GuidedNet: A General CNN Fusion Framework via High-resolution Guidance for Hyperspectral Image Super-resolution

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

Hyperspectral image super-resolution (HISR) is about fusing a low-resolution hyperspectral image (LR-HSI) and a high-resolution multispectral image (HR-MSI) to generate a high-resolution hyperspectral image (HR-HSI). Recently, convolutional neural network (CNN)-based techniques have been extensively investigated for HISR yielding competitive outcomes. However, existing CNN-based methods often require a huge amount of network parameters leading to a heavy computational burden, thus limiting the generalization ability. In this paper, we fully consider the characteristic of the HISR, proposing a general CNN fusion framework with high-resolution guidance, called GuidedNet. This framework consists of two branches, including 1) the high-resolution guidance branch (HGB) that can decompose the high-resolution guidance image into several scales; 2) the feature reconstruction branch (FRB) that takes the low-resolution image and the multi-scaled high-resolution guidance images from the HGB to reconstruct the high-resolution fused image. GuidedNet can effectively predict the high-resolution residual details that are added to the upsampled HSI to simultaneously improve spatial quality and preserve spectral information. The proposed framework is implemented using recursive and progressive strategies, which can promote high performance with a significant network parameter reduction, even ensuring network stability by supervising several intermediate outputs. Additionally, the proposed approach is also suitable for other resolution enhancement tasks, such as remote sensing pansharpening and single image super-resolution (SISR). Extensive experiments on simulated and real datasets demonstrate that the proposed framework generates state-of-the-art outcomes for several applications (*i.e.*, HISR, pansharpening, and SISR). Finally, an ablation study and more discussions assessing, *e.g.*, the network generalization, the low computational cost, and the fewer network parameters are provided to the readers.

***Index Terms*—Convolutional neural network (CNN), High-resolution guidance, Image fusion, Hyperspectral image super-resolution, Pansharpening, Single image super-resolution.**

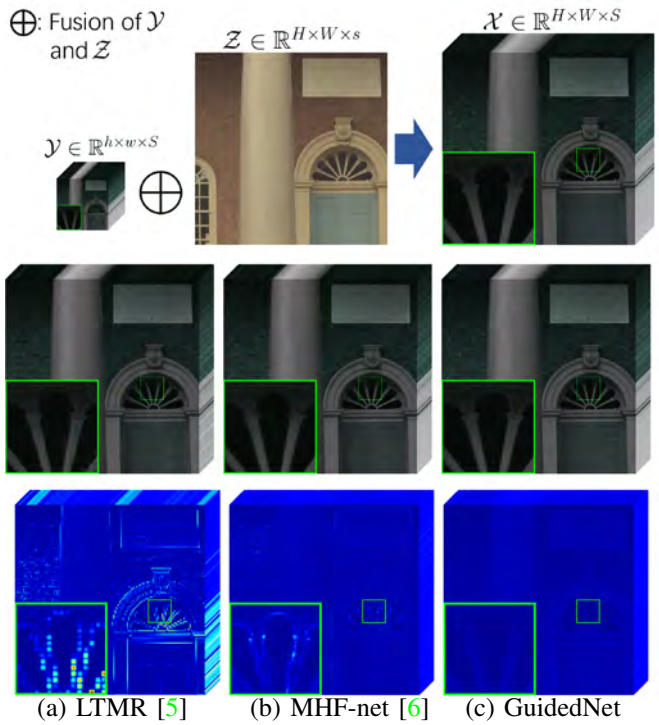
This research is supported by NSFC (12271083, 62203089, 12001446), Natural Science Foundation of Sichuan Province (2022NSFSC0501, 2022NSFSC0507, 2022NSFSC1798), Key Projects of Applied Basic Research in Sichuan Province (Grant No. 2020YJ0216), and National Key Research and Development Program of China (Grant No. 2020YFA0714001). *Corresponding author: Liang-Jian Deng.

R. Ran, L. -J. Deng and J. -F. Hu are with the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China (e-mails: ranran@std.uestc.edu.cn; liangjian.deng@uestc.edu.cn; hujf0206@163.com).

T. -X. 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, Sichuan, 610074, China (e-mail: taixiangjiang@gmail.com).

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

G. Vivone is with the National Research Council - Institute of Methodologies for Environmental Analysis, CNR-IMAA, 85050 Tito Scalo, Italy (e-mail: gemine.vivone@imaa.cnr.it).



(a) LTMR [5] (b) MHF-net [6] (c) GuidedNet

Fig. 1. First row: the schematic diagram of HISR. The image on the right is the HR-HSI $\mathcal{X}$ (*i.e.*, the ground-truth). Second row: the visual results (8× scale) of (a) the subspace-based low tensor multi-rank regularization approach (LTMR) [5] (PSNR = 41.88dB), (b) the MS/HS fusion net (MHF-net) [6] (PSNR = 43.36dB), and (c) the proposed GuidedNet (PSNR = 44.75dB). Note that all the images are displayed with a pseudo-color RGB format using R = 17th band, G = 30th band and B = 27th band. Third row: the related error maps. From a visual point of view, our GuidedNet result is the closest to the ground-truth.

I. Introduction

Recently, hyperspectral image super-resolution (HISR), as shown in Fig. 1, has become a fundamental issue in computer vision since it can significantly improve the spatial resolution of LR-HSI and the spectral information of HR-MSI to finally yield a fused hyperspectral image with both high spatial and spectral resolutions. Many applications can benefit from the fused HISR image, *e.g.*, several remote sensing data analysis [1], environment detection [2], classification [3], and recognition [4].

In general, HISR approaches could be roughly classified into two categories. Namely, variational optimization (VO)

approaches and deep learning (DL) approaches. The approach proposed in this work falls within the latter class.

VO-based methods are mainly about formulating an optimization model by considering proper regularizers and fidelity terms to solve computer vision problems [7]–[15], thus accurately representing the main properties of the HISR issue at hand [5], [16]–[22]. Afterward, some practical algorithms are designed for efficiently solving the given model, estimating the final super-resolved images. Although these VO-based methods produced satisfactory SR results, they need prior information before reconstructing the high-resolution HSI. This information is usually scene-dependent, requiring a fine adjustment to be adapted to different real scenarios. Moreover, the computational burden for this class is usually heavy.

In the last decade, DL-based algorithms have been considered to solve several image processing tasks such as super-resolution [23]–[26], image classification [27], and visual question answering [28]. Mainly, convolutional neural network (CNN), as core technique of DL-based approaches, has been applied to HISR [29]–[37], getting promising results. These deep learning methods can learn the relationship between the hyperspectral image and the ground-truth. They showed satisfactory performance in the HISR task. However, these methods still have some drawbacks. Firstly, some methods have a complex network structure and a considerable amount of network parameters to severely consume computing resources taking a long time for training and execution. Secondly, previous methods generally utilize the entire MSI without extracting multi-scale spatial features. The HR-HSI's features significantly differ from LR-HSI's features leading to obstructions in fusion and reconstruction on a large scale. Thirdly, some DL-based methods cannot easily extend to other image SR problems (*e.g.*, pansharpening, or SISR) with satisfying results. Hence, the above-mentioned issues motivate us to further improve DL-based HISR approaches.

