Multistage Dual-Attention Guided Fusion Network for Hyperspectral Pansharpening

Peiyan Guan and Edmund Y. Lam, *Fellow, IEEE*

***Abstract*— Deep learning, especially the convolutional neural network, has been widely applied to solve the hyperspectral pansharpening problem. However, most do not explore the intraimage characteristics and the interimage correlation concurrently due to the limited representation ability of the networks, which may lead to insufficient fusion of valuable information encoded in the high-resolution panchromatic images (HR-PANs) and low-resolution hyperspectral images (LR-HSIs). To cope with this problem, we develop a hyperspectral pansharpening method called multistage dual-attention guided fusion network (MDA-Net) to fully extract the important information and accurately fuse them. It employs a three-stream structure, which enables the network to incorporate the intrinsic characteristics of each input and correlation among them simultaneously. In order to combine as much information as possible, we merge the features extracted from three streams in multiple stages, where a dual-attention guided fusion block (DAFB) with spectral and spatial attention mechanisms is utilized to fuse the features efficiently. It identifies the useful components in both spatial and spectral domains, which are beneficial to improving the fusion accuracy. Moreover, we design a multiscale residual dense block (MRDB) to extract dense and hierarchical features, which improves the representation power of the network. Experiments are conducted on both real and simulated datasets. The evaluation results validate the superiority of the MDA-Net.**

***Index Terms*— Convolutional neural network (CNN), dual-attention guided fusion, hyperspectral pansharpening, multistage fusion.**

I. Introduction

HYPERSPECTRAL imaging technique is capable of acquiring hyperspectral images (HSIs) with hundreds of narrow spectral bands that cover a wide spectral range [1], [2]. Each spatial pixel of the HSI contains abundant spectral information of the object. Because of the high discriminative capability in the spectral domain, hyperspectral imaging has been widely used in various remote sensing applications, such as geochemical exploration [3], land-cover classification [4], and environment monitoring [5]. However, due to the hardware limitation, tradeoffs always exist between spatial and spectral resolutions while acquiring HSIs [6]. Specifically, the HSIs with rich spectral information have a rather low spatial resolution, which hinders their practical applications. On the other hand, panchromatic images (PANs) with one single band usually have a much higher spatial resolution. Considering that the high-resolution PAN (HR-PAN) and low-resolution HSI (LR-HSI) contain abundant spatial information and spectral information separately, one way to acquire an HR-HSI is to fuse the HR-PAN and LR-HSI of the same scene. This process is referred to as hyperspectral pansharpening [7]. It greatly promotes the application of HSI in accurate scene classification and object identification.

Recently, deep learning has drawn much attention and demonstrates impressive performance in image processing and computer vision areas [8]–[11]. In particular, a convolutional neural network (CNN) is applied for various image restoration tasks, such as image superresolution [12], denoising [13], and inpainting [14]. In hyperspectral imaging, researchers also apply CNN to deal with the pansharpening problem, and among them, the pansharpening network (PNN) [15], PanNet [16], and the two-stream fusion network [17] are some representative methods. Nevertheless, there still exist several challenges.

1) Due to the large gap between the LR-HSI and HR-PAN in spectral range and spatial resolution, it is difficult to adaptively merge the spatial details in HR-PAN into all bands of LR-HSI according to the spectral characteristics. Thus, how to fully extract the essential spectral information and spatial information from the two inputs and fuse them together accurately still remains a critical challenge for the representation capability of the networks.
2) Most CNN methods integrate the two images either on the raw or the feature level. Such fusion strategy fails to explore the interimage correlation and intraimage characteristics simultaneously, which may cause insufficient integration of information embedded in the two images.
3) The existing methods treat the extracted representations at different locations or bands equally during the fusion procedure, which neglects to differentiate the valuable versus redundant information. Fusing the raw features without reinforcement may cause serious spectral and spatial distortion in the reconstructed result.

To overcome these issues, we propose a novel hyperspectral pansharpening method, called multistage dual-attention guided fusion network (MDA-Net), to efficiently extract and fuse the valuable spatial information and spectral information in

Manuscript received May 22, 2021; revised July 14, 2021 and August 16, 2021; accepted September 6, 2021. Date of publication October 5, 2021; date of current version January 26, 2022. This work was supported in part by the Research Grants Council of Hong Kong (GRF) under Grant 17201818, Grant 17200019, and Grant 17201620; and in part by The University of Hong Kong under Grant 104005864. *(Corresponding author: Peiyan Guan.)*

The authors are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, SAR, China (e-mail: pyguan@eee.hku.hk; elam@eee.hku.hk).

Digital Object Identifier 10.1109/TGRS.2021.3114552

LR-HSI and HR-PAN step by step. The main contributions of this article can be summarized as follows.

1) We implement our method with a three-stream CNN architecture to sufficiently learn both interimage and intraimage information concurrently for pansharpening. The entire fusion procedure is completed in multiple stages, where the feature maps extracted by the three streams are fused on various levels. This design enables the network to extract adequate valuable contents of different inputs and combine them together more completely. We use the three-stream architecture and multistage fusion to address the first challenge on learning essential useful information and the second one on insufficient integration, respectively.
2) We propose the dual-attention guided fusion block (DAFB), which incorporates both spectral and spatial attention mechanisms, to guide the fusion of features extracted by the three streams. This block uses a dual-attention mechanism to highlight the desired information and suppress the redundant components of the extracted features before fusing them, which greatly improves the spatial and spectral prediction accuracy. The DAFB overcomes the third challenge by differentiating the useful versus redundant contents while fusing the extracted features.
3) We develop the multiscale residual dense block (MRDB) to extract rich, dense, and multiscale hierarchical features of the input images. This block enhances the representation capability of the MDA-Net by enabling it to extract spatial details over large areas and spectral correlation across a wide spectral range. The MRDB addresses the first challenge of improving the representation ability of the network by extracting features of multiple scales.

The rest of this article is organized as follows. Section II introduces the related work on hyperspectral pansharpening and the attention mechanism used in our method. Section III describes the network architecture and detailed design of the DAFB and MRDB. Section IV presents the experimental results on collected datasets and ablation study of the MDA-Net. We conduct the experiments on both simulated and real datasets, and the evaluation results demonstrate the superiority of our method in hyperspectral pansharpening. Conclusions are drawn in Section V.

II. Related Work

A. Traditional Methods

Over the years, many methods have been developed for hyperspectral pansharpening. They can be categorized into four classes: component substitution (CS), multiresolution analysis (MRA), Bayesian methods, and matrix factorization. CS methods substitute the spatial component of the LR-HSI with the corresponding HR-PAN. Then, inverse transformation is employed to restore the fused image. These methods are often easy and fast to implement. They can also preserve spatial information, but the drawback is that they typically suffer from severe spectral distortion. Example methods include principal component analysis (PCA) [18], Gram–Schmidt adaptive (GSA) technique [19], and intensity–hue–saturation (IHS) [20]. MRA methods first perform a multiresolution decomposition on the HR-PAN and then inject the spatial structure information into the upsampled LR-HSI to acquire the HR-HSI. Compared to CS methods, MRA gives better performance in preserving spectral information but is more susceptible to coregistration error between two images. Examples of the MRA methods are the wavelet transform methods [21], Indusion [22], smoothing filter intensity modulation (SFIM) [23], and modulation transfer function (MTF) methods [24], [25]. Bayesian methods regard fusing images as an optimization problem and use some prior constraints as regularization. Representative Bayesian methods include Bayesian naive Gaussian prior [26], Bayesian posterior probability [27], and Bayesian HySure [28]. Matrix factorization methods employ spectral unmixing to transform the observed image into signal subspace. By fusing the learned spectral end-members and spatial abundance maps, these methods manage to reconstruct the HSIs with abundant spatial information. One typical method of this type is the coupled nonnegative matrix factorization (CNMF) [29]. The Bayesian and matrix factorization methods both require optimization. They perform well on restoring both spatial information and spectral information but require huge computation costs. In addition, the inadequate representation capability of the models usually limits the performance of these methods.

B. CNN Methods

Recently, deep learning methods have shown notable achievements in the image super-resolution area [12], [30], which is quite related to hyperspectral pansharpening. Existing works treat the HR-PAN as an extra band of the LR-HSI and concatenate them to form the input to the network. These early fusion methods combine the two input images on the raw level, which generally does not explore the distinct characteristics of HR-PAN and LR-HSI separately. One early example is the pansharpening neural network (PNN), which utilizes three convolutional layers to learn the mappings between the HR-HSI and the two stacked images [15]. Similarly, Palsson *et al.* [31] employ a 3-D-convolutional layer to improve the feature extraction capability of the network. In addition, the multiscale and multidepth CNN (MSDCNN) improves the PNN by employing a multiscale convolution block to enhance the representation power [32]. However, these methods are still incapable of extracting deep features with spectral or spatial information from the corresponding LR-HSI and HR-PAN separately.

