Hyperspectral Image Super-Resolution via Deep Spatiospectral Attention Convolutional Neural Networks

Jin-Fan Hu, Ting-Zhu Huang, *Member, IEEE*, Liang-Jian Deng, *Member, IEEE*, Tai-Xiang Jiang, *Member, IEEE*, Gemine Vivone, *Senior Member, IEEE*, and Jocelyn Chanussot, *Fellow, IEEE*

***Abstract*— Hyperspectral images (HSIs) are of crucial importance in order to better understand features from a large number of spectral channels. Restricted by its inner imaging mechanism, the spatial resolution is often limited for HSIs. To alleviate this issue, in this work, we propose a simple and efficient architecture of deep convolutional neural networks to fuse a low-resolution HSI (LR-HSI) and a high-resolution multispectral image (HR-MSI), yielding a high-resolution HSI (HR-HSI). The network is designed to preserve both spatial and spectral information thanks to a new architecture based on: 1) the use of the LR-HSI at the HR-MSI's scale to get an output with satisfied spectral preservation and 2) the application of the attention and pixelShuffle modules to extract information, aiming to output high-quality spatial details. Finally, a plain mean squared error loss function is used to measure the performance during the training. Extensive experiments demonstrate that the proposed network architecture achieves the best performance (both qualitatively and quantitatively) compared with recent state-of-the-art HSI super-resolution approaches. Moreover, other significant advantages can be pointed out by the use of the proposed approach, such as a better network generalization ability, a limited computational burden, and the robustness with respect to the number of training samples. Please find the source code and pretrained models from https://liangjiandeng.github.io/Projects_Res/HSRnet_2021tnnls.html.**

***Index Terms*— Attention module (AM), deep convolutional neural network (CNN), hyperspectral image (HSI) super-resolution, image fusion, pixelShuffle (PS).**

I. Introduction

TRADITIONAL multispectral images (MSIs, e.g., RGB images) usually contain a limited number of spectral bands providing a limited spectral information. Since hyperspectral imaging obtains more spectral bands containing more information of the spectral structure, it has become a non-negligible technology that can capture the intrinsic properties of different materials. However, due to the physical limitation of imaging sensors, there is a trade-off between the spatial resolution and the spectral resolution in an HSI [1]. Therefore, it is burdensome to obtain a HSI with a high spatial resolution. In this condition, hyperspectral image (HSI) super-resolution by fusing a low-resolution hyperspectral image (LR-HSI) with a high-resolution multispectral image (HR-MSI) is a promising way to address the problem.

Many researchers have focused on HSI super-resolution to increase the spatial resolution of LR-HSI proposing several algorithms. Many of them consider the following linear model:

$$\mathbf{Y} = \mathbf{XBS}, \quad \mathbf{Z} = \mathbf{RX} \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{S\times hw}$, $\mathbf{Z} \in \mathbb{R}^{s\times HW}$ and $\mathbf{X} \in \mathbb{R}^{S\times HW}$ represent the mode-3 unfolding matrices of LR-HSI ($\mathcal{Y} \in \mathcal{R}^{h\times w\times S}$), HR-MSI ($\mathcal{Z} \in \mathcal{R}^{H\times W\times s}$) and the latent high-resolution hyperspectral image (HR-HSI) ($\mathcal{X} \in \mathcal{R}^{H\times W\times S}$), respectively, h and w represent the height and width of LR-HSI, H and W denote the height and width of HR-MSI, s and S denote the spectral band number of HR-MSI and LR-HSI, respectively. Additionally, $\mathbf{B} \in \mathbb{R}^{HW\times HW}$ is the blur matrix, $\mathbf{S} \in \mathbb{R}^{HW\times hw}$ denotes the downsampling matrix, and $\mathbf{R} \in \mathbb{R}^{s\times S}$ represents the spectral response matrix. It is worth to be remarked that coherently with the notation adopted above, in this article, we denote scalar, matrix, and tensor in nonbold case, bold upper case, and calligraphic upper case letters, respectively.

Based on the models in (1), many related approaches have been proposed. Different prior knowledge or regularization

Manuscript received November 1, 2020; revised February 2, 2021 and April 20, 2021; accepted May 24, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61772003, Grant 61702083, Grant 12001446, and Grant 61876203; in part by the National Key Research and Development Program of China under Grant 2020YFA0714001; in part by the Key Projects of Applied Basic Research in Sichuan Province under Grant 2020YJ0216; and in part by the Fundamental Research Funds for the Central Universities under Grant JBK2102001; and in part by MIAI@Grenoble Alpes under Grant ANR-19-P3IA-0003. *(Corresponding authors: Ting-Zhu Huang; Liang-Jian Deng.)*

Jin-Fan Hu, Ting-Zhu Huang, and Liang-Jian Deng are with the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: hujf0206@163.com; tingzhuhuang@126.com; liangjian.deng@uestc.edu.cn).

Tai-Xiang Jiang is with the FinTech Innovation Center, Financial Intelligence and Financial Engineering Research Key Laboratory of Sichuan Province, School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu 610074, China (e-mail: taixiangjiang@gmail.com).

Gemine Vivone is with the National Research Council—Institute of Methodologies for Environmental Analysis, CNR-IMAA, 85050 Tito Scalo, Italy (e-mail: gvivone@unisa.it).

Jocelyn Chanussot is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with Université Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France (e-mail: jocelyn.chanussot@grenoble-inp.fr).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3084682.

Digital Object Identifier 10.1109/TNNLS.2021.3084682

terms are integrated into those methods. However, the spectral response matrix **R** is usually unknown, thus the traditional methods need to select or estimate the matrix **R** and other involved parameters. Additionally, the related regularization parameters used in these kinds of approaches are often image-dependent.

Recently, with the tremendous development of neural networks, deep learning has become a promising way to deal with the pansharpening and HSI super-resolution problems. In the research of [2], a model of HSI fusion problems is proposed. It integrates the convolutional neural network (CNN) modules into the traditional framework, which can utilize the automatic feature learning ability and keep the advantages of the traditional model. Dian *et al.* [4] mainly focus on the spatial detail recovery and learn image priors via a CNN. These learned priors have been included in a traditional regularization model to improve the outcomes getting better image features than traditional regularization model-based methods. Xie *et al.* [3] proposed a model-enlightened deep learning method for HSI super-resolution. This method has exhibited an ability to preserve spectral information and spatial details, thus obtaining state-of-the-art HSI super-resolution results. Liu *et al.* [5] design a structure with a shallow network and a deep network that can capture features in different levels for the pan-sharpening problem. It is proposed to obtain the spatial details with minimal spectral distortion, and then those details are merged into the MSIs.

However, deep learning-based approaches for HSI super-resolution also encounter some challenges. First of all, these methods sometimes have *complicated architectures* with millions of parameters to estimate. Second, due to the complicated architecture and large-scale training data, *expensive computation and storage* are usually involved. Third, deep learning-based methods are data-dependent, which usually holds a *weak network generalization*. Thus, the model trained on a specific dataset could poorly perform on a different kind of dataset. Instead, the proposed network architecture is an improvement on these above-mentioned drawbacks.

In this article, the proposed network architecture (called HSRnet from hereon) can be decomposed into two parts. One part is to preserve the spectral information of HR-HSI by upsampling the LR-HSI. The other part is mainly to get the spatial details of HR-HSI by training a CNN with the HR-MSI and LR-HSI as inputs. By imposing a similarity between the network output and the reference [ground-truth (GT)] image, we can efficiently estimate the parameters involved in the network. In summary, this article mainly consists of the following contributions.

1) An effective CNN with a plain architecture and few parameters is proposed for the fusion of HSIs and MSIs. In the proposed CNN, the spectral and spatial information is well preserved and fused, yielding an HR-HSI.
2) The channel attention (CA) and spatial attention (SA) modules are designed and incorporated for refining the spectral and spatial details. A cross-scale operation conducted on the lower and the original scales is exploited in the network architecture, aiming to reduce the computation and benefit the multiscale information. The pixelShuffle (PS) module is introduced to capture details during the upsampling process. Meanwhile, the leaky ReLU is adopted for a more accurate estimation of the negative part of the detail.
3) Experiments on different datasets illustrate the superior of our method. More discussions are conducted to illustrate that, compared with state-of-the-art CNN-based methods (see Fig. 1), our network: 1) is robust to the number of training samples; 2) consumes less time in both the training and testing stage; and 3) has better promising generalization ability to yield competitive results for different datasets even though the network is trained only on a specific dataset.

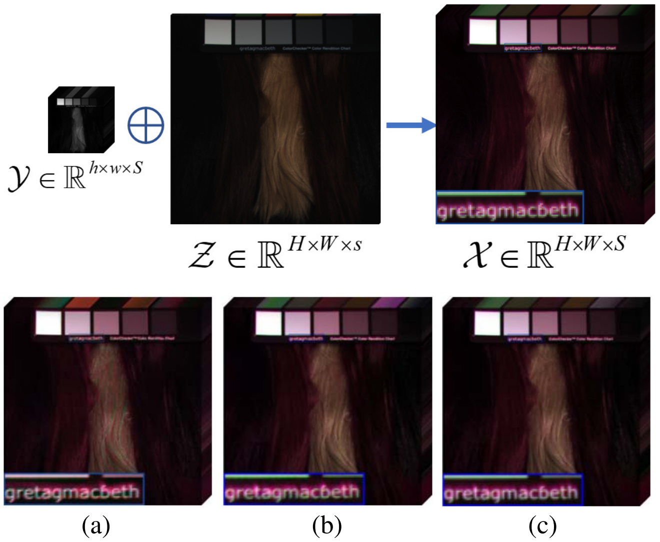

$\mathcal{Y} \in \mathbb{R}^{h\times w\times S}$
$\mathcal{Z} \in \mathbb{R}^{H\times W\times s}$
$\mathcal{X} \in \mathbb{R}^{H\times W\times S}$
gretagmacbeth
(a)
(b)
(c)

Fig. 1. First row: the schematic of HSI resolution on a test image from the CAVE dataset (h and w represent the height and width of LR-HSI, H and W denote the height and width of HR-MSI, s, and S denote the spectral band number of HR-MSI and LR-HSI, respectively). The right image is the GT HR-HSI, $\mathcal{X}$. Second row: the results obtained by (a) CNN-FUS (PSNR = 43.77 dB), (b) MHFnet (PSNR = 46.53 dB), and (c) proposed HSRnet (PSNR = 47.78 dB), where PSNR stands for the peak signal-to-noise ratio. Note that all the images are displayed with pseudocolor red, green, and blue (RGB) format using R = first band, G = ninth band, and B = second band. From the visual analysis, our HSRnet achieves the best result, while CNN-FUS shows wrong colors and the result of MHFnet depicts some artifacts.

The rest of this article is outlined as follows. Section II presents the related works about the hyperspectral super-resolution problem. Section III introduces the proposed network architecture. In Section IV, extensive experiments are conducted to assess the effectiveness of the proposed architecture. Furthermore, some discussions about the image spectral response, the network generalization, the computational burden, the benefit of the leaky ReLU activation function, and the use of AM and PS modules are provided to the readers.

II. Related Works

HSI super-resolution is a popular topic, which is receiving more and more attention. In particular, the combination of hyperspectral data with higher spatial resolution MSIs represents a fruitful scheme leading to satisfying results. Recent fusion or super-resolution approaches can be roughly categorized into two families: model-based approaches and deep learning-based methods.

Model-based approaches are classic solutions. Indeed, many works have already been published [1], [2], [6]–[42] for super-resolution and pansharpening problems. For instance, Li *et al.* [17] utilized the tensor theory [43], which shows a significant improvement in many image processing and computer vision tasks [29], [30], [44]–[46]. They consider the HR-HSI as a 3-D tensor exploiting the sparsity of the core tensor. It is decomposed into a three-mode dictionary of sparse core tensor multiplication. Following the baseline of tensor decomposition, Xu *et al.* [18] further employ the sparsity and the piecewise smoothness along with the width, height, and spectral mode of the GT HR-HSI. Both of them use the classical Tucker decomposition [47] to reformulate the fusion problem by estimating a sparse core tensor and coefficient dictionaries of the three modes. The super-resolution can then be regarded as an optimization problem, which has a satisfying solution under the well-known alternating direction multipliers minimization (ADMM) [48] framework with various constraints. However, some parameters in these tensor factorization or matrix factorization approaches are sensitive to the scene under test, *i.e.*,, different scenarios have their own unique optimal parameters setting.