In this paper, we propose the so-called GuidedNet introducing two crucial branches (mainly for the application to HISR). The first one is the high-resolution guidance branch (HGB) decomposing an image into several scales that are fully exploited into the subsequent fusion branch. The other is the feature reconstruction branch (FRB), which can fuse the LR input and the multi-scale information from the HGB to produce the final HR output. Besides, recursive blocks are also integrated into the proposed network architecture, leading to fewer network parameters and less computational time while maintaining high-quality outcomes.

In summary, the main contributions are as follows:

1) A general CNN fusion framework is proposed in this paper. We successfully applied it to multiple image resolution enhancement problems, such as HISR, pansharpening, and SISR, in the meanwhile obtaining state-of-the-art (SOTA) performance for each task.
2) Two novel network branches, *i.e.*, the FRB and the HGB, are designed to utilize multi-scale information of high-resolution guidance images and reconstruct the fused high-resolution output. In particular, the two developed branches have the following characteristics, *i.e.*, multi-scale information fusion, progressive feature injection, and gradual feature reconstruction. Rich structural information can be captured more accurately by using a receptive field from wide to fine in a multi-scale framework. Compared with direct upsampling, which leads to difficulties in learning mapping functions and blur effects for large scaling factors, a progressive structure can better address the problem by adapting it to large-scale super-resolution. Moreover, intermediate results predicted by GuidedNet are supervised, aiding network stability. Thanks to these characteristics, GuidedNet can easily obtain promising outcomes for resolution enhancement.
3) GuidedNet has several advantages with respect to previously developed approaches: SOTA performance thanks to the designed network architecture, fewer network parameters thanks to the usage of recursive blocks, a remarkable ability to upsample to several scales, and good adaptability to other image resolution enhancement tasks (verified in the experimental section).

The organization of this paper is as follows. Sect. II will briefly introduce the related works. In Sect. III, we will describe the proposed network architecture, including the two designed network branches, the recursive blocks, and the training details. In Sect. IV, we conduct extensive experiments to assess the effectiveness of the proposed network for HISR. Finally, conclusions are drawn in Sect. V.

II. Related Works

A. Related Works

In general, the relationship between the HR-HSI, the LR-HSI, and the HR-MSI can be expressed by the following linear models [38]:

$$\begin{aligned} \mathbf{Y} &= \mathbf{XBS} + \mathbf{N_Y}, \\ \mathbf{Z} &= \mathbf{RX} + \mathbf{N_Z}, \end{aligned} \tag{1}$$

where $\mathbf{Z} \in \mathbb{R}^{HW \times s}$, $\mathbf{Y} \in \mathbb{R}^{hw \times S}$, and $\mathbf{X} \in \mathbb{R}^{HW \times S}$ represent the input HR-MSI, LR-HSI and the target HR-HSI, respectively. H and W are the height and width of the target resolution, *i.e.*, the height and width of HR-MSI and HR-HSI, and h, w are the height and width of the input LR-HSI. S is the number of the spectral bands of the hyperspectral image, and s is the number of the spectral bands of the LR-MSI. $\mathbf{B} \in \mathbb{R}^{HW \times HW}$ represents the circular convolution operator, $\mathbf{S} \in \mathbb{R}^{HW \times hw}$ represents the downsampling operator, and $\mathbf{R} \in \mathbb{R}^{s \times S}$ is the spectral response matrix of the HR-MSI. $\mathbf{N_Y}$ and $\mathbf{N_Z}$ are the noises related to the LR-HSI and the HR-MSI, respectively.

Based on the above models, many studies have been proposed with effective solutions for the HSI super-resolution problem, see *e.g.*, [5], [17]–[19], [39]. For instance, in [17], spectral unmixing and sparse coding ideas have been studied to enhance the resolution of HSIs. Yokoya *et al.* developed in [18] a coupled nonnegative matrix factorization (CNMF) unmixing algorithm using a linear spectral mixture model, which can effectively and efficiently obtain competitive HISR results. Dian *et al.* in [19] clustered the HR-MSI and the HR-HSI, respectively, applying a low tensor-train rank (LTTR)

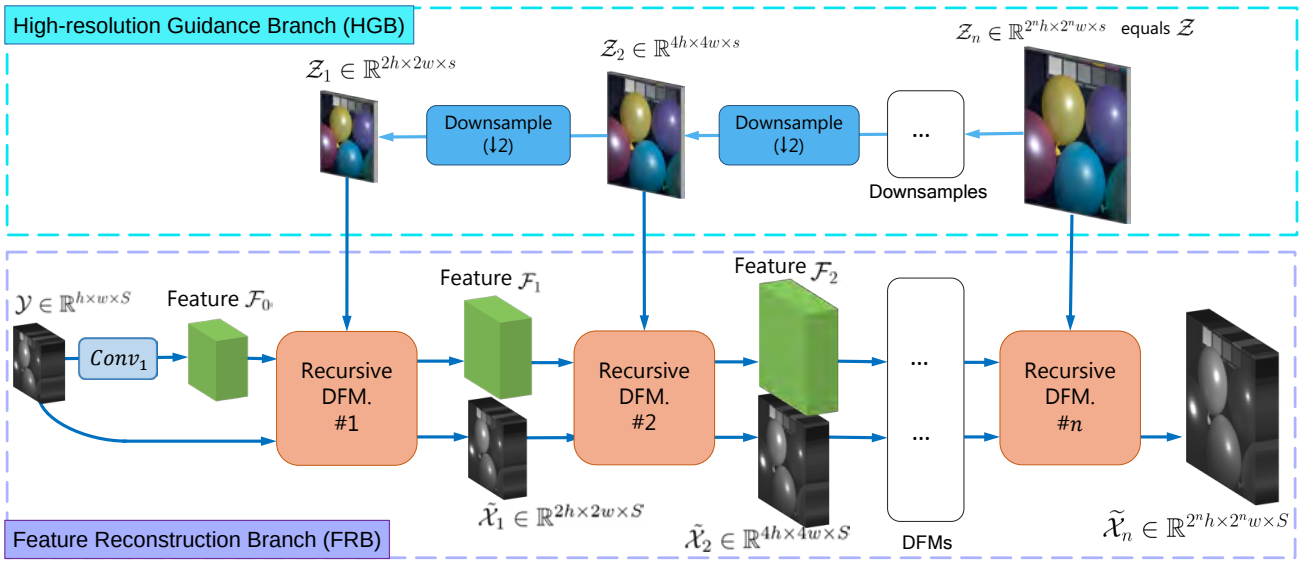


Fig. 2. The architecture of the proposed GuidedNet. LR-HSI, $\mathcal{Y}$, and HR-MSI, $\mathcal{Z}_n$, are the inputs, $\tilde{\mathcal{X}}_1$, $\tilde{\mathcal{X}}_2$ denote the intermediate scale outputs, and $\tilde{\mathcal{X}}_n$ is the final output. Note that $\mathcal{Z}_n$ is equal to the aforementioned $\mathcal{Z}$. This framework consists of two branches: the high-resolution guidance branch (HGB) that can generate several scales guidances, and the feature reconstruction branch (FRB) that fuses the low-resolution image and the multi-scaled high-resolution guidance images to reconstruct the high-resolution output. Note that the shown architecture includes n detail fusion modules (DFMs) to perform n scale HISR tasks, and all the parameters in the DFM of each layer are shared. Details can be found in Sect. III-D.

constraint to transform the HISR into an optimization problem, thus achieving excellent outcomes. Dian *et al.* [21] exploited a CNN denoiser to regularize the fusion procedure, achieving excellent fusion performance without needing additional HSIs and MSIs for the pre-training stage.