In order to overcome this drawback, other researchers propose to fuse the HR-PAN and LR-HSI on the feature level, which is called late fusion. Shao and Cai [33] propose to use a two-branch architecture to extract high-level features of two inputs separately and fuse the extracted features by elementwise addition. This method is referred to as a remote sensing image fusion network (RSIFNN). The two branches enable the network to explore the intrinsic characteristics

of each input image. Following this manner, Liu *et al.* [17] propose to use two subnetworks to extract spectral and spatial features, respectively, and integrate the features by using a fusion network with several convolutional layers. In addition to combining the features of the two inputs, some methods propose to fuse one raw image with the deep features extracted from the other input. Zhou *et al.* [34] propose a pyramid fully CNN (PFCNN) to integrate the HR-PAN of various scales into the features extracted from LR-HSI of the same scale. He *et al.* [35] propose a spectrally predictive CNN (HyperPNN) to extract the spectral features with two layers from the LR-HSI first and fuse them with the HR-PAN for the following reconstruction. These methods succeed in extracting different features of each input image but do not learn the correlation between the LR-HSI and HR-PAN during the feature extraction procedure. These methods suffer from a misalignment of features extracted by different branches, which may cause significant spectral and spatial distortion.

C. Attention Mechanism

Attention mechanism has been widely used to enhance the deep learning models for image processing, such as image super-resolution, identification, and classification. The principle behind the attention mechanism is to automatically highlight the most informative components and suppress the insignificant ones for better computational efficiency. Hu *et al.* [36] first propose to employ channel attention to improve the performance of the network on object identification. Roy *et al.* [37] propose to generate the spatial attention map by using a convolutional layer with 1×1 kernel. Woo *et al.* [38] propose the convolutional block attention module (CBAM), which generates 3-D attention mask by merging spectral and spatial attention. Inspired by these works, we propose to incorporate an attention mechanism into our fusion block, DAFB, which greatly improves the fusion accuracy and information fidelity.

III. Proposed Method

A. Problem Statement

HR-PAN has abundant spatial information, while LR-HSI, despite the poor spatial quality, contains rich spectral information. The objective of this work is to fuse the complementary spectral and spatial contents to generate HSIs with high spatial resolution.

To accomplish this goal, we propose a deep learning method to adaptively fuse the spatial information and spectral information in multiple stages and reconstruct the HR-HSI from the fused feature maps. Let $\mathbf{Y} \in \mathbb{R}^{w \times h \times C}$ be the LR-HSI with $w \times h$ pixels and C spectral bands and $\mathbf{P} \in \mathbb{R}^{W \times H \times 1}$ be the HR-PAN, where $W \times H$ is the spatial dimension. Generally speaking, the spectral band number of LR-HSI is much larger than that of HR-PAN, while the spatial resolution of HR-PAN is much higher than that of LR-HSI. That is, $W \gg w$, $H \gg h$, and $C \gg 1$. Then, the corresponding HR-HSI is indicated as $\mathbf{X} \in \mathbb{R}^{W \times H \times C}$. By fusing $\mathbf{Y}$ and $\mathbf{P}$, we obtain $\mathbf{X}$ with high spatial and spectral resolutions.

B. Motivation and Model Overview

As mentioned above, the early fusion CNN methods do not make use of the intraimage characteristics of each image, while the late-fusion methods do not explore the correlation between LR-HSI and HR-PAN. Our work aims to learn the intraimage and interimage information of the input images concurrently. We use a parallel three-stream structure to extract features of various levels of LR-HSI, HR-PAN, and their concatenation. In order to fully merge the useful spatial information and spectral information, we combine the features extracted by the three distinct streams in multiple stages. In each stage, features of equivalent levels from the three streams are fused via a DAFB, into which the spectral and spatial attention operations are incorporated. The attention mechanism is employed to recognize the informative contents in the extracted features. By using the attention mask, the original feature maps are rescaled adaptively to focus on the relevant information.

This framework manages to extract abundant multilevel features of intraimage characteristics and interimage correlation concurrently, while the extracted features are fully fused in multiple steps. The dual attention mechanism defines the more important components in both spatial and spectral domains, which further improves the fusion accuracy. The details of the network structure are discussed in the following.

C. Network Design

The workflow of the MDA-Net is depicted in Fig. 1, which consists of two phases. In the first phase, three parallel streams, which refer to the three subnetworks that take HR-PAN, upsampled LR-his, and their concatenation as input, respectively, explore the intrinsic characteristics of each kind of image and the correlation among them. Specifically, LR-HSI is upscaled to the same size as HR-PAN via bicubic interpolation. Extracted features from the three streams are then merged in multiple stages. The fused feature of the last stage is used as the input of the next phase, where a reconstruction module learns to map the merged features into the desired HR-HSI.

The streams taking HR-PAN or upsampled LR-HSI as input extract multilevel features of the corresponding image, while the other one exploits their correlation. As shown in Fig. 1, the three streams have a similar structure, except for three additional DAFBs in the middle stream. At the beginning of each stream, one convolutional layer maps the input image into features with the same 64 channels. Such features are used as inputs of the first fusion stage. Then, MRDBs extract rich and hierarchical features of different inputs, which are then integrated by a DAFB. The new fusion feature is propagated to the next stage. The DAFB uses both spectral and spatial attentions to merge the extracted features under the guidance of learned intraimage/interimage information, where the predicted attention maps refine the input features for more accurate fusion. The structures of MRDB and DAFB are discussed in detail in the following.

The feature extraction and fusion procedure are repeated for L stages. In the end, the fusion feature merges useful information extracted by the three streams from all previous L stages. After the multistage feature fusion, a reconstruction

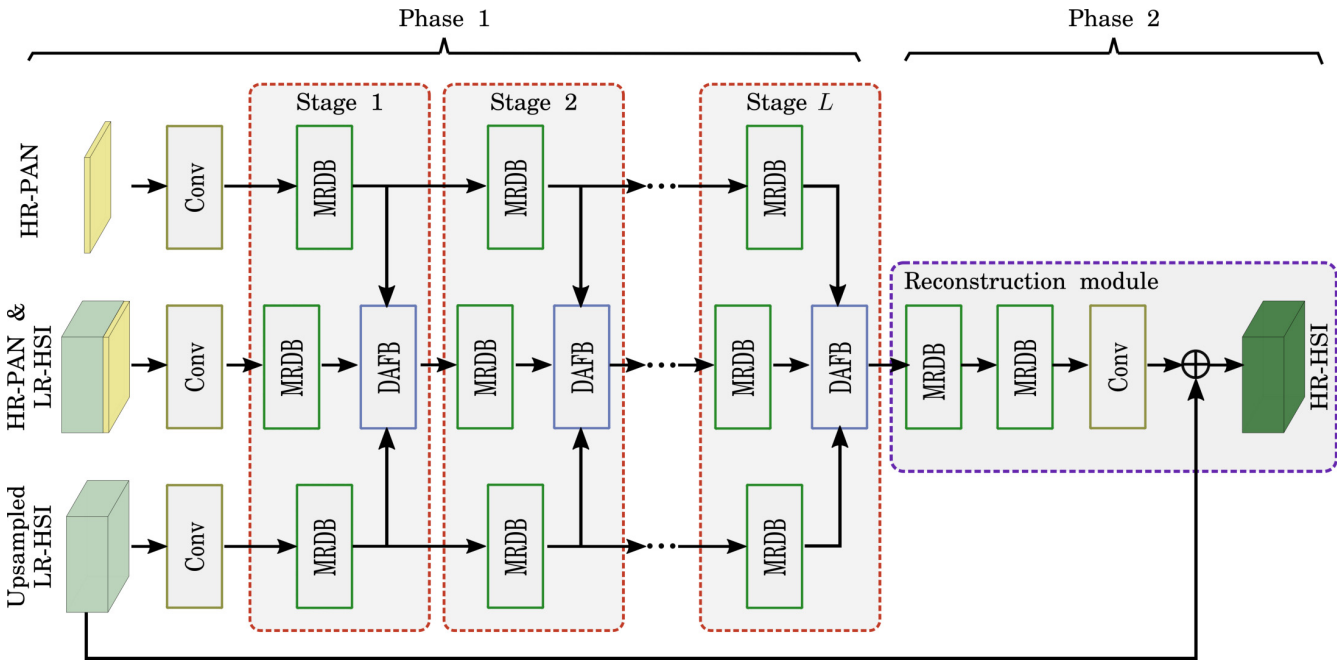

Fig. 1. Detailed schematics of the MDA-Net.

module with two MRDBs followed by one convolutional layer is applied to estimate the latent residual HR-HSI based on the fusion feature. To avoid learning redundant information, we conduct global residual learning by creating a skip connection between the upsampled LR-HSI and the reconstructed residual, while the latent HR-HSI can be obtained by adding them. The residual learning strategy makes the network converge faster and solves the gradient vanishing problem in a deep network.

1) Dual-Attention Guided Fusion Block: Let $F_{(y,l)}$, $F_{(p,l)}$, and $F_{(c,l)}$ denote the feature maps produced by the MRDBs of the three streams, whose inputs are upsampled LR-HSI, HR-PAN, and their concatenation, respectively, at stage l. The three feature maps have identical size of $W \times H \times 64$. The first two, $F_{(y,l)}$ and $F_{(p,l)}$, either lack spatial or spectral information, while $F_{(c,l)}$ has both. Thus, we use $F_{(c,l)}$ as the guidance to enhance $F_{(y,l)}$ and $F_{(p,l)}$ separately.