Deep learning-based methods have recently showed exceptional performance in the field of image super-resolution, see [2]–[4], [49]–[65]. A powerful example is provided by the so-called PanNet developed in [53]. Yang *et al.* designed a new architecture training the deep-learning network with high-pass filtered details rather than original images. This is done in order to simultaneously preserve the spatial and spectral structures. Thanks to the use of high-pass filters, a greater generalization ability is observed. However, the PanNet roughly absorbs the features obtained from the MSI and panchromatic image by plain and straightforward ResNet blocks; some deeper and more abstract features of those images are ignored. Also, the spatial and spectral characteristics extracted by the ResNet blocks are not differentiated. Another instance of deep learning-based methods for solving the HSI super-resolution issue is provided in [3], where a model-based deep learning method is proposed. The method exhibits a great ability to preserve structures and details obtaining state-of-the-art results. Unlike other deep learning-based methods that mainly regard the image super-resolution issue as a simple regression problem, this approach is driven by the generation mechanism of the HSI and the MSI to build a novel fusion model. It adopts the low rankness knowledge along with the spectral mode of the HR-HSI under analysis. Instead of solving the model by traditional alternating iterative algorithms, the authors design a deep network learning the proximal operators and model parameters by exploiting CNNs. Nevertheless designing a universal low rankness of various scenarios is quite challenging. The network performance on different cases is related to their affinity to this prior knowledge, thus reducing the generalization ability. In addition, this approach takes up massive computing resources, *i.e.*,, a longer time for training and greater memory requirements for storage.

III. Proposed HSRnet

In this section, we introduce first the regularization-based model for the HSI super-resolution problem. Motivated by the above-mentioned model, we propose our network architecture that will be detailed in Section III-B.

A. Problem Formulation

Estimating the HR-HSI from LR-HSI and HR-MSI is an ill-posed inverse problem. Thus, prior knowledge is introduced exploiting regularization terms under the maximum *a posteriori* (MAP) framework. Those methods can be formulated as

$$\min_{\mathbf{X}} \ \mathcal{L} = \lambda_1 f_1(\mathbf{X}, \mathbf{Y}) + \lambda_2 f_2(\mathbf{X}, \mathbf{Z}) + R(\mathbf{X}) \tag{2}$$

where $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ are the mode-3 unfolding matrices of tensor HR-HSI, LR-HSI, and HR-MSI, respectively, which have been introduced in Section I. λ_1 and λ_2 represent two regularization parameters, f_1 and f_2 force the spatial and spectral consistency, respectively, and R stands for the regularization term depending on the prior knowledge. In general, f_1 and f_2 are defined based on the relations in (1), that is

$$\begin{aligned} f_1(\mathbf{X}, \mathbf{Y}) &= \|\mathbf{Y} - \mathbf{XBS}\|_F^2 \\ f_2(\mathbf{X}, \mathbf{Z}) &= \|\mathbf{Z} - \mathbf{RX}\|_F^2 \end{aligned} \tag{3}$$

where $\|\mathbf{X}\|_F = (\sum\sum x_{ij}^2)^{1/2}$ is the Frobenius norm. In particular, the regularization term R is crucial for regularization-based methods.

Deep learning can be viewed as an estimation problem of a function mapping input data with GT (labeled) data. In our case, starting from the input images (i.e., LR-HSI and HR-MSI), we can estimate the mapping function f by minimizing the following expression:

$$\min_{\Theta} \ \mathcal{L} = \| f_{\Theta}(\mathbf{YZ}) - \mathbf{X}\|_F^2 \tag{4}$$

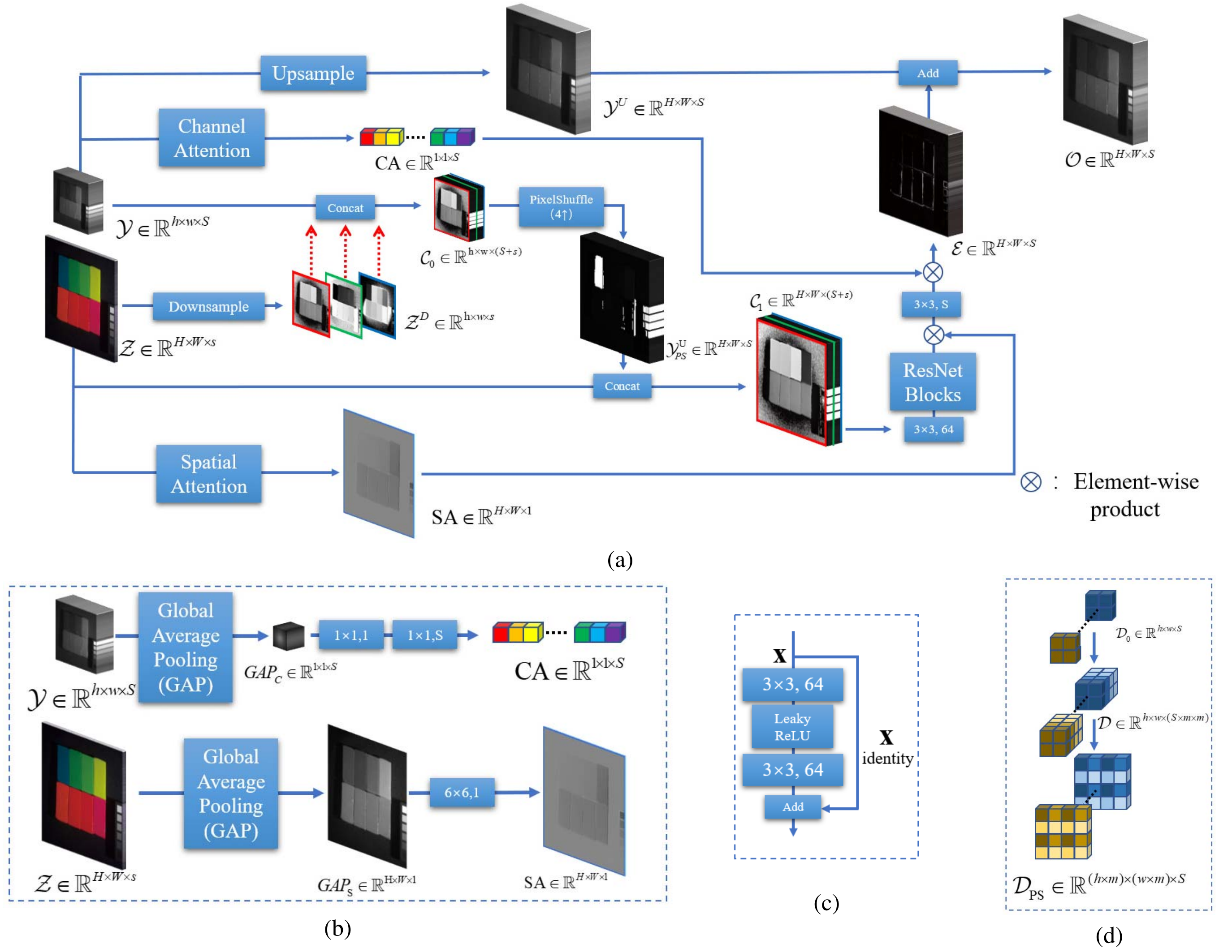
where $\mathbf{Y}$ and $\mathbf{Z}$ are the LR-HSI and the HR-MSI, respectively, and $\mathbf{X}$ is the reference (GT) HR-HSI. The mapping function f can be viewed as a deep CNN, thus Θ represents the parameters of the network. Besides, the prior knowledge can be viewed as being implicitly expressed by the learned parameters. In Section III-B, we will present the network architecture recasting the problem as in (4), where the function f is estimated thanks to several examples provided to the network during the training phase.

B. Network Architecture

Fig. 2 shows the proposed HSRnet for the HSI super-resolution problem. From the figure, it is easy to see that we decompose the network into two parts, such that the two parts can preserve the most crucial characteristics of an HSI, i.e., the spectral information and the spatial details.

1) Spectral Preservation: The LR-HSI $\mathcal{Y} \in \mathbb{R}^{h\times w\times S1}$ has the same spectral band number as the GT HR-HSI $\mathcal{X} \in \mathbb{R}^{H\times W\times S}$. Indeed, most of the spectral information

[1] We use three coordinates format to better represent the 3-D HSI, i.e., $h \times w \times S$.


Upsample
$\mathcal{Y}^U \in \mathbb{R}^{H\times W\times S}$
Add
$\mathcal{O} \in \mathbb{R}^{H\times W\times S}$
Channel Attention
$\text{CA} \in \mathbb{R}^{1\times1\times S}$
$\mathcal{Y} \in \mathbb{R}^{h\times w\times S}$
Concat
$\mathcal{C}_0 \in \mathbb{R}^{h\times w\times(S+s)}$
PixelShuffle (4↑)
$\mathcal{E} \in \mathbb{R}^{H\times W\times S}$
Downsample
$\mathcal{Z}^D \in \mathbb{R}^{h\times w\times s}$
$\mathcal{Z} \in \mathbb{R}^{H\times W\times s}$
$\mathcal{C}_1 \in \mathbb{R}^{H\times W\times(S+s)}$
3×3, S
$\mathcal{Y}^U_{PS} \in \mathbb{R}^{H\times W\times S}$
ResNet Blocks
Concat
3×3, 64
Spatial Attention
$\text{SA} \in \mathbb{R}^{H\times W\times 1}$
⊗ : Element-wise product
Global Average Pooling (GAP)
1×1,1
1×1,S
$GAP_C \in \mathbb{R}^{1\times1\times S}$
$\text{CA} \in \mathbb{R}^{1\times1\times S}$
$\mathcal{Y} \in \mathbb{R}^{h\times w\times S}$
Global Average Pooling (GAP)
6×6,1
$\mathcal{Z} \in \mathbb{R}^{H\times W\times s}$
$GAP_S \in \mathbb{R}^{H\times W\times 1}$
$\text{SA} \in \mathbb{R}^{H\times W\times 1}$
x
3×3, 64
Leaky ReLU
3×3, 64
Add
x identity
$\mathcal{D}_0 \in \mathbb{R}^{h\times w\times S}$
$\mathcal{D} \in \mathbb{R}^{h\times w\times(S\times m\times m)}$
$\mathcal{D}_{\text{PS}} \in \mathbb{R}^{(h\times m)\times(w\times m)\times S}$

(a)

(b)

(c)

(d)

Fig. 2. Flowchart of the proposed network architecture (HSRnet). (a) Architecture of our HSRnet. LR-HSI $\mathcal{Y}$ and HR-MSI $\mathcal{Z}$ are the two inputs, and the $\mathcal{O}$ is the final output. (b) Schematic of the CA and SA modules. $\text{GAP}_c \in \mathbb{R}^{1\times1\times S}$ is obtained by the global average pooling (GAP) of the two spatial dimensions, while $\text{GAP}_s \in \mathbb{R}^{H\times W\times s}$ is obtained by the GAP along the spectral dimension. (c) Diagram of one ResNet block with two layers and 64 kernels (size 3 × 3) for each layer and the activation function is replaced by the leaky ReLU. (d) Illustration of the PS for upsampling m times.

of the HR-HSI is contained in the LR-HSI (the remaining part is due to the spectral information of the high-resolution spatial details). In order to corroborate it, we plot the sampled spectral signatures obtained by the GT HR-HSI $\mathcal{X}$ and by the corresponding upsampled LR-HSI $\mathcal{Y}^U \in \mathbb{R}^{H\times W\times S}$ in Fig. 3. It is easy to be noted that the plots are very close to each other, indicating that $\mathcal{Y}^U$ holds most of the spectral content of $\mathcal{X}$. Therefore, to guarantee spectral preservation, we simply upsample $\mathcal{Y}$ getting $\mathcal{Y}^U$ by bicubic interpolation [as shown in the top part of Fig. 2(a)].

