A Deep-Shallow Fusion Network with Multi-Detail Extractor and Spectral Attention for Hyperspectral Pansharpening

Yu-Wei Zhuo, Tian-Jing Zhang, Jin-Fan Hu, Hong-Xia Dou, Ting-Zhu Huang, Liang-Jian Deng *Member, IEEE*

***Abstract*—Hyperspectral (HS) pansharpening aims at fusing a low-resolution HS (LRHS) image with a high-resolution panchromatic (PAN) image to obtain a hyperspectral image with both higher spectral and spatial resolutions. However, existing HS pansharpening algorithms are mainly based on multispectral (MS) pansharpening approaches, which cannot perfectly restore much spectral information and more high-frequency spatial details in the continuous spectral bands and much broader spectral range, leading to spectral distortion and spatial blur. In this paper, we develop a new hyperspectral pansharpening network architecture (called Hyper-DSNet) to fully preserve latent spatial details and spectral fidelity via a deep-shallow fusion structure with multi-detail extractor and spectral attention. Specifically, the proposed architecture mainly consists of three parts. First, to solve the problem of spatial ambiguity and exploit the potential information, five types of high-pass filter templates are used to fully extract the spatial details of the PAN image, constructing a so-called multi-detail extractor (MDE). Then, after passing a multi-scale convolution module, a deep-shallow fusion (DSF) structure, which reduces parameters by decreasing the number of output channels as the network goes deeper, is utilized sequentially. In final, a spectral attention (SA) module is conducted to preserve the spectrum for a wealth of spectral information of HS images. Visual and quantitative experiments on three commonly used simulated datasets and one full-resolution dataset demonstrate the effectiveness and robustness of the proposed Hyper-DSNet against the recent state-of-the-art hyperspectral pansharpening techniques. Ablation studies and discussions further verify our contributions, *e.g.*, better spectral preservation and spatial detail recovery.**

***Index Terms*—Hyperspectral Pansharpening, Convolutional Neural Network, Deep-Shallow Architecture, Spectral Attention, Multi-Detail Extractor**

I. Introduction

Hyperspectral (HS) images have hundreds of narrow continuous bands in the same scene simultaneously [3], which contain rich spectral information, making HS images widely applied in many fields such as military surveillance [4], environmental monitoring [5], mineral exploration [6], [7], agriculture [8], [9] and change detection in commercial products [10]. However, due to the physical limitations of sensors, expanding the spectral range also brings a reduction in spatial resolution. When compared to panchromatic (PAN) images, HS images typically have a lower spatial resolution, which may be insufficient in some practical applications where both high spatial and spectral resolutions are desired [11]. Therefore, hyperspectral pansharpening, aiming to merge the HS and PAN images to generate a fused HS image with both higher spectral and spatial resolution, is of great significance from many perspectives, also receiving great attention from

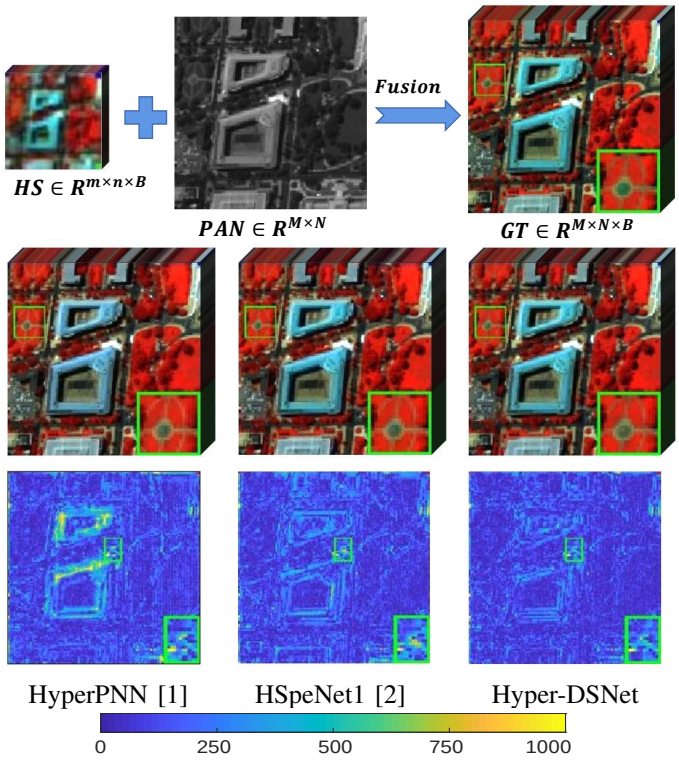

Fig. 1: First row: schematic diagram of hyperspectral pansharpening on an example from Washington DC dataset. Second row: visual results of three compared methods, *i.e.*, HyperPNN (CC/SAM/ERGAS=0.946/4.42/4.96), HSpeNet1 (0.955/4.43/4.27) and Hyper-DSNet (**0.965/4.07/3.75**). Third row: the corresponding error maps yielded by the compared methods.

The work is supported by National Natural Science Foundation of China grants 12171072, 61876203, and Key Projects of Applied Basic Research in Sichuan Province (Grant No. 2020YJ0216), and National Key Research and Development Program of China (Grant No. 2020YFA0714001).

Y.-W. Zhuo, T.-J. Zhang is with the Yingcai Honors College, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China (e-mails: yuuweii@yeah.net; zhangtianjinguestc@163.com).

H.-X. Dou is with the School of Science, Xihua Univercity, Chengdu, 610039, China (e-mail: hongxia.dou@mail.xhu.edu.cn).

J.-F. Hu, T.-Z. Huang and L.-J. Deng is with the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China (e-mails: hujf0206@163.com; tingzhuhuang@126.com; liangjian.deng@uestc.edu.cn).

the remote sensing and image processing communities [12].

In the recent decade, a number of data fusion techniques have been developed to improve the spatial resolution of HS imagery. They can be roughly classified into five categories: component substitution (CS), multiresolution analysis (MRA), Bayesian, matrix factorization and deep learning (DL) based approaches.

The CS approach, which relies on the substitution of a component of the HS images by the PAN image, contains algorithms such as principal component analysis (PCA) [13]–[15], intensity-hue-saturation (IHS) [16]–[19], Gram-Schmidt (GS) spectral sharpening [20], [21] and guided filtering (GF) [22]. They perform well in terms of spatiality and particularly resist co-registration problems, but may cause spectral distortion. The MRA approach first takes spatial features from the PAN images and then injects them into the HS images in a multiresolution way, including wavelet transform based method [23]–[25], Laplacian pyramid based method [26], smoothing filter-based intensity modulation (SFIM) [27], modulation transfer function (MTF) generalized Laplacian pyramid method (MTF-GLP) [28] and MTF-GLP with high-pass modulation (MTF-GLP-HPM) [29]. Such methods could well preserve spectral information but mainly suffer from spatial distortions. Besides, there are also some hybrid methods that use both component substitution and multiscale decomposition, such as guided filter PCA (GFPCA) [30].

The Bayesian approach depends on the usage of the posterior distribution of the required high-resolution HS (HRHS) image for the given LRHS and PAN images [12]. Wherein, Gaussian prior (Bayesian sparse) [31], Bayesian naive Gaussian prior (Bayesian naive) [32], and Bayesian HySure [33] are typical Bayesian approaches. Moreover, the matrix factorization based method is to utilize an optimization tool to factorize the related matrices after first modeling the observed data with a signal subspace representation, including a representative method called the coupled non-negative matrix factorization (CNMF) [34]. Besides, there are other typical variational methods that also belong to VO-based methods [35]–[41]. The Bayesian and matrix factorization based methods are often constrained by the insufficient representation ability, and serious quality degradation may occur if the prior assumptions do not fit the situation. Furthermore, the majority of available fusion model optimization strategies are solved iteratively, which is time-consuming and inefficient.

Over recent years, deep learning (DL) based methods, particularly convolutional neural network (CNN) based DL techniques, have achieved significant advances in image processing fields, $e.g.$, image resolution reconstruction [42]–[50], image classification [51]–[53], image denoising [54], image fusion [55]–[61], $etc.$ Therefore, many methods [1], [2], [62]–[75] based on deep learning have also been applied to solve the pansharpening problem. Dong et al. [42] originally introduce a shallow three-layer CNN (SRCNN) to learn the mapping between LR and HR patches for single image super-resolution. Based on the effective residual learning technique, Ledig et al. [43] employs a residual network (ResNet) to build a deeper network for image SR. Especially, CNNs have shown promising results not only in single image super-resolution but also in MS pansharpening. More recently, more researchers have made attempts to employ CNN in HS pansharpening. Masi et al. [62] develop a three-layer CNN architecture for pansharpening, utilizing pre-interpolated low-resolution MS images stacked with PAN images as input. This is the first work utilizing CNN for MS pansharpening, inspired by the SRCNN. Besides, Yang et al. [63] propose a deep network (PanNet) for the pansharpening problem whose main contribution is adding up-sampled multispectral images to the network output to propagate the spectral information and training parameters in the high-pass filtering domain rather than the image domain. He et al. [1] introduce spectrally predictive structure (HyperPNN) to strengthen the spectral prediction capability of the CNN for the task of HS pansharpening. Moreover, HS pansharpening is also handled as a restricted minimization problem with extra priors learned by the CNN by Xie et al. [64]. Furthermore, He et al. [2] develop new spectral-fidelity CNN architecture (HSpeNet) for HS pansharpening to keep the fidelity of the pansharpened image, focusing on the decomposability of HS details and meanwhile introducing a spectral-fidelity loss. Recently, Some works have achieved good results by directly using no-reference loss without downsampling to simulate training data. Xiong et al. [76] first designed a loss function that does not need the reference fused image. Based on this, Li et al. [77] combined CNN with transformer block to design a CNN+ pyramid Transformer network with no-reference loss.

However, in some of these approaches, the particularity of remote sensing images, especially hyperspectral images, is ignored due to all features extracted from input images being treated identically, further restricting the ability to employ relevant information selectively. Besides, for the characteristics of a wider spectral range of the hyperspectral image than the multispectral one, most networks are not designed for the special spectral preservation, which fails to consider the importance and sensitivity of spectral information and leads to spectral distortion easily. Besides, for PAN images, pioneer works often feed them directly into the network together with HS images or use a fixed high-frequency template for preprocessing, which will inevitably lose some spatial information. Moreover, when it comes to a deep network structure, researchers often only pay attention to the results after multilayer convolution and ignore the importance of the shallow feature. In addition, the features extracted from the deep and shallow layers in the network are different, and the shallow features usually contain more texture details.

To tackle the problems mentioned above, we propose a so-called Hyper-DSNet, containing a deep-shallow fusion structure with multi-detail extractor and spectral attention, for the task of HS pansharpening. To summarize, the main contributions of the work include four aspects listed as follows.

1) For the challenging of spectral preservation in the HS pansharpening, we appropriately and specially used a spectral attention module generating different channel weights to distinctively preserve the HS image's rich and sensitive spectral information. It delivers the impact of reducing spectral distortion and improving the network's spectral fidelity.

2) We give a multi-detail extractor (MDE) module that contains several distinct high-pass filtering templates for extracting different spatial details from the PAN image and injecting them into the network alongside the PAN image. Abundant and diverse high-frequency information with other characteristics promotes better use of the spatial information of the PAN image.
3) After passing a multi-scale convolution, extracted features will go into a specifically designed deep-shallow fusion (DSF) module, not only connecting the deep and shallow features but also reducing network parameters, for better spatial information recovery.

Experimental results on three benchmark HS datasets demonstrate the superiority of the proposed Hyper-DSNet over recent state-of-the-art (SOTA) HS pansharpening techniques, as shown in Fig. 1. What's more, the best evaluation results at full resolution prove the robustness of our method.

II. Related works

In this section, a brief review of several DL-based methods for HS pansharpening, some works related to the proposed architecture and our motivation will be presented.

A. CNN-based HS Pansharpening Framework

Recently, CNNs have been widely used in the field of image processing and computer vision. They are mainly proposed for processing regular matrices by continuous sliding window (kernel) convolution. In the training process, each parameter of the convolution kernel is continuously updated and optimized via forward and backpropagation to minimize the loss function. The main mathematical formulation for CNN can be summarized as follows:

$$\mathbf{O}_l = f(\mathbf{W}_l * \mathbf{O}_{l-1} + \mathbf{b}_l), \tag{1}$$

where $*$ is the convolutional operation, $\mathbf{O}_l$ represents the output feature map on the l-th layer, $\mathbf{W}_l$ and $\mathbf{b}_l$ stand for the network parameters and biases on this layer, respectively, and $f(\cdot)$ means an activation function.

