0.7549	**0.2822**	0.4035
FLOPs(G)	–	–	–	1599	8019	–	4390	**421**	1411

FLOPs, with the DUNet method achieving the best performance. The MDA-DUNet is a lightweight framework with less running time and the low FLOPs, demonstrating the effectiveness and efficiency of the proposed method.

The conclusion that can be drawn from the above experimental results is that the method in this article has good spatial and spectral reconstruction capabilities on the simulated dataset.

E. Ablation Studies

1) Function of Each Component of the Proposed MDA-DUNet: The role of the different parts of the MDA-DUNet, i.e., different variants, are trained on the same training data, i.e., the CAVE dataset. Fig. 15(a) shows UNet-1, Fig. 15(b) shows UNet-2, Fig. 15(c) shows UNet-3, Fig. 15(d) shows UNet-4, and Fig. 15(e) shows UNet-5. To verify the effectiveness of

TABLE VII
PERFORMANCE COMPARISON OF ABLATION EXPERIMENTS WITH THE CAVE DATASET

method	UNet-1	UNet-2	UNet-3	UNet-4	UNet-5	MDA-DUNet
MPSNR↑	46.8404	46.2961	47.0014	46.9978	46.8064	**47.1945**
RMSE↓	1.3815	1.4418	1.3547	1.3492	1.3840	**1.3292**
ERGAS↓	0.6296	0.6472	0.6074	0.6063	0.6250	**0.5976**
SAM↓	3.7076	3.6430	3.5200	3.4639	3.6233	**3.4568**
UIQI↑	0.9287	0.9288	0.9321	0.9334	0.9296	**0.9343**
MSSIM↑	0.9916	0.9920	0.9926	0.9927	0.9919	**0.9929**

TABLE VIII
COMPARISON OF THE PERFORMANCE OF AFFM AND AFSAM ON THE CAVE DATASET

module	MPSNR↑	RMSE↓	ERGAS↓	SAM↓	UIQI↑	MSSIM↑
AFFM	45.5680	1.5745	0.7071	4.0148	0.9174	0.9902
AFSAM	**47.1945**	**1.3292**	**0.5976**	**3.4568**	**0.9343**	**0.9929**

the proposed attention module for asymmetric feature selective, the first variant removes AFSAM compared to the original network, denoted by UNet-4. The second variant cancels the wavelet detail injection, denoted by UNet-3. The third variant removes the residual module, denoted by UNet-5, while the fourth removes wavelet detail injection and AFSAM, namely UNet-2. Besides, the original UNet is set as a baseline for comparison with networks of different structures, demonstrating the effectiveness of each component, mainly because the number of parameters of the original UNet compared with the proposed method is relatively small. The structures of the five variants are shown in Fig. 15. The performances of these methods on the CAVE dataset are shown in Table VII.

From the ablation experiments, we can see that the wavelet detail extraction module and AFSAM proposed in this article are facilitative to the network, and therefore, we consider them to be effective.

2) Comparison of the Proposed AFSAM With the AFFM: From Table VIII, it can be seen that the performance of the AFSAM proposed in this article is improved by 1.6 dB in terms of MPSNR compared to the AFFM. This demonstrates the effectiveness of using spatial and spectral attention selection mechanisms to extract important features from asymmetric features.

IV. REAL DATA EXPERIMENT

In the following, the real dataset is used to further verify the proposed method's effectiveness. We use the LR-HSI acquired by Hyperion sensors using the Eo-1 satellite and the HR-MSI acquired using the Sentinel-2 satellite. The Hyperion HSI has a spectral range of 400–2500 nm, including 242 bands, with a spatial resolution of 30 m. After removing the water vapor and noise bands from the HSI, 89 bands remain. The MSI S2 has a total of 13 bands, from which four bands of 490, 560, 665, and 842 nm are selected as the HR-MSI with a spatial resolution of 10 m.

This section aims to fuse 30-m HSI and 10-m MSI data to obtain 10-m HSI. Since the proposed network is supervised learning, 10-m HSI data are required as a reference image, but there are no 10-m HSI data in the real scene. Therefore, we use the strategy in [54] and [55] to convert 30-m HSI and 10-m MSI through downsampling. The original 30-m HSI is used as a reference image for training, while in the testing phase, the original 30-m HSI and 10-m MSI are used to obtain 10-m HSI data.


(a) (b)

Fig. 18. Pseudocolor images of (a) RGBs and (b) MSIs (band 5, 3, 2).