Admittedly, $\mathcal{Y}^U$ is able to preserve most of the spectral information, but some spatial details are lost (which can retain part of the spectral information). Instead, the proposed HSRnet can learn the HR-HSI's spectral information, even preserving the spatial counterpart. Note that the CA weights are obtained by the original LR-HSI $\mathcal{Y}$, and the global average pooling (GAP) is along the spatial dimensions. It forces the CA module to focus on the spectral relation. Thanks to this module, the network can obtain the different missing spectral information of each band. As a result, the final outcome of the proposed HSRnet clearly shows an almost perfect spectral preservation (see Fig. 3).

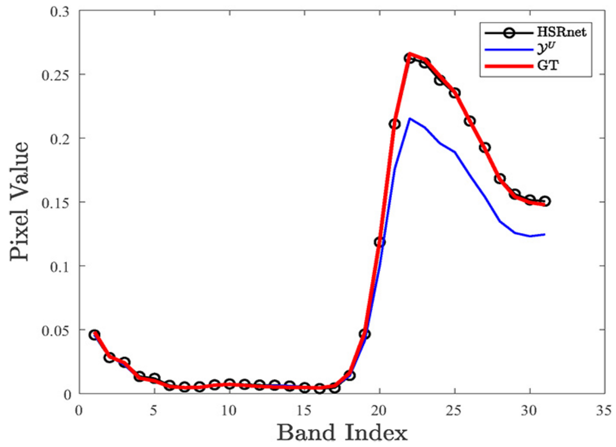
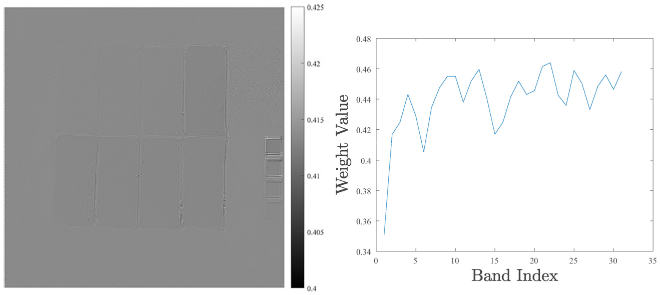


Fig. 3. Sampled spectral signatures for the *clay* at pixel (175, 400) as obtained by the (GT) HR-HSI, the upsampled LR-HSI $\mathcal{Y}^U$, and the estimated version of the high resolution HSI exploiting the proposed HSRnet.



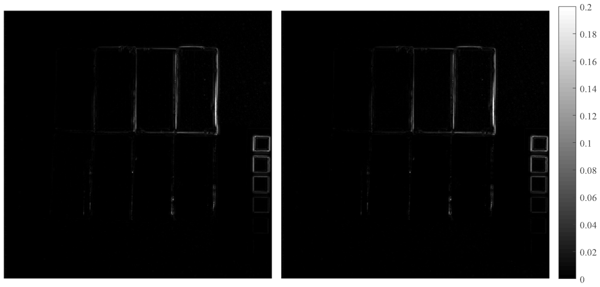
(a) (b)

Fig. 4. (a) Spatial weight map learned by the SA module of *clay*. (b) Channel weight learned by the CA module.

2) Spatial Preservation: Since the HR-MSI $\mathcal{Z} \in \mathbb{R}^{H\times W\times s}$ contains high spatial resolution information, we aim to use $\mathcal{Z}$ to extract spatial details injecting them into the final hyperspectral super-resolution image. Moreover, $\mathcal{Y}$ still contains some spatial details. Thus we also consider employing $\mathcal{Y}$ to extract them. However, we do not simply concatenate $\mathcal{Z}$ and $\mathcal{Y}$ together taking them into the network. We calculate first the spatial information at the LR-HSI scale. In particular, we add other details at the same scale by extracting them from the HR-MSI $\mathcal{Z}$. The downsampled HR-MSI $\mathcal{Z}^D$ is obtained by convoluting HR-MSI $\mathcal{Z}$ with a learnable kernel of size 6 × 6 and setting stride as downsampling factor of 4 [see Fig. 2(a) again]. Finally, we concatenate this information, i.e., $\mathcal{Y}$ and $\mathcal{Z}^D$, to get $\mathcal{C}_0 \in \mathbb{R}^{h\times w\times(S+s)}$.

In order to acquire adequate information, the spatial details at the HR-MSI scale are extracted, which can be concatenated with $\mathcal{Y}^U_{\text{PS}}$. While the $\mathcal{Y}^U_{\text{PS}}$ is properly convoluted and upsampled to the HR-MSI scale by the PS [66]. It designs convolution filters for every single feature map capturing more details during the upsampling process [see Fig. 2(d)]. Thus, $\mathcal{C}_1 \in \mathbb{R}^{H\times W\times(S+s)}$ indicates the concatenation of the data at two different scales (the LR-HSI one and the HR-MRI one). This represents the input of the ResNet implementing the well-known concept of multiresolution analysis often considered in previously developed researches (*e.g.* [67]–[71]) either by designing diverse kernel sizes for convolution [67], [68] or extracting different spatial resolutions by filtering input data [69]–[71]. Moreover, the concatenation operator is about adding the multispectral bands with a high spatial resolution (three bands, RGB image) into the hyperspectral bands (as shown in Fig. 2). In this work, the red, the green, and the blue slices of $\mathcal{Z}^D$ and $\mathcal{Z}$ are inserted as the head, the middle, and the tail frontal slices to complement the spectral information of the HSI.

It is worth to be remarked that we do not expect the output details are directly learned by the input LR-HSI $\mathcal{Y}$ and HR-MSI $\mathcal{Z}$, so we bring the AMs [72], [73] in the architecture. They have the ability to distinguish more noteworthy information from the raw data, which is consistent with what we are trying to achieve, i.e., the additional details. The SA module helps the network to focus on noteworthy areas. See Fig. 4(a), the edges with higher weights mean that the network needs to pay more attention to them while the background parts with lower weights represent lesser importance. However, the SA $\in \mathbb{R}^{H\times W\times 1}$ just expresses a single attention map in space. The CA module is introduced in the network to make up for this deficiency. Different channel weights can lead to different interest levels for each band. Fig. 4(b) shows the channel weights of the image *clay*. HR-MSI $\mathcal{Z} \in \mathbb{R}^{H\times W\times s}$ carries more accurate and abundant spatial information of the scenario, thus we expect to learn the SA map from the HR-MSI, while the LR-HSI $\mathcal{Y} \in \mathbb{R}^{h\times w\times S}$ keeps the spectral signatures, so the CA weights are obtained by the LR-HSI [see Fig. 2(b)].



(a) (b)

Fig. 5. Residual maps at the 15th band. (a) $\mathcal{E} = \mathcal{O} - \mathcal{Y}^U$ and (b) $\mathcal{E}_{\text{gt}} = \mathcal{X} - \mathcal{Y}^U$.

Fig. 5 shows a comparison between $\mathcal{E}$ and $\mathcal{E}_{\text{gt}}$. From the figure, it is clear that $\mathcal{E}$ (i.e., the details extracted by the proposed approach) and $\mathcal{E}_{\text{gt}}$ (i.e., the details extracted by using the reference image) are very close to each other validating the effectiveness of the proposed network design.

3) Loss Function: After obtaining the spectral preserved $\mathcal{Y}^U$ image and the spatial preserved $\mathcal{E}$ image from the ResNet fed by the image cube $\mathcal{C}_1$, we subsequently add the two outputs together to get the outcome. Thus, the loss function exploited during the training phase to drive the estimation of the function mapping in (4) can be defined as

$$\min_{\Theta}\ \mathcal{L} = \|f_\Theta(\mathcal{Y}\mathcal{Z}) + \mathcal{Y}^U - \mathcal{X}\|_F^2 \tag{5}$$

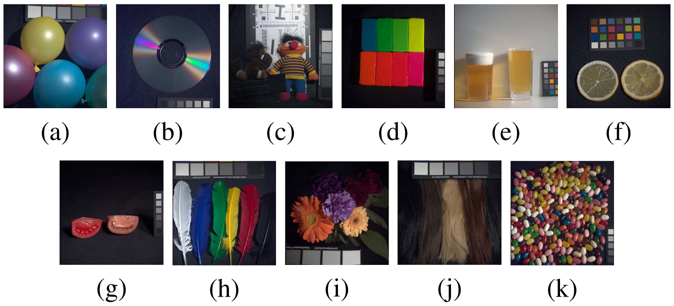
where $f_\Theta(\cdot)$ is the mapping function that has as input the details at the two different scales used to estimate the spatial preserved image $\mathcal{E}$ and the upsampled LR-HSI $\mathcal{Y}^U$. The loss function imposes the similarity between the network output $f_\Theta(\mathcal{Y}, \mathcal{Z}) + \mathcal{Y}^U$ and the reference (GT) $\mathcal{X}$ image.

C. Network Training

1) Training Data: In the work, we mainly use the CAVE dataset [74] for training the network. It contains 32 HSIs with size 512 × 512 and 31 spectral bands. Additionally, each HSI also has a corresponding RGB image with size 512 × 512 and three spectral bands (i.e., the HR-MSI image). We selected 20 images for training the network, and the other 11 images to be considered for testing,[2] as done for the MHFnet in [3]. The CAVE test images are shown in Fig. 6.

2) Data Simulation: We extracted 3920 overlapped patches with a size of 64 × 64 × 31 from the 20 images of the CAVE dataset used as GT, thus forming the HR-HSI patches. Accordingly, the LR-HSI patches are generated starting from the HR-HSI by applying a Gaussian blur with kernel size

[2] One image, i.e., “Watercolors," is discarded as it is unavailable for use.


(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j) (k)

Fig. 6. 11 testing images from the CAVE dataset. (a) *balloons*, (b) *cd*, (c) *chart and stuffed toy*, (d) *clay*, (e) *fake and real beers*, (f) *fake and real lemon slices*, (g) *fake and real tomatoes*, (h) *feathers*, (i) *flowers*, (j) *hairs*, and (k) *jelly beans*.

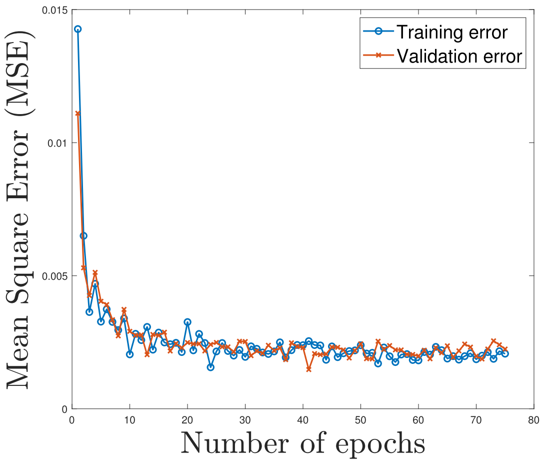


Fig. 7. Training and validation errors for the proposed HSRnet.

equal to 3 × 3 and standard deviation equal to 0.5 and then downsampling the blurred patches to the size of 16 × 16, i.e., with a downsampling factor of 4. Moreover, the spectral response function of the Nikon D700 camera (the same camera spectral response function in [2], [3], [18], [29], and [30]) is used to generate the RGB patches. Thus, 3920 patches of size of 64 × 64 × 3 are available to represent the HR-MSI. Following these indications, the patches for the training phase are the 80% of the whole dataset and the rest (i.e., the 20%) is used for the validation.

3) Training Platform and Parameters Setting: The proposed network is trained on Python 3.7.4 with Tensorflow 1.14.0 and Windows operating system with NVIDIA GPU GeForce GTX 2080Ti. We use Adam optimizer with a learning rate equal to $1e-4$ in order to minimize the loss function (5) by 75 epochs, and each one has 2000 iterations. 32 batches are trained at the same time in one iteration. The ResNet block in our network architecture is crucial. Indeed, we use six ResNet blocks. (Each one with two layers and 64 kernels of size 3×3 for each layer. See Fig. 2.) Fig. 7 shows the training and validation errors of the proposed HSRnet confirming the convergence of the proposed CNN using the above-mentioned parameters setting.