Consider the case of HS pansharpening, CNN-based framework accepts the observed HS image and the PAN image as input, and finally outputs an HRHS image. The PAN image with the size $L \times W$ is denoted as $\mathbf{P}_0 \in \mathbb{R}^{L\times W\times 1}$, while the LRHS image with $l \times w$ pixels and B spectral bands is indicated as $\mathbf{H}_0 \in \mathbb{R}^{l\times w\times B}$. The expected HRHS output is $\mathbf{H} \in \mathbb{R}^{L\times W\times B}$ and the fused output of the CNN-based framework can be written as $\hat{\mathbf{H}}$ with the same dimension, *i.e.*,

$$\hat{\mathbf{H}} = M(\mathbf{P}, \mathbf{H}_0; \theta), \tag{2}$$

where $M(\cdot;\theta)$ means the mapping from input to output with all parameters θ to be optimized. In final, the network parameters of CNN-based HS pansharpening can be generally updated by minimizing the following ℓ_2 loss function,

$$\begin{aligned}\mathcal{L}(\theta) &= \|\hat{\mathbf{H}} - \mathbf{H}\|_2^2 \\ &= \|M(\mathbf{P}, \mathbf{H}_0; \theta) - \mathbf{H}\|_2^2,\end{aligned} \tag{3}$$

where $\|\cdot\|_2$ refers to the ℓ_2 norm. Once $M(\cdot;\theta)$ is learned, and the new observed PAN and HS images $\mathbf{P}_0$ and $\mathbf{H}_0$ are input into the mapping again, the predicted HRHS image can be obtained.

Compared with the general multispectral (MS) pansharpening problem, HS pansharpening is faced with greater challenges. One is that the spectrum range of HS image (191 bands from 400 nm to 2400 nm of HYDICE sensor) is wider than the range of MS image (8 bands from 400 nm to 1040 nm of WorldView-3 sensor), causing a larger spectral gap between the HS image and the PAN image; the other is that more details in continuous bands with high spectral resolution need to be reconstructed at the same time. These challenges make HS pansharpening more prone to problems such as spectral distortion and have higher requirements on the accuracy of the algorithm and the ability to predict and reconstruct the spectrum.

In view of the characteristics of HS images, many corresponding solutions have been proposed. For instance, HyperPNN [1] adds spectrally predictive layers to strengthen the spectral prediction ability of the network, and composes a spectral prediction sub-network and a spatial-spectral inference sub-network. Both HSpeNet1 and HSpeNet2 [2] assume the decomposability of HS details and accordingly synthesize those details progressively. Specifically, HSpeNet1 reconstructs HS details from bottom level to top level, and HSpeNet2 synthesizes those details in a manner of band group-wise reconstruction. Besides, FusionNet [78] focuses attention on traditional CS and MRA frameworks and directly extracts details by differencing the single PAN image with each MS band.

B. Image Differential Operator

For the MS pansharpening task, Yang et al. [63] propose a deep network (called PanNet) that uses up-sampled multispectral images to the network and training parameters in the high-pass filtering domain rather than the image domain. However, they only use one predefined high-pass template, which may cause the loss of some detailed information. Based on this idea, we expect to use more different high-pass templates to extract more types of high-frequency details for a better fusion process. In this part, some high-pass image differential operators that we will use are first introduced.

The first one is the simplest first-order difference operator. For two-dimensional images, it contains differences in two directions, *i.e.*, x-axis and y-axis, which can be represented by the following kernels,

$$\begin{bmatrix}-1\\+1\end{bmatrix}, \quad \begin{bmatrix}-1 & +1\end{bmatrix}. \tag{4}$$

Also, we can use the following 2-D kernels to describe the difference between the two diagonal directions, *i.e.*, Roberts operator,

$$\begin{bmatrix}-1 & 0\\0 & +1\end{bmatrix}, \quad \begin{bmatrix}0 & -1\\+1 & 0\end{bmatrix}. \tag{5}$$

However, this kind of operator is not very convenient in practice because there is no center pixel, thus we intend to use the operator of 3×3 such as the Prewitt operator. When calculating the gradient of the center position, unlike the

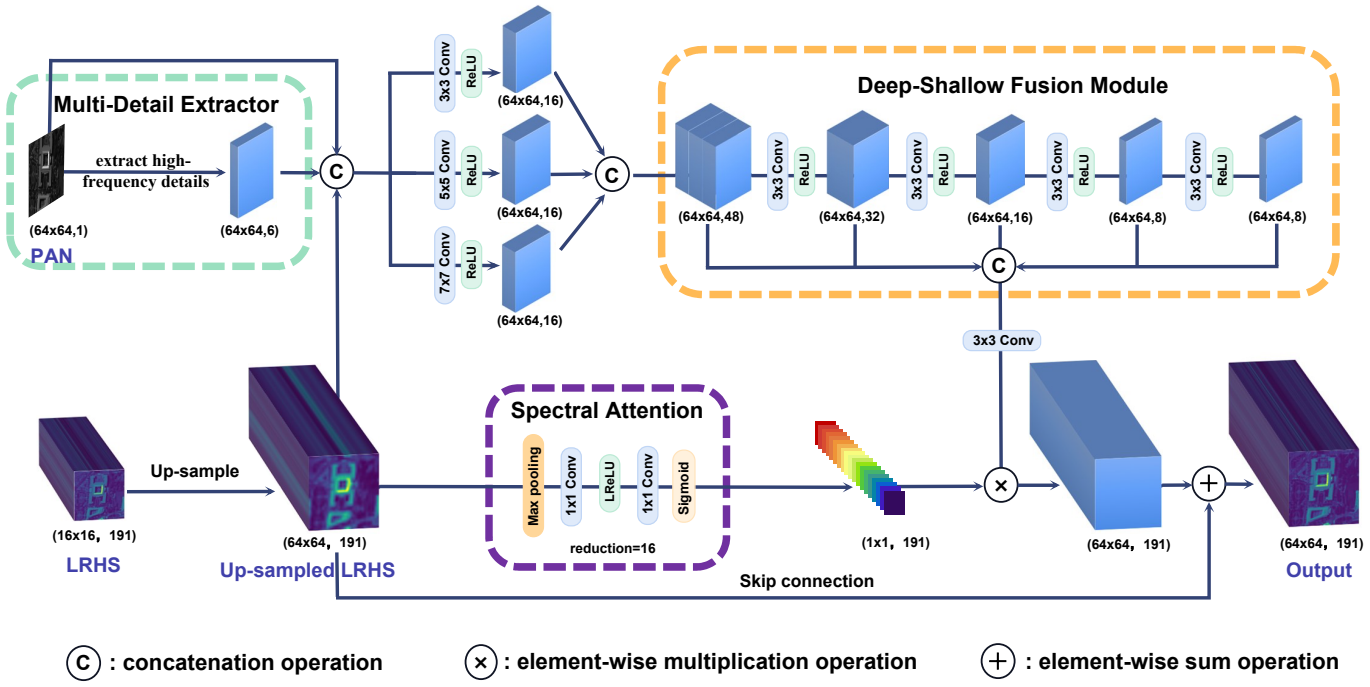

Fig. 2: The overall architecture of the proposed Deep-Shallow Fusion Network (Hyper-DSNet). Under the each cube block, the height-width size and the channel number (H×W, C) is shown.

previous 2×2, which uses the positive and negative deviations of only one pair of pixels, 3×3 expands outward into three pairs to make it more sensitive to specific directions.

$$\begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ +1 & +1 & +1 \end{bmatrix}, \quad \begin{bmatrix} -1 & 0 & +1 \\ -1 & 0 & +1 \\ -1 & 0 & +1 \end{bmatrix}. \tag{6}$$

On this basis, the Sobel operator performs a certain weighting to make the nearest pair of pixels have a higher weight, which is beneficial to reduce the influence of noise, see the following operators,

$$\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix}, \quad \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix}. \tag{7}$$

In addition, the Laplacian operator is a second-order differential operator that often appears in image enhancement. Compared with the first-order operator, the second-order differential has a stronger edge positioning ability and a better sharpening effect. The Laplacian operator is defined as the result of performing the gradient operation ∇ on the function g first, and then the divergence operation $\nabla\cdot\nabla$, see as follows,

$$\Delta g = \nabla^2 g = \nabla\cdot\nabla g, \tag{8}$$

where g is a second-order differential function and Δ is the Laplacian operator.

C. Motivations

As mentioned before, the HS pansharpening method must deal with two key issues, *i.e.*, the substantial spectral coverage disparity between the HS and PAN images, as well as the necessity to recover features in numerous continuous narrow bands simultaneously. Although the methodologies discussed above presented numerous empirical approaches to realize these challenges, some constraints have yet to be addressed:

1) The PAN image is an important basis for restoring spatial details, but it is usually directly used as the input of the network. Therefore, the high frequency information in PAN images cannot be fully utilized. It motivates us to give multiple high-pass filters for constructing a so-called MDE module for better detail extraction.
2) Second, few methods take into account the particularity of the more continuous spectra HS bands, which makes the spectrum information critical and sensitive. Spectrum preservation operations should be specially designed, motivating us to utilize spectral attention for spectral preservation.
3) Third, a large number of spectral bands also brings an increase in the number of parameters, leading to the difficulty of training. Additionally, low-level feature information needs to be valued more in the image fusion task. Therefore, a special module with reduced parameters can be appropriately designed and embedded or replaced in other networks, which motivates us to develop a DSF module with the reduction of channel numbers.

Taking these considerations together, we design our Hyper-DSNet, which will be introduced in detail in what follows.

III. Proposed Methods

Based on the above analysis and motivation, we will introduce each part of our proposed Hyper-DSNet detailedly in the section, including the detailed main architecture shown in Fig. 2 and the corresponding loss function.

In general, our Hyper-DSNet contains three sub-modules, that is MDE module, DSF module and spectral attention (SA) module, which will be described one by one in the following sections.

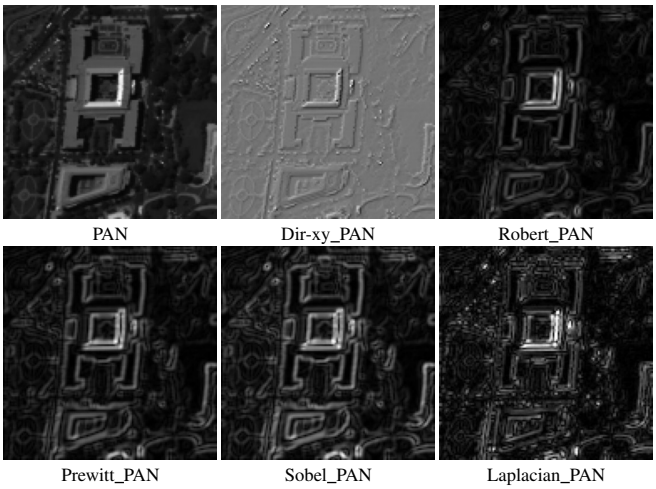

Fig. 3: Multi-template operator results in MDE module. Dir-xy_PAN, Robert_PAN, Prewitt_PAN, Sobel_PAN and Laplacian_PAN respectively refer to the image obtained after applying the corresponding operator on the PAN image.

A. Multi-detail Extractor

In the field of image super-resolution, the reconstruction quality of high-frequency information (*e.g.*, edges, contours and textures) is pretty crucial for the performance. Thus, we expect to extract and utilize those rich high-frequency details of the PAN image instead of training with the original image. We believe that the artificial extraction and intervention process will bring better efficiency and effects. Furthermore, PanNet [63] have noticed the importance of features on the high-pass filtering domain, but only one type of high-pass filter is integrated for extracting one single level of detail. It inspires us to adopt a more comprehensive detail extraction method. We believe that the multilevel high-pass information could favor a better performance, thus proposing the so-called MDE module.