The Hyperion HSI contains a spatial size of 2350×990, and the size of the S2 MSI is 7050×2970. For the experiment, 200×200 pixels from the HSI data and 600×600 pixels from the MSI data are cropped as the test set. At the same time, the rest of these data are used as the training dataset. Furthermore, the test images are shown in Fig. 16.

The training image is divided into patches with the $4 \times 4 \times 89$, $12 \times 12 \times 4$, and $12 \times 12 \times 89$ for the LR-HSI, HR-MSI, and HR-HSI, respectively.

From Fig. 17, it can be seen that there is not only much noise in the fusion image generated by the CNMF method but also a lot of color distortion in the river part. In the image generated by the HySure method, not only the green part is distorted, but most of the white area has disappeared, and other areas also have distortion. The green part of the fusion image generated by the CSU method is distorted into gray. The fusion result of the DHSIS method has not only spectral distortion in the white area, but also has striped noise in the image. There is more noise in the image obtained by the DBIN method. CNN-Fus method fusion image retains only rough outlines and blurred details. There was a small amount of noise in the images fused by the MoG-DCN and DUNet methods. Compared with other methods, the results obtained by the method proposed in this article have less distortion.

Furthermore, we also validate the performance of the proposed method on the real-world MSI dataset WV2.[14] This dataset consists of pairs of real LR-MSI with eight bands and HR-RGB images. We reconstruct the HR-MSI from a pair of LR-MSI and HR-RGB. Since the data only contain a pair of images, we first take the 100×100 RGB and 400×400 MSI data as the test set, and the remaining data as the training dataset. The test set data are shown in Fig. 18. Similarly, we use the block method to prepare the training dataset. The sizes of LR-MSI, HR-RGB, and HR-MSI image blocks are $3 \times 3 \times 8$, $12 \times 12 \times 3$, and $12 \times 12 \times 8$, respectively.

[14]https://www.harrisgeospatial.com/Data-Imagery/Satellite-Imagery/High-Resolution/WorldView-2

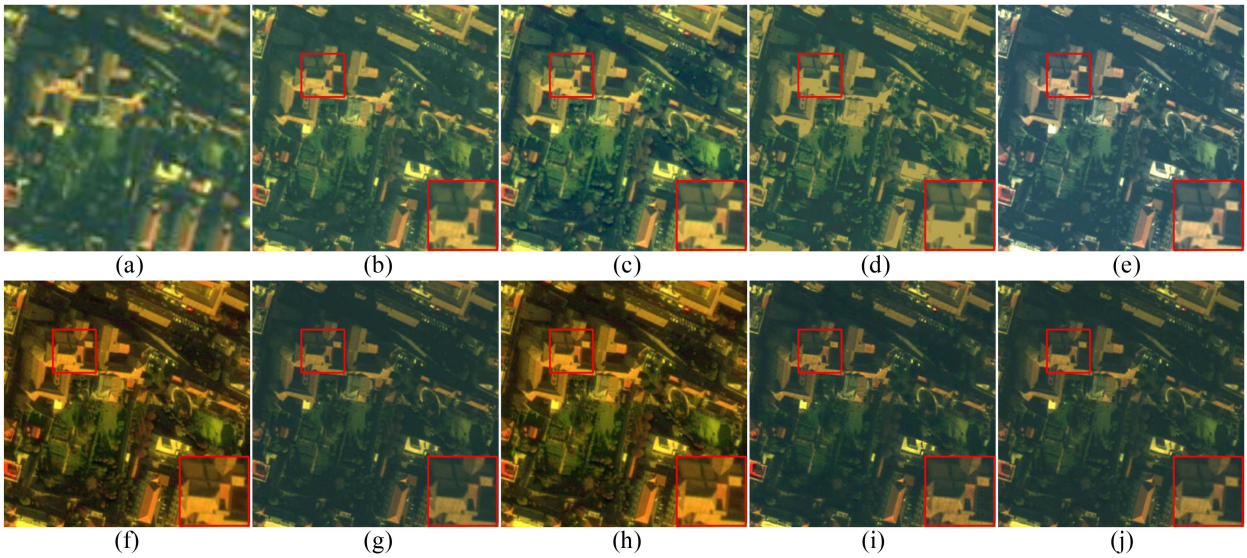

Fig. 19. Fusion result of Worldview-2's pseudocolor composite map (bands 5,3,2), size of the image is 400×400. (a) Original LR-HSI. (b) CNMF [18]. (c) HySure [48]. (d) CSU [19]. (e) DHSIS [26]. (f) DBIN [29]. (g) CNN-Fus [30]. (h) MoG-DCN [31]. (i) DUNet [36]. (j) MDA-DUNet.