IV. Experimental Results

In this section, we compare the proposed HSRnet with several state-of-the-art methods for the hyperspectral super-resolution problem. In particular, the benchmark consists of the CSTF method[3] [17], the FUSE approach[4] [75], the GLP-HS method[5] [15], the UTV technique[6] [18], the CNN-FUS approach[7] [2], the MHFnet[8] [3], and the proposed HSRnet approach. For a fair comparison, the MHFnet is trained on the same training data as the proposed approach. Furthermore, the batch size and the training iterations of the MHFnet are set to 32 and 150 000, respectively. Three widely used benchmark datasets, i.e., CAVE database[9] [74], Harvard database[10] [76] and Chikusei database[11] [77], are selected.

TABLE I

Average QIs and Related Standard Deviations of the Results on 50 Patches Extracted From the Testing Images on the CAVE Dataset. The Best Values Are Highlighted in Boldface

Method	PSNR	SAM	ERGAS	SSIM
FUSE	33.87±4.8	5.49±3.0	3.84±3.0	0.953±0.04
GLP-HS	32.92±4.4	4.62±2.7	3.96±2.8	0.955±0.04
CSTF	39.40±4.3	7.83±7.4	2.49±1.9	0.970±0.04
UTV	39.34±4.6	6.75±5.0	2.45±1.7	0.968±0.05
CNN-FUS	38.84±4.4	5.97±5.5	2.67±2.6	0.976±0.02
MHFnet	40.57±4.8	3.77±3.3	2.04±2.3	0.987±0.01
HSRnet	**42.38**±4.3	**2.12**±1.3	**1.38**±1.2	**0.993**±0.00
Best value	+∞	0	0	1

TABLE II

Average QIs and Related Standard Deviations of the Results on 11 Testing Images on the CAVE Datasets. The Best Values Are Highlighted in Boldface

Method	PSNR	SAM	ERGAS	SSIM
FUSE	39.72±3.5	5.83±2.0	4.18±3.1	0.975±0.02
GLP-HS	37.81±3.1	5.36±1.8	4.66±2.7	0.972±0.01
CSTF	42.14±3.0	9.92±4.1	3.08±1.6	0.964±0.03
UTV	42.31±3.2	11.71±4.4	3.50±1.7	0.961±0.02
CNN-FUS	42.66±3.5	6.44±2.3	2.95±2.2	0.982±0.01
MHFnet	46.32±2.7	4.33±1.8	1.74±1.2	0.992±0.00
HSRnet	**47.82**±2.7	**2.66**±0.9	**1.34**±0.8	**0.995**±0.00
Best value	+∞	0	0	1

For quantitative evaluation, we adopt four quality indexes (QIs), i.e., the peak signal-to-noise ratio (PSNR), the spectral angle mapper (SAM) [78], the erreur relative globale adimensionnelle de synthèse (ERGAS) [79], and the structure similarity (SSIM) [80]. The SAM measures the average angle between the spectral vectors of the target and of the reference image. Instead, the ERGAS represents the fidelity of the image based on the weighted sum of mean squared errors. The ideal value in both cases is zero. The lower the index, the better the quality. Finally, PSNR and SSIM are widely used to evaluate the similarity between the target and the reference image. The higher the index, the better the quality. The ideal value for SSIM is one.

[3] https://github.com/renweidian/CSTF

[4] http://wei.perso.enseeiht.fr/publications.html

[5] http://openremotesensing.net/knowledgebase/hyperspectral-and-multispectral-data-fusion/

[6] https://liangjiandeng.github.io/

[7] https://github.com/renweidian/CNN-FUS

[8] https://github.com/XieQi2015/MHF-net

[9] http://www.cs.columbia.edu/CAVE/databases/multispectral/

[10] http://vision.seas.harvard.edu/hyperspec/download.html

[11] http://naotoyokoya.com/Download.html

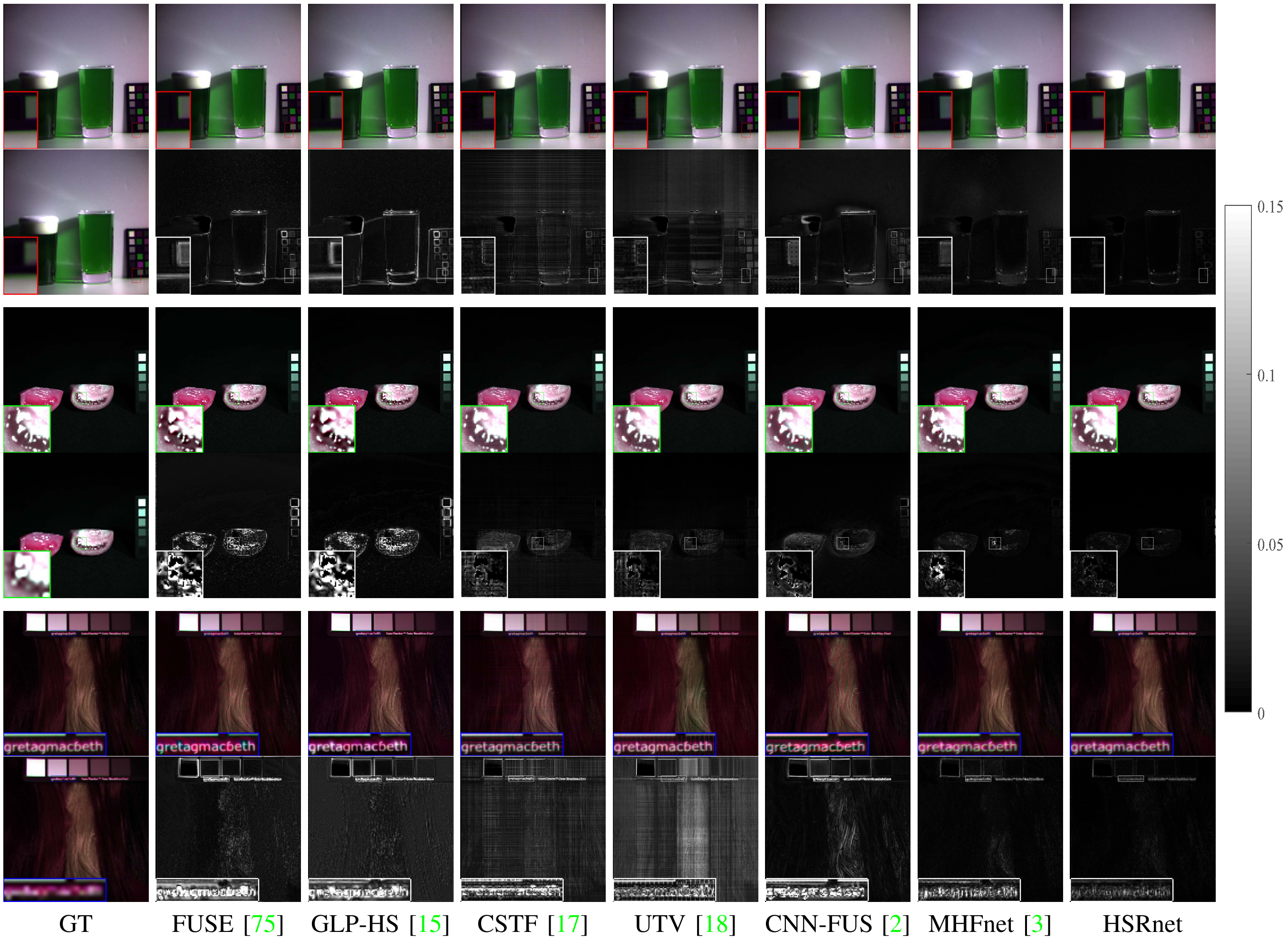


Fig. 8. First column: the true pseudocolor images from the original CAVE dataset and the corresponding LR-HSI images of *fake and real beers* (R-3, G-13, B-2) (first and second rows), *fake and real tomatoes* (R-22, G-19, B-20) (third and fourth rows), and *hairs* (R-1, G-9, B-2) (fifth and sixth rows). Second–eighth columns: the true pseudocolor fused products and the corresponding residuals for the different methods in the benchmark pointing out some close-ups to facilitate the visual analysis.

A. Results on CAVE Dataset

In order to point out the effectiveness of all the methods on different kinds of scenarios and local areas, we divide first the remaining 11 testing images on the CAVE dataset into small patches of size 128 × 128. Then, 50 patches are randomly selected. We exhibit the average QIs and corresponding standard deviations of the results for the different methods on these patches in Table I. From Table I, we can find that the proposed HSRnet significantly outperforms the compared methods. In particular, the SAM value of our method is much lower than that of the compared approaches (about the half with respect to the best compared method). This is in agreement with our previously developed analysis, namely that the proposed HSRnet is able to preserve the spectral features of the acquired scene.

Afterward, we conduct the experiments on the whole 11 testing images. Table II presents the average QIs on the 11 testing images. To ease the readers' burden, we only show the results on *fake and real beers*, *fake and real tomatoes*, and *hairs*. Table III lists the specific QIs of the results on these three images for the different methods. The proposed method outperforms the compared approaches. Furthermore, the running time of the HSRnet is also the lowest one. In Fig. 8, we display the pseudocolor images of the fusion results and the corresponding error maps on three images. From the error maps in Fig. 8, it can be observed that the proposed HSRnet approach has a better reconstruction of the high-resolution details with respect to the compared methods, thus clearly reducing the errors in the corresponding error maps. In particular, CSTF and UTV have strict requirements on the parameters for different images. It is easy to observe there are prominent stripes in the *fake and real beers* and *hairs* from the visual analysis, while they work pretty well on the *fake and real tomatoes* with the same parameter setting. Spectral fidelity is of crucial importance when the fusion of HSIs is considered. In order to illustrate the spectral reconstruction provided by the different methods, we plot the spectral vectors for three exemplary cases (see Fig. 9). It is worth to be remarked that the spectral vectors estimated by our method and the GT ones are very close to each other.

B. Results on Harvard Dataset

The Harvard dataset is a public dataset that has 77 HSIs of indoor and outdoor scenes including different kinds

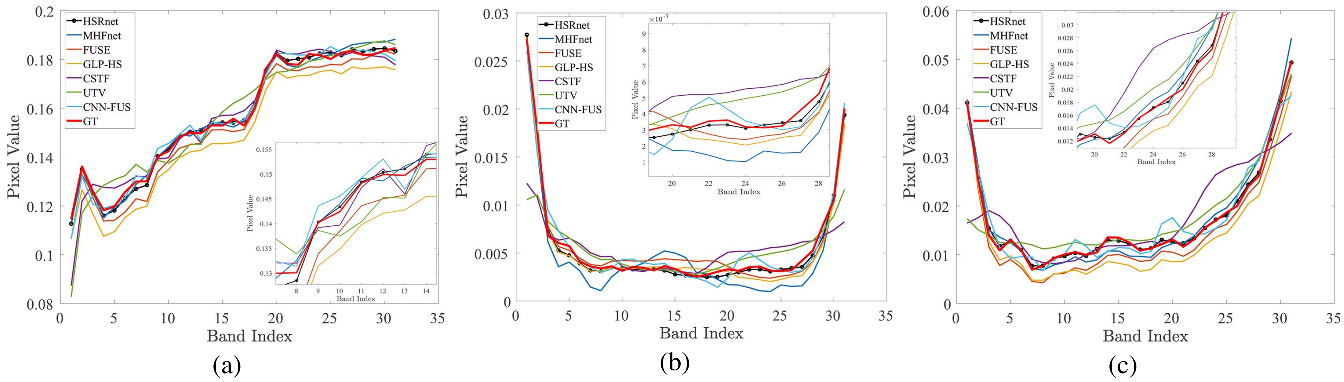


Fig. 9. Selected spectral vectors for the outcomes coming from the different fusion methods and the GT. The indications of the specific dataset and the location of the pixel under analysis are also provided. (a) Spectral vectors in *fake and real beers* located at (19 272). (b) Spectral vectors in *fake and real tomatoes* located at (156 128). (c) Spectral vectors in *hairs* located at (195 165).