For the MDE module, PAN image $\mathbf{P}_0 \in \mathbb{R}^{L\times W\times 1}$ first goes through five high-pass operators to extract multilevel high-frequency information which will be then concatenated with PAN image itself to construct the input feature. The five high-pass operators, *i.e.*, first-order difference operator, Robert operator, Prewitt operator, Sobel operator, Laplacian operator, have been shown in Eq. (4)-(8) in turn, and here we denote them as α_{dir}, α_{robert}, $\alpha_{prewitt}$, α_{sobel}, $\alpha_{laplacian}$, respectively, thus the input high-pass feature $\mathbf{O}_P \in \mathbb{R}^{L\times W\times 7}$ is as follows,

$$\begin{aligned}\mathbf{O}_P =&[\alpha_{dir-x} * \mathbf{P}, \alpha_{dir-y} * \mathbf{P}, \alpha_{robert} * \mathbf{P}, \alpha_{prewitt} * \mathbf{P},\\ &\alpha_{sobel} * \mathbf{P}, \alpha_{laplacian} * \mathbf{P}, \mathbf{P}].\end{aligned} \tag{9}$$

We show the results of using these five high-pass operators on PAN image in Fig. 3. As we can see that each extracts significantly different high-frequency information, some are smoother, and some are more delicate, which meets our expectations.

B. Deep-Shallow Fusion Module

In this section, we mainly present the structure of detail extraction which could be divided into two parts, *i.e.*, multi-scale convolution module and DSF module whose goal is to extract effective and crucial spatial-spectral information. Before this, the HS image will be first up-sampled to the same size as PAN by a polynomial kernel [78]. The output of the MDE module and the up-sampled HS image ($\text{LRHS}^U \in \mathbb{R}^{L\times W\times B}$) are concatenated along the spectral dimension as the input of the structure of detail extraction.

The multi-scale convolution module, first introduced in MSDCNN by Yuan et al. [68], is used here to extract multi-scale information. Three different sizes of convolution kernels are followed to perform feature extraction in diverse receptive fields. This process can be formulated as:

$$\begin{cases}\mathbf{O}_3 = \delta(\mathbf{W}_3 * [\mathbf{O}_P, \text{LRHS}^U] + \mathbf{b}_3)\\ \mathbf{O}_5 = \delta(\mathbf{W}_5 * [\mathbf{O}_P, \text{LRHS}^U] + \mathbf{b}_5)\\ \mathbf{O}_7 = \delta(\mathbf{W}_7 * [\mathbf{O}_P, \text{LRHS}^U] + \mathbf{b}_7)\\ \mathbf{O}_b = [\mathbf{O}_3, \mathbf{O}_5, \mathbf{O}_7],\end{cases} \tag{10}$$

where $\mathbf{W}_i$ and $\mathbf{b}_i$ respectively represent the kernel weights and biases, $\mathbf{O}_i$ is the output of the response convolutional layer, the subscript i ($i = 3, 5, 7$) means the size of the convolutional kernel, $\mathbf{O}_b$ is the output of this multi-scale convolution module, and $\delta(\cdot)$ standards for an activation function of Rectified Linear Unit (ReLU) [79]. Here the channel number of output feature maps at each layer is set to 16 for the aim of parameters reduction.

After the multi-scale convolution module, it is followed by a DSF module. In general, the shallow convolutions are mainly used to focus on local region with small receptive field yielding fine-grained features, which lacks contextual information. In comparison, the deep layer has larger receptive fields obtaining abstract features with semantic information. However, it may be too abstract to utilize in the field of low-level vision task that focuses on pixel reconstruction instead of understanding the image content. So the shallow and deep features are both important in our HS pansharpening task. In previous methods, the result of deep convolution is often used directly as the final output, which will result in only paying attention to the deep information, may lose part of the low-level features. Here each shallow and deep convolution result will be concatenated to maintain those two types of critical information in each step.

First focus on the first layer of convolution, which could be viewed as a weighting of the three different sizes of

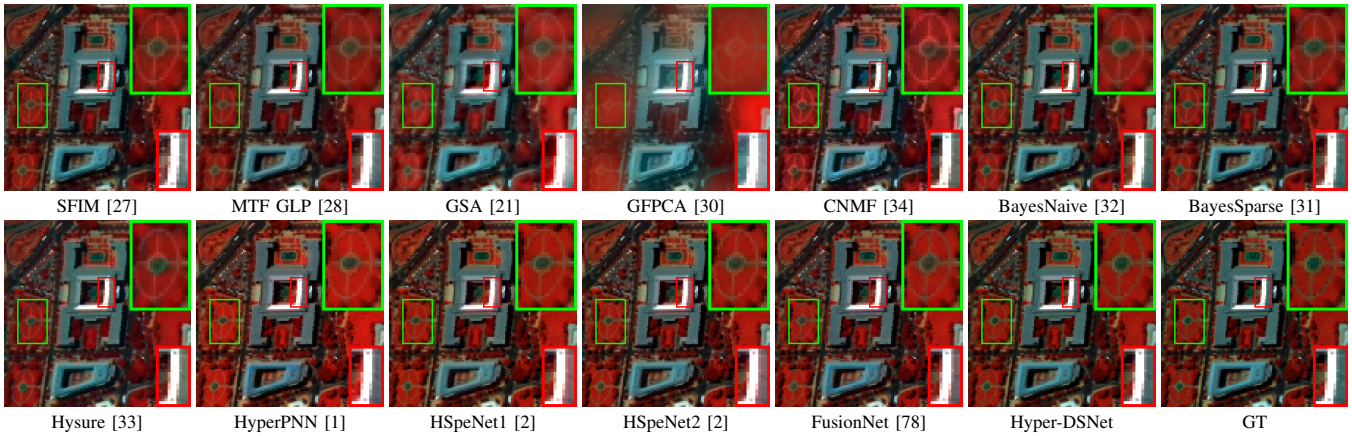

Fig. 4: The visual comparisons of fusion results obtained by different methods on a reduced-resolution WDC dataset obtained by Hydice (shown by bands: 20, 40 and 60).

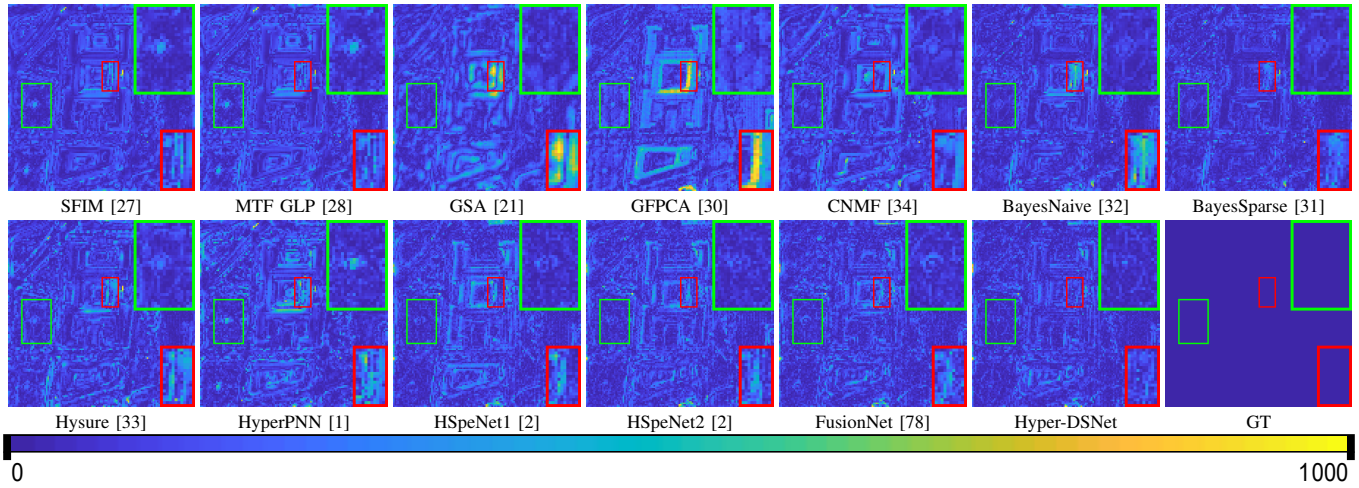

Fig. 5: The visual comparisons of the correspongding residual maps using the GT image as reference. Please note that here we select the 3-rd spectral band for better observation.

convolutions in the front. Then the following several deep convolutions can be mathematically represented as

$$\begin{cases} \mathbf{O}_{b1} = \delta(\mathbf{W}_{31} * \mathbf{O}_b + \mathbf{b}_{31}) \\ \mathbf{O}_{b2} = \delta(\mathbf{W}_{32} * \mathbf{O}_{b1} + \mathbf{b}_{32}) \\ \quad = \delta(\mathbf{W}_{32} * \delta(\mathbf{W}_{31} * \mathbf{O}_b + \mathbf{b}_{31}) + \mathbf{b}_{32}) \\ \quad \cdots \\ \mathbf{O}_{b4} = \delta(\mathbf{W}_{34} * \mathbf{O}_{b3} + \mathbf{b}_{34}) \\ \quad = \delta(\mathbf{W}_{34} * \delta(\mathbf{W}_{33} * \delta(\mathbf{W}_{32} * \delta(\mathbf{W}_{31} * \mathbf{O}_b + \mathbf{b}_{31}) \\ \quad + \mathbf{b}_{32}) + \mathbf{b}_{33}) + \mathbf{b}_{34}), \end{cases} \tag{11}$$

where $\mathbf{O}_{bi}$ means the i-th convolution's output, $\mathbf{W}_{3i}$ and $\mathbf{b}_{3i}$ represent the weights and bias of the i-th 3×3 convolution in this part.

As mentioned before, we concatenate each shallow and deep convolution results in the channel dimension to keep useful key information in each step:

$$\mathbf{O}_c = [\mathbf{O}_b, \mathbf{O}_{b1}, \mathbf{O}_{b2}, \mathbf{O}_{b3}, \mathbf{O}_{b4}], \tag{12}$$

where $\mathbf{O}_c$ represents the output of DSF module.

Furthermore, the low-level spatial information obtained by shallow convolution needs more attention in the pixel-wise vision task. Shallow and deep convolution kernels with the same number of features will bring a certain amount of information redundancy. Thus, more feature maps are set to describe the low-level information to avoid the redundancy problem. With the deepening of the convolutional layer, the number of feature maps decreases from high to low. More clearly, the number of channels in the DSF module is set to $[48, 32, 16, 8, 8]$ in order as shown in Fig. 2, which will be further introduced in Sect. IV-E2.

C. Spectral Attention Module

Compared with other panchromatic sharpening fusion tasks, the biggest challenge of hyperspectral pansharpening lies in the spectral information that is rich and sensitive, which places higher demands on the spectral fidelity of HS images. For this reason, we argue that a dedicated module is needed to guarantee spectral information in super-resolution.

The feature maps extracted from the previous detail extraction module attach equal importance to each feature channel,

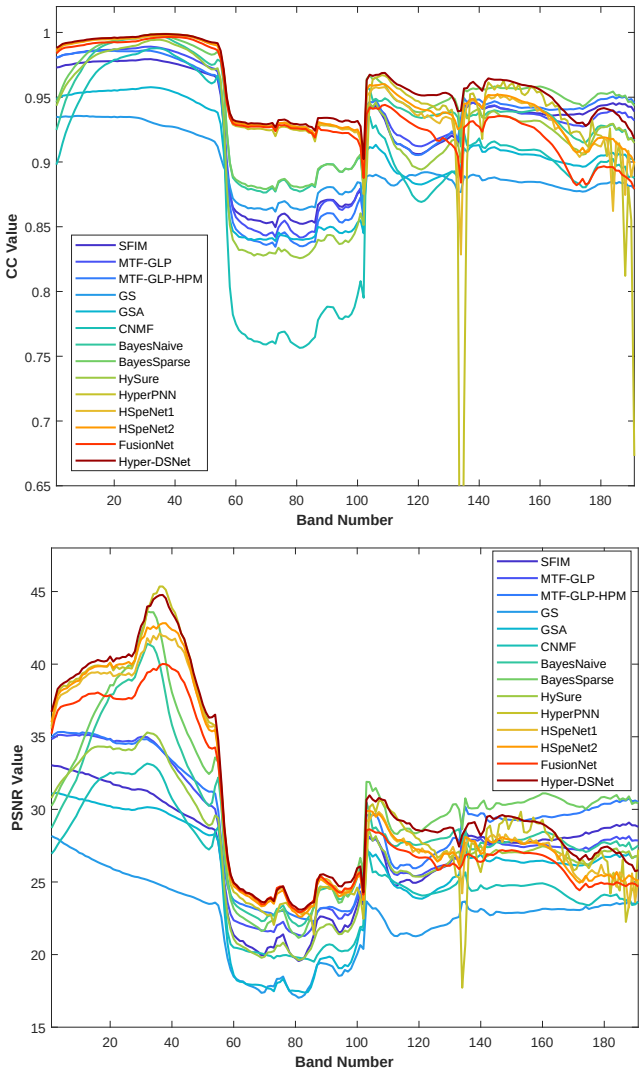

Fig. 6: CC and PSNR curves of the Washington DC dataset as functions of the spectral bands for different methods.

ignoring the different degrees of spectral contribution, which needs some attention to help call out different channels' importance and remove the information redundancy. Among many attention mechanisms, we give the so-called SA module that is actually based on the channel attention mechanism proposed in [80] for hyperspectral pansharpening, due to its competitive abilities of cost-effective property and spectral preservation. Thus, a SA module is constructed to characterize the relationship among channels.