As shown in Fig. 2, we feed the concatenation of $F_{(y,l)}$ and $F_{(c,l)}$ into the spatial attention module to generate a spatial attention map for the former. With the guidance of rich information on texture details and correlation among two images encoded in $F_{(c,l)}$, this spatial attention map could enhance $F_{(y,l)}$ over various spatial locations. In a similar way, we obtain a spectral attention mask for $F_{(p,l)}$ by using the spectral attention module. This mask is able to reinforce $F_{(p,l)}$ over various spectral channels. This process can be defined as

$$A_{\text{spa}} = f_{\text{spa}}([F_{(y,l)}, F_{(c,l)}]; \theta_{\text{spa}}) \tag{1}$$

$$A_{\text{spe}} = f_{\text{spe}}([F_{(p,l)}, F_{(c,l)}]; \theta_{\text{spe}}) \tag{2}$$

where $f_{\text{spa}}(\cdot; \theta_{\text{spa}})$ and $f_{\text{spe}}(\cdot; \theta_{\text{spe}})$ represent the functions of the spatial and spectral attention modules, respectively. θ_{spa} and θ_{spe} are the parameters of corresponding attention modules. $[F_{(y,l)}, F_{(c,l)}]$ and $[F_{(p,l)}, F_{(c,l)}]$ denote the concatenation

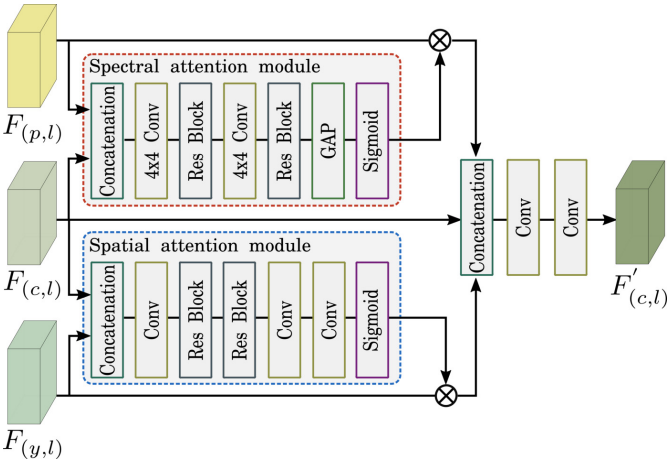

Fig. 2. Architecture of the DAFB.

of the corresponding input features. A_{spa} and A_{spe} indicate the generated attention maps, which are of size $W \times H \times 1$ and $1 \times 1 \times 64$, respectively. Details of these two attention modules will be given in the following. The attention masks are used to refine the features $F_{(y,l)}$ and $F_{(p,l)}$ via

$$F'_{(y,l)} = A_{\text{spa}} \circ F_{(y,l)} \tag{3}$$

$$F'_{(p,l)} = A_{\text{spe}} \circ F_{(p,l)} \tag{4}$$

where $\circ$ denotes elementwise multiplication, During the multiplication, attention maps A_{spa} and A_{spe} are broadcasted to the size of $F_{(y,l)}$ and $F_{(p,l)}$, respectively. $F'_{(y,l)}$ and $F'_{(p,l)}$ represent the refined features with attention guidance. We concatenate the refined features $F'_{(y,l)}$, $F'_{(p,l)}$, and $F_{(c,l)}$ and encode the stacked features into a more compact representation by using

two convolutional layers as

$$F'_{(c,l)} = W_2 * \left(\delta\left(W_1 * \left[F'_{(y,l)}, F'_{(p,l)}, F_{(c,l)}\right]\right)\right) \tag{5}$$

where $\delta(\cdot)$ indicates the ReLU activation function and $*$ denotes the convolution operation. $F'_{(c,l)}$ with the same size as $F_{(c,l)}$ represents the output of DAFB, which is propagated to the next stage. W_1 and W_2 are weights of the two convolutional layers, respectively, where the bias term is omitted for simplicity.

DAFB employs spatial and spectral attention mechanisms to adaptively recalibrate $F_{(y,l)}$ and $F_{(p,l)}$ under the guidance of $F_{(c,l)}$. The dual-attention mechanism enhances the channelwise and locationwise contents in the extracted features, which greatly reduces spectral and spatial distortion in the fusion process. The effectiveness of the DAFB is further studied in Section IV-D.

1) *Spatial Attention Module:* As shown in Fig. 2, the spatial attention module takes the concatenation of $F_{(y,l)}$ and $F_{(c,l)}$ as input. It consists of one convolutional layer, two residual blocks, and two following convolutional layers. The residual block adopts the design in [39] without the batch-normalization layers. In the end, a sigmoid function is used to generate spatial attention maps for $F_{(y,l)}$ with values between 0 and 1. Under the guidance of spatial–spectral information in $F_{(c,l)}$, this module highlights the informative spatial components in $F_{(y,l)}$, which is beneficial to improving spatial fidelity.
2) *Spectral Attention Module:* The structure of the spectral attention module is depicted in Fig. 2. It takes the stacked $F_{(p,l)}$ and $F_{(c,l)}$ as input. There are two convolutional layers, and each of them is followed by a residual block. The two convolutional layers comprise 4×4 kernels with a stride of 4 to aggregate the spatial information gradually. The residual blocks have the same structure as those in the spatial attention module. Then, a global average pooling (GAP) layer is employed to squeeze features into a vector, the values of which are then rescaled into $[0, 1]$ by the following sigmoid function. The size of the attention vector is identical with the channel number of $F_{(p,l)}$. This module refines $F_{(p,l)}$ channelwise by using the interimage information in $F_{(c,l)}$, which reduces spectral distortion in the fusion result.

2) Multiscale Residual and Dense Block: The conventional residual dense block (RDB) usually employs multiple convolutional layers with small kernels to extract deep and dense features [12]. However, it does not exploit the nonlocal similarities in the spatial domain due to the small receptive field of the used filters. Besides, target objects in HSIs are usually of various scales, while information of multiple scales is required for accurate reconstruction. Thus, we introduce multiscale convolution into the RDB to extract more hierarchical features.

To give an example, we take the MRDB in the stream with HR-PAN as input. As illustrated in Fig. 3, it takes $F_{(p,l-1)}$ as input and then generates the output $F_{(p,l)}$ at stage l. The MRDB utilizes four branches comprising of convolutional layers of different kernel sizes to separately extract multiscale

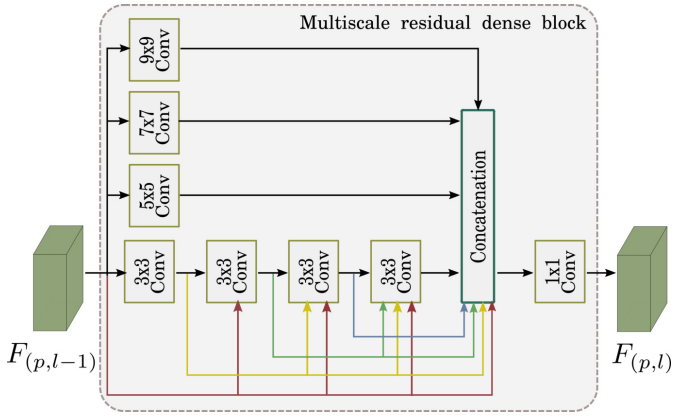

Fig. 3. Structure of the MRDB.

information of the input feature. The bottom branch stacks four 3×3 layers to extract dense small-scale features. It adopts the classical dense design in RDB, where each layer takes the concatenation of outputs from the previous layers as input. In contrast, in order to reduce computation complexity, only one convolutional layer with larger filters is used for each of the other three branches. Specifically, filters in these three branches are set to 5×5, 7×7, and 9×9, respectively. Such design allows us to extract both dense and multiscale features of inputs efficiently. Feature maps generated by these four branches are then stacked and fed into one local fusion layer to learn a more compact representation. The local fusion layer, which has one 1×1 convolutional layer, reduces the number of concatenated features to 64.

D. Loss Function

In addition to the network structure, the loss function, which guides the training of a deep network, is also a key factor. Most previous methods for hyperspectral pansharpening utilize the ℓ_2 norm as the loss function to optimize the parameters of the networks. However, the ℓ_2 norm generally leads to blurry and oversmooth images. Instead, we use the ℓ_1 norm as the loss function to train the MDA-Net. For a training set $\{\mathbf{Y}_i, \mathbf{P}_i, \mathbf{X}_i\}_{i=1}^{k}$, where $\mathbf{Y}_i$, $\mathbf{P}_i$ and $\mathbf{X}_i$ refer to the ith LR-HSI, HR-PAN, and HR-HSI, respectively, the ℓ_1 loss function can be computed as

$$\ell_1 = \frac{1}{k}\sum_{i=1}^{k} |\Phi(\mathbf{Y}_i, \mathbf{P}_i; \Theta) - \mathbf{X}_i| \tag{6}$$

where $\Phi(\cdot; \Theta)$ represents the function of the MDA-Net and Θ denotes all the parameters involved in the network.