TABLE III

QIs of the Results by Different Methods and the Running Times on *Fake and Real Beers*, *Fake and Real Tomatoes*, and *Hairs* on the CAVE Dataset. G Indicates That the Method Is Running on the GPU Device, While C Denotes the Use of the CPU. The Best Values Are Highlighted in Boldface

fake and real beers (512 × 512 × 31)							
Method	FUSE	GLPHS	CSTF	UTV	CNN-FUS	MHFnet	HSRnet
PSNR	42.61	40.72	44.14	42.48	44.33	47.02	**49.38**
SAM	1.73	1.32	1.96	2.12	1.88	1.33	**0.81**
ERGAS	0.79	0.98	1.19	1.16	0.85	0.42	**0.30**
SSIM	0.985	0.982	0.972	0.971	0.988	0.993	**0.994**
fake and real tomatoes (512 × 512 × 31)							
Method	FUSE	GLPHS	CSTF	UTV	CNN-FUS	MHFnet	HSRnet
PSNR	40.65	40.46	45.96	47.29	47.15	49.33	**50.23**
SAM	6.40	5.34	12.99	15.66	6.88	6.53	**2.87**
ERGAS	12.71	11.89	4.84	4.89	4.53	3.25	**2.75**
SSIM	0.981	0.984	0.977	0.978	0.991	0.992	**0.997**
hairs (512 × 512 × 31)							
Method	FUSE	GLPHS	CSTF	UTV	CNN-FUS	MHFnet	HSRnet
PSNR	39.99	37.52	43.57	44.84	43.77	46.53	**47.78**
SAM	5.59	6.25	9.55	10.68	8.40	4.87	**3.18**
ERGAS	2.96	3.86	2.64	4.13	2.07	1.29	**1.07**
SSIM	0.988	0.979	0.986	0.983	0.987	0.994	**0.996**
Average time(s)	1.9(C)	4.6(C)	31.8(C)	487.7(C)	6.7(C+G)	0.2(G)	**0.1**(G)

TABLE IV

Average QIs and Related Standard Deviations of the Results for Ten Testing Images on the Harvard Dataset. The Best Values Are Highlighted in Boldface

Method	PSNR	SAM	ERGAS	SSIM
FUSE	42.06±2.9	3.23±0.9	3.14±1.5	0.977±0.01
GLP-HS	40.14±3.2	3.52±1.0	3.74±1.4	0.966±0.01
CSTF	42.97±3.5	3.30±1.2	**2.43**±1.1	0.972±0.02
UTV	42.64±3.3	3.78±1.0	2.68±1.1	0.969±0.02
CNN-FUS	43.61±4.7	3.32±1.2	2.78±1.6	0.978±0.02
MHFnet	40.37±3.7	4.64±1.8	24.17±46.7	0.966±0.01
HSRnet	**44.29**±3.0	**2.66**±0.7	2.45±0.8	**0.984**±0.01
Best value	+∞	0	0	1

TABLE V

QIs of the Results for the Different Methods and the Running Times on *Window*, *Tree*, and *Backpack* for the Harvard Dataset. G Indicates That the Method Is Running on the GPU Device, While C Denotes the Use of the CPU. The Best Values Are Highlighted in Boldface

window (1000 × 1000 × 31)							
Method	FUSE	GLPHS	CSTF	UTV	CNN-FUS	MHFnet	HSRnet
PSNR	39.83	36.20	40.79	40.19	41.86	41.66	**43.07**
SAM	2.61	2.92	2.64	3.44	2.33	2.62	**2.10**
ERGAS	1.90	2.97	**1.62**	2.02	1.67	2.87	1.64
SSIM	0.977	0.959	0.974	0.965	0.984	0.981	**0.985**
tree (1000 × 1000 × 31)							
Method	FUSE	GLPHS	CSTF	UTV	CNN-FUS	MHFnet	HSRnet
PSNR	36.69	34.95	36.88	36.36	35.86	36.65	**38.96**
SAM	3.94	4.11	4.80	5.40	5.41	3.89	**2.77**
ERGAS	3.07	3.85	2.90	3.05	3.52	5.85	**2.66**
SSIM	0.953	0.935	0.937	0.930	0.935	0.957	**0.971**
backpack (1000 × 1000 × 31)							
Method	FUSE	GLPHS	CSTF	UTV	CNN-FUS	MHFnet	HSRnet
PSNR	42.75	41.10	39.52	39.70	44.45	40.92	**45.02**
SAM	2.88	3.01	5.31	5.88	3.06	4.08	**2.23**
ERGAS	4.94	4.99	4.41	4.46	3.12	10.22	**2.80**
SSIM	0.976	0.967	0.923	0.919	0.980	0.967	**0.986**
Average time(s)	12.0(C)	30.7(C)	25.7(C)	446.6(C)	25.1(C+G)	0.9(G)	**0.3**(G)

of objects and buildings. Every HSI has a spatial size of 1392 × 1040 with 31 spectral bands, and the spectral bands are acquired at an interval of 10 nm in the range of 420–720 nm. We select the top left part of the image (1000 × 1000), then ten images are randomly selected for testing.

As in the previous settings, the original data are regarded as the GT HR-HSI. The LR-HSI data are simulated as in Section III-C. The HR-MSI is also obtained by applying the spectral response of the Nikon D700 camera as in Section III-C.

We would like to remark that both our method and the MHFnet are trained on the CAVE dataset, and we directly test them on the Harvard dataset without any retraining or fine-tuning. Thus, the performance on the Harvard dataset of these two methods could reflect their generalization abilities.

Table IV records the average QIs and the corresponding standard deviations for the different methods using the ten testing images. Note that the MHFnet is unstable; the ERGAS value shows the worst result compared with other methods because of its poor generalization ability. Table V gives the QIs and the running times for three specific images of the Harvard dataset. The proposed method ranks first with the lowest running time. Finally, considering the details in the pseudocolor images in Fig. 10, we can see that the results of our method get the highest qualitative performance, thus obtaining very dark error maps (i.e., with errors that tend to zero everywhere).

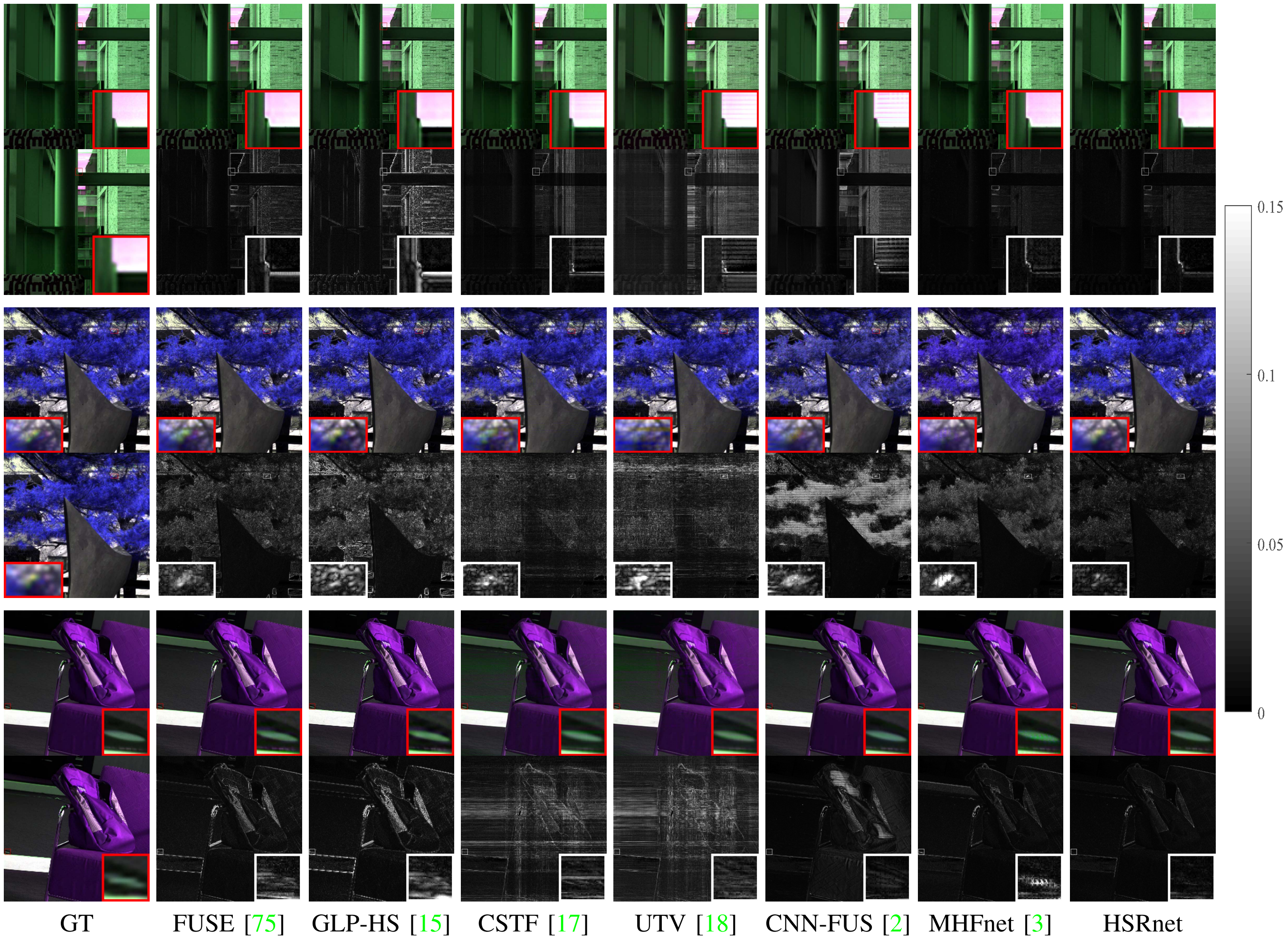


Fig. 10. First column: the true pseudocolor images from the original Harvard dataset and the corresponding LR-HSI images of *window* (R-29, G-22, B-28) (first and second rows), *trees* (R-27, G-25, B-30) (third and fourth rows), and *backpack* (R-29, G-22, B-31) (fifth and sixth rows). Second to Eighth columns: the true pseudocolor fused products and the corresponding residuals for the different methods in the benchmark pointing out some close-ups to facilitate the visual analysis.

C. Results on Chikusei Dataset

In order to present our HSRnet's performance on remote sensed HSIs, we conduct an experiment on Chikusei dataset, which is taken over agricultural and urban areas in Chikusei, Ibaraki, Japan. It consists of 2517 × 2335 pixels and has 128 bands in the spectral range from 363 to 1018 nm. We regard the original data as the GT HR-HSI and simulate the LR-HSI in the same way as the previous experiments. As for the HR-MSI, the corresponding RGB image is obtained by Canon EOS 5D Mark II together with the HR-HSI. Afterward, we select the top-left area with the spatial size 1000 × 2200 for training and crop 64 × 64 overlapped patches from the training part as the GT HR-HSI patches. Moreover, the input HR-MSI and LR-HSI patches are of size 64 × 64 × 3 and 16 × 16 × 128, respectively. As for the testing data, we extract six nonoverlap 680 × 680 × 128 area from the remaining part of the Chikusei dataset.

Table VI shows the average QIs and corresponding standard deviations for the testing images on all methods. It is clear that our HSRnet outperforms the other comparing methods on each metric. We also display the pseudocolor images of the obtained outcomes and the corresponding error maps for a visual comparison in Fig. 11. Obviously, the fused results of our HSRnet are the most satisfactory and the error maps are the darkest. Table VII gives the corresponding QIs and the running times for three selected areas. Since the spectral channels of the Chikusei dataset are considerable, highly increasing the computational cost of the MHFnet, we use the PCA prior in [52] in the training process of the MHFnet to address the issue as Xie *et al.* [3] did in the original work. Note that the MHFnet gets the lowest running time on the three testing images as our HSRnet.