Specifically, the LRHS^U image is as input of the SA module. First, a global average pooling layer is adopted to aggregate spatial information more conveniently, which will output a vector $v_b \in \mathbb{R}^{1\times1\times B}$:

$$v_b = \frac{1}{L \times W} \sum_{i=1}^{L} \sum_{j=1}^{W} I_b(i,j), \tag{13}$$

where $I_b(i,j)$ is the value at the position (i,j) in the b-th channel of the LRHS^U image, v_b means the b-th value of the output vector. Following this, the global spectral information is squeezed into a B-length vector. To properly and fully capture channel-specific dependencies, here we employ a simple gating mechanism with a sigmoid activation,

$$s = \sigma\left(\mathbf{W}_2.\delta\left(\mathbf{W}_1 v_b\right)\right), \tag{14}$$

where output $s \in \mathbb{R}^B$, $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ are the weights of two fully connected convolution layers with the kernel size of 1×1 and σ means the sigmoid actvation. In order to reduce the amount of calculation, the number of channels is first reduced with a ratio r and then expanded back to B successively through two consecutive layers of convolution:

$$\mathbf{O}_{SA} = \mathbf{F}_{scale}(\mathbf{O}_c, s) = [\mathbf{O}_{c1}s_1, \mathbf{O}_{c2}s_2, \cdots, \mathbf{O}_{cb}s_b]. \tag{15}$$

By applying this SA module, the final output is obtained by rescaling the detailed extracted output, and skipping connection to add the initial LRHS^U as the residual part. It is believed that the target ground truth can be seen as adding more detailed information on the basis of LRHS^U. As a result, employing the initial LRHS^U as a skip connection can preserve its original spectral information, avoid overfitting, prevent degradation as the network depth increases and speed up convergence, allowing the network to train better and more quickly to achieve the desired effect, which is respired by He et al. [81] and proved by other pansharpening methods [63], [78].

D. Loss Function

To depict the difference between the network output and the ground-truth (GT), we adopt ℓ_1 loss function to optimize the proposed network in the training process. The loss function can be expressed as follows,

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{O}_{SA} + \text{LRHS}^U - \text{GT} \right\|_1, \tag{16}$$

where GT is the GT image, N represents the number of training samples and $\|\cdot\|_1$ means the ℓ_1 norm.

IV. Experimental Results and Analysis

This section is devoted to experimental evaluation to demonstrate the effectiveness of the given Hyper-DSNet. The proposed method will be compared with some recent SOTA HS pansharpening approaches on benchmark datasets obtained by different sensors.

A. Experimental Setup

This part introduces the details of experimental datasets, including data simulation, experimental platform and hyperparameter settings. To evaluate the effectiveness of our Hyper-DSNet for remote sensing pansharpening, a series of experiments are conducted on three simulated HS datasets, *i.e.*, Washington DC, Pavia Center and Botswana, and one full-resolution dataset, *i.e.*, FR1, which is described in detail as follows. The various features of the dataset are displayed in Table I for easier comparison. Since the number of bands in

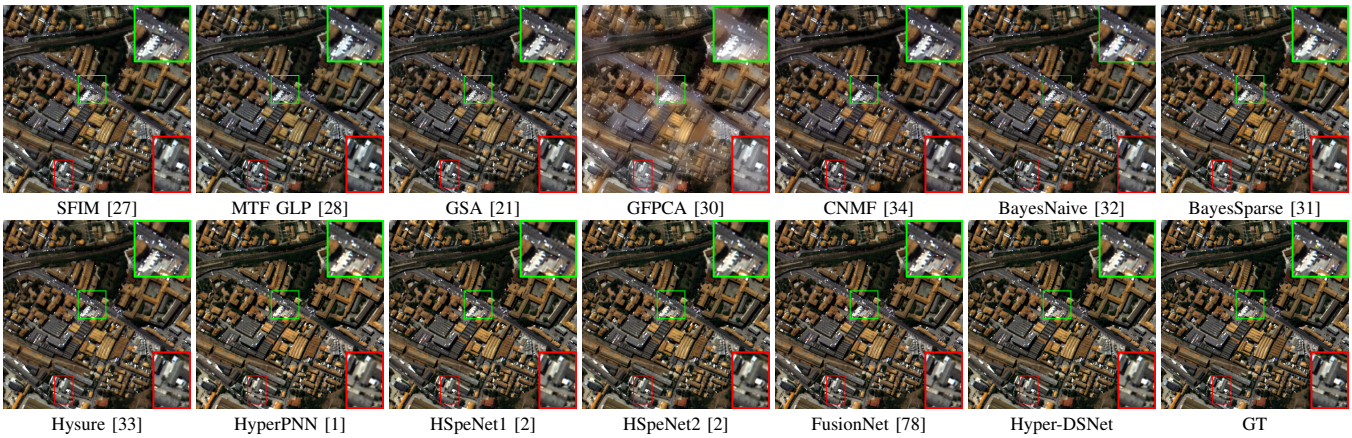

Fig. 7: The visual comparisons of fusion results obtained by different methods on a reduced-resolution Pavia dataset obtained by ROSIS (shown by bands: 20, 40 and 60).

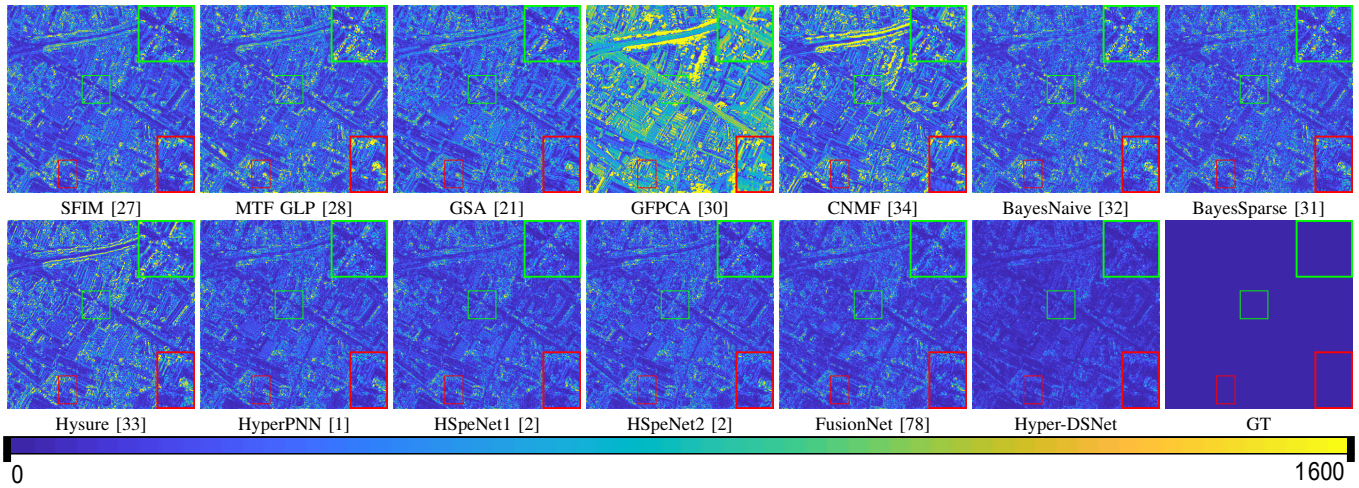

Fig. 8: The visual comparisons of the correspongding residual maps using the GT image as reference. Please note that here we select the 98-th band for better observation.

TABLE I: Information about the datasets.

	The number of bands	*The spetral range*	*The spatial resolution*	*The size of image*	*The type of land covers*
Washington DC	191	0.4 - 2.4 μm	1 m	1208 × 307	Roofs, Streets
Pavia Center	102	0.4 - 0.9 μm	1.3 m	1096 × 715	Water, Trees
Botswana	145	0.4 - 2.5 μm	30 m	1496 × 256	Seasonal Swamps
FR1	69	0.4 - 2.5 μm	30 m	2400 × 2400	Roofs, Streets

each dataset is different, we retrain different CNNs for all different datasets.

1) *Washington DC Mall Dataset* is gathered by the Hyperspectral Digital Imagery Collection Experiment (HYDICE) sensor, which contains a total of 210 bands (191 bands were retained after removing some unusable bands) ranging from 0.4 to 2.4 μm visible light and near-infrared, and the data size is 1208 × 307. Feature categories include roofs, streets, gravel roads, grass, trees, water, and shadows.
2) *Pavia Center Dataset* is acquired by the Reflective Optics System Imaging sensor (ROSIS) which records data in the spectral range from 0.4 to 0.9 μm using 115 bands (102 bands are retained after processed), and the data size is 1096 × 715.
3) *Botswana Dataset* is collected by the National Aeronautics and Space Administration Earth Observing-1 (EO-1) Hyperion satellite in Botswana from 2001 to 2004, which has a spatial dimension of 1496 × 256 and obtained data in the spectral range from 0.4 to 2.5 μm with 10 nm intervals using 242 bands (145 bands are remained after removing noise bands). This dataset consists of observations from 14 identified classes representing the land cover types in seasonal swamps, occasional swamps, and drier woodlands located in the distal portion of the Delta.

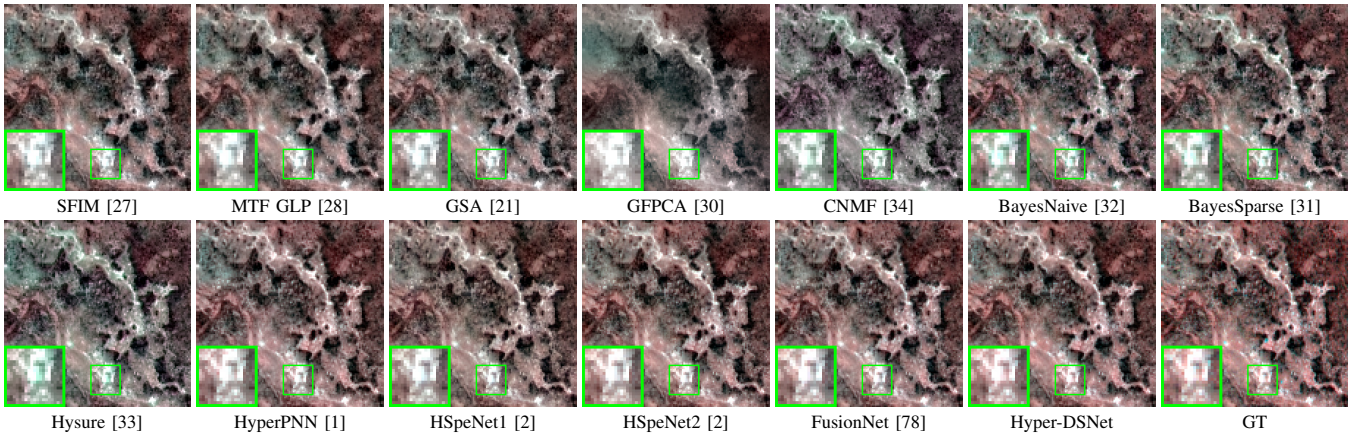

Fig. 9: The visual comparisons of fusion results obtained by different methods on a reduced-resolution Botswana dataset obtained by EO-1 (shown by bands: 10, 15 and 70).