As shown in Fig. 19, the fused images generated by CNMF and HySure methods are blurred and too bright. The spectral distortion occurs in CSU, with red roofs and gray background. The colors of the images generated by the DHSIS and CNN-Fus methods are distorted, while the image generated by the DBIN and DUNet methods is blurry. The images generated by the MoG-DCN method has mesh noise, whereas the proposed method has less spectral distortion and less noise.

From the above analysis, it can be concluded that the MDA-DUNet has better spatial and spectral reconstruction capabilities on real datasets.

V. Conclusion

In contrast to the present CNN-based approaches, the suggested fusion technique can aid the CNN in sufficiently exploring the HR-MSI's spatial information and incorporating the extracted spatial information into the latent image to rebuild the HR-HSI in a global-to-local pattern progressively. This article proposes a dual-ended UNet to improve the spatial resolution of HSI. The first branch is the detail extraction network, which is an encoding–decoding whose main purpose is to extract different spatial features of MSI. The other branch is the spectio-spectral fusion module, which aims to inject the features of the detail extraction network into the HSI to better reconstruct the HSI. Moreover, this network uses an asymmetric attention to focus on essential features at different scales. The experimental results on simulated and real data indicate that the proposed models are qualitatively and quantitatively outperforming the existing state-of-the-art methods.

Since transformer [56], [57] can effectively mine the nonlocal correlation of images, it has been widely used in the direction of image restoration and classification. The literature [58], [59] also introduces the transformer to HSI fusion. However, this method does not fully exploit the multiscale information of HSI and MSI. Therefore, the combination of transformer and UNet is used, both multiscale and nonlocal information of HSI can be exploited simultaneously.

References

[1] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 193–200.

[2] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[3] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.

[4] N. Durand et al., "Ontology-based object recognition for remote sensing image interpretation," in *Proc. IEEE 19th IEEE Int. Conf. Tools Artif. Intell.*, 2007, vol. 1, pp. 472–479.

[5] A. Asokan and J. Anitha, "Change detection techniques for remote sensing applications: A survey," *Earth Sci. Informat.*, vol. 12, no. 2, pp. 143–160, 2019.

[6] V. Walter, "Object-based classification of remote sensing data for change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 58, no. 3–4, pp. 225–238, 2004.

[7] C. Van Westen, "Remote sensing for natural disaster management," *Int. Arch. Photogrammetry Remote Sens.*, vol. 33, no. B7/4; PART 7, pp. 1609–1617, 2000.

[8] W. Turner, S. Spector, N. Gardiner, M. Fladeland, E. Sterling, and M. Steininger, "Remote sensing for biodiversity science and conservation," *Trends Ecol. Evol.*, vol. 18, no. 6, pp. 306–314, 2003.

[9] T. Akgun, Y. Altunbasak, and R. M. Mersereau, "Super-resolution reconstruction of hyperspectral images," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1860–1875, Nov. 2005.

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

[11] P. Kwarteng and A. Chavez, "Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis," *Photogramm. Eng. Remote Sens.*, vol. 55, no. 1, pp. 339–348, 1989.

[12] N. Koutsias, M. Karteris, and E. Chuvico, "The use of intensity-hue-saturation transformation of Landsat-5 thematic mapper data for burned land mapping," *Photogrammetric Eng. Remote Sens.*, vol. 66, no. 7, pp. 829–840, 2000.