However, since it is generally necessary to assume some subjective priors, traditional methods are sensitive to the change of scenario showing difficulties when applied to different scenes. Recently, deep learning methods based on convolutional neural networks (CNNs) have been widely exploited for various low-level vision tasks [23], [40]–[42]. For example, Lim *et al.* designed EDSR [40] using residual networks and achieved competitive single image super-resolution outcomes. Zeng *et al.* in [41] learned the intrinsic representations of LR and HR image blocks via a proposed coupled deep autoencoder (CDA) holding outstanding performance for single image super-resolution. CNN-based methods *e.g.* [6], [29], [30], [37] can solve the HISR problem without relying upon subjective priors. Dian *et al.* [29] proposed a novel deep CNN-based HSI and MSI fusion method, which considers the imaging model of the HSI and MSI and achieves superior fusion performance. In [30], Palsson *et al.* proposed a 3-D CNN network using a principal component analysis to fuse HR-MSI and LR-HSI. This method significantly reduces the computational cost and has stronger robustness to noise. Zhu *et al.* in [33] proposed a lightweight progressive zero-centric residual network. Xie *et al.* designed a HISR model according to (1) in [31], then constructed the solving algorithm using the approximate gradient method. After that, a new fusion network, called MHF-net [6], is designed by expanding this solving algorithm. Benefiting from excellent preservation of the spectral and spatial details, the MHF-net outperforms other DL-based approaches, currently representing a state-of-the-art HISR method.

HISR is closely related to the multispectral image pansharpening task. In this work, we also extend our method to the pansharpening task. The pansharpening problem reconstructs the HR-MSI by fusing an LR-MSI and an HR panchromatic image. Traditional pansharpening approaches are represented by both component substitution (CS) and multi-resolution analysis (MRA) based methods. CS-based methods, such as the band dependent on spatial detail (BDSD) [43] and the BDSD with physical constraints (BDSD-PC) [44], can produce acceptable spatial fidelity outcomes but introducing spectral distortion. The class of MRA-based methods contains the generalized Laplacian pyramid (GLP) [45] and the GLP at full resolution for regression-based (GLP-Reg) [46].

Many deep learning-based methods have been designed for the pansharpening problem yielding competitive performance, see *e.g.*, [47]–[53]. In [48], Masi *et al.* adapted a simple three-layer convolution network for pansharpening. In [47], Yang *et al.* proposed a deep network structure (PanNet) that focuses on spectral and spatial preservation by training the network in the high-pass domain through a high-pass filter. In [52], Deng *et al.* combined the traditional CS and MRA fusion schemes developing a deep network (FusionNet) that extracts high-quality details, achieving competitive performance. However, pansharpening reaching high spatial resolutions can generate significant spectral distortion. The introduction and the full use of progressive and multi-scale architectures in pansharpening can alleviate this issue.

In what follows, we will present the proposed general fusion framework in more detail.

III. The Proposed GuidedNet

In this section, we present the motivation under the developing of the proposed method, the designed network, including the network architecture consisting of the two proposed

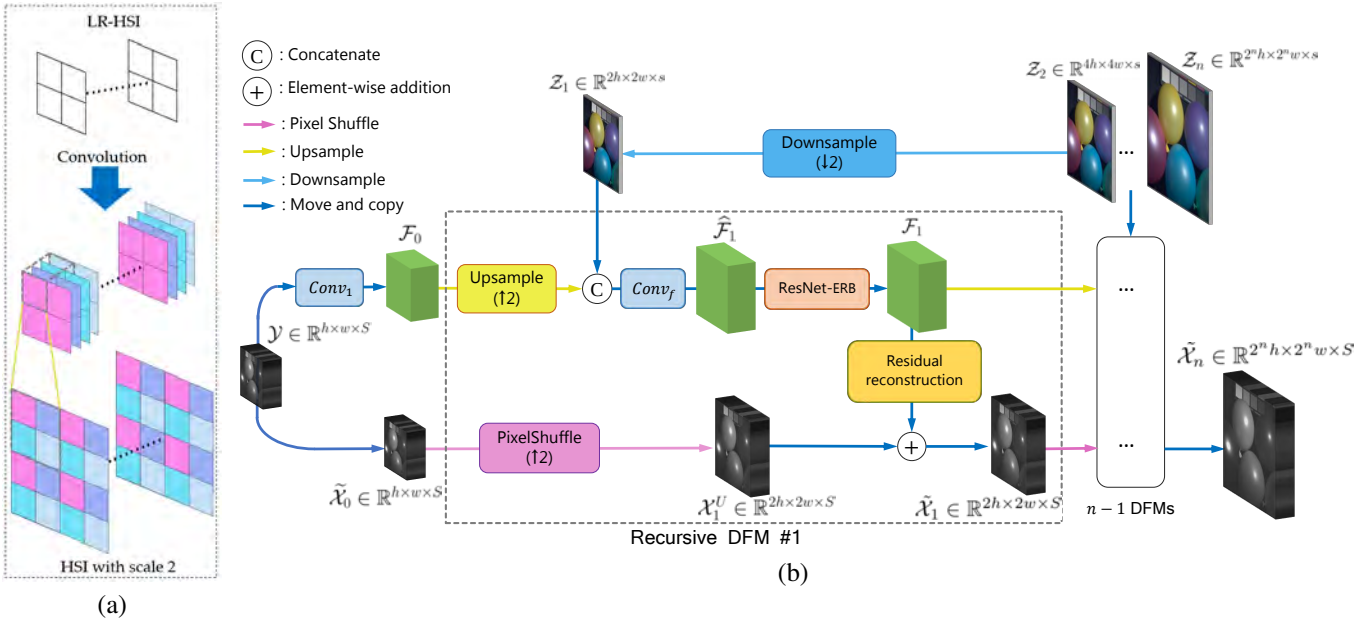


Fig. 3. The detailed network architecture of the proposed GuidedNet. (a) Illustration of the pixel shuffle (PS) with an upsampling scale factor of 2. (b) Architecture of the GuidedNet. $\mathcal{Y}$ and $\mathcal{Z}_n$ are the two inputs, and $\widetilde{\mathcal{X}}_1, \ldots, \widetilde{\mathcal{X}}_n$ are the outputs. $\mathcal{F}_0, \ldots, \mathcal{F}_n$ and $\mathcal{X}_1^U, \ldots, \mathcal{X}_n^U$ denote image features and upsampled images by the PS for several resolutions. $\widetilde{\mathcal{X}}_1, \ldots, \widetilde{\mathcal{X}}_n$ are high-resolution outputs of progressive generation. Note that the part enclosed by dotted lines is the DFM for the first layer in Fig. 2, and $\widetilde{\mathcal{X}}_0 = \mathcal{Y}$. Other DFM modules are similar as the first one.

branches, the recursive mechanism for parameter reduction, the loss function for multi-scale training, and some network training details.