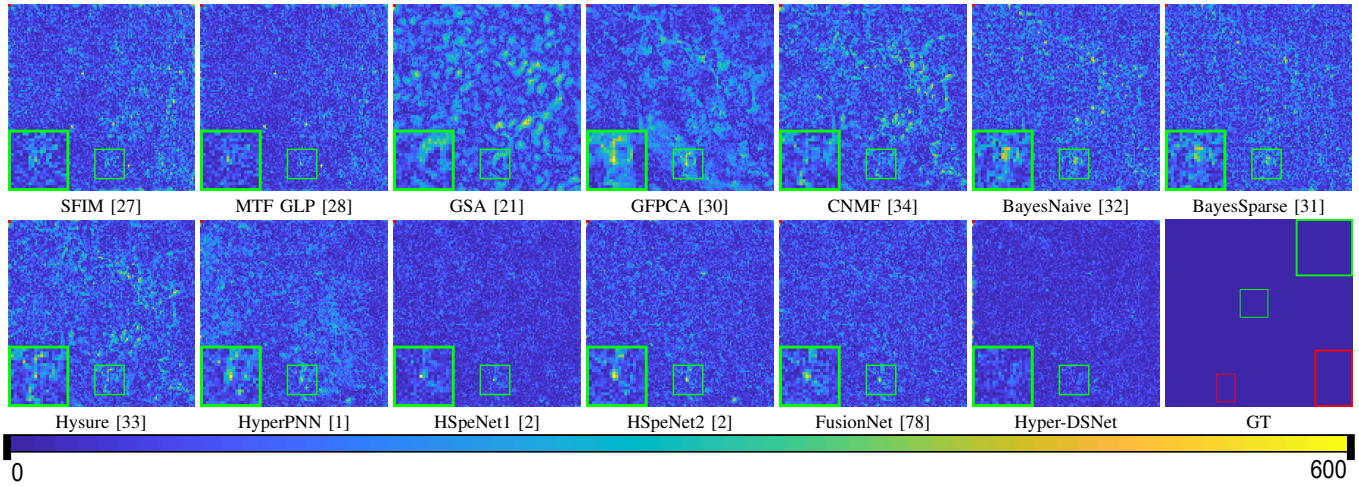

Fig. 10: The visual comparisons of the correspongding residual maps using the GT image as reference. Please note that here we select the 26-th band for better observation.

TABLE II: Average quantitative on 4 reduced-resolution WDC examples. The best performance is shown in bold, and the second place is underlined.

Method	*CC*	*SAM*	*RMSE*	*ERGAS*	*SSIM*	*PSNR*
SFIM [27]	0.934	6.504	469.59	5.220	0.740	27.01
GLP [28]	0.934	6.546	468.14	5.109	0.760	27.94
GLP HP [29]	0.938	6.451	456.27	4.883	0.762	28.68
GS [20]	0.883	7.427	640.47	7.154	0.615	21.66
GSA [21]	0.906	7.846	567.79	6.078	0.671	24.46
PCA [18]	0.750	12.09	813.19	8.213	0.501	21.06
GFPCA [30]	0.811	10.49	830.73	8.935	0.416	18.37
CNMF [34]	0.889	8.443	605.40	6.691	0.676	24.60
BayesNaive [32]	0.936	6.362	441.72	5.063	0.769	27.79
BayesSparse [31]	0.947	5.831	408.74	4.578	0.795	28.93
Hysure [33]	0.914	7.232	530.02	5.800	0.717	25.60
HyperPNN [1]	0.945	4.050	317.11	5.748	0.860	29.25
HSpeNet1 [2]	0.959	4.039	299.54	4.265	0.870	29.63
HSpeNet2 [2]	0.960	4.008	300.67	4.260	0.871	29.70
FusionNet [78]	0.958	3.917	297.37	4.338	0.866	29.69
Hyper-DSNet	**0.967**	**3.709**	**292.79**	**3.795**	**0.886**	**30.83**
Ideal value	**1**	**0**	**0**	**0**	**1**	∞

4) *FR1 Dataset* is distributed for the PRISMA contest, for pansharpening at the full spatial resolution, which can be downloaded from the website[1]. The images PAN and HS of FR1 have been obtained by extracting a $12km \times 12km$ portion (2400×2400 pixels for PAN and $400 \times 400 \times 69$ pixels for HS) from the original $30km \times 30km$ PRISMA acquisition, after accurate co-registration.

According to the Wald's protocol [82], the original HS images from three datasets serve as the reference (REF) images, and the LRHS images are gained by applying a Gaussian blur and then downsampling the result by selecting one out of every 4 pixels in both the horizontal and vertical directions. The simulated PAN image is obtained by multiplying the reference HS image on the left of the original HS images, by a suitably chosen spectral response vector. Next, we use the down-sampled LRHS image and the simulated PAN map to obtain the estimated super-resolution result images through various hyperspectral super-resolution methods. Finally, the estimated HS images will be compared with the original HS images to obtain quantitative quality measures. The specific

[1]https://openremotesensing.net/hyperspectral-pansharpening-challenge/

TABLE III: Average quantitative on 4 reduced-resolution Pavia examples. The best performance is shown in bold, and the second place is underlined.

Method	*CC*	*SAM*	*RMSE*	*ERGAS*	*SSIM*	*PSNR*
SFIM [27]	0.935	5.759	273.56	5.013	0.720	31.38
GLP [28]	0.935	6.098	278.28	4.909	0.748	31.94
GLP HP [29]	0.936	5.727	274.91	4.851	0.745	32.32
GS [20]	0.842	6.807	399.31	8.120	0.564	26.67
GSA [21]	0.937	6.281	268.08	4.978	0.722	31.50
PCA [18]	0.780	8.013	470.63	8.861	0.529	26.01
GFPCA [30]	0.825	8.163	443.99	8.618	0.429	21.18
CNMF [34]	0.890	7.175	359.31	6.310	0.654	31.22
BayesNaive [32]	0.940	6.156	271.23	4.773	0.770	34.86
BayesSparse [31]	0.946	5.617	251.75	4.485	0.780	35.11
Hysure [33]	0.920	6.254	303.81	5.469	0.730	32.24
HyperPNN [1]	0.963	4.566	203.47	3.749	0.826	33.44
HSpeNet1 [2]	0.963	4.689	200.10	3.721	0.823	33.78
HSpeNet2 [2]	0.962	4.642	205.01	3.818	0.818	33.59
FusionNet [78]	0.966	4.520	191.87	3.539	0.835	34.25
Hyper-DSNet	**0.969**	**4.294**	**184.79**	**3.434**	**0.849**	**34.56**
Ideal value	**1**	**0**	**0**	**0**	**1**	∞

TABLE IV: Average quantitative on 4 reduced-resolution Botswana examples. The best performance is shown in bold, and the second place is underlined.

Method	*CC*	*SAM*	*RMSE*	*ERGAS*	*SSIM*	*PSNR*
SFIM [27]	0.944	1.410	90.06	1.323	0.814	31.47
GLP [28]	0.951	1.382	83.47	1.207	0.837	32.55
GLP HP [29]	0.951	1.384	83.64	1.201	0.837	32.63
GS [20]	0.930	1.471	113.3	1.640	0.794	29.17
GSA [21]	0.938	1.388	91.15	1.385	0.828	31.73
PCA [18]	0.930	1.469	114.1	1.636	0.793	29.11
GFPCA [30]	0.858	1.845	160.2	2.400	0.498	24.88
CNMF [34]	0.917	1.928	104.4	1.717	0.787	30.20
BayesNaive [32]	0.945	1.592	87.07	1.343	0.829	32.02
BayesSparse [31]	0.950	1.483	82.08	1.237	0.842	32.67
Hysure [33]	0.928	1.741	96.31	1.583	0.795	30.58
HyperPNN [1]	0.960	1.365	66.96	1.195	0.873	33.11
HSpeNet1 [2]	0.959	1.339	66.42	1.209	0.867	32.99
HSpeNet2 [2]	0.960	1.330	65.96	1.198	0.868	33.07
FusionNet [78]	0.959	1.328	65.34	1.209	0.865	33.06
Hyper-DSNet	**0.964**	**1.305**	**64.70**	**1.120**	**0.876**	**33.54**
Ideal value	**1**	**0**	**0**	**0**	**1**	∞

simulation process refers to the MATLAB toolbox[2] of Loncan et al. [12].

For fair comparisons, all DL-based methods are retrained in Python 3.8.5 with Pytorch 1.9.0 on a Linux system with NVIDIA GeForce GTX 3080Ti. We set 2000 epochs for our Hyper-DSNet training with an initial learning rate of 0.0001. We use Adam [83] optimizer to minimize the ℓ_1 loss function (16) and the weight_decay is set to 1×10^{-7}. Besides, our network approach takes around 6 hours to train.

B. Compared Methods and Quantitative Metrics

For comparison, we select eleven competitive traditional fusion approaches, including SFIM [27], MTF GLP [28], MTF GLP HP [29], GS [20], GSA [21], PCA [18], GFPCA [30], CNMF [34], BayesNaive [32], BayesSparse [31], Hysure [33]. The implementation codes of these methods can be found from the public MALTAB toolbox from Loncan et al [12]. In addition, four recent benchmark deep convolutional networks for hyperspectral/multispectral pansharpening are used for comparisons, including HyperPNN [1], HSpeNet1, HSpeNet2 [2], FusionNet [78]. All codes are implemented with pytorch according to the network and strictly refer to the reported parameters in the corresponding papers.

[2]http://openremotesensing.net

TABLE V: Average quantitative on two full-resolution FR1 images. The best performance is shown in bold, and the second place is underlined.

Method	*QNR*	D_λ	D_s
SFIM [27]	0.9069	0.0318	0.0634
MTF GLP [28]	0.8986	0.0448	0.0593
MTF GLP HP [29]	0.8901	0.0479	0.0652
GS [20]	0.8339	0.1294	0.0422
GSA [21]	0.9482	0.0240	0.0285
PCA [18]	0.8909	0.0628	0.1023
GFPCA [30]	0.9546	**0.0173**	0.0286
CNMF [34]	0.9320	0.0217	0.0473
BayesNaive [32]	0.6687	0.1297	0.2317
BayesSparse [31]	0.6704	0.1293	0.2300
Hysure [33]	0.9385	0.0173	0.0450
HyperPNN [1]	0.9549	0.0274	0.0182
HSpeNet1 [2]	0.9607	0.0248	0.0149
HSpeNet2 [2]	0.9644	0.0233	0.0126
FusionNet [78]	0.9392	0.0319	0.0299
Hyper-DSNet	**0.9676**	0.0229	**0.0097**
Ideal value	**1**	**0**	**0**

Several quantitative assessments are carried out to evaluate different HS pansharpening methods with reference images. In this work, we consider four of the most often used metrics to assess the quality of the results, including cross-correlation (CC), spectral angle mapper (SAM), root mean squared error (RMSE), erreur relative globale adimensionnelle de synthèse (ERGAS) [12], structural similarity index (SSIM) [84] and peak signal-to-noise ratio (PSNR) [84]. Wherein CC, SSIM and PSNR give the measurement of spatial distortion, characterizing the geometric distortion by the average CC for each image band. SAM is a spectral index defined as the angle between the reference and fused images. As global indices, RMSE and ERGAS calculate the ℓ_2 norm between the estimated and reference images, aiming to evaluate the spatial fidelity.

In addition, to evaluate the performance of all involved methods on full-resolution, the *QNR*, D_λ, and D_s [85], [86] indexes are applied. The QNR has an ideal value of 1, instead D_λ and D_s has an ideal value of 0.

C. Experimental Results on Reduced-Resolution Datasets

This section tests the performance of all compared approaches on the three simulated datasets where the simulated way as mentioned before.