[13] J. Cheng, H. Liu, T. Liu, F. Wang, and H. Li, "Remote sensing image fusion via wavelet transform and sparse representation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 104, pp. 158–173, 2015.
[14] J. Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *Int. J. Remote Sens.*, vol. 21, no. 18, pp. 3461–3472, 2000.
[15] 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.
[16] N. Akhtar, F. Shafait, and A. Mian, "Bayesian sparse representation for hyperspectral image super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3631–3640.
[17] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Bayesian fusion of multi-band images," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 6, pp. 1117–1127, Sep. 2015.
[18] 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.
[19] 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.
[20] N. Akhtar, F. Shafait, and A. Mian, "Sparse spatio-spectral representation for hyperspectral image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 63–78.
[21] J. Liu, Z. Wu, L. Xiao, J. Sun, and H. Yan, "A truncated matrix decomposition for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 8028–8042, Jul. 2020.
[22] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang, "Hyperspectral image super-resolution with optimized RGB guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11661–11 670.
[23] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5344–5353.
[24] Y. Xu, Z. Wu, J. Chanussot, P. Comon, and Z. Wei, "Nonlocal coupled tensor CP decomposition for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 348–362, Jan. 2020.
[25] Y. Xu, Z. Wu, J. Chanussot, and Z. Wei, "Hyperspectral images super-resolution via learning high-order coupled tensor ring representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4747–4760, Nov. 2020.
[26] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5345–5355, Nov. 2018.
[27] A. Khader, L. Xiao, and J. Yang, "A model-guided deep convolutional sparse coding network for hyperspectral and multispectral image fusion," *Int. J. Remote Sens.*, vol. 43, no. 6, pp. 2268–2295, 2022.
[28] 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.
[29] 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.
[30] 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.
[31] W. Dong, C. Zhou, F. Wu, J. Wu, G. Shi, and X. Li, "Model-guided deep hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 5754–5768, May 2021.
[32] W. Wei, J. Nie, L. Zhang, and Y. Zhang, "Unsupervised recurrent hyperspectral imagery super-resolution using pixel-aware refinement," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 5500315.
[33] J. Liu, X. Fan, J. Jiang, R. Liu, and Z. Luo, "Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 105–119, Jan. 2022.
[34] F. Zhou, R. Hang, Q. Liu, and X. Yuan, "Pyramid fully convolutional network for hyperspectral and multispectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 5, pp. 1549–1558, May 2019.
[35] S. Xu, O. Amira, J. Liu, C.-X. Zhang, J. Zhang, and G. Li, "HAM-MFN: Hyperspectral and multispectral image multiscale fusion network with RAP loss," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4618–4628, Jul. 2020.
[36] J. Xiao, J. Li, Q. Yuan, and L. Zhang, "A dual-UNet with multistage details injection for hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5515313.
[37] M. Stephane, *A Wavelet Tour of Signal Processing*. Amsterdam, The Netherlands: Elsevier, 1999.
[38] W. Bae, J. Yoo, and J. Chul Ye, "Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 145–153.
[39] E. Kang, W. Chang, J. Yoo, and J. C. Ye, "Deep convolutional framelet denosing for low-dose CT via wavelet residual network," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1358–1369, Jun. 2018.
[40] C. E. Heil and D. F. Walnut, "Continuous and discrete wavelet transforms," *SIAM Rev.*, vol. 31, no. 4, pp. 628–666, 1989.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[42] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[43] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2015, pp. 234–241.
[44] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4641–4650.
[45] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.
[46] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[47] S. W. Zamir et al., "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14821–14831.
[48] 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.
[49] R. H. Yuhas, J. W. Boardman, and A. F. Goetz, "Determination of semi-arid landscape endmembers and seasonal trends using convex geometry spectral unmixing techniques," in *Proc. JPL, Summaries 4th Annu. JPL Airborne Geosci. Workshop*, 1993, vol. 1.
[50] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[51] L. Wald, *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions*. Paris, France: Presses Des MINES, 2002.
[52] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.
[53] W. He, Y. Chen, N. Yokoya, C. Li, and Q. Zhao, "Hyperspectral super-resolution via coupled tensor ring factorization," *Pattern Recognit.*, vol. 122, 2022, Art. no. 108280.
[54] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "MHF-Net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1457–1473, Mar. 2022.
[55] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5443–5457, Sep. 2018.
[56] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
[57] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12299–12310.
[58] 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, Jul. 2022, Art. no. 6012305.
[59] Q. Ma, J. Jiang, X. Liu, and J. Ma, "Learning a 3D-CNN and transformer prior for hyperspectral image super-resolution," 2021, *arXiv:2111.13923*.

Jian Fang received the B.S. degree in computer science and technology from Hefei University, Hefei, China, in 2017. She is currently working toward the Ph.D. degree in computer science with the Nanjing University of Science and Technology, Nanjing, China.

His research interests include deep learning and remote sensing image processing.

Jingxiang Yang (Member, IEEE) received the Ph.D. degree in control theory and control engineering from Northwestern Polytechnical University, Xi'an, China, in 2019, and the Ph.D. degree in engineering science from 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.

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

His research interests include deep learning and hyperspectral image super-resolution.

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 2006 to 2008, he was a Postdoctoral Research Fellow with the Pattern Recognition Laboratory, NJUST. From 2009 to 2010, he was a Postdoctoral Fellow with Rensselaer Polytechnic Institute, Troy, NY, USA. Since 2013, he has been the Deputy Director of the Jiangsu Key Laboratory of Spectral Imaging Intelligent Perception, Nanjing. Since 2014, he has been the second Director of the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, NJUST, where he is a Professor with the School of Computer Science and Engineering. His research interests include remote sensing image processing, image modeling, computer vision, machine learning, and pattern recognition.