A. Motivation

Some above-mentioned issues, such as progressive feature injection, gradual feature reconstruction, and parameter sharing, have motivated us to develop a general CNN fusion framework, which can fully consider, in a simple manner, a progressive multi-scale structure (PMS) for the HISR problem. Besides, we also expect to achieve promising outcomes with a significant network parameters reduction. Meanwhile, we hope that the proposed architecture can easily be extended to multiple image fusion tasks promoting the design of a general fusion framework. Thus, we need to design two branches for the two inputs of the fusion task, guaranteeing sufficient information exchange and communication from the different inputs. In addition, spatial information is fused into the feature domain. Therefore, in the reconstruction branch, the network has a dual data stream (DDS) coming from the feature and the image domains, which are connected through a residual learning module.

B. Overall Network Architecture

This work aims to formulate a general fusion framework for image fusion tasks while fully exploiting multi-scale information, progressive feature injection, and gradual feature reconstruction. To reach this goal, we design a general CNN fusion framework via high-resolution guidance for image fusion, *i.e.*, the proposed GuidedNet. The overall and detailed architectures are shown in Fig. 2 and Fig. 3(b), respectively. In the following, we will introduce first the two branches of the GuidedNet. To illustrate the given network architecture, we refer to the HISR as an application. Note that the architecture can easily be extended to other image fusion tasks, *e.g.*, pansharpening and SISR.

1) High-resolution Guidance Branch

Since there is a high-resolution input in the fusion tasks, fully utilizing this high-resolution input and injecting the image details into the low-resolution input is crucial. Besides, the high-resolution input on lower scales still holds high-frequency information, which can be integrated into the low-resolution input. Motivated by the two above-mentioned points, we designed a high-resolution guidance branch (HGB) to inject the high-resolution details from different scales into the low-resolution input branch, see the top side in Fig. 2. The proposed GuidedNet introduces a two-branches strategy to regard spatial details as a guided term to drive the injection of high-resolution information into the feature domain. Compared with previously developed networks based on the two-branches strategy, such as the efficient bidirectional pyramid network (BDPN) for the pansharpening in [54] and the deep multi-scale guidance network (MSGNet) for the depth map super-resolution in [55], GuidedNet shows several differences in the fusion mode exploiting a gradual feature reconstruction, see Sect. III-B3 for details.

The generation of the multi-scale high-resolution guidance image can be expressed as follows:

$$\mathcal{Z}_k = Downsample(\mathcal{Z}_{k+1}, \mathbf{\Theta}_d), \tag{2}$$

where $Downsample$ represents the downsampling network consisting of 2D convolutions, $\mathbf{\Theta}_d$ indicates the network parameters to be trained, $\mathcal{Z}_k$ is the guidance image of size $2^k h \times 2^k w \times s$ at the k-th stage of the HGB with $k = 1, 2, \cdots, n-1$, and n is the total number of layers.

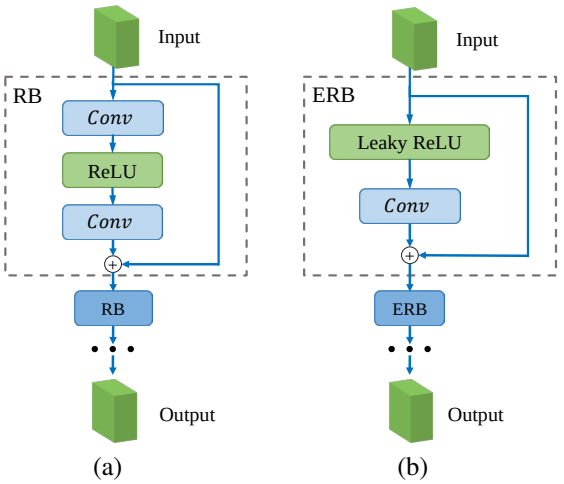


Fig. 4. (a) Structure of the conventional ResNet with the residual block (RB). (b) Structure of the ResNet with the efficient residual block (ResNet-ERB), which is used in the GuidedNet.

2) Feature Reconstruction Branch

The feature reconstruction branch (FRB) is about progressively injecting the high-frequency details from the high-resolution input (*i.e.*, the HR-MSI) on different scales into the LR-HSI, see the bottom side in Fig. 2.

2.1) FRB Flow

The LR-HSI feature $\mathcal{F}_0$ is extracted first through a convolutional layer $Conv_1$ with parameters indicated as $\mathbf{\Theta}_1$:

$$\mathcal{F}_0 = Conv_1(\mathcal{Y}, \mathbf{\Theta}_1), \tag{3}$$

then the extracted LR-HSI feature is considered by the designed detail fusion module (DFM) to complete the spatial feature reconstruction by incorporating the high-resolution input on the smallest scale (*i.e.,* $\mathcal{Z}_1$). The reconstructed HR-HSI comes from the previous level[1]. The details about DFM can be found in Sect. III-B2. When the fusion procedure by the recursive DFM is ended, we obtain two outputs, *i.e.*, the reconstructed HR-HSI feature $\mathcal{F}_1$ and the HR-HSI image $\tilde{\mathcal{X}}_1$ at a finer scale. Afterwards, the two obtained outputs and the high-resolution input at a finer scale, are considered in the next DFM. This structure of two data in parallel is called dual data stream (DDS), and after repeating this process several times, the final HR-HSI is yielded by the FRB.

2.2) Detail Fusion Module (DFM)

This section is devoted to presenting the detail fusion module (DFM). This module incorporates three inputs (*i.e.*, the high-resolution input from the HGB, the reconstructed HR-HSI feature, and the HR-HSI image from the previous step) into a designed convolutional module for gradually injecting high-frequency information into the hyperspectral image. This module considers first a feature upsampling consisting of a convolution operation and a deconvolution strategy to increase the feature size at a finer scale (corresponding to the scale of the high-resolution input provided by the HGB). Then, the upsampled HSI feature is concatenated with the high-resolution guidance from the HGB, seen as a new feature

[1]Note that the reconstructed HR-HSI for the starting (first) level is the LR-HSI, *i.e.*, $\mathcal{Y}$.

Feature $\mathcal{F}_1$ Feature $\mathcal{F}_2$ Feature $\mathcal{F}_3$

Fig. 5. Visual comparison of $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$ extracted from the 1st DFM, the 2nd DFM, and the 3rd DFM, respectively. Note that the images are scaled to the same size for a better visualization.

with detailed information. The number of channels of the new feature is restored by a simple convolutional layer:

$$\widehat{\mathcal{F}}_k = Conv_f(Upsample(\mathcal{F}_{k-1}), \mathcal{Z}_k, \mathbf{\Theta}_f), \tag{4}$$

where $\widehat{\mathcal{F}}_k$ represents a feature with size $2^k h \times 2^k w \times C$, $\widehat{\mathcal{F}}_{k-1}$ is another feature with size $2^{k-1}h \times 2^{k-1}w \times C$, $\mathcal{Z}_k$ is the high-resolution guidance from the HGB, and $\mathbf{\Theta}_f$ indicates the parameters to be trained.