1) *Dataset of Washington DC Mall:* Washington DC Mall (WDC) dataset has 191 channels and the test data consists of four 128×128 images clipped from the original image, the rest is used to train the network parameters. For the training

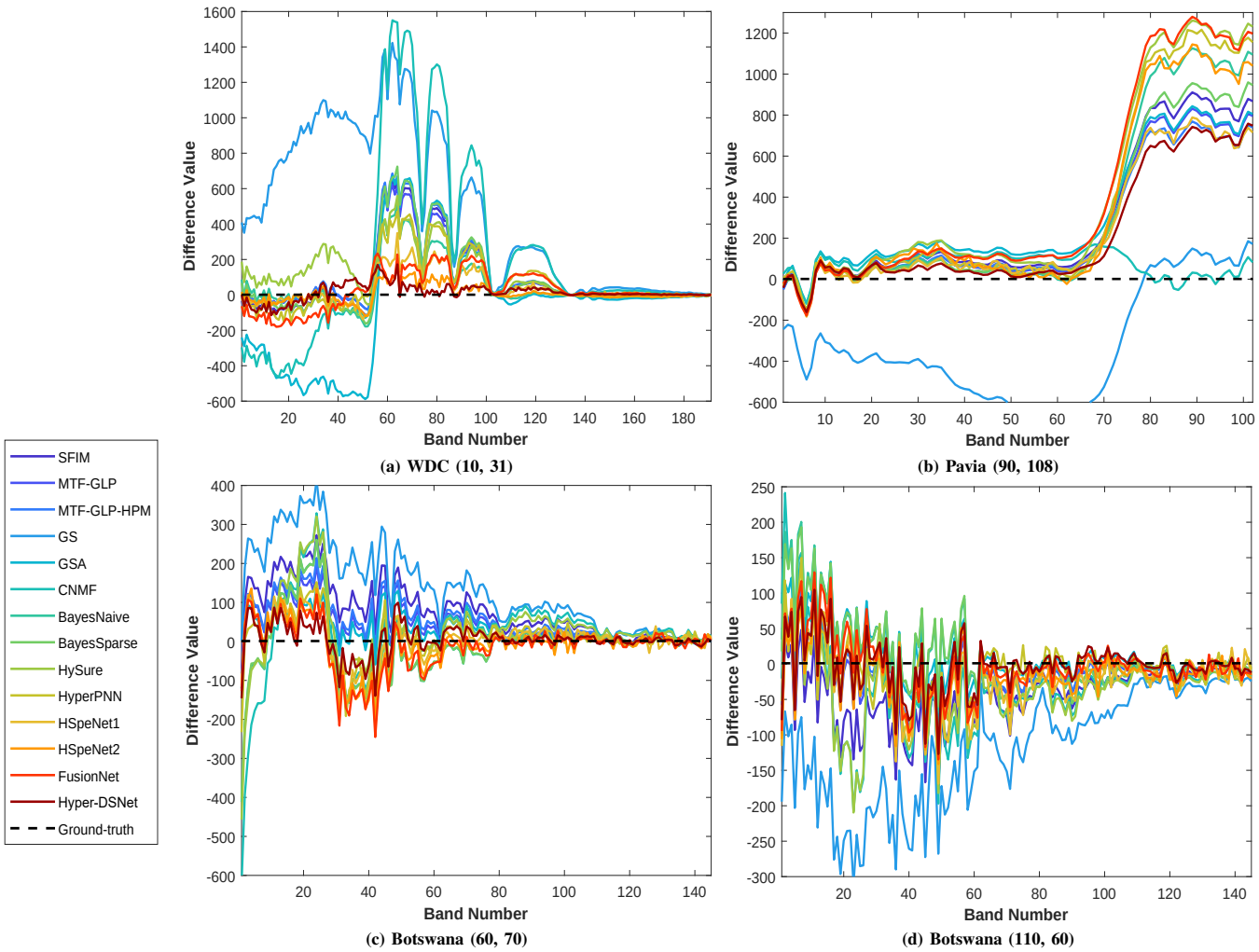

Fig. 11: Difference values between the ground-truth spectrum and the HS pansharpening results of four locations. (a) Pixel located at (10, 31) in Fig. 4. (b) Pixel located at (90, 108) in Fig. 7. (c) Pixel located at (60, 70) in Fig. 9. (d) Pixel located at (110, 60) in Fig. 9.

part, the original PAN and HS images are divided into 921 small patch pairs of 64×64 PAN patches and 16×16 HS patches, respectively. For validation, we leave 103 patch pairs from the simulated patches.

For the testing, Table II shows the average quantitative assessment of different methods on the HYDICE Washington DC Mall dataset. The best performance is shown in bold and the second is underlined. As shown in Table II, all DL-based methods show better results than traditional techniques, and far exceed in the SAM, RMSE, ERGAS metrics. Moreover, our method also surpasses the other four DL-based methods in all indicators, which verifies the effectiveness of our spectral preservation and the better extraction of spatial details.

To show a visual comparison of all methods, Fig. 4 shows the pansharpened outcomes with the pseudo-color images by selecting three bands from all the 191 image bands. It can be seen that our Hyper-DSNet method is closer to the GT map, especially the edges and corners of the building in the enlarged part. At the same time, the residual map has shown in Fig. 5. In the magnified region that we specially present, the bright spots in most traditional methods can be seen clearly, while that in other DL-based methods are obviously reduced, but there are still visible remnants. Obviously, our method has more dark blue and less yellow, which means that our error map is closer to 0.

In addition, to perform band-dependent quality evaluations of the fused HS images on the Washington DC dataset, the CC and PSNR curves as functions of the spectral bands for different methods are presented in Fig. 6. Our results in dark red show better performance overall.

2) *Pavia Center Dataset:* Pavia Center Dataset has 102 channels and the test data consists of two 400×400 images clipped from the original image, the rest is used to train the network parameters. For the training part, the big PAN and REF images are divided into 1512 small patches of 64×64 with overlapping. For validation, 168 patch pairs are left from the simulated patches.

For the testing, Table III lists the average quantitative assessment of different methods on the Pavia datasets. As shown that our Hyper-DSNet method takes first place under CC, SAM, RMSE and ERGAS metrics. For visual inspection, Fig. 7 shows the HS pansharpened outcomes with the pseudo-color images by different methods. In the enlarged green part, details such as houses and roofs are more clearly restored

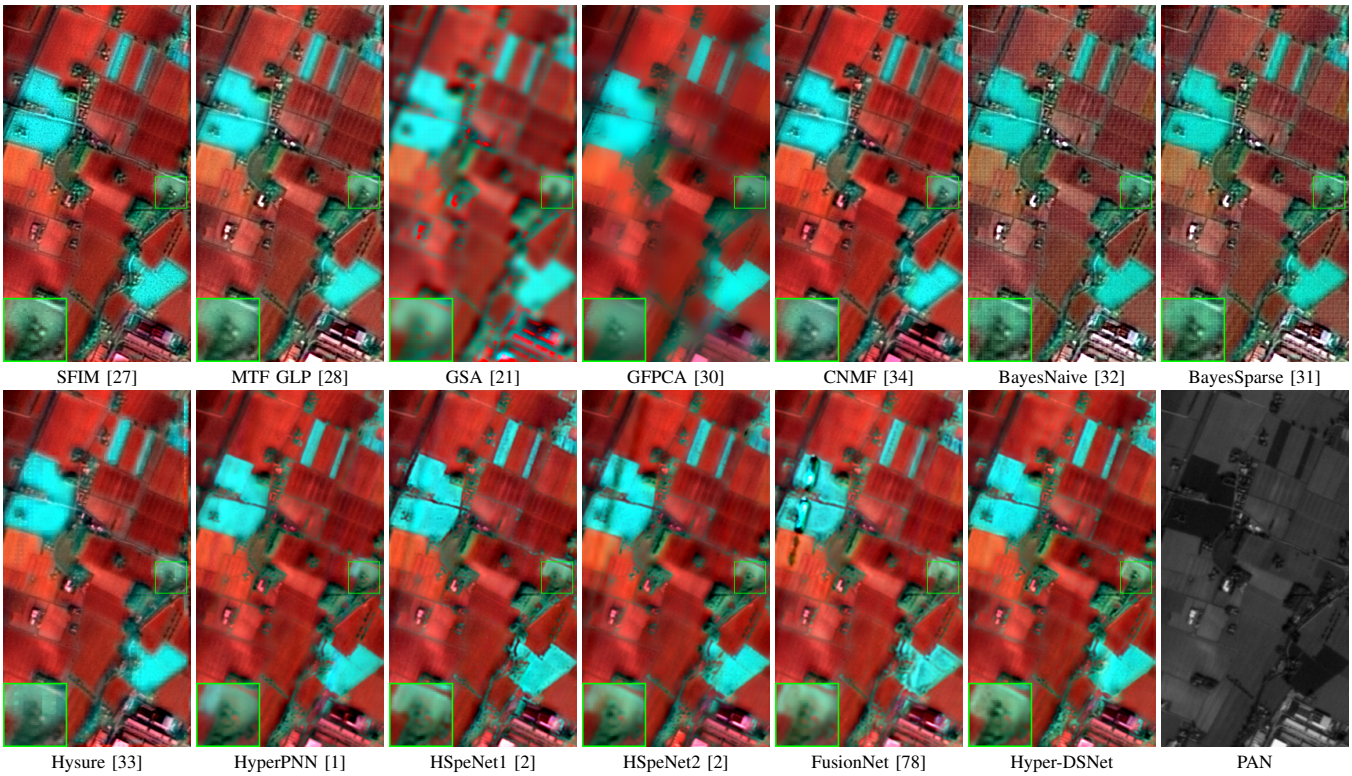

Fig. 12: The visual comparisons of fusion results obtained by different methods on a full-resolution FR1 dataset obtained by PRISMA (shown by bands: 20, 30 and 40).

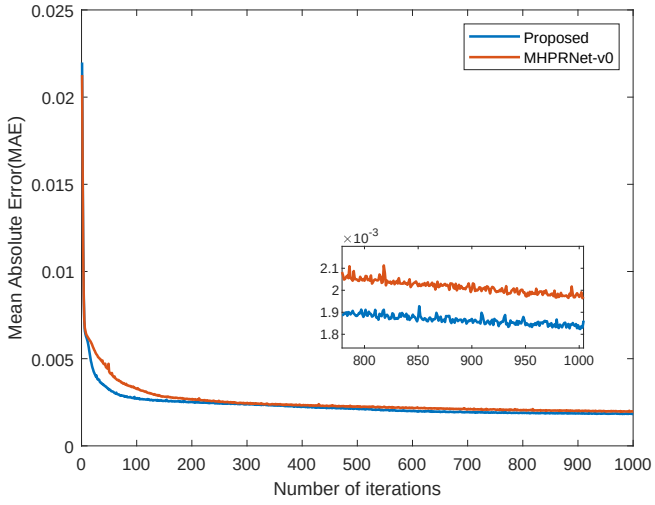

Fig. 13: Training loss of using Multi-detail Extractor module or not. The MHPRNet-V0 has defined in detail in Table. VI, the most primitive method without adding any high-pass template.

in our Hyper-DSNet result. Similarly, the error maps of one chosen channel are present in Fig. 8, which also confirms the superiority of our method.

3) *Botswana Dataset:* Botswana dataset has 102 channels and the test data consists of four 128×128 images clipped from the original image. Similar to the previously mentioned, the original PAN and HS images are divided into 799 small patches of 64×64 with overlapping in the training part, while

TABLE VI: Experimental Settings on Multi-detail Extractor module.

Method	Dir-xy	Robert	Prewitt	Sobel	Laplacian	PAN	Total
Hyper-DSNet-a1		✓	✓	✓	✓	✓	6
Hyper-DSNet-a2	✓		✓	✓	✓	✓	6
Hyper-DSNet-a3	✓	✓		✓	✓	✓	6
Hyper-DSNet-a4	✓	✓	✓		✓	✓	6
Hyper-DSNet-a5	✓	✓	✓	✓		✓	6
Hyper-DSNet-a6	✓	✓	✓	✓	✓		6
Hyper-DSNet	✓	✓	✓	✓	✓	✓	7
Hyper-DSNet-b1	×6					×1	7
Hyper-DSNet-b2		×6				×1	7
Hyper-DSNet-b3			×6			×1	7
Hyper-DSNet-b4				×6		×1	7
Hyper-DSNet-b5					×6	×1	7
Hyper-DSNet-b6						×7	7
Hyper-DSNet-v0						×1	1

168 patch pairs for simulation.

For the testing, Table IV, Fig. 9 and Fig. 10 respectively display the results of average quantitative evaluation, visual presentation and residual analysis. On this different dataset and sensor, we can still achieve the best results compared to other methods, further confirming the reliability and popularity of our proposed method. In Fig. 9, first judging from the overall color perception, the traditional method has an obvious color difference compared to the GT image. Near the pink ripple, the red of our method is more vivid and the color contrast is more obvious, which is closer to GT. At the same time, we have almost no bright spots in the error map of Fig. 10.