IV. Experimental Results and Discussion

A. Dataset

Experiments are conducted on three simulated hyperspectral datasets and one real hyperspectral dataset to validate the performance of the MDA-Net. The collected datasets are described as follows.

1) Pavia University: The image of Pavia University is captured by the reflective optics system imaging spectrometer (ROSIS) during a flight over Pavia, Italy. The spatial resolution of this image is 1.3 m. After removing noisy bands, we have an image with a size of 610 × 610 and 103 spectral bands in the spectral range of 430–860 nm. Since one area in the right contains no information, we only use the top left corner with 560 × 336 pixels of each band in our experiment. It is then subdivided into 15 nonoverlapping patches of size 112 × 112 × 103 to form the PaviaU dataset. Ten patches are randomly chosen to form the training dataset, while the remaining five patches are used for testing.

2) Houston: It is captured over the campus and neighboring urban area of the University of Houston. There are 144 spectral bands covering the spectral region of 380–1050 nm, each of which contains 1905 × 349 pixels. We only use the top left area of size 1680 × 336 for our experiment. The selected region is then subdivided into five patches of size 336 × 336 without overlapping to constitute the dataset. For the Houston dataset, three patches are selected randomly to train the network, and the remaining two patches are used for validation.

3) Washington DC Mall: It is acquired by the hyperspectral digital imagery collection experiment (HYDICE) over the Washington DC mall area. The water absorption bands are discarded, and there remain 191 bands ranging from 400 to 2500 nm, each of which has a size of 1280 × 307. For the experiment, the top left area with 1216 × 304 pixels is cropped and then partitioned into 16 nonoverlapping patches to form the dataset. We randomly select 11 patches for the network training, and the remaining five patches comprise the testing dataset.

4) WorldView-4: This dataset consists of image pairs of LR-HSI and HR-PAN of the same scene over Acapulco in Mexico, which are collected by the WorldView-4 (formerly GeoEye2) satellite. It is used for our real hyperspectral pansharpening experiment. This satellite is capable of capturing LR-HSI with a spatial resolution of 0.31 m and HR-PAN with a spatial resolution of 1.24 m simultaneously. In other words, the resolution ratio between these two kinds of images is 1 : 4. We discard the water area on the left. The remaining parts are subdivided into 16 image patch pairs without overlapping. The LR-HSI patch contains 128 × 128 pixels in each band, while the dimension of the corresponding HR-PAN patch is 512 × 512.

The Pavia University, Houston, and Washington DC Mall datasets are all simulated hyperspectral datasets. For these three datasets, the original HSIs are considered as the reference HSIs, from which the LR-HSI and HR-PAN pairs can be simulated. Specifically, the LR-HSI is obtained by Gaussian blurring the reference HSI with an 8 × 8 filter and then downsampling the blurred image by a certain scale factor. The scale factor is set to 8 for the Washington DC Mall dataset and 4 for the PaviaU and Houston datasets. To create the corresponding HR-PAN, we average the visible bands of the same reference HSI.

For the experiment on real data, ground-truth HR-HSIs are not available in the WorldView-4 dataset. Therefore, we follow Wald's protocol to generate the training data [40]. Based on this method, the original LR-HSI and HR-PAN are first downsampled by the resolution ratio between them separately. Then, the downsampled image pairs are used as the input to train the network, while the original LR-HSI is considered as the reference image. After the training procedure, the proposed method is assessed on the image pairs of the original LR-HSI and HR-PAN.

B. Evaluation Metrics

In order to evaluate the performance of the MDA-Net on the simulated hyperspectral datasets objectively, five widely used metrics are employed, i.e., root mean square error (RMSE), peak signal-to-noise ratio (PSNR), spectral angle mapper (SAM) [41], structure similarity index (SSIM) [42], and erreur relative globale adimensionnelle de synthèse (ERGAS) [43]. RMSE measures the intensity difference between the reconstructed image and ground truth. The results in terms of RMSE are all absolute values. As a function of RMSE, PSNR reflects the reconstruction quality of the fused image. SAM is used to assess the spectral fidelity of the generated image. SSIM evaluates the spatial structure similarity between the fused result and the reference image. ERGAS assesses the global quality of the produced image. For RMSE, SAM, and ERGAS, a lower value implies better reconstruction quality. In contrast, the higher the PSNR and SSIM, the better the performance.

Since there is no reference image in the experiment on real data, we employ the no-reference HSI quality assessment method proposed in [44] to evaluate the pansharpening performance of different methods. This metric uses some pristine HSIs as training data to acquire the benchmark features. Then, it extracts a set of features from the test image and predicts the quality by comparing the extracted features with those from the training data. In our experiment, to ensure the accuracy of assessment, we follow the suggestions in [45] to use the original LR-HSIs of the WorldView-4 satellite to train the evaluation model. For this metric, a smaller value represents better image quality.

C. Implementation Details

Here, we provide the details of the model parameters, network design, and training settings. If not specified otherwise, the default convolutional layers are set to contain 64 kernels of size 3 × 3 with a stride of 1 followed by an ReLU function in the MDA-Net. Meanwhile, zero padding is applied in every convolutional layer to retain the size of feature maps by default. As such, the last convolutional layer has the same number of filters as the spectral bands of reference HR-HSI. In our implementation, the growth rate of the bottom branch of MRDBs is set to 32.

For the network training, the reference HR-HSI, the pixel intensities of which are normalized to $[0, 1]$, is partitioned into patches of size 64 × 64 with a stride of 32 to generate the training pairs by using the described above simulation. The batch size is set to 16 in our experiment. The training process is completed after 300 epochs. The ℓ_1 loss function is minimized by using Adam optimizer [46] with

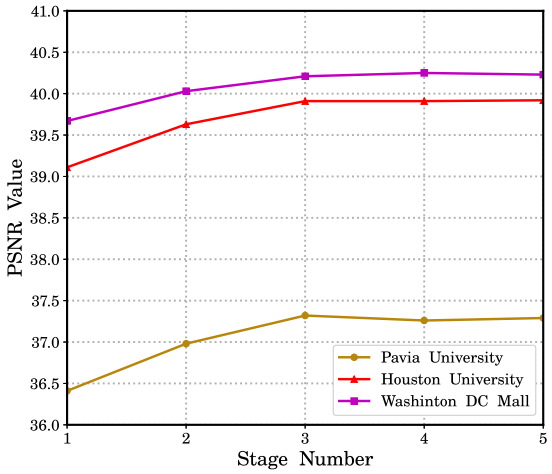

Fig. 4. Quantitative evaluation results of the number of stage L.

TABLE I

QUANTITATIVE RESULTS OF ABLATION STUDY OF THE MDA-NET ON THE HOUSTON DATASET. THE BEST VALUE IS IN BOLD

Methods	PSNR	SAM	SSIM	ERGAS	RMSE
MDA-Net (one stage)	39.37	5.145	0.9543	2.663	0.0108
MDA-Net (one stream)	39.45	5.036	0.9553	2.626	0.0107
MDA-Net (two streams)	39.34	5.162	0.9541	2.666	0.0108
MDA-Net w/o attention	39.30	5.134	0.9538	2.670	0.0108
MDA-Net w/o spectral att	39.44	5.035	0.9552	2.640	0.0107
MDA-Net w/o spatial att	39.48	5.012	0.9555	2.628	0.0106
MDA-Net (res block)	39.28	5.125	0.9536	2.675	0.0109
MDA-Net (RDB)	39.39	5.095	0.9546	2.654	0.0107
MDA-Net	**39.91**	**4.747**	**0.9596**	**2.495**	**0.0101**

$\beta_1 = 0.9$, while the learning rate is set to 2×10^{-4} initially and decreases 10% for every 30 epochs. The network is implemented based on Pytorch, and parameters are computed on a Linux OS computer with an Intel Core i5-6500 CPU (3.2 GHz and four cores), 16-GB physical memory, and an Nvidia RTX 2060s graphics card. The code is available at https://github.com/pyguan88/MDA-Net.

The network extracts the useful information of various inputs and integrates them in multiple stages. The fusion performance is highly dependent on the number of stages denoted by L. To investigate how this value affects performance, we construct several variants of the MDA-Net, each containing a different number of stages, while the other settings of these variants remain the same. The experiment is conducted on the three simulated datasets, and PSNR is employed to evaluate the performance of the variants quantitatively. The results are presented in Fig. 4. Generally speaking, the network gives better fusion performance as L increases. In particular, the MDA-Net achieves the best performance on all the three datasets when $L = 3$ and does not deliver significant performance improvement, while more stages are employed. However, the computational complexity of the network increases greatly as it goes deeper. Taking both the pansharpening performance and the network complexity into consideration, the number of stage L is set to 3 in the subsequent experiments.

D. Ablation Study

We conduct an ablation study of the MDA-Net on the Houston dataset to assess the effectiveness of the network structure, DAFB, and MRDB.