TABLE VI

AVERAGE QIs AND RELATED STANDARD DEVIATIONS OF THE RESULTS FOR SIX TESTING IMAGES ON THE CHIKUSEI DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLDFACE

Method	PSNR	SAM	ERGAS	SSIM
FUSE	27.76±1.5	4.80±1.2	7.22±0.5	0.882±0.02
GLP-HS	31.60±1.3	3.29±0.3	5.69±0.3	0.919±0.01
CSTF	30.36±0.9	4.58±0.5	5.91±0.6	0.824±0.02
UTV	28.06±1.3	4.62±0.3	8.30±0.5	0.874±0.01
CNN-FUS	31.83±1.7	4.76±0.9	5.25±0.9	0.918±0.01
MHFnet	33.19±1.0	3.18±0.4	6.24±0.4	0.927±0.01
HSRnet	**36.95**±1.1	**2.08**±0.2	**3.60**±0.3	**0.952**±0.01
Best value	$+\infty$	0	0	1

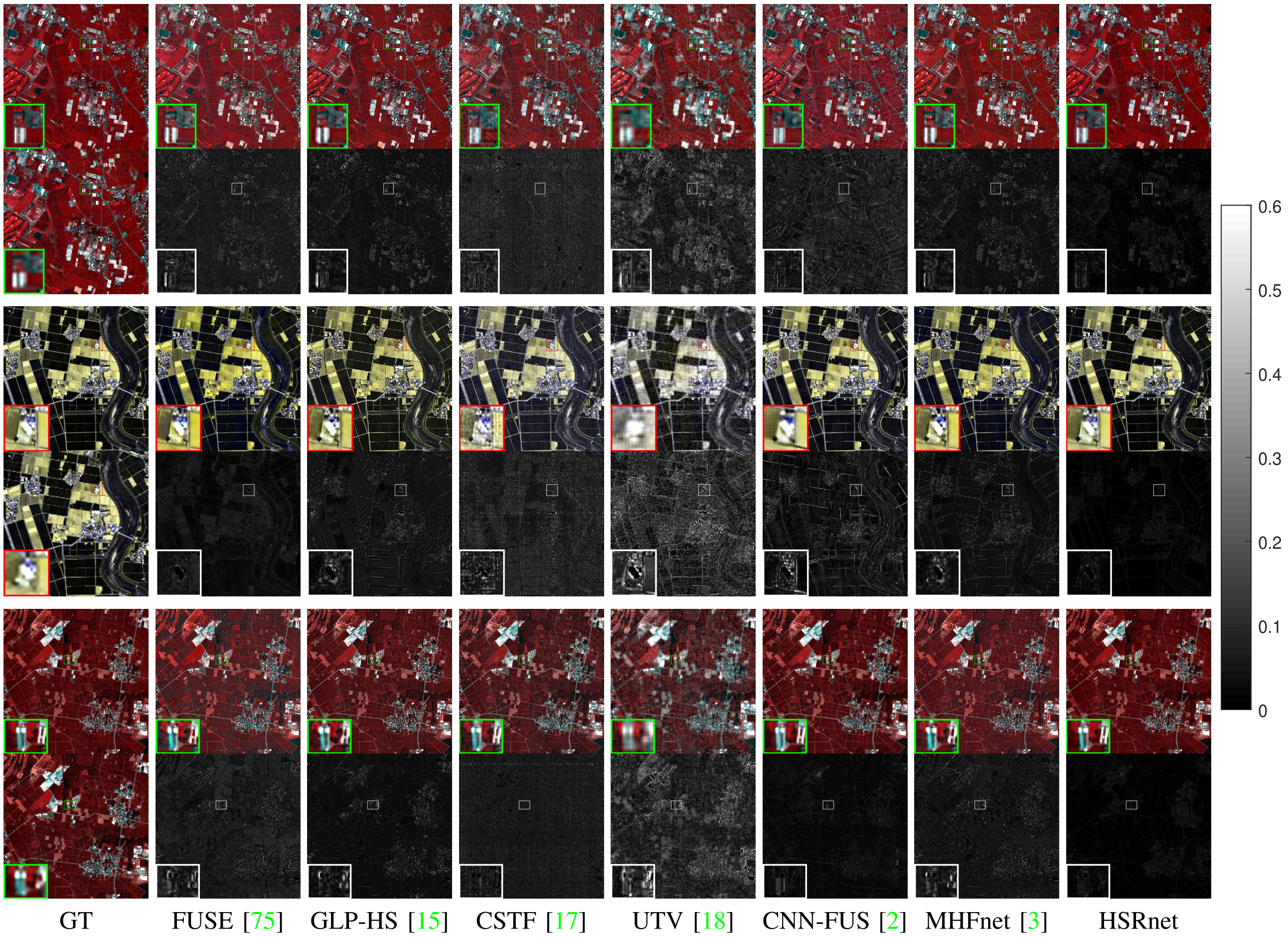


Fig. 11. First column: true pseudocolor images from the Chikusei dataset and the corresponding LR-HSI images of *area 1* (R-69, G-9, B-8) (first and second rows), *area 2* (R-64, G-58, B-16) (third and fourth rows), and *area 3* (R-67, G-15, B-13) (fifth and sixth rows). Second to eighth columns: true pseudocolor fused products and the corresponding residuals for the different methods in the benchmark pointing out some close-ups to facilitate the visual analysis.

D. Ablation Study

1) AM : In order to investigate the effects of the use of AMs, we compare our HSRnet with its variant that is similar to the original HSRnet but without any AM. The network is trained on the same training data of the HSRnet with the same training settings. Table VIII presents the average QIs of these two networks on the 11 testing images from the CAVE dataset and the ten testing images from the Harvard dataset. As we can see from the CAVE dataset results in Table VIII, the mean values and standard deviations of the proposed network are much better than that of the one without AM. These modules do help the network to focus on more significant features. In the Harvard dataset, however, the application of AM has slightly weakened the generalization capability of the network. We believe that it is still acceptable due to the improvement in the CAVE testing images brought by the AM.

2) PS Module: In previous research studies for super-resolution problems of images or videos, low-resolution input is usually upscaled by the bicubic interpolation or simple transposed convolution. In comparison, the PS module helps the network structure to gain useful information of every single feature map. The data cube $\mathcal{C}_0$ in Fig. 2 is fed to the PS, and then $\mathcal{Y}_{\text{PS}}^{U}$ is yielded.

TABLE VII

QIs of the Results for the Different Methods and the Running Times on Areas 1–3 for the Chikusei Dataset. G Indicates That the Method Is Running on the GPU Device, While C Denotes the Use of the CPU. The Best Values Are Highlighted in Boldface

			area 1 (680 × 680 × 128)				
Method	FUSE	GLPHS	CSTF	UTV	CNN-FUS	MHFnet	HSRnet
PSNR	26.08	30.46	30.17	27.16	30.96	32.22	**35.43**
SAM	4.85	3.31	4.33	4.59	3.97	3.10	**2.19**
ERGAS	8.10	6.38	6.14	8.95	5.19	7.08	**4.14**
SSIM	0.855	0.902	0.817	0.856	0.902	0.910	**0.941**
			area 2 (680 × 680 × 128)				
Method	FUSE	GLPHS	CSTF	UTV	CNN-FUS	MHFnet	HSRnet
PSNR	29.91	33.99	32.33	30.68	35.37	34.52	**38.98**
SAM	4.42	3.02	5.42	4.22	4.12	2.96	**1.78**
ERGAS	6.87	5.29	7.00	7.41	5.35	5.69	**3.34**
SSIM	0.908	0.934	0.794	0.894	0.926	0.942	**0.965**
			area 3 (680 × 680 × 128)				
Method	FUSE	GLPHS	CSTF	UTV	CNN-FUS	MHFnet	HSRnet
PSNR	29.34	32.77	30.43	28.47	30.39	33.38	**37.38**
SAM	4.08	3.12	3.83	4.56	6.55	2.98	**2.02**
ERGAS	7.47	5.76	5.63	8.86	7.02	6.38	**3.61**
SSIM	0.887	0.913	0.826	0.863	0.907	0.920	**0.948**
Average time(s)	4.4(C)	43.4(C)	25.8(C)	504.8(C)	10.2(C+G)	**0.3**(G)	**0.3**(G)

To prove the strength of this module, we compare our original HSRnet and the simpler architecture that only uses

TABLE VIII

AVERAGE QIs AND RELATED STANDARD DEVIATIONS OF THE RESULTS ON THE CAVE AND THE HARVARD DATASETS USING THE PROPOSED METHOD WITH AND WITHOUT THE AMs. THE BEST VALUES ARE HIGHLIGHTED IN BOLDFACE

		CAVE		
Method	PSNR	SAM	ERGAS	SSIM
w/o AM	47.74±2.7	2.70±0.9	1.37±0.8	**0.995**±0.00
HSRnet	**47.82**±2.7	**2.66**±0.9	**1.34**±0.8	**0.995**±0.00
		Harvard		
Method	PSNR	SAM	ERGAS	SSIM
w/o AM	**44.31**±3.0	**2.64**±0.7	**2.34**±0.8	**0.984**±0.01
HSRnet	44.29±3.0	2.66±0.7	2.45±0.8	**0.984**±0.01

TABLE IX

AVERAGE QIs AND RELATED STANDARD DEVIATIONS OF THE RESULTS ON THE CAVE AND THE HARVARD DATASETS USING THE PROPOSED METHOD WITH AND WITHOUT THE PS. THE BEST VALUES ARE HIGHLIGHTED IN BOLDFACE

		CAVE		
Method	PSNR	SAM	ERGAS	SSIM
w/o PS	47.816±2.7	**2.66**±0.9	1.35±0.8	**0.995**±0.00
HSRnet	**47.824**±2.7	**2.66**±0.9	**1.34**±0.8	**0.995**±0.00
		Harvard		
Method	PSNR	SAM	ERGAS	SSIM
w/o PS	44.069±3.2	**2.66**±0.7	**2.39**±0.8	**0.984**±0.01
HSRnet	**44.285**±3.0	**2.66**±0.7	2.45±0.8	**0.984**±0.01

the transposed convolution. The results of the two compared approaches are reported in Table IX. The QI values show the necessity of the PS module in our HSRnet. It is relevant to the improvement in performance measured by PSNR, especially on Harvard testing images.

3) Activation Function Selection: The activation function plays a vital role in deep learning researches. The ReLU [81] is the most common one. A neuron only has an output when the input is greater than 0. Otherwise, this output will be 0. Thus, neural networks exploiting the ReLU tend to be very sparse. Only a part of neurons will be involved in the computation. The sparsity brought by the ReLU is similar to how the human brain neurons activate, and it avoids gradient explosion and vanishing gradient problems. However, the ReLU cannot work well in our network; it kills each neuron's negative inputs. While the $\mathcal{Y}^U$ is upsampled by the nearest interpolation, the detail injection $\mathcal{E}$ learned by the network will inevitably contain positive and negative parts.

Therefore, the network will behave inadequately in absorbing negative inputs if we select the normal ReLU function. Thus, we integrate the leaky ReLU as our main activation function to collect those abundant details, that is,

$$f(x) = \max(0.2x, x). \tag{6}$$

This activation function keeps a part of these negative inputs and still maintains the nonlinear mapping. The results of different activation functions are listed in Table X. From the table, it is clear that those negative information delivered by leaky ReLU deeply improves the performance of the detail reconstruction phase.

4) GAP: To show the difference between the GAP and the global max pooling (GMP) in AM, the network replacing the GAP with GMP is retrained. We conducted experiments on 11 CAVE testing images and ten Harvard testing images.