In order to further evaluate the spectral preservation capability of different HS pansharpening methods, the spectral

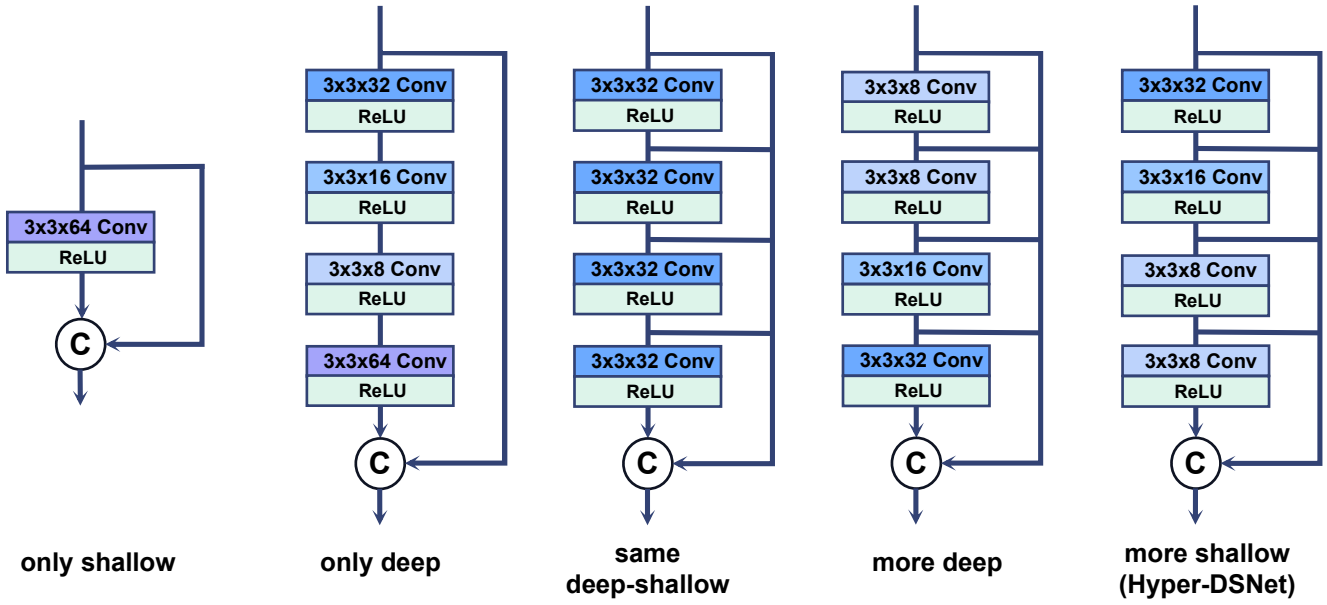

Fig. 14: Comparison experiment diagram on deep-shallow fusion module.

TABLE VII: Ablation experiment on Multi-detail Extractor module. The suffixes a1-a6, b1-b6, v0 has defined in detail in Table. VI.

Method	*CC*	*SAM*	*RMSE*	*ERGAS*
Hyper-DSNet-a1	0.96099	4.1972	317.00	4.1486
Hyper-DSNet-a2	0.96400	4.0154	308.61	4.0707
Hyper-DSNet-a3	0.96520	3.9072	305.94	4.0220
Hyper-DSNet-a4	0.96491	3.9660	306.43	3.9188
Hyper-DSNet-a5	0.96453	3.9320	305.92	3.9930
Hyper-DSNet-a6	0.96337	4.0368	309.32	3.9982
Hyper-DSNet	**0.96763**	**3.7094**	**292.79**	**3.7950**
Hyper-DSNet-b1	0.96503	3.8955	304.10	3.9460
Hyper-DSNet-b2	0.96097	4.2751	325.26	4.1626
Hyper-DSNet-b3	0.96466	3.9102	305.44	3.8722
Hyper-DSNet-b4	0.96478	4.0382	312.19	3.8560
Hyper-DSNet-b5	0.96360	4.1600	318.21	4.0068
Hyper-DSNet-b6	0.96294	3.9361	306.95	4.0463
Hyper-DSNet-v0	0.96175	4.1232	316.06	4.2611
Ideal value	**1**	**0**	**0**	**0**

different value curves of four random pixels in the previous three datasets are shown in Fig. 11. Apparently, our Hyper-DSNet provides lower spectral differences in most bands, which also shows that our algorithm can better reconstruct the details caused by the large spectral gap.

D. Experimental Results on Full-Resolution Datasets

We also test the performance of all compared approaches on the full-resolution dataset FR1. The dataset FR1 has 69 channels and the test data consists of two images (240×240 for HS and 60×60 for PAN) clipped from the original image, while the rest is trained after the downsampling simulation mentioned earlier. Similarly, we divided the training part into 734 small patch pairs of 60×60 PAN patches and 10×10 HS patches, respectively. For validation, we leave 82 patch pairs from the simulated patches.

The quantitative results in terms of all indicators are reported in Table V. Furthermore, through the visual experiment of Fig. 12, the advantages and disadvantages of each strategy can be represented more naturally. It can be seen that our proposed Hyper-DSNet can achieve better results at the full resolution, which also shows the effectiveness and robustness of the proposed method.

E. Ablation Study

1) Multi-detail Extractor Module: In this section, we illustrate the effectiveness of the proposed MDE module on the Washington DC Mall dataset. In our method, five types of high-pass operators are concatenated with PAN images as the input of the network. Here we test the effect of each high-pass operator. The specific experimental settings are shown in Table VI and the average quantitative results are presented in Table VII correspondingly.

Hyper-DSNet represents our proposed method and the suffix v0 means that only the PAN image is concatenated like most common methods. From the first six suffixes a1-a6 of Table VI, we reduced one operator in turn based on the original operators, *i.e.*, Dir-xy, Robert, Prewitt, Sobel, Laplacian operator and the PAN image, to test their effects. Furthermore, we test that only one operator is selected at a time, while keeping the same dimension as the original for fairness, or not using a high-pass module at all, which are defined as suffixes b1-b6.

As can be seen from the results in Table VII, all results with high-pass templates are much better than those without using a high-pass operator. Hyper-DSNet-v0, the most primitive method without adding any high-pass template, has the worst ERGAS and second-worst CC value in the result. While in

TABLE VIII: Ablation study results on Deep-shallow Fusion Module module.

Method	*CC*	*SAM*	*RMSE*	*ERGAS*	*Parameters*
only shallow	0.95842	4.7397	355.03	4.1267	294467
only deep	0.96011	4.1346	312.13	4.2008	316795
same deep-shallow	0.96608	3.7505	295.66	3.8607	342803
more deep	0.96177	4.1467	315.00	4.1320	298835
Hyper-DSNet	**0.96763**	**3.7094**	**292.79**	**3.7950**	309203
Ideal value	**1**	**0**	**0**	**0**	

TABLE IX: Other experiment results on multi-scale and spectral attention module.

Method	*CC*	*SAM*	*RMSE*	*ERGAS*
Without Multi-scale	0.96219	4.0994	310.15	4.0574
Without SA	0.95520	4.1150	316.31	4.8998
Hyper-DSNet	**0.96763**	**3.7094**	**292.79**	**3.7950**
Ideal value	**1**	**0**	**0**	**0**

all methods that add high-pass operators, Hyper-DSNet has achieved the best results. It is worth noting that the evaluation indicators will also slightly decrease in the a6 group without the PAN image. In addition, the training loss comparison of whether to use a high-pass module is shown in Fig. 13. The proposed method has lower loss and converges faster.

2) Deep-shallow Fusion Module: To evaluate the advantage of the DSF module, we replace this module with the following forms in Fig. 14. We set only the shallow layer and only the deep layer to prove the advantages of the deep-shallow fusion module. Furthermore, we believe that in the fusion task, more attention should be paid to shallow texture information rather than deep semantics. Therefore, we specially set up three experiments about the same number of feature maps, more shallow layers and more deep layers. We summarize the results and the corresponding module parameters in Table VIII.

It is obvious that the effects of the last three with both deep and shallow layers are better than the first two, which means that the deep and shallow layers both have the information we need. It is also noticed that the SAM and RMSE metrics deteriorate significantly in the only shallow network. In addition, the effect of more shallow layers is better than more deep layers, which also shows that the detailed information in the shallow layers may be more important. Compared with the same number of feature maps, setting the numbers of channels to decrease with depth can not only reduce the parameters, but also maintain a fairly better effect.

3) Multi-scale convolution module and SA module: Finally, we discuss the role of the multi-scale convolution module and the SA module. On the basis of the original network, we set up two sets of experiments by removing the corresponding part. For example, the multi-scale convolution module is replaced with a general 3x3 convolution with the same number of feature maps. The result is shown in Table IX which indicates the improved effect of adding these two modules, especially the SA module. If discarding the SA module, the ERGAS and CC indicators have dropped significantly. In other words, the SA module is indispensable for spectral preservation.

TABLE X: The number of parameters (NoPs) and test time of DL-based methods on Pavia dataset.

Method	HyperPNN	HSpeNet1	HSpeNet2	Hyper-DSNet
NoPs	1.3×10^5	1.8×10^5	1.1×10^5	1.8×10^5
Time	0.4382	0.4879	0.5276	0.4713

F. Parameter Numbers

The number of parameters (NoPs) of all the compared DL-based methods and corresponding test time on the Pavia dataset are presented in Table. X. It can be seen that the amount of parameters of Hyper-DSNet has not increased much than the other compared DL-based methods, but achieved the best results, which proves our method can fully mine and utilize information.

V. Conclusions

In this article, we propose a new framework named Hyper-DSNet for the two challenges in HS pansharpening, *i.e.*, spectral distortion by the wider spectral range between HS and PAN image, and spatial information loss in continuous spectral bands. Specifically, our Hyper-DSNet mainly consists of three parts, *i.e.*, MDE module, DSF module and SA module. Plenty of experiments on three benchmark datasets and one full-resolution dataset acquired by multiple sensors demonstrate that our method has both good quantitative indicators and visual outcomes, surpassing the previous traditional and SOTA CNN-based techniques. We emphatically examined the importance of the MDE module and DSF module, which can also be widely embedded in other networks. Also, sufficient ablation studies are given to verify the effectiveness of multiple high-pass operators in the task of HS pansharpening.

References

[1] Lin He, Jiawei Zhu, Jun Li, Antonio Plaza, Jocelyn Chanussot, and Bo Li, "Hyperpnn: Hyperspectral pansharpening via spectrally predictive convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 8, pp. 3092–3100, 2019.

[2] Lin He, Jiawei Zhu, Jun Li, Deyu Meng, Jocelyn Chanussot, and Antonio Plaza, "Spectral-fidelity convolutional neural networks for hyperspectral pansharpening," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5898–5914, 2020.

[3] Weisheng Dong, Fazuo Fu, Guangming Shi, Xun Cao, Jinjian Wu, Guangyu Li, and Xin Li, "Hyperspectral image super-resolution via non-negative structured sparse representation," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2337–2352, 2016.

[4] Alp Ertürk, Marian-Daniel Iordache, and Antonio Plaza, "Sparse unmixing with dictionary pruning for hyperspectral change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 1, pp. 321–330, 2016.

[5] Richard J Ellis and Peter W Scott, "Evaluation of hyperspectral remote sensing as a means of environmental monitoring in the st. austell china clay (kaolin) region, cornwall, uk," *Remote Sensing of Environment*, vol. 93, no. 1-2, pp. 118–130, 2004.

[6] Thais Andressa Carrino, Alvaro Penteado Crósta, Catarina Labouré Bemfica Toledo, and Adalene Moreira Silva, "Hyperspectral remote sensing applied to mineral exploration in southern peru: A multiple data integration approach in the chapi chiara gold prospect," *International Journal of Applied Earth Observation and Geoinformation*, vol. 64, pp. 287–300, 2018.

[7] Charlotte A Bishop, Jian Guo Liu, and Philippa J Mason, "Hyperspectral remote sensing for mineral exploration in pulang, yunnan province, china," *International Journal of Remote Sensing*, vol. 32, no. 9, pp. 2409–2426, 2011.

[8] S Mahesh, DS Jayas, J Paliwal, and NDG White, "Hyperspectral imaging to classify and monitor quality of agricultural materials," *Journal of Stored Products Research*, vol. 61, pp. 17–26, 2015.

[9] Clement Chion, Jacques-André Landry, and Luis Da Costa, "A genetic-programming-based method for hyperspectral data information extraction: Agricultural applications," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 8, pp. 2446–2457, 2008.

[10] Carlos Souza Jr, Laurel Firestone, Luciano Moreira Silva, and Dar Roberts, "Mapping forest degradation in the eastern amazon from spot 4 through spectral mixture models," *Remote Sensing of Environment*, vol. 87, no. 4, pp. 494–506, 2003.

[11] Yun Zhang, "Understanding image fusion," *Photogrammetric Engineering and Remote Sensing*, vol. 70, no. 6, pp. 657–661, 2004.

[12] Laetitia Loncan, Luis B De Almeida, José M Bioucas-Dias, Xavier Briottet, Jocelyn Chanussot, Nicolas Dobigeon, Sophie Fabre, Wenzhi Liao, Giorgio A Licciardi, Miguel Simoes, et al., "Hyperspectral pansharpening: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, no. 3, pp. 27–46, 2015.