A unique ResNet accounting for efficient residual blocks (ERBs), called ResNet-ERB, is designed to fuse details and reconstruct high-resolution HSI features. Generally, the ResNet consists of two convolution layers and an activation function in the middle, as shown in Fig. 4(a). However, as the depth of the ResNet increases, the gradient information tends to vanish when it reaches the end because of a significant amount of redundancy in the deep ResNet [56]. Too many convolutions with limited benefits can increase the computational burden, thus suggesting the simplification of the network by removing the superfluous layers. For image spatial enhancement tasks, feature propagation can be strengthened by creating short paths from early to later layers. Therefore, the proposed DFM utilizes a ResNet, including an efficient residual block (ERB). In the proposed ERB, just a LeakyReLU activation function and a convolutional layer are adopted to simplify the network structure, improving efficiency. The structure is shown in Fig. 4(b); several blocks are connected in a row to form the final ResNet-ERB module. Thus, the network reduces the number of parameters thanks to the more straightforward structure of the block. Furthermore, this block structure can extract features more effectively, reducing the difficulty of the network in the learning phase (preventing gradient exploding). ResNet-ERB is represented in our network as:

$$\mathcal{F}_k = ResNet_{ERB}(\widehat{\mathcal{F}}_k, \mathbf{\Theta}_e), \tag{5}$$

where $ResNet_{ERB}$ is the ResNet-ERB function, $\mathcal{F}_k$ is the output feature, and $\mathbf{\Theta}_e$ is the set of parameters to be learned.

Through this design, the spatial details of the guidances are gradually injected into the HSI features in the DFM related to the different layers. Fig. 5 shows a visual comparison of the features $\mathcal{F}_k$ $(k \in \{1, 2, 3\})$ of the DFM for the different layers. Specifically, in the *chart and stuffed toy* test case from the CAVE dataset, we selected the 31st, the 54th, and the 15th bands of the feature maps as R, G, and B, respectively, and the images are sampled to reach the same size for visualization purposes. The figure shows that the spatial detail information in the three features increases.

After generating the reconstructed high-resolution features as output of the ResNet-ERB, the residual image is predicted by a residual reconstruction module consisting of a convolutional layer to adjust channels. In the other stream, the LR-HSI is upsampled with a factor of 2 by an upsampling block, *i.e.*, the pixel shuffle (called sub-pixel convolution)[2]. The operation is shown in Fig. 3(a). Finally, the upsampled image is added to the residual image to reconstruct the final HR-HSI. Thus, we have:

$$\mathcal{X}_k^U = PS(\widetilde{\mathcal{X}}_{k-1}, \mathbf{\Theta}_p), \tag{6}$$

where PS is the pixel shuffle function providing the upsampling of the input data, $\mathcal{X}_k^U$ is the upsampled image at the k-th level, $\mathbf{\Theta}_p$ indicates the set of the parameters to be trained,

$$\widetilde{\mathcal{X}}_k = resRecon(\mathcal{F}_k, \mathbf{\Theta}_r) + \mathcal{X}_k^U, \tag{7}$$

$resRecon(\cdot)$ represents the residual reconstruction module to predict a residual image from $\mathcal{F}_k$, and $\mathbf{\Theta}_r$ indicates the set of the parameters. It is worth to be remarked that if $k = 1$, $\widetilde{\mathcal{X}}_{k-1}$ is equal to $\mathcal{Y}$.

2.3) Recursive Mechanism for DFMs

After determining the DFM, the GuidedNet approach repeats the DFM several times to reach the desired resolution of the HSI. Since the DFMs for different scales hold the same network structure, we can use the recursive mechanism for each DFM to significantly reduce the network parameters. Besides, thanks to the repetitive usage of the DFM, the proposed network can theoretically get fusion outcomes with any scaling factor power of 2. For example, we tested the performance of our GuidedNet considering the HISR application with scaling factors of 4, 8, 16, and 32, see Sect. III-B2.

3) Comparison with Previous Works

The GuidedNet is related to previous works, such as the BDPN [54] and the MSGNet [55], which use different resolutions for bidirectional multi-scale feature enhancement. Compared with the BDPN, the GuidedNet strongly focuses on fusing features into two branches to extract details, while BDPN just uses a simple addition operator to solve this problem. The guidance images must be mapped into the feature space when fusing the low-resolution image. Hence, we attach importance in the GuidedNet to the mapping learning among image features. Comparing with the MSGNet, the GuidedNet generates intermediate results many times and adopts multi-scale loss training to ensure spectral preservation and stability. In addition, the fusion step for the MSGNet approach is just based on a simple convolution, reducing the task's efficiency. Finally, the GuidedNet achieves multi-scale fusion and reconstruction into the feature domain, and thanks to its sharing strategy, it can significantly reduce the network parameters.

C. Loss Function

In the GuidedNet, several intermediate outputs, *i.e.*, $\tilde{\mathcal{X}}_i$, $i = 1, 2, \cdots, L$, are generated by the recursive DFMs. These outputs can progressively generate the final HR output with the desired scale through the specially designed network architecture. For a better supervision of the network learning, it is better to enforce a mean square error (MSE) loss between the output of a given scale and the corresponding downsampled ground-truth (GT) image. Thus, the final loss function is a multiple-loss one, which is defined as follows:

$$\begin{aligned} \mathcal{L}(\mathcal{Y}, \mathcal{Z}_n, \mathcal{X}_n; \mathbf{\Theta}) &= \sum_{k \in K} \alpha_k \mathcal{L}_k(\mathcal{Y}, \mathcal{Z}_k, \mathcal{X}_k; \mathbf{\Theta}), \\ \mathcal{L}_k(\mathcal{Y}, \mathcal{Z}_k, \mathcal{X}_k; \mathbf{\Theta}) &= \left\| \tilde{\mathcal{X}}_k - \mathcal{X}_k \right\|_F^2, \end{aligned} \tag{8}$$

where k represents the layer number of the reconstructed HSI, $\mathbf{\Theta}$ involves all the related network parameters to be learned, $\mathcal{Y}$ and $\mathcal{Z}_n$ are the LR-HSI and the maximum resolution guidance image in input, respectively, $\mathcal{X}_k$ represents the HR-HSI at the k-th layer, K indicates all the layer numbers (with $K = \{1, 2, \ldots, n\}$), and α_k is the weight of each sub-loss function at the k-th layer.