1) Effect of the Network Structure: One major contribution of this work is the architecture of the network. Its superiority mainly lies in two aspects: three parallel streams and multistage fusion. In this experiment, we evaluate the effect of the three-stream and multistage fusion architecture, which are proposed in our first contribution. The three-stream structure enables the network to explore the intraimage characteristics and interimage correlation concurrently, while the multistage-fusion design fuses the valuable information of different inputs more completely. To demonstrate the effect of the network structure, we employ three variants of the MDA-Net, each of which has fewer streams or fusion stages. The first one is referred to as "MDA-Net (one stream)," which keeps the stream that takes the concatenation of HR-PAN and upsampled LR-HSI as input and eliminates the other two. It has a similar structure to the early fusion methods. The second one, which is called "MDA-Net (two streams)," keeps the two streams that take HR-PAN and upsampled LR-HSI as inputs, respectively, and removes the other one. This variant can be categorized as a late-fusion method. The last one is termed "MDA-Net (one stage)," where the first two DAFBs are removed and the extracted features from the three streams are only merged in the last stage.

The experimental results are given in Table I. As can be seen, variants with one or two streams both incur obvious performance drop in terms of all the evaluation metrics. This is because they do not exploit the characteristics of each kind of image and the correlation among them simultaneously. The performance of "MDA-Net (one stage)" also drops, which is due to the insufficient integration of various features. The MDA-Net outperforms all the three variants in all the objective indices, which proves the effectiveness of the three-stream structure and the multistage-fusion design proposed in our method.

2) Effect of the DAFB: In this experiment, we assess the effect of the DAFB proposed in the second contribution. We incorporate both spectral and spatial attention mechanisms into the DAFB. The dual attentions identify and enhance the valuable components of features from different streams under the guidance of learned intraimage/interimage knowledge. By refining the features of different inputs before integrating them together, the fusion accuracy could be further improved. To verify the significance of the two attention modules in the DAFB, we construct three variants of the MDA-Net, which are termed "MDA-Net w/o attention," "MDA-Net w/o spectral att," and "MDA-Net w/o spatial att." The first one removes both spatial and spectral attention modules from the DAFB and fuses the raw features directly. The second removes the spectral attention module while keeping the spatial

Fig. 5. Visual comparison of different methods on the selected patch "PU_10" of the Pavia University dataset with a scale factor of 4. The yellow box represents the zoomed-in ROI.

attention one. The last one works the opposite way to the second one.

The quantitative evaluation results of these variants are presented in Table I for comparison. The "MDA-Net w/o attention" comes across obvious performance drop without attention assistance. In contrast, incorporating a single spatial or spectral attention module improves the fusion accuracy noticeably. Moreover, our DAFB with two attention modules gives the best performance compared to these three variants, which demonstrates the effectiveness of the dual-attention mechanism for feature fusion.

3) Effect of the MRDB: In this experiment, we evaluate the effect of the MRDB developed in our third contribution. We apply multiple convolutional layers with kernels of various sizes in the MRDB. Layers with small filters extract the deep features of local information, while those with large filters exploit the global spatial similarities. Such design gives the MRDB strong representation capability. All these layers jointly extract the multiscale information of the input feature maps, which is beneficial to reconstructing targets of various sizes. To illustrate the effectiveness of the MRDB, we compare MDA-Net with two variants that are called "MDA-Net (res block)" and "MDA-Net (RDB)." These two variants replace the MRDBs in the MDA-Net with residual blocks, which adopt the same design as those in DAFB, and conventional RDBs, which have the identical structure with the bottom branch in our MRDBs, respectively.

We present the quantitative comparison of these variants in Table I. It can be observed that "MDA-Net (res block)" with residual blocks gives the lowest scores on the evaluation metrics. Replacing MRDBs with conventional RDBs also causes an obvious performance drop. Our method with the MRDBs outperforms the two variants in terms of all indices, which verifies the powerful representation ability of our MRDB.

E. Experiments on Simulated Datasets

In order to evaluate the performance of the MDA-Net, we compare it with eight pansharpening methods belonging to different categories, namely, PCA [18], GSA [19], SFIM [23], Indusion [22], and four state-of-the-art CNN methods, including MSDCNN [32], HyperPNN [35], TFNet [17], and DARN [47]. We reimplement the CNN methods by Pytorch. These methods are trained with the same simulated datasets as the MDA-Net for a fair comparison. For HyperPNN, the best HyperPNN2 model with skip connection is selected for comparison. In addition, the residual variation ResTFNet of TFNet with better performance is used in our experiment. We also employ four CSA ResBlocks, which are the same as the original paper, to build the DARN model for the evaluation. Except for the CNN methods, the four traditional methods are implemented with the released official MATLAB codes. The parameters of all the comparison methods are set to follow the suggestions of the original papers to give the best performance. Experiments are conducted on three simulated datasets: Pavia University, Washington DC Mall, and Houston datasets. Results on each dataset are presented as follows.

1) Pavia University Dataset: The MDA-Net is compared to the aforementioned methods on the test subset of the Pavia University dataset with a scale factor of 4. The quantitative evaluation results are given in Table II, where the best values are marked in bold. CNN methods outperform the traditional methods on the objective metrics clearly. Among them, TFNet does a better job than MSDCNN and HyperPNN because it has a deeper structure and explores the intrinsic characteristics of HR-PAN and LR-HSI before fusing them together. DARN gives better results due to the attention mechanism used in the residual blocks. The MDA-Net, which explores the intraimage characteristics and interimage correlation simultaneously, outperforms other methods significantly.

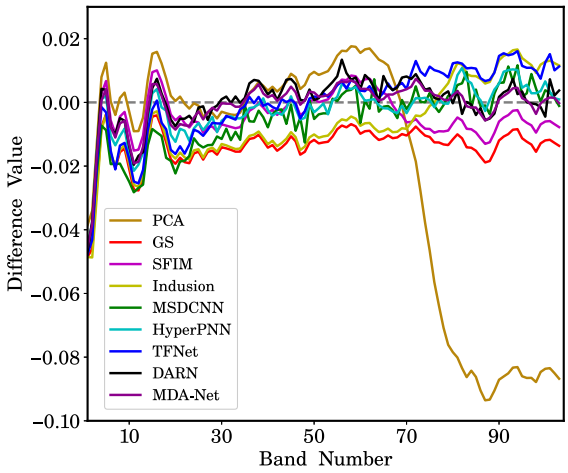

Fig. 6. Comparison of the spectral reflectance difference curve of a selected point (30, 80) of the patch "PU_10."

In addition to the above numerical assessment, the visual comparison on a randomly selected image patch "PU_10" is presented in Fig. 5. One region of interest (ROI) is also zoomed in for each image patch to better demonstrate the reconstruction results. PCA and GSA produce very blurry images with serious spectral distortion, while SFIM, Indusion, MSDCNN, and HyperPNN obtain reconstruction results with severe ringing artifacts. TFNet and DARN generate relatively clearer images but recover unfaithful edges and texture details. As shown in the zoomed-in regions in Fig. 5, MDA-Net succeeds in recovering the accurate contour and edge of the lawn, while all the other methods fail. This demonstrates the superior capability of our method in restoring faithful spatial details and textures.

In addition, we compare these methods on spectral information preservation. One location is randomly selected from patch "PU_10" to exhibit the spectral reflectance difference values between each fused image and ground truth. As shown in Fig. 6, one dotted line serves as the optimal curve for the different values. On the whole, the MDA-Net produces the best spectral preservation result.

2) Houston Dataset: For the experiment on the test subset of the Houston dataset with a scale factor of 4, Table III presents the numerical evaluation results that are similar to those of the Pavia University dataset. As can be seen, the CNN methods outperform the traditional ones again, while the MDA-Net ranks the first in all the quantitative metrics and leads by a large margin.

For the qualitative evaluation of the Houston dataset, patch "Houston_5" is selected to present the performance of different approaches. As shown in Fig. 7, severe ringing artifacts occur in the generated images by SFIM and Indusion. PCA, GSA, and MSDCNN yield blurry and off-color images, indicating failure of reconstructing spatial details and spectral correlations. HyperPNN and TFNet produce clearer images. However, there still exists a noticeable spectral distortion in the reconstruction results. For example, the color of the athletic field of their fused images changes greatly in comparison with the reference image. By contrast, DARN performs relatively better

TABLE II

QUANTITATIVE RESULTS OF THE DIFFERENT METHODS ON THE PAVIA UNIVERSITY DATASET. THE BEST VALUE IS IN BOLD

Methods	PSNR	SAM	SSIM	ERGAS	RMSE
PCA [18]	21.04	19.49	0.4961	16.70	0.0887
GSA [19]	27.09	7.047	0.8196	7.569	0.0442
SFIM [23]	26.57	9.395	0.8068	9.707	0.0469
Indusion [22]	25.54	8.883	0.7685	8.401	0.0528
MSDCNN [32]	30.65	5.084	0.9070	5.915	0.0293
HyperPNN [35]	31.42	5.376	0.9186	5.262	0.0269
TFNet [17]	34.01	4.613	0.9383	3.918	0.0199
DARN [47]	34.45	4.133	0.9488	3.860	0.0190
MDA-Net	**37.32**	**3.360**	**0.9639**	**2.801**	**0.0136**

on reducing both spatial and spectral distortions. Moreover, the MDA-Net gives the slightest chromatic aberration and recovers sharper edges, as shown in the enlarged area in Fig. 7, which illustrates the strong ability of our method on restoring spectral information and spatial information.