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

[14] Vittala K Shettigara, "A generalized component substitution technique for spatial enhancement of multispectral images using a higher resolution data set," *Photogrammetric Engineering and Remote Sensing*, vol. 58, no. 5, pp. 561–567, 1992.

[15] Vijay P Shah, Nicolas H Younan, and Roger L King, "An efficient pan-sharpening method via a combined adaptive pca approach and contourlets," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 5, pp. 1323–1335, 2008.

[16] Wjoseph Carper, Thomasm Lillesand, and Ralphw Kiefer, "The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data," *Photogrammetric Engineering and Remote Sensing*, vol. 56, no. 4, pp. 459–467, 1990.

[17] Te-Ming Tu, Shun-Chi Su, Hsuen-Chyun Shyu, and Ping S Huang, "A new look at ihs-like image fusion methods," *Information Fusion*, vol. 2, no. 3, pp. 177–186, 2001.

[18] Pats Chavez, Stuart C Sides, Jeffrey A Anderson, et al., "Comparison of three different methods to merge multiresolution and multispectral data-landsat tm and spot panchromatic," *Photogrammetric Engineering and Remote Sensing*, vol. 57, no. 3, pp. 295–303, 1991.

[19] Te-Ming Tu, Ping Sheng Huang, Chung-Ling Hung, and Chien-Ping Chang, "A fast intensity-hue-saturation fusion technique with spectral adjustment for ikonos imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 1, no. 4, pp. 309–312, 2004.

[20] Craig A Laben and Bernard V Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," Jan. 4 2000, US Patent 6,011,875.

[21] Bruno Aiazzi, Stefano Baronti, and Massimo Selva, "Improving component substitution pansharpening through multivariate regression of MS + Pan data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3230–3239, 2007.

[22] Jiahui Qu, Yunsong Li, and Wenqian Dong, "Hyperspectral pansharpening with guided filter," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 11, pp. 2152–2156, 2017.

[23] Stephane G Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," in *Fundamental Papers in Wavelet Theory*, pp. 494–513. Princeton University Press, 2009.

[24] Guy P Nason and Bernard W Silverman, "The stationary wavelet transform and some statistical applications," in *Wavelets and statistics*, pp. 281–299. Springer, 1995.

[25] Mark J Shensa et al., "The discrete wavelet transform: wedding the a trous and mallat algorithms," *IEEE Transactions on Signal Processing*, vol. 40, no. 10, pp. 2464–2482, 1992.

[26] Peter J Burt and Edward H Adelson, "The laplacian pyramid as a compact image code," in *Readings in Computer Vision*, pp. 671–679. Elsevier, 1987.

[27] JG Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *International Journal of Remote Sensing*, vol. 21, no. 18, pp. 3461–3472, 2000.

[28] Bruno Aiazzi, L Alparone, Stefano Baronti, Andrea Garzelli, and Massimo Selva, "Mtf-tailored multiscale fusion of high-resolution ms and pan imagery," *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 5, pp. 591–596, 2006.

[29] Gemine Vivone, Rocco Restaino, Mauro Dalla Mura, Giorgio Licciardi, and Jocelyn Chanussot, "Contrast and error-based fusion schemes for multispectral image pansharpening," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 5, pp. 930–934, 2013.

[30] Wenzhi Liao, Xin Huang, Frieke Van Coillie, Sidharta Gautama, Aleksandra Pižurica, Wilfried Philips, Hui Liu, Tingting Zhu, Michal Shimoni, Gabriele Moser, et al., "Processing of multiresolution thermal hyperspectral and digital color data: Outcome of the 2014 IEEE GRSS data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 6, pp. 2984–2996, 2015.

[31] Qi Wei, José Bioucas-Dias, Nicolas Dobigeon, and Jean-Yves Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3658–3668, 2015.

[32] Qi Wei, Nicolas Dobigeon, and Jean-Yves Tourneret, "Fast fusion of multi-band images based on solving a sylvester equation," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4109–4121, 2015.

[33] Miguel Simoes, José Bioucas-Dias, Luis B Almeida, and Jocelyn Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3373–3388, 2014.

[34] Naoto Yokoya, Takehisa Yairi, and Akira Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 2, pp. 528–537, 2011.

[35] Michael Moeller, Todd Wittman, and Andrea L Bertozzi, "A variational approach to hyperspectral image fusion," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV*. International Society for Optics and Photonics, 2009, vol. 7334, p. 73341E.

[36] Faming Fang, Fang Li, Chaomin Shen, and Guixu Zhang, "A variational approach for pan-sharpening," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2822–2834, 2013.

[37] Joan Duran, Antoni Buades, Bartomeu Coll, and Catalina Sbert, "A nonlocal variational model for pansharpening image fusion," *SIAM Journal on Imaging Sciences*, vol. 7, no. 2, pp. 761–796, 2014.

[38] Liang-Jian Deng, Gemine Vivone, Weihong Guo, Mauro Dalla Mura, and Jocelyn Chanussot, "A variational pansharpening approach based on reproducible kernel hilbert space and heaviside function," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4330–4344, 2018.

[39] Zi-Yao Zhang, Ting-Zhu Huang, Liang-Jian Deng, Jie Huang, Xi-Le Zhao, and Chao-Chao Zheng, "A framelet-based iterative pan-sharpening approach," *Remote Sensing*, vol. 10, no. 4, pp. 622, 2018.

[40] Ting Xu, Ting-Zhu Huang, Liang-Jian Deng, Xi-Le Zhao, and Jie Huang, "Hyperspectral image superresolution using unidirectional total variation with tucker decomposition," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4381–4398, 2020.

[41] Liang-Jian Deng, Minyu Feng, and Xue-Cheng Tai, "The fusion of panchromatic and multispectral remote sensing images via tensor-based sparse modeling and hyper-laplacian prior," *Information Fusion*, vol. 52, pp. 76–89, 2019.

[42] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[43] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690.

[44] Lianru Gao, Danfeng Hong, Jing Yao, Bing Zhang, Paolo Gamba, and Jocelyn Chanussot, "Spectral superresolution of multispectral imagery with joint sparse and low-rank learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 2269–2280, 2020.

[45] Ye-Tao Wang, Xi-Le Zhao, Tai-Xiang Jiang, Liang-Jian Deng, Yi Chang, and Ting-Zhu Huang, "Rain streaks removal for single image via kernel-guided convolutional neural network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 8, pp. 3664–3676, 2021.

[46] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[47] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*. Springer, 2014, pp. 184–199.

[48] Renwei Dian and Shutao Li, "Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 5135–5146, 2019.

[49] Renwei Dian, Shutao Li, and Leyuan Fang, "Learning a low tensor-train rank representation for hyperspectral image super-resolution," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2672–2683, 2019.

[50] Jin-Fan Hu, Ting-Zhu Huang, Liang-Jian Deng, Tai-Xiang Jiang, Gemine Vivone, and Jocelyn Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[51] Xiangyong Cao, Jing Yao, Zongben Xu, and Deyu Meng, "Hyperspectral image classification with convolutional neural network and active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4604–4616, 2020.

[52] Danfeng Hong, Lianru Gao, Jing Yao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[53] Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[54] Xiangyong Cao, Xueyang Fu, Chen Xu, and Deyu Meng, "Deep spatial-spectral global reasoning network for hyperspectral image denoising," *IEEE Transactions on Geoscience and Remote Sensing*, DOI: 10.1109/TGRS.2020.3016820.

[55] Shutao Li, Renwei Dian, Leyuan Fang, and José M Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4118–4130, 2018.

[56] Yong Yang, Hangyuan Lu, Shuying Huang, and Wei Tu, "Remote sensing image fusion based on fuzzy logic and salience measure," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 11, pp. 1943–1947, 2019.

[57] Zi-Rong Jin, Liang-Jian Deng, Tian-Jing Zhang, and Xiao-Xu Jin, "Bam: Bilateral activation mechanism for image fusion," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4315–4323.

[58] Jin-Fan Hu, Ting-Zhu Huang, Liang-Jian Deng, Tai-Xiang Jiang, Gemine Vivone, and Jocelyn Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021.

[59] Zhong-Cheng Wu, Ting-Zhu Huang, Liang-Jian Deng, Jin-Fan Hu, and Gemine Vivone, "Vo+net: An adaptive approach using variational optimization and deep learning for panchromatic sharpening," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2021.

[60] Zhong-Cheng Wu, Ting-Zhu Huang, Liang-Jian Deng, Gemine Vivone, Jia-Qing Miao, Jin-Fan Hu, and Xi-Le Zhao, "A new variational approach based on proximal deep injection and gradient intensity similarity for spatio-spectral image fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 6277–6290, 2020.

[61] Renwei Dian, Shutao Li, and Xudong Kang, "Regularizing hyperspectral and multispectral image fusion by cnn denoiser," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 3, pp. 1124–1135, 2020.

[62] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa, "Pansharpening by convolutional neural networks," *Remote Sensing*, vol. 8, no. 7, pp. 594, 2016.

[63] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley, "Pannet: A deep network architecture for pan-sharpening," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5449–5457.

[64] Weiying Xie, Jie Lei, Yuhang Cui, Yunsong Li, and Qian Du, "Hyperspectral pansharpening with deep priors," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1529–1543, 2019.

[65] Xueyang Fu, Wu Wang, Yue Huang, Xinghao Ding, and John Paisley, "Deep multiscale detail networks for multiband spectral image sharpening," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2090–2104, 2020.

[66] Xueyang Fu, Zihuang Lin, Yue Huang, and Xinghao Ding, "A variational pan-sharpening with local gradient constraints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10265–10274.

[67] Renwei Dian, Shutao Li, Anjing Guo, and Leyuan Fang, "Deep hyperspectral image sharpening," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5345–5355, 2018.

[68] Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, and Liangpei Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 978–989, 2018.

[69] Peixian Zhuang, Qingshan Liu, and Xinghao Ding, "Pan-ggf: A probabilistic method for pan-sharpening with gradient domain guided image filtering," *Signal Processing*, vol. 156, pp. 177–190, 2019.

[70] Penghao Guo, Peixian Zhuang, and Yecai Guo, "Bayesian pan-sharpening with multiorder gradient-based deep network constraints," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 950–962, 2020.

[71] Yong Yang, Lei Wu, Shuying Huang, Weiguo Wan, Wei Tu, and Hangyuan Lu, "Multiband remote sensing image pansharpening based on dual-injection model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1888–1904, 2020.

[72] Yong Yang, Hangyuan Lu, Shuying Huang, and Wei Tu, "Pansharpening based on joint-guided detail extraction," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 389–401, 2020.

[73] Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, and Tian-Jing Zhang, "Dynamic cross feature fusion for remote sensing pansharpening," in *Proceedings of the IEEE International Conference on Computer Vision*, October 2021, pp. 14687–14696.

[74] Yudong Wang, Liang-Jian Deng, Tian-Jing Zhang, and Xiao Wu, "Ss-conv: Explicit spectral-to-spatial convolution for pansharpening," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4472–4480.

[75] Cheng Jin, Liang-Jian Deng, Ting-Zhu Huang, and Gemine Vivone, "Laplacian pyramid networks: A new approach for multispectral pansharpening," *Information Fusion*, vol. 78, pp. 158–170, 2022.

[76] Zhangxi Xiong, Qing Guo, Mingliang Liu, and An Li, "Pan-sharpening based on convolutional neural network by using the loss function with no-reference," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 897–906, 2020.

[77] Sijia Li, Qing Guo, and An Li, "Pan-sharpening based on cnn+ pyramid transformer by using no-reference loss," *Remote Sensing*, vol. 14, no. 3, pp. 624, 2022.

[78] Liang-Jian Deng, Gemine Vivone, Cheng Jin, and Jocelyn Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 8, pp. 6995–7010, 2020.

[79] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010.

[80] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[81] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[82] Lucien Wald, Thierry Ranchin, and Marc Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogrammetric Engineering and Remote Sensing*, vol. 63, no. 6, pp. 691–699, 1997.

[83] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[84] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[85] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A Licciardi, Rocco Restaino, and Lucien Wald, "A critical comparison among pansharpening algorithms," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2565–2586, 2014.

[86] Gemine Vivone, Mauro Dalla Mura, Andrea Garzelli, Rocco Restaino, Giuseppe Scarpa, Magnus Orn Ulfarsson, Luciano Alparone, and Jocelyn Chanussot, "A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods," *IEEE Geoscience and Remote Sensing Magazine*, 2020.