The weights can be set in several ways. An attempt is to set them considering the approximation degree to the final result, *i.e.*, the weight gradually becomes more significant as the scale increases. For instance, if the SR ratio is 8, we set $K = \{1, 2, 3\}$, and α_1, α_2, and α_3 are set to 1, 2, and 4, respectively. However, the network's stability could be reduced, and the prediction results could appear highly distorted under this setting. This is because the intermediate results are not significantly supervised. Another possibility is to increase the weights of the intermediate results to solve this problem. Thus, α_1, α_2, and α_3 can be set to 4, 2, and 1, respectively, to improve stability and accuracy. Indeed, the network architecture is trained progressively. Thus, if we take larger weights for the initial and intermediate loss functions (*i.e.*, layers 1 and 2), we can have a better final HR image reconstruction, even if we use a smaller weight for the final layer (*i.e.*, layer 3), see also the ablation study in Sect. IV-C.

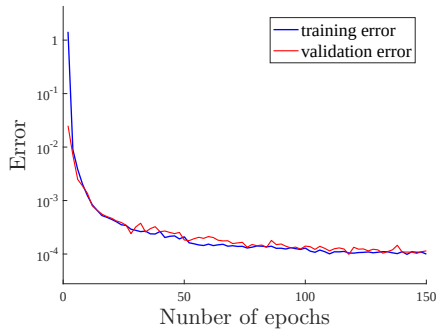


Fig. 6. Training and validation errors of our GuidedNet.

D. Network Training Details

Network details: This section is devoted to showing more network details. More specifically, the number of channels C for all the features is set to 64, the sizes of all the convolutional kernels are 3×3, the sizes of all the downsampling convolution and deconvolution kernels are 6×6, and the padding type of all the convolutions is set as "SAME". Additionally, all the

[2]The pixel shuffle approach expands, by a convolutional layer, the LR-HSI with size $h \times w \times S$ to reach the size of $rh \times rw \times S$, where r is the scaling factor.

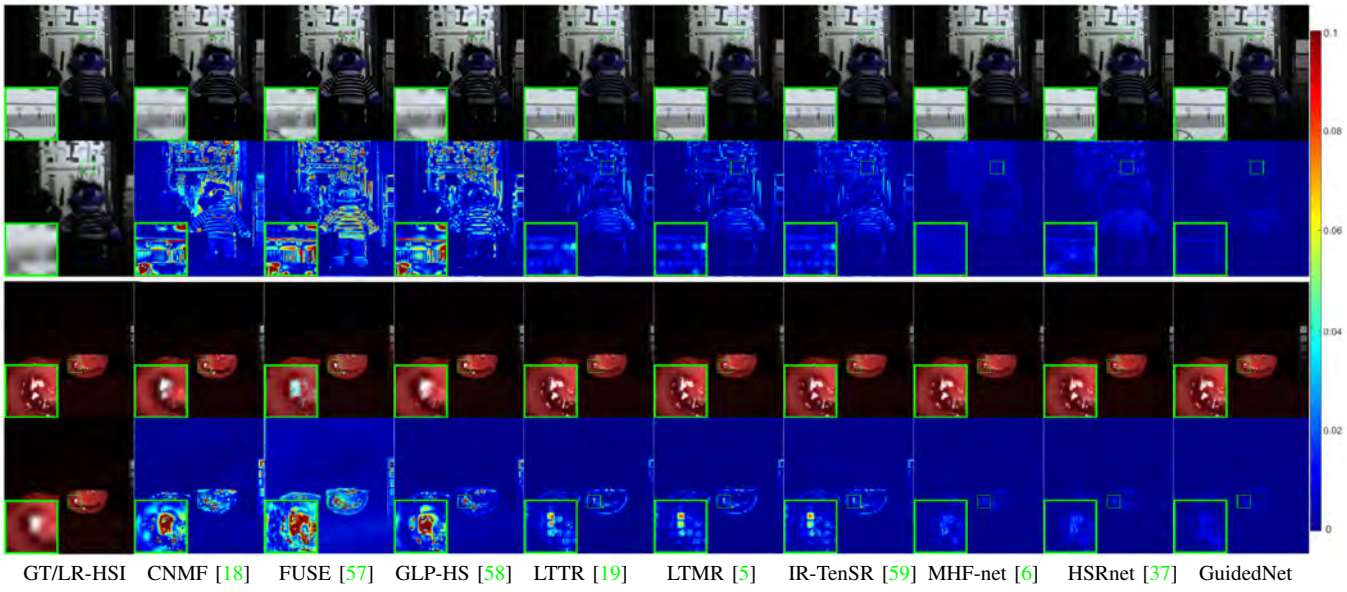


Fig. 7. The first column: the ground-truths and the corresponding LR-HSI images (in pseudo-colors) for the *chart and stuffed toy* (R-16, G-15, B-21) (1st-2nd rows) and the *fake and real tomatoes* (R-31, G-15, B-16) (3rd-4th rows) test cases from the CAVE dataset. The 2nd-8th columns: the visual results and the related error maps for all the compared approaches. A zoomed area has been added to aid the visual inspection.

related activation functions use the LeakyReLU with a slope of 0.2 when $x < 0$. In particular, the number of ERBs for a scale in the ResNet-ERB is 10, and the ERB structure is shown in Fig. 4(b).

Training data: We use the CAVE dataset [60] to train and test all the compared methods. This dataset consists of 32 hyperspectral images (HSIs) with a size of $512 \times 512 \times 31$ and corresponding RGB images with a size of $512 \times 512 \times 3$ (viewed as multispectral images, MSIs), which are generated by a general spectral response function $\mathbf{R}$ to simulate the Nikon D700 camera. This dataset has also been used in [5], [19], [31]. We selected 21 HSIs as the training set and 11 HSIs as the testing set. To reduce the storage cost, we crop the original HSIs (HR-HSIs) and MSIs (HR-MSIs) into sizes of $80 \times 80 \times 31$ and $80 \times 80 \times 3$, respectively. Then, we simulate LR-HSIs ($10 \times 10 \times 31$, scale = 8) by adopting a Gaussian blur with a kernel of 3×3 and a standard deviation of 0.5 to HR-HSIs and taking $8\times$ bicubic downsampling. Moreover, the involved intermediate ground-truth HSIs (GT-HSIs), $\tilde{\mathcal{X}}$, are also obtained by bicubic downsampling. Note that all the simulated LR-HSIs, HR-MSIs, HR-HSIs, and GT-HSIs from the 21 samples are divided into two parts: training set (90%) and validation set (10%).

Similar to the CAVE dataset, the Harvard dataset [61] consists of 77 hyperspectral images in indoor and outdoor scenes with a spatial resolution of 1024×1392 and 31 spectral bands with a size of $1024 \times 1392 \times 31$. We selected 20 images as training set. Moreover, we randomly selected 10 HSIs from the Harvard dataset cutting the upper left 1000×1000 side of the image as testing set. The data simulation process is the same as that of the CAVE dataset.

Training details: For fairness, all the DL-based approaches are implemented and trained in Tensorflow 1.8 framework on an NVIDIA GeForce GTX 2080Ti (11G RAM) and 2.90GHz Intel i5-9400F (32G Memory). The Adam Optimizer [62] trained our GuidedNet with a learning rate of 0.00001. Furthermore, the training epochs are set to 150, and the mini-batch size is 32. Fig. 6 plots the errors of the proposed GuidedNet on the training and validation datasets at each epoch separately, demonstrating its good convergence. For the other compared DL methods (*e.g.*, the MHF-net and HSRnet), we consider the available source codes for both training and testing, thus ensuring a fair comparison.