To assess the spectral preservation ability of each method, similar to the previous experiment, we calculate the spectral signature difference between the fused image and ground truth. The results of one random point in patch "Houston_5" are presented in Fig. 8. As can be seen, the curve of the difference value of the MDA-Net is closest to the reference dotted line, which demonstrates that our method produces the least spectral distortion. The superior performance of our method on preserving spectral fidelity also verifies the quantitative results on SAM in Table III.

3) Washington DC Mall Dataset: For the Washington DC Mall dataset, Table IV exhibits the quantitative assessment results of all methods with a scale factor of 8. It can be seen that CNN methods present better performance compared to the traditional ones. In addition, the MDA-Net surpasses all the comparison approaches on the evaluation metrics.

Fig. 9 presents the visual comparison of different methods on the image patch "WDC_15." Serious ringing artifacts also occur in the images produced by SFIM and Indusion. As can be seen, there exist obvious color variations in the grassland of the PCA and GSA fusion images. Besides, the color in the roof parts of the fused images of HyperPNN and TFNet differs from that of the reference image greatly, which indicates a serious loss of spectral information. MSDCNN and DARN present better performance in preserving spectral information. However, the reconstruction results of these methods are still blurry. For the grass area (pointed by the red arrow) in the enlarged subimages in Fig. 9, only our method can recover it accurately, while all the other methods fail to reconstruct it faithfully to the ground truth. In other words, a fused image of our method is closest to the ground truth, which demonstrates that our model performs better in spatial restoration and spectral preservation.

In addition, we draw the spectral reflectance difference curves for various methods to compare their spectral preservation ability. The comparison results of one randomly

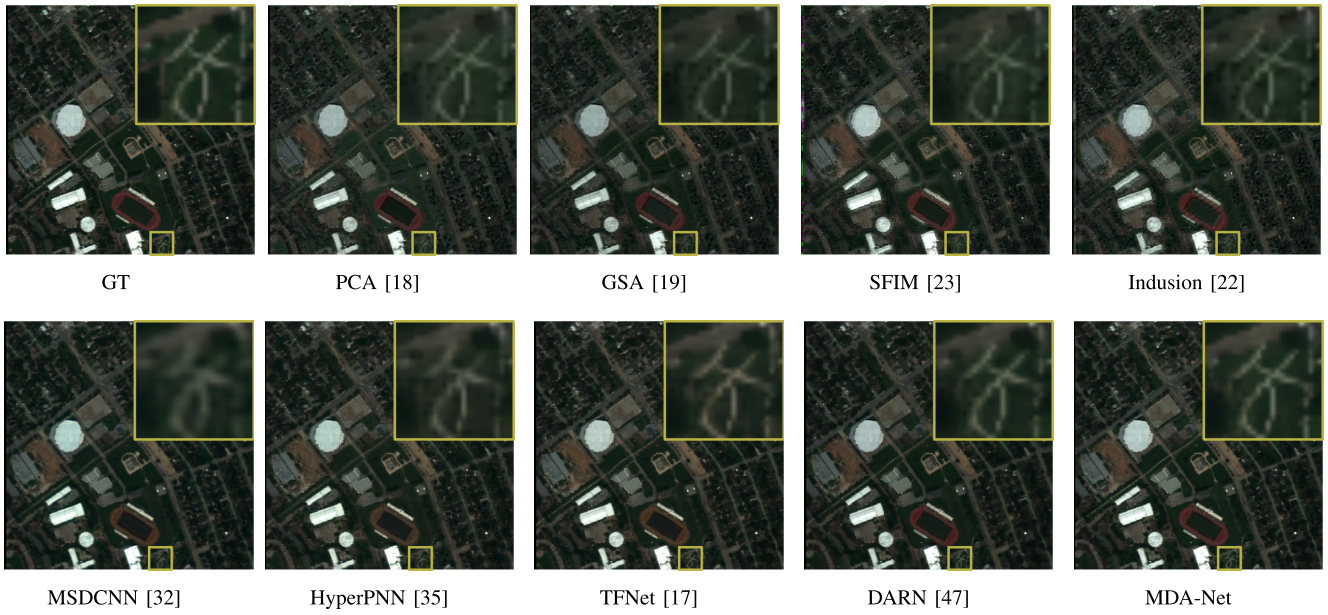

Fig. 7. Visual comparison of different methods on the selected patch "Houston_5" of the Houston dataset with a scale factor of 4. Yellow boxes represent the zoomed-in ROIs.

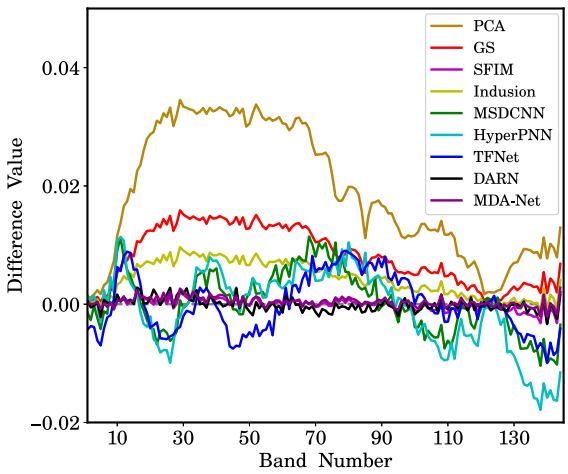

Fig. 8. Comparison of the spectral reflectance difference curve of a selected point (221, 129) of the patch "Houston_5."

chosen point in patch "WDC_15" are presented in Fig. 10. As can be seen, the MDA-Net curve is closest to the dotted reference line, which demonstrates the best effectiveness of our method on preserving spectral information.

F. Experiments on Real Dataset

To evaluate the generality of the MDA-Net, we perform experiments on the real hyperspectral dataset, i.e., the WorldView-4 dataset. Since there is no ground-truth HR-HSI available, we follow the no-reference assessment method in [44] to quantitatively evaluate the pansharpening performance of different methods, which is shown in Table V. As can be seen, the MDA-Net obtains the best result.

The visual comparison on the selected patch "WV4_9" is depicted in Fig. 11. We use the upsampled LR-HSI via bicubic interpolation as the reference image. It can be observed that there exist obvious artifacts in the reconstruction results of Indusion and SFIM. PCA and GSA produce images that are

TABLE III

QUANTITATIVE RESULTS OF THE DIFFERENT METHODS ON THE HOUSTON UNIVERSITY DATASET. THE BEST VALUE IS IN BOLD

Methods	PSNR	SAM	SSIM	ERGAS	RMSE
PCA [18]	33.54	10.11	0.8623	5.097	0.0210
GSA [19]	34.07	9.327	0.8721	4.756	0.0198
SFIM [23]	32.01	9.438	0.8582	6.133	0.0251
Indusion [22]	32.91	9.588	0.8503	5.473	0.0226
MSDCNN [32]	35.01	7.156	0.8905	4.324	0.0178
HyperPNN [35]	36.23	7.149	0.9117	3.833	0.0154
TFNet [17]	37.74	6.179	0.9357	3.248	0.0130
DARN [47]	38.39	5.725	0.9432	2.962	0.0120
MDA-Net	**39.91**	**4.747**	**0.9596**	**2.495**	**0.0101**

TABLE IV

QUANTITATIVE RESULTS OF THE DIFFERENT METHODS ON THE WASHINGTON DC MALL DATASET. THE BEST VALUE IS IN BOLD

Methods	PSNR	SAM	SSIM	ERGAS	RMSE
PCA [18]	32.73	12.06	0.8774	7.259	0.0243
GSA [19]	36.52	6.988	0.9393	4.674	0.0152
SFIM [23]	36.98	6.638	0.9449	4.332	0.0143
Indusion [22]	35.47	7.578	0.9277	5.162	0.0171
MSDCNN [32]	37.96	5.666	0.9537	4.100	0.0128
HyperPNN [35]	38.10	5.827	0.9583	3.997	0.0126
TFNet [17]	38.26	5.592	0.9583	3.856	0.0124
DARN [47]	38.57	5.479	0.9597	3.672	0.0119
MDA-Net	**40.21**	**4.265**	**0.9719**	**2.969**	**0.0099**