TABLE X

AVERAGE QIs AND RELATED STANDARD DEVIATIONS OF THE RESULTS ON THE CAVE AND THE HARVARD DATASETS USING THE PROPOSED METHOD WITH DIFFERENT ACTIVATION FUNCTION. THE BEST VALUES ARE HIGHLIGHTED IN BOLDFACE

		CAVE		
Method	PSNR	SAM	ERGAS	SSIM
with ReLU	47.18±2.6	2.67±0.8	1.46±0.8	**0.995**±0.00
HSRnet	**47.82**±2.7	**2.66**±0.9	**1.34**±0.8	**0.995**±0.00
		Harvard		
Method	PSNR	SAM	ERGAS	SSIM
with ReLU	43.76±2.7	2.68±0.7	2.46±0.8	0.983±0.01
HSRnet	**44.29**±3.0	**2.66**±0.7	**2.45**±0.8	**0.984**±0.01

TABLE XI

AVERAGE QIs AND RELATED STANDARD DEVIATIONS OF THE RESULTS ON THE CAVE AND THE HARVARD DATASETS USING THE PROPOSED METHOD WITH GAP AND GMP. THE BEST VALUES ARE HIGHLIGHTED IN BOLDFACE

		CAVE		
Method	PSNR	SAM	ERGAS	SSIM
GMP	**47.91**±2.6	2.67±0.9	**1.34**±0.8	**0.995**±0.00
HSRnet	47.82±2.7	**2.66**±0.9	**1.34**±0.8	**0.995**±0.00
		Harvard		
Method	PSNR	SAM	ERGAS	SSIM
GMP	**44.32**±3.0	**2.66**±0.7	**2.44**±0.9	**0.984**±0.01
HSRnet	44.29±3.0	**2.66**±0.7	2.45±0.8	**0.984**±0.01

TABLE XII

AVERAGE QIs AND RELATED STANDARD DEVIATIONS OF THE RESULTS ON THE CAVE AND THE HARVARD DATASETS USING THE PROPOSED METHOD WITH DIFFERENT AM ORDER (CA → SA INDICATES THAT CA IS BEFORE THE SA). THE BEST VALUES ARE HIGHLIGHTED IN BOLDFACE

		CAVE		
Method	PSNR	SAM	ERGAS	SSIM
CA→SA	**47.92**±2.6	2.68±0.9	**1.33**±0.8	**0.995**±0.00
HSRnet	47.82±2.7	**2.66**±0.9	1.34±0.8	**0.995**±0.00
		Harvard		
Method	PSNR	SAM	ERGAS	SSIM
CA→SA	44.01±2.9	2.72±0.7	2.63±1.0	**0.984**±0.01
HSRnet	**44.29**±3.0	**2.66**±0.7	**2.45**±0.8	**0.984**±0.01

The average results of the QIs are reported in Table XI. The use of the GMP instead of the GAP does not significantly affect the results, thus concluding that both the strategies can be considered acceptable.

5) AM's Order: In the proposed HSRnet, the SA is conducted on the outcome coming from the ResNet block, and then the CA is applied to the data after a simple convolution layer. To assess the performance varying the order between the above-mentioned two modules, we put the SA after the CA retraining the network. The quantitative comparison is given in Table XII. The new structure shows slightly poorer performance than the original one, but, generally, the order of the two modules does not significantly affect the performance.

E. Comparison With MHFnet

To our knowledge, the MHFnet developed by Xie *et al.* [3] outperforms the state of the art of the model-based and the deep learning-based methods, actually representing the best way to address the HSI super-resolution problem. Due to the fact that the MHFnet and our HSRnet are both deep

TABLE XIII

Results of the Two Deep Learning-Based Methods Varying the Number of the Training Samples. The Best Values Are Highlighted in Boldface

Datasets	# training data	Methods	PSNR	SAM	ERGAS	SSIM
CAVE	3136	MHFnet	46.32	4.33	1.74	0.992
		HSRnet	**47.82**	**2.66**	**1.34**	**0.995**
	2000	MHFnet	46.47	4.31	1.71	0.992
		HSRnet	**47.89**	**2.67**	**1.34**	**0.995**
	1000	MHFnet	46.16	4.36	1.77	0.992
		HSRnet	**47.76**	**2.70**	**1.35**	**0.995**
	500	MHFnet	45.90	4.41	1.83	0.991
		HSRnet	**47.31**	**2.74**	**1.41**	**0.995**
Harvard	3136	MHFnet	40.37	4.64	24.17	0.966
		HSRnet	**44.28**	**2.66**	**2.45**	**0.984**
	2000	MHFnet	40.49	4.55	15.03	0.967
		HSRnet	**44.33**	**2.66**	**2.38**	**0.984**
	1000	MHFnet	40.34	4.50	15.98	0.967
		HSRnet	**43.73**	**2.72**	**2.71**	**0.983**
	500	MHFnet	40.18	4.60	15.71	0.963
		HSRnet	**43.96**	**2.72**	**2.55**	**0.983**

TABLE XIV

Average QIs and Related Standard Deviations of the Results for the Networks Trained on the Harvard Dataset. The Best Values Are Highlighted in Boldface

		CAVE		
Method	PSNR	SAM	ERGAS	SSIM
MHFnet	35.70±2.7	11.95±3.6	5.60±2.4	0.935±0.02
HSRnet	**40.88**±2.6	**4.93**±1.6	**2.79**±1.4	**0.976**±0.02
		Harvard		
Method	PSNR	SAM	ERGAS	SSIM
MHFnet	43.10±3.9	2.76±0.8	3.28±1.5	0.977±0.01
HSRnet	**45.01**±3.0	**2.56**±0.7	**2.11**±0.8	**0.985**±0.01

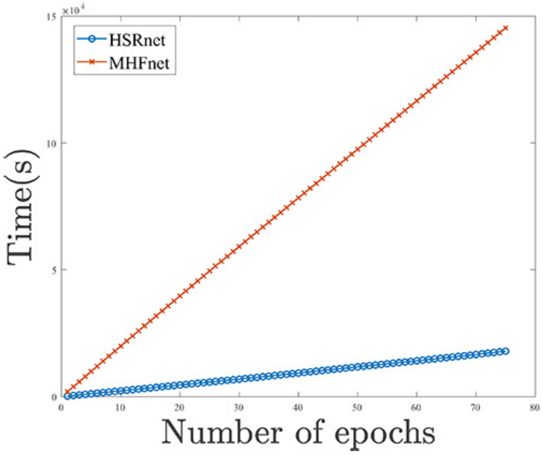


Fig. 12. Comparison of the training times for the MHFnet and the proposed HSRnet.

learning-based methods, in this subsection, we keep on discussing about the HSRnet comparing it with the MHFnet.

1) Sensitivity to Number of Training Samples: We train the MHFnet and our HSRnet with different numbers of training samples to illustrate their sensitivity with respect to this parameter. We randomly select 500, 1000, 2000, and 3136 samples from the training data. Testing data consists of 11 testing images on the CAVE dataset and ten testing images on the Harvard dataset. Table XIII reports the average QIs of the results obtained by the MHFnet and by our HSRnet varying the number of the training samples. From the results on the CAVE dataset in Table XIII, we can note that our method steadily outperforms the MHFnet in every case. Instead, from the results on the Harvard dataset, we can remark that the generalization ability of our method is robust with respect to changes in the numbers of the training samples. Whereas the MHFnet shows poor performance due to its manual predefined parameters that are sensitive to scene changes. Noted that HSRnet and MHFnet both get the best performance in the case of 2000 training samples, but we still select the most fully trained models of them. This does not prejudice the fairness of the comparison.

2) Network Generalization: In the above content, MHFnet and our HSRnet are both trained with CAVE data. We can find that our HSRnet outperforms the MHFnet in all the experiments on the testing data provided by the Harvard dataset. This shows the remarkable generalization ability of our network. To further corroborate it, we retrain these two networks on training samples provided by the Harvard dataset. Namely, we extract from the Harvard dataset 3136 training samples, in which the HR-MSI is of size 64 × 64, and the LR-HSI is of size 16 × 16. As previously done, we select the same 11 images from the CAVE dataset and the same ten images from the Harvard dataset to build the testing set. We show the QIs of the results for these two networks trained on the Harvard dataset in Table XIV. It can be seen that the generalization ability of the MHFnet is still limited. Instead, the proposed approach still shows better generalization ability when used on CAVE data but trained on the Harvard samples.

3) Parameters and Training Time: MHFnet contains 3.6 million parameters; however, only 1.9 million parameters have to be learned by our HSRnet. In Fig. 12, we plot the training time with respect to the epochs. We can find that our network needs much less training time than MHFnet. The MHFnet needs about 40 h, while our HSRnet just needs 5 h. Actually, from Tables III and V, the testing time of our HSRnet is also less than that of the MHFnet. Indeed, fewer parameters result in less training and testing times, making our method more practical.

V. Conclusion

In this article, a simple and efficient deep network architecture has been proposed for addressing the HSI super-resolution issue. The network architecture consists of two parts: 1) a spectral preservation module and 2) a spatial preservation module that aims to reconstruct the image's spatial details with AM and PS modules. The combination of these two parts is performed to get the final network output. This latter is compared with the reference (GT) image under the Frobenius norm-based loss function. This is done with the aim of estimating the network parameters during the training phase.

Extensive experiments demonstrated the superiority of our HSRnet with respect to recent state-of-the-art HSI super-resolution approaches. Additionally, advantages of our HSRnet have also been reported from other points of view, such as the network generalization, the limited computational burden, and the robustness with respect to the number of training samples.

References

[1] R. Dian, S. Li, B. Sun, and A. Guo, "Recent advances and new guidelines on hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 69, pp. 40–51, May 2021.

[2] R. Dian, S. Li, and X. Kang, "Regularizing hyperspectral and multispectral image fusion by CNN denoiser," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 3, pp. 1–12, Mar. 2020.

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

[4] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5345–5355, Nov. 2018.

[5] L. Liu *et al.*, "Shallow–deep convolutional network and spectral-discrimination-based detail injection for multispectral imagery pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1772–1783, Mar. 2020.

[6] F. Fang, F. Li, C. Shen, and G. Zhang, "A variational approach for pan-sharpening," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2822–2834, Jul. 2013.

[7] T. Wang, F. Fang, F. Li, and G. Zhang, "High-quality Bayesian pan-sharpening," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 227–239, Jan. 2019.

[8] X. Fu, Z. Lin, Y. Huang, and X. Ding, "A variational pan-sharpening with local gradient constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10265–10274.

[9] C. I. Kanatsoulis, X. Fu, N. D. Sidiropoulos, and W.-K. Ma, "Hyperspectral super-resolution: Combining low rank tensor and matrix structure," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3318–3322.

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

[11] R. A. Borsoi, T. Imbiriba, and J. C. M. Bermudez, "Super-resolution for hyperspectral and multispectral image fusion accounting for seasonal spectral variability," *IEEE Trans. Image Process.*, vol. 29, pp. 116–127, 2020.

[12] Y. Xing, M. Wang, S. Yang, and K. Zhang, "Pansharpening with multiscale geometric support tensor machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2503–2517, May 2018.

[13] 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.

[14] X. Han, J. Luo, J. Yu, and W. Sun, "Hyperspectral image fusion based on non-factorization sparse representation and error matrix estimation," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2017, pp. 1155–1159.

[15] M. Selva, B. Aiazzi, F. Butera, L. Chiarantini, and S. Baronti, "Hyper-sharpening: A first approach on SIM-GA data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 3008–3024, Jun. 2015.

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

[17] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4118–4130, Aug. 2018.

[18] T. Xu, T.-Z. Huang, L.-J. Deng, X.-L. Zhao, and J. Huang, "Hyperspectral image superresolution using unidirectional total variation with tucker decomposition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4381–4398, 2020.

[19] Z.-W. Pan and H.-L. Shen, "Multispectral image super-resolution via RGB image fusion and radiometric calibration," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1783–1797, Apr. 2019.

[20] K. Zhang, M. Wang, and S. Yang, "Multispectral and hyperspectral image fusion based on group spectral embedding and low-rank factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1363–1371, Mar. 2017.

[21] 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.

[22] Y. Zhang, C. Liu, M. Sun, and Y. Ou, "Pan-sharpening using an efficient bidirectional pyramid network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5549–5563, Aug. 2019.

[23] R. Dian, S. Li, L. Fang, and Q. Wei, "Multispectral and hyperspectral image fusion with spatial-spectral sparse representation," *Inf. Fusion*, vol. 49, pp. 262–270, Sep. 2019.

[24] K. Rong, L. Jiao, S. Wang, and F. Liu, "Pansharpening based on low-rank and sparse decomposition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 12, pp. 4793–4805, Dec. 2014.

[25] K. Rong, S. Wang, X. Zhang, and B. Hou, "Low-rank and sparse matrix decomposition-based pan sharpening," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2012, pp. 2276–2279.

[26] H. A. Aly and G. Sharma, "A regularized model-based optimization framework for pan-sharpening," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2596–2608, Jun. 2014.

[27] P. Liu, L. Xiao, and T. Li, "A variational pan-sharpening method based on spatial fractional-order geometry and spectral–spatial low-rank priors," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1788–1802, Mar. 2018.