IV. Experiments

This section analyzes first the qualitative and quantitative performance of HISR. Then, extensive discussions on the super-resolution ability of the proposed GuidedNet are provided to the readers. After that, we extend the given method to a remote sensing fusion task, *i.e.*, the multispectral pansharpening. Finally, we show that the proposed network architecture can be viewed as a general framework that can enhance spatial resolution only if there is a high-resolution branch as guidance. Thus, the given framework is also extended to another super-resolution problem, *i.e.*, the single image super-resolution (SISR), adding HR guidance.

More in detail, we assess the performance of the proposed network by exploiting several state-of-the-art HISR methods, such as the fusion using coupled nonnegative matrix factorization unmixing (CNMF) [18], the fast fusion based on Sylvester equation (FUSE) approach [57], the generalized Laplacian pyramid (GLP) approach for hypersharpening (GLP-HS) [58], the low tensor-train rank (LTTR) based approach [19], the subspace-based low tensor multi-rank regularization (LTMR) approach [5], the iterative regularization based on tensor subspace representation (IR-TenSR) [59], the MS/HS fusion network (MHF-net) [6], and the HSRnet [37], on the CAVE dataset [60] and the Harvard dataset [61]. Four widely used quality indexes (QIs) for HISR are utilized to evaluate the

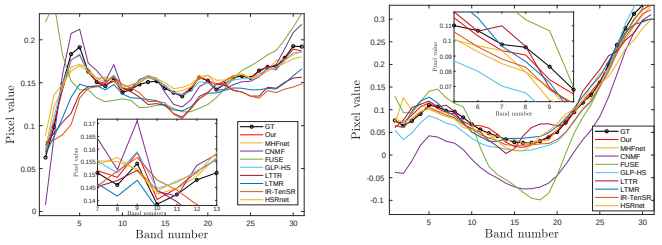
performance quantitatively, *i.e.*, the peak signal-to-noise ratio (PSNR), the spectral angle mapper (SAM [63]), and the erreur relative globale adimensionnelle de synthèse (ERGAS [64]), and the structure similarity (SSIM [65]). The higher the values of the PSNR and the SSIM, the better the performance. Conversely, the smaller the values of the SAM and the ERGAS, the better the performance. For a fair comparison, all the compared methods are tested on the same GPU or CPU (see the training details in Sect.III-D).

TABLE I
AVERAGE QUALITY INDICES WITH RELATED STANDARD DEVIATIONS OF THE RESULTS PROVIDED BY ALL THE COMPARED METHODS ON 11 TESTING IMAGES FROM THE CAVE DATASET. THE BEST RESULTS ARE HIGHLIGHTED.

Method	PSNR	SAM	ERGAS	SSIM
CNMF [18]	32.97±2.6	10.98±3.8	4.27±2.9	0.909±0.04
FUSE [57]	29.21±2.4	23.04±10.2	6.04±4.5	0.791±0.08
GLP-HS [58]	32.25±2.2	10.15±3.6	3.99±2.2	0.916±0.03
LTTR [19]	37.56±2.8	5.35±1.9	2.21±1.0	0.970±0.02
LTMR [5]	37.56±2.7	5.36±1.8	2.15±1.0	0.970±0.02
IR-TenSR [59]	37.58±2.7	7.44±2.7	2.12±0.9	0.959±0.02
MHF-net [6]	45.00±3.1	4.88±1.9	0.99±0.7	0.989±0.01
HSRnet [37]	44.88±3.5	**3.74**±1.4	0.98±0.6	0.991±0.00
GuidedNet	**45.41**±3.6	4.03±1.4	**0.97**±0.7	**0.991**±0.00

A. Experiments on CAVE Dataset

We conduct simulated experiments on the CAVE image dataset to verify the effectiveness of the proposed GuidedNet. We generate the HR-MSI by combining all the GT-HSI bands according to the spectral response function, $\mathbf{R}$. Then, we simulate the LR-HSI by downsampling the GT-HSI with a factor of 8. This process is described in Sect. IV-B. The testing dataset is formed by 11 hyperspectral images from the CAVE dataset with a size of $512 \times 512 \times 31$.



(a) *chart and stuffed toy* (b) *fake and real tomatoes*

Fig. 8. Spectral vectors analysis of the ground-truth (GT) and coming from the outcomes of the compared approaches for the *chart and stuffed toy* located at (251,255) and the *fake and real tomatoes* located at (230, 241).

The average quality indexes and the corresponding standard deviations, calculated on all the testing data, are shown in Tab. I. Tab. II reports the quality indexes for some specific test cases, *i.e.*, the *chart and stuffed toy* and the *fake and real tomatoes*, and the average running times on all the testing data. These tables clearly show that the GuidedNet outperforms the other methods, even requiring less computational burden. In Fig. 7, to show the visual comparison, we draw some pseudo-color images of the HSI super-resolution results and the corresponding error maps on the *chart and stuffed toy* and the *fake and real tomatoes* test cases, from the CAVE dataset. It can be observed from the error maps that CNMF, FUSE, GLP-HS, LTTR, LTMR, and IR-TenSR introduce artifacts. Conversely, the MHF-net and HSRnet, belonging to the DL class, perform better than the above-mentioned traditional methods but still performing unsatisfactorily on the reproduction of details. Instead, the residual map between our method and the GT image contains fewer errors for the compared approaches, thus showing a better spatial detail reconstruction. Another analysis is about spectral fidelity. In Fig. 8, to better compare the effects of spectral preservation, we plot the spectral vectors for all the compared approaches on the two mentioned test cases by fixing a pixel to provide this analysis. The spectral vectors generated by the proposed GuidedNet and the ground-truth are very similar, demonstrating the ability of GuidedNet to reduce spectral distortion.

TABLE II
QUALITY INDEXES VALUES AND THE AVERAGE RUNNING TIMES FOR THE COMPARED METHODS ON TWO TEST CASES FROM THE CAVE DATASET. G MEANS THAT THE METHOD EXPLOITS THE GPU, INSTEAD, C MEANS THAT IT EXPLOITS THE CPU. THE BEST RESULTS ARE HIGHLIGHTED.

Method	*chart and stuffed toy*				*fake and real tomatoes*				Time
	PSNR	SAM	ERGAS	SSIM	PSNR	SAM	ERGAS	SSIM	s
CNMF	30.35	9.23	2.80	0.936	41.54	6.38	12.68	0.964	14.5(C)
FUSE	29.14	12.53	3.33	0.890	38.65	7.80	8.41	0.967	4.1(C)
GLP-HS	29.52	8.33	3.02	0.930	38.32	6.33	8.21	0.974	5.2(C)
LTTR	35.45	6.03	1.62	0.964	42.50	5.53	3.69	0.987	1543.6(C)
LTMR	35.78	6.47	3.08	0.965	42.33	5.51	6.96	0.987	812.6(C)
IR-TenSR	35.80	6.22	2.94	0.964	42.55	5.60	4.64	0.987	211.9(C)
MHF-net	43.02	5.17	0.72	0.991	48.73	6.79	1.69	0.991	0.75(G)
HSRnet	43.13	4.95	0.74	0.992	48.65	5.07	1.66	0.994	0.31(G)
GuidedNet	**44.15**	**4.19**	**0.59**	**0.993**	**49.70**	**5.01**	**1.48**	**0.995**	**0.26**(G)

B. Experiments on Harvard Dataset

Tab. III reports the quality indexes and the corresponding standard deviations for all the compared methods on the Harvard testing data. We observe that the GuidedNet outperforms all the compared approaches considering the PSNR, SAM, and SSIM as metrics. For the ERGAS metric, GuidedNet ranks second. Again, we show the visual comparison displaying pseudo-color images and the related error maps on two specific test cases, see Fig. 9. The GuidedNet yields better visual results in agreement with the quantitative analysis.