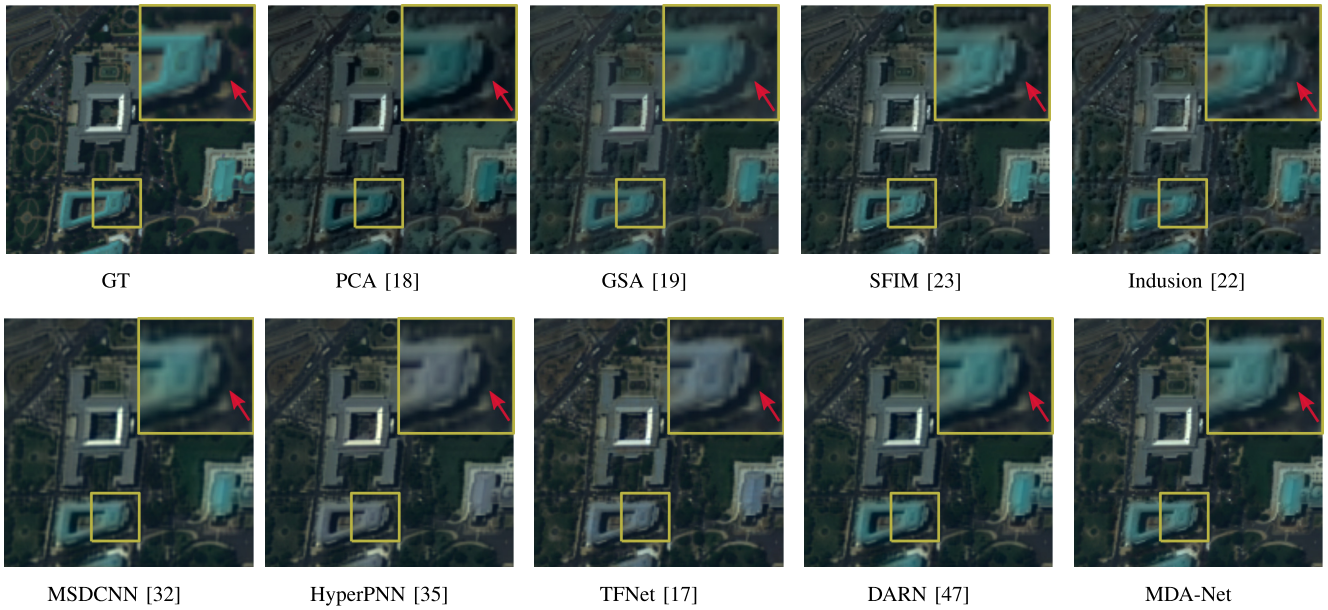

Fig. 9. Visual comparison of different methods on the selected patch "WDC_15" of the Washington DC Mall dataset with a scale factor of 8. The yellow box represents the zoomed-in ROI.

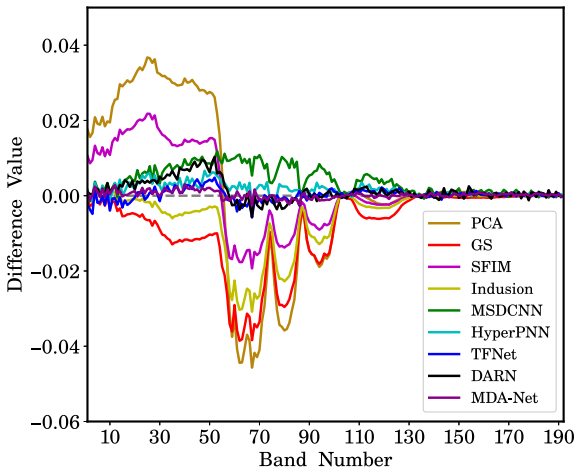

Fig. 10. Comparison of the spectral reflectance difference curve of a selected point (45, 20) of the patch "WDC_15."

oversmooth and contain severe spectral distortion. MSDCNN and TFNet generate relatively clearer fused images. However, chromatic aberration is still not solved for these methods. By comparison, our method gives the best performance in texture details restoration and spectral information preservation. The qualitative and quantitative evaluation results depicted in Fig. 11 and Table V both illustrate the superior performance of our MDA-Net on pansharpening real HSIs.

G. Analysis of the Computational Complexity

Computational complexity is also an important indicator for the evaluation of the pansharpening methods. In this section, we compute the average test runtime of each method on all the test images of the Houston dataset. The experiment is conducted on the same computing platform as above. Each test image has a size of $336 \times 336 \times 114$.

As can be seen in Table VI, CNN methods run noticeably faster than the traditional ones because the former only perform feedforward computation in the test phase. Among all the comparison methods, HyperPNN and MSDCNN rank first and second, respectively. These two methods implement a relatively shallow and light network. However, their fusion performance is quite poor as shown in the previous experiments. The MDA-Net is comparable to the TFNet in terms of speed while costs much less time than DARN. Taking both the runtime and fusion performance into consideration, the MDA-Net is an effective and appropriate choice for hyperspectral pansharpening.

H. Sensitivity Analysis of the Network Parameters

Many parameters are used in our method to adjust the performance. For some of them, such as the β_1 for Adam optimizer and the growth rate of MRDBs, we follow the commonly used settings that have been proven effective in various applications and over different datasets. The other parameters, such as the learning rate and the batch size, need to be determined according to the characteristics of the specific input data and the employed network. In this section, we perform the sensitivity analysis of the MDA-Net over these parameters, i.e., the learning rate and batch size. The experiment is conducted on the three simulated datasets, and PSNR is used to evaluate the fusion results quantitatively.

As shown in Table VII, the experimental results do not present significant variations as the batch size and learning rate change, and a similar situation occurs in all three datasets. This demonstrates that the MDA-Net is not very sensitive to these parameters when they lie in a reasonable range, which is empirically determined based on the specific dataset and

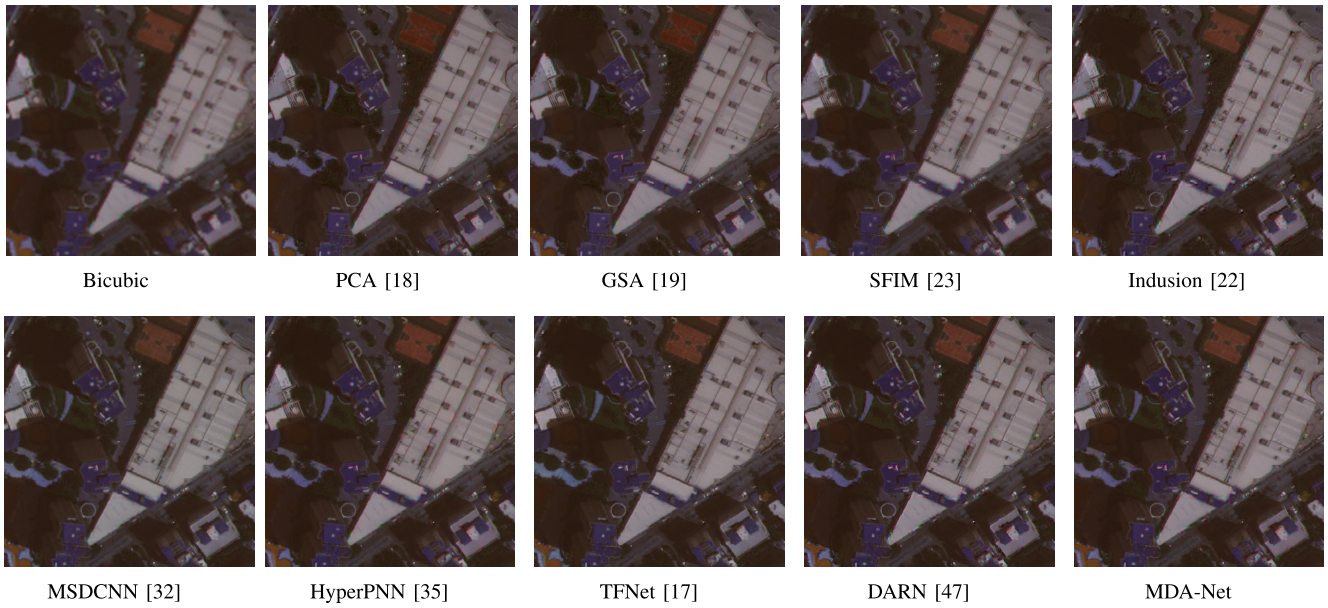

Fig. 11. Visual comparison of different methods on the selected patch "WV4_9" of the real WorldView-4 dataset with a scale factor of 4.

TABLE V
NO-REFERENCE HSIs ASSESSMENT SCORE OF DIFFERENT METHODS ON THE WORLDVIEW-4 DATASET. THE BEST VALUE IS IN BOLD

Methods	PCA [18]	GSA [19]	SFIM [23]	Indusion [22]	MSDCNN [32]	HyperPNN [35]	TFNet [17]	DARN [47]	MDA-Net
Score	13.01	12.74	12.96	14.22	12.28	11.33	11.54	11.08	**10.89**

TABLE VI
AVERAGE RUNTIME OF DIFFERENT METHODS ON THE TEST IMAGES OF THE HOUSTON DATASET. THE BEST VALUE IS IN BOLD

Methods	PCA [18]	GSA [19]	SFIM [23]	Indusion [22]	MSDCNN [32]	HyperPNN [35]	TFNet [17]	DARN [47]	MDA-Net
Runtime (s)	2.638	0.312	0.567	2.131	0.004	**0.002**	0.008	58.51	0.011

TABLE VII
PSNR INDICES OF THE SENSITIVITY ANALYSIS OF THE MDA-NET OVER THE LEARNING RATE AND THE BATCH SIZE. THE BEST VALUE IS IN BOLD

	Parameter setting	Pavia University	Houston University	Washington DC Mall
Learning rate	0.0003	39.88	37.23	**40.22**
	0.0002	**39.91**	**37.32**	40.21
	0.0001	39.74	37.02	39.98
	0.00005	39.66	37.11	40.01
Batch size	24	39.87	37.08	40.08
	16	39.91	**37.32**	**40.21**
	8	**39.93**	37.29	40.15

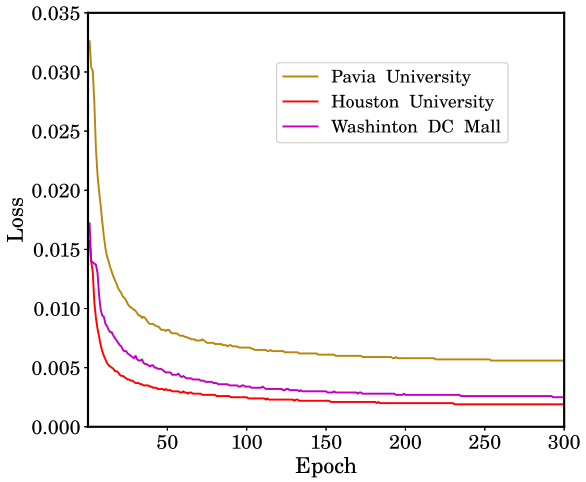

Fig. 12. Loss values of the proposed MDA-Net over different datasets during training.

network design. Thus, we set the learning rate to 2×10^{-4}, which decreases 10% for every 30 epochs, and the batch size to 16 for all testing datasets in our experiments.