[28] M. A. Veganzones, M. Simoes, G. Licciardi, N. Yokoya, J. M. Bioucas-Dias, and J. Chanussot, "Hyperspectral super-resolution of locally low rank images from complementary multisource data," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 274–288, Jan. 2016.

[29] R. Dian and S. Li, "Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 5135–5146, Oct. 2019.

[30] R. Dian, S. Li, and L. Fang, "Learning a low tensor-train rank representation for hyperspectral image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2672–2683, Sep. 2019.

[31] Z.-C. Wu *et al.*, "A new variational approach based on proximal deep injection and gradient intensity similarity for spatio-spectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 6277–6290, 2020, doi: 10.1109/JSTARS.2020.3030129.

[32] X. Fu, W. Wang, Y. Huang, X. Ding, and J. Paisley, "Deep multiscale detail networks for multiband spectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 1–15, May 2020.

[33] L.-J. Deng, M. Feng, and X.-C. Tai, "The fusion of panchromatic and multispectral remote sensing images via tensor-based sparse modeling and hyper-Laplacian prior," *Inf. Fusion*, vol. 52, pp. 76–89, Dec. 2019.

[34] Z.-Y. Zhang, T.-Z. Huang, L.-J. Deng, J. Huang, and H.-X. Dou, "Pan-sharpening via RoG-based filtering," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 2790–2793.

[35] J. Liu, "A Rudin-Osher-Fatemi model-based pansharpening approach using RKHS and AHF representation," *East Asian J. Appl. Math.*, vol. 9, no. 1, pp. 13–27, Jun. 2019.

[36] L.-J. Deng, G. Vivone, W. Guo, M. D. Mura, and J. Chanussot, "A variational pansharpening approach based on reproducible kernel Hilbert space and heaviside function," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4330–4344, Sep. 2018.

[37] G. Vivone *et al.*, "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

[38] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3418–3431, Jul. 2018.

[39] G. Vivone, R. Restaino, M. D. Mura, and J. Chanussot, "Multi-band semiblind deconvolution for pansharpening applications," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 41–44.

[40] G. Vivone *et al.*, "Pansharpening based on semiblind deconvolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1997–2010, Apr. 2015.

[41] 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.

[42] R. Dian, S. Li, L. Fang, T. Lu, and J. M. Bioucas-Dias, "Nonlocal sparse tensor factorization for semiblind hyperspectral and multispectral image fusion," *IEEE Trans. Cybern.*, vol. 50, no. 10, pp. 4469–4480, Oct. 2020.

[43] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Aug. 2009.

[44] T.-X. Jiang, M. K. Ng, X.-L. Zhao, and T.-Z. Huang, "Framelet representation of tensor nuclear norm for third-order tensor completion," *IEEE Trans. Image Process.*, vol. 29, pp. 7233–7244, 2020.

[45] Y. Xu, Z. Wu, J. Chanussot, and Z. Wei, "Hyperspectral images super-resolution via learning high-order coupled tensor ring representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 1–14, Nov. 2020.

[46] W. Hu, D. Tao, W. Zhang, Y. Xie, and Y. Yang, "The twist tensor nuclear norm for video completion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 12, pp. 2961–2973, Dec. 2017.

[47] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, Sep. 1966.

[48] J. M. Bioucasdias and M. A. Figueiredo, "Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing," in *Proc. Workshop Hyperspectral Image Signal Process., Evol. Remote Sens.*, Jun. 2010, pp. 1–4.

[49] S. Mei, X. Yuan, J. Ji, S. Wan, J. Hou, and Q. Du, "Hyperspectral image super-resolution via convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4297–4301.

[50] J. Hu, Y. Li, X. Zhao, and W. Xie, "A spatial constraint and deep learning based hyperspectral image super-resolution method," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 5129–5132.

[51] Q. Huang, W. Li, T. Hu, and R. Tao, "Hyperspectral image super-resolution using generative adversarial network and residual learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3012–3016.

[52] 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.

[53] 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. 1753–1761.

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

[55] 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.

[56] Y. Li, J. Hu, X. Zhao, W. Xie, and J. Li, "Hyperspectral image super-resolution using deep convolutional neural network," *Neurocomputing*, vol. 266, pp. 29–41, Nov. 2017.

[57] W. Yao, Z. Zeng, C. Lian, and H. Tang, "Pixel-wise regression using U-Net and its application on pansharpening," *Neurocomputing*, vol. 312, pp. 364–371, Oct. 2018.

[58] S. Vitale, "A CNN-based pansharpening method with perceptual loss," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 3105–3108.

[59] F. Palsson, J. Sveinsson, and M. Ulfarsson, "Sentinel-2 image fusion using a deep residual network," *Remote Sens.*, vol. 10, no. 8, p. 1290, Aug. 2018.

[60] X. Han, J. Yu, J. Luo, and W. Sun, "Hyperspectral and multispectral image fusion using cluster-based multi-branch BP neural networks," *Remote Sens.*, vol. 11, no. 10, p. 1173, May 2019.

[61] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Inf. Fusion*, vol. 42, pp. 158–173, Jul. 2018.

[62] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, "A new pan-sharpening method with deep neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 1037–1041, May 2015.

[63] Y. Rao, L. He, and J. Zhu, "A residual convolutional neural network for pan-shaprening," in *Proc. Int. Workshop Remote Sens. With Intell. Process. (RSIP)*, May 2017, pp. 1–4.

[64] X. Liu, Y. Wang, and Q. Liu, "PSGAN: A generative adversarial network for remote sensing image pan-sharpening," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 873–877.

[65] L.-J. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, early access, Oct. 27, 2020, doi: 10.1109/TGRS.2020.3031366.

[66] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[67] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 517–532.

[68] G. Tang, L. Zhao, R. Jiang, and X. Zhang, "Single image dehazing via lightweight multi-scale networks," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 154–169.

[69] H. Zhang, V. Sindagi, and V. M. Patel, "Multi-scale single image dehazing using perceptual pyramid deep network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 902–911.

[70] C.-H. Yeh, C.-H. Huang, and L.-W. Kang, "Multi-scale deep residual learning-based single image haze removal via image decomposition," *IEEE Trans. Image Process.*, vol. 29, pp. 3153–3167, 2020.

[71] X. Wang, H. Ma, X. Chen, and S. You, "Edge preserving and multi-scale contextual neural network for salient object detection," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 121–134, Jan. 2018.

[72] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[73] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5659–5667.

[74] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010.

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

[76] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. CVPR*, Jun. 2011, pp. 193–200.

[77] N. Yokoya and A. Iwasaki, "Airborne hyperspectral data over chikusei," Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27, May 2016.

[78] R. Yuhas, J. Boardman, and A. Goetz, "Determination of semi-arid landscape endmembers and seasonal trends using convex geometry spectral unmixing techniques," in *Proc. 4th Annu. JPL Airborne Geosci. Workshop*, 1993, p. 22.

[79] L. Wald, *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions*. Paris, France: Presses des MINES, 2002.

[80] 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.

[81] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, Jun. 2011, pp. 315–323.

Jin-Fan Hu received the B.S. degree in computational mathematics from the School of Mathematical Sciences, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2019, where he is currently pursuing the M.S. degree in mathematics.

His current research interests include image processing and deep learning.

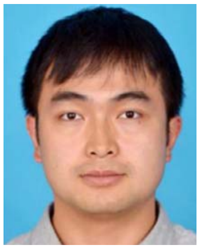
Ting-Zhu Huang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computational mathematics from the Department of Mathematics, Xi'an Jiaotong University, Xi'an, China, in 1986, 1992, and 2001, respectively.

He is currently a Professor with the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, China. His research interests include scientific computation and applications, numerical algorithms for image processing, numerical linear algebra, preconditioning technologies, and matrix analysis with applications.

Dr. Huang is an Editor of the *Scientific World Journal*, *Advances in Numerical Analysis*, the *Journal of Applied Mathematics*, the *Journal of Pure and Applied Mathematics: Advances in Applied Mathematics*, and the *Journal of Electronic Science and Technology*, China.

Liang-Jian Deng (Member, IEEE) received the B.S. and Ph.D. degrees in applied mathematics from the School of Mathematical Sciences, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2010 and 2016, respectively.

He is currently a Research Fellow with the School of Mathematical Sciences, UESTC. From 2013 to 2014, he was a joint-training Ph.D. student with the Case Western Reserve University, Cleveland, OH, USA. In 2017, he was a Post-Doctoral Fellow with Hong Kong Baptist University (HKBU), Hong Kong. In addition, he also stayed at the Isaac Newton Institute for Mathematical Sciences, University of Cambridge, Cambridge, U.K., and HKBU, for short visits. His research interests include using partial differential equations (PDEs), optimization modeling, deep learning to address several tasks in image processing, and computer vision, e.g., resolution enhancement and restoration.



Tai-Xiang Jiang (Member, IEEE) received the Ph.D. degree in mathematics from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2019. From 2017 to 2018, he was a co-training Ph.D. student with the University of Lisbon, Lisbon, Portugal, supervised by Prof. Jose M. Bioucas-Dias.

He was a Research Assistant with the Hong Kong Baptist University, Hong Kong, supported by Prof. Michael K. Ng in 2019. He is currently an Associate Professor with the School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu. His research interests include sparse and low-rank modeling and tensor decomposition for multidimensional image processing, especially on the low-level inverse problems for multidimensional images.

Gemine Vivone (Senior Member, IEEE) received the B.Sc. *(summa cum laude)*, the M.Sc. *(summa cum laude)*, and the Ph.D. degrees in information engineering from the University of Salerno, Salerno, Italy, in 2008, 2011, and 2014, respectively.

He is a Researcher at the National Research Council, Rome, Italy. In 2019, he was an Assistant Professor at the University of Salerno, Fisciano, Italy. In 2014, he joined the North Atlantic Treaty Organization (NATO), Science and Technology Organization (STO), Centre for Maritime Research and Experimentation (CMRE), La Spezia, Italy, as a Scientist. He was a Visiting Professor with the Grenoble Institute of Technology (INPG), Grenoble, France. His main research interests focus on statistical signal processing, detection of remotely sensed images, data fusion, and tracking algorithms.

Dr. Vivone is the Leader of the Image and Signal Processing Working Group of the IEEE Image Analysis and Data Fusion Technical Committee. He received the IEEE-GRSS Early Career Award in 2021, the Symposium Best Paper Award at the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) in 2015 and the Best Reviewer Award of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING in 2017. He is currently an Associate Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL) and he is an Editorial Board Member of *MDPI Remote Sensing*, *MDPI Sensors*, and *MDPI Encyclopedia*. He served as a guest associate editor for several special issues.

Jocelyn Chanussot (Fellow, IEEE) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from the Université de Savoie, Annecy, France, in 1998.

Since 1999, he has been with Grenoble INP, where he is currently a Professor of signal and image processing. He has been a Visiting Scholar at Stanford University, Stanford, CA, USA, KTH, Stockholm, Sweden, and NUS, Singapore. Since 2013, he has been an Adjunct Professor with the University of Iceland, Reykjavik, Iceland. From 2015 to 2017, he was a Visiting Professor at the University of California at Los Angeles (UCLA), Los Angeles, CA, USA. He holds the AXA Chair in remote sensing and is an Adjunct Professor at the Chinese Academy of Sciences, Aerospace Information Research Institute, Beijing, China. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence.

Dr. Chanussot was the founding President of the IEEE Geoscience and Remote Sensing French Chapter from 2007 to 2010, which received the 2010 IEEE GRS-S Chapter Excellence Award. He has received multiple outstanding paper awards. From 2017 to 2019, he was the Vice-President of the IEEE Geoscience and Remote Sensing Society, in charge of meetings and symposia. He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing (WHISPERS). He was the Chair (2009–2011) and the Co-Chair of the GRS Data Fusion Technical Committee from 2005 to 2008. He was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society from 2006 to 2008 and the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing in 2009. He is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the PROCEEDINGS OF THE IEEE. He was the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2011 to 2015. In 2014, he served as a Guest Editor for the *IEEE Signal Processing Magazine*. He is a member of the Institut Universitaire de France from 2012 to 2017 and has been a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters) since 2018.