In addition, we study the effect of the training epoch over different datasets. As can be seen in Fig. 12, all three loss curves decline rapidly in the first 50 epochs and level off after 250 epochs. Thus, we set the training epoch to 300, which is adequate for the network to give an optimal performance over different datasets.

TABLE VIII
PERFORMANCE OF EACH METHOD UNDER DIFFERENT LEVELS OF NOISE ON THE PAVIA UNIVERSITY DATASET. THE BEST VALUE IS IN BOLD

Methods	25dB	30dB	35dB
PCA [18]	19.71	20.47	20.89
GSA [19]	25.37	25.83	26.72
SFIM [23]	24.36	25.42	26.18
Indusion [22]	23.63	24.61	25.23
MSDCNN [32]	29.14	29.93	30.48
HyperPNN [35]	30.13	30.55	31.07
TFNet [17]	32.07	32.92	33.69
DARN [47]	32.26	33.59	34.32
MDA-Net	**35.20**	**36.37**	**37.04**

I. Analysis of the Generality to Different Levels of Noise

In this section, we evaluate the performance of the MDA-Net under the noise of different levels on the simulated Pavia University dataset. Three different levels of additive Gaussian noise are added to generate the LR-HSIs, while the signal-to-noise ratio (SNR) is set to be 25, 30, and 35 dB, respectively. The numerical results in terms of PSNR of all comparison methods are presented in Table VIII. As can be seen, the MDA-Net still gives the best performance under different levels of noise, which demonstrates the strong noise robustness of our method.

V. CONCLUSION

In this work, we proposed a novel method called MDA-Net for hyperspectral pansharpening. This approach applies a three-stream network, which is able to learn the intraimage information and the correlation among LR-HSI and HR-PAN simultaneously. To merge the useful information as fully as possible, we implement the fusion procedure in multiple stages. In each fusion stage, a DAFB is used to guide the combination of features from different streams. It incorporates spectral and spatial attention mechanisms to enhance the valuable location and channel components. Moreover, we develop an MRDB to extract multiscale deep features of the input images, which improves the representation capability of the network. The experimental results on both simulated and real hyperspectral datasets demonstrate that the MDA-Net outperforms other state-of-the-art methods significantly in terms of qualitative and quantitative evaluations.

REFERENCES

[1] 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.

[2] N. Akhtar and A. Mian, "Hyperspectral recovery from RGB images using Gaussian processes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 100–113, Jan. 2020.

[3] T. A. Carrino, A. P. Crósta, C. L. B. Toledo, and A. M. Silva, "Hyperspectral remote sensing applied to mineral exploration in southern Peru: A multiple data integration approach in the Chapi Chiara gold prospect," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 64, pp. 287–300, Feb. 2018.

[4] X. Tong, H. Xie, and Q. Weng, "Urban land cover classification with airborne hyperspectral data: What features to use?" *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 10, pp. 3998–4009, Oct. 2014.

[5] M. B. Stuart, A. J. McGonigle, and J. R. Willmott, "Hyperspectral imaging in environmental monitoring: A review of recent developments and technological advances in compact field deployable systems," *Sensors*, vol. 19, no. 14, p. 3071, Jul. 2019.

[6] 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.

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

[8] Z. Ren, H. K.-H. So, and E. Y. Lam, "Fringe pattern improvement and super-resolution using deep learning in digital holography," *IEEE Trans. Ind. Informat.*, vol. 15, no. 11, pp. 6179–6186, Nov. 2019.

[9] N. Meng, T. Zeng, and E. Y. Lam, "Spatial and angular reconstruction of light field based on deep generative networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4659–4663.

[10] N. Meng, H. K.-H. So, X. Sun, and E. Y. Lam, "High-dimensional dense residual convolutional neural network for light field reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 873–886, Mar. 2021.

[11] Y. Zhu, C. Hang Yeung, and E. Y. Lam, "Digital holographic imaging and classification of microplastics using deep transfer learning," *Appl. Opt.*, vol. 60, no. 4, pp. A38–A47, Feb. 2021.

[12] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Dec. 2018, pp. 2472–2481.

[13] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[14] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4471–4480.

[15] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, p. 594, Jul. 2016.

[16] 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. (ICCV)*, Oct. 2017, pp. 5449–5457.

[17] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Inf. Fusion*, vol. 55, pp. 1–15, Mar. 2020.

[18] V. P. Shah, N. H. Younan, and R. L. King, "An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1323–1335, May 2008.

[19] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of MS + pan data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Sep. 2007.

[20] T.-M. Tu, P. S. Huang, C.-L. Hung, and C.-P. Chang, "A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 309–312, Oct. 2004.

[21] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli, "Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2300–2312, Oct. 2002.

[22] M. M. Khan, J. Chanussot, L. Condat, and A. Montanvert, "Indusion: Fusion of multispectral and panchromatic images using the induction scaling technique," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 1, pp. 98–102, Jan. 2008.

[23] J. G. Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *Int. J. Remote Sens.*, vol. 21, no. 18, pp. 3461–3472, Dec. 2000.

[24] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and PAN imagery," *Photogramm. Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, May 2006.

[25] G. Vivone, R. Restaino, M. D. Mura, G. Licciardi, and J. Chanussot, "Contrast and error-based fusion schemes for multispectral image pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 5, pp. 930–934, May 2014.

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

[27] D. Fasbender, J. Radoux, and P. Bogaert, "Bayesian data fusion for adaptable image pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1847–1857, Jun. 2008.

[28] 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.

[29] 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.
[30] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 286–301.
[31] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-D-convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 639–643, May 2017.
[32] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 978–989, Mar. 2018.
[33] Z. Shao and J. Cai, "Remote sensing image fusion with deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1656–1669, May 2018.
[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] L. He, J. Zhu, J. Li, A. Plaza, J. Chanussot, and B. Li, "HyperPNN: Hyperspectral pansharpening via spectrally predictive convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 3092–3100, Aug. 2019.
[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[37] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Sep. 2018, pp. 421–429.
[38] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2018, pp. 3–19.
[39] S. Gross and M. Wilber, "Training and investigating residual nets," Feb. 2016. [Online]. Available: http://torch.ch/blog/2016/02/04/resnets.html
[40] Y. Zeng, W. Huang, M. Liu, H. Zhang, and B. Zou, "Fusion of satellite images in urban area: Assessing the quality of resulting images," in *Proc. IEEE 18th Int. Conf. Geoinform.*, Jun. 2010, pp. 1–4.
[41] R. Yuhas, A. F. H. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, vol. 1, Jun. 1992, pp. 147–149.
[42] 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.
[43] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?" in *Proc. 3rd Conf. Fusion Earth Data, Merging Point Meas., Raster Maps Remotely Sensed Images*, 2000, pp. 99–103.
[44] J. Yang, Y. Zhao, C. Yi, and J. C.-W. Chan, "No-reference hyperspectral image quality assessment via quality-sensitive features learning," *Remote Sens.*, vol. 9, no. 4, p. 305, Mar. 2017.
[45] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network," *Remote Sens.*, vol. 10, no. 5, p. 800, May 2018.
[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, May 2015, pp. 1–15.
[47] Y. Zheng, J. Li, Y. Li, G. Jie, X. Wu, and J. Chanussot, "Hyperspectral pansharpening using deep prior and dual attention residual network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8059–8076, Nov. 2020.

Peiyan Guan received the B.S. degree in communication engineering from Nanjing University, Nanjing, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong.

His research interests include hyperspectral image analysis, pattern recognition, and machine learning.

Edmund Y. Lam (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1995, 1996, and 2000, respectively.

From 2010 to 2011, he was a Visiting Associate Professor with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. He is currently a Professor of electrical and electronic engineering with The University of Hong Kong, Hong Kong, where he also serves as the Computer Engineering Program Director. His main research is in computational imaging.

Dr. Lam is also a fellow of the Optical Society of America (OSA), the Society of Photo-Optical Instrumentation Engineers (SPIE), the Society for Imaging Science and Technology (IS&T), and the Hong Kong Institution of Engineers (HKIE). He was a recipient of the IBM Faculty Award.