C. Ablation Study

This section is about several ablation studies to assess the effectiveness of the GuidedNet, mainly concerning pixel shuffle, DFM, and loss function.

1) Pixel Shuffle

The GuidedNet employs pixel shuffle to upsample the LR-HSI to a larger image size. To verify the effectiveness of pixel shuffle compared with traditional methods, we change the upsampling of the recursive DFMs to deconvolution while keeping the remaining network structure to conduct comparative experiments on both CAVE and Harvard training datasets. The average quality indexes, shown in Tab. IV, demonstrate that the current setting is the best choice for the HISR task.

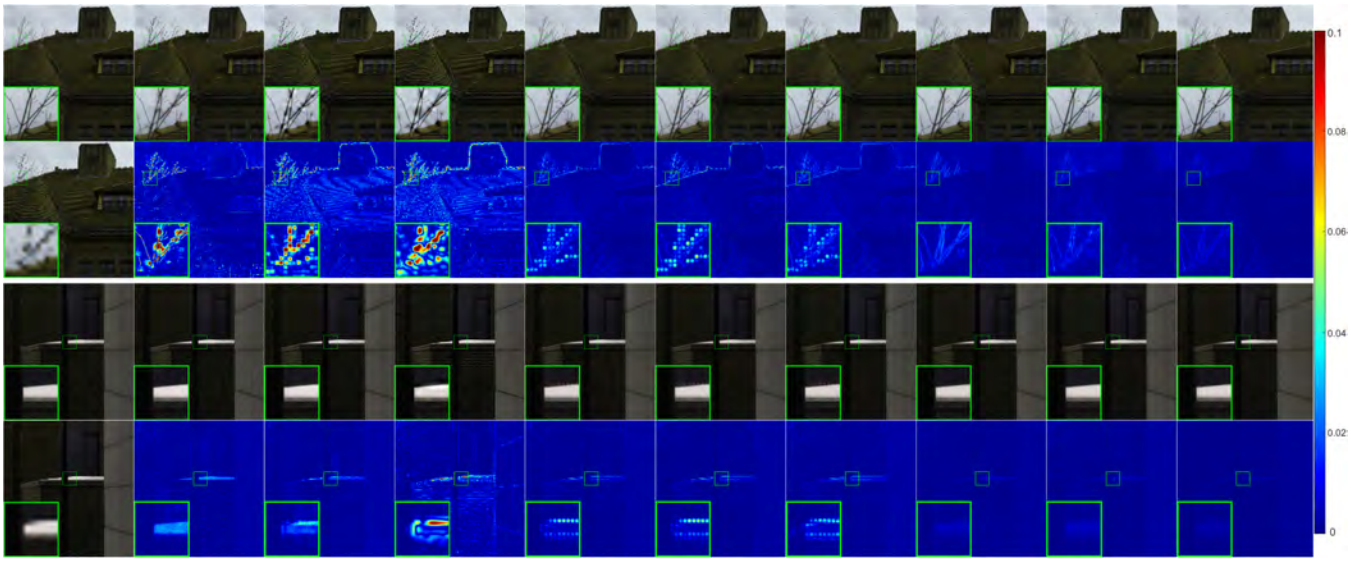


Fig. 9. The first column: the ground-truths and the corresponding LR-HSI images (in pseudo-colors) for the *house* (R-23, G-18, B-14) (1st-2nd rows) and the *fence* (R-20, G-21, B-13) (3rd-4th rows) test cases from the Harvard dataset. The 2nd-8th columns: the visual results and the related error maps for all the compared approaches. A zoomed area has been added to aid the visual inspection.

TABLE III
AVERAGE QUALITY INDEXES WITH THE RELATED STANDARD DEVIATIONS OF THE RESULTS PROVIDED BY ALL THE COMPARED METHODS ON 10 TESTING IMAGES FROM THE HARVARD DATASET. THE BEST RESULTS ARE HIGHLIGHTED.

Method	PSNR	SAM	ERGAS	SSIM
CNMF [18]	39.54±5.0	3.33±1.0	1.71±0.9	0.974±0.02
FUSE [57]	38.04±5.2	4.11±1.5	1.69±0.8	0.969±0.02
GLP-HS [58]	38.97±4.4	3.96±1.3	2.14±0.8	0.960±0.02
LTTR [19]	38.38±5.0	3.81±1.4	2.06±0.8	0.966±0.02
LTMR [5]	39.56±4.4	3.54±1.3	1.66±1.1	0.970±0.02
IR-TenSR [59]	38.84±4.9	3.97±1.5	1.87±0.7	0.966±0.02
MHF-net [6]	41.60±5.9	3.51±1.2	1.29±0.6	0.977±0.02
HSRnet [37]	41.52±6.1	2.96±1.0	**1.18**±0.4	0.980±0.02
GuidedNet	**41.64**±6.3	**2.85**±1.0	1.20±0.5	**0.981**±0.02

TABLE IV
AVERAGE QUALITY INDEXES WITH THE RELATED STANDARD DEVIATIONS OF THE RESULTS PROVIDED BY THE GUIDEDNET APPROACH USING DECONVOLUTION OR PIXEL SHUFFLE (PS) FOR LR-HSI UPSAMPLING.

CAVE				
Method	PSNR	SAM	ERGAS	SSIM
Deconv	44.87±3.6	4.17±1.4	1.01±0.7	0.991±0.01
PS	**45.41**±3.6	**4.03**±1.4	**0.97**±0.7	**0.991**±0.00
Harvard				
Method	PSNR	SAM	ERGAS	SSIM
Deconv	36.43±7.5	6.03±3.8	3.86±2.9	0.945±0.06
PS	**37.96**±6.8	**4.48**±2.0	**3.52**±2.5	**0.961**±0.03

TABLE V
AVERAGE QUALITY INDEXES WITH THE RELATED STANDARD DEVIATIONS OF THE RESULTS ON CAVE AND HARVARD DATASET BY OUR METHOD USING THE TRADITIONAL RESBLOCK (RB) AND THE PROPOSED EFFICIENT RESIDUAL BLOCK (ERB).

CAVE					
Method	PSNR	SAM	ERGAS	SSIM	Time(s)
RB	45.05±3.6	**3.93**±1.3	0.980±0.7	**0.992**±0.00	0.37
ERB	**45.41**±3.6	4.03±1.3	**0.969**±0.7	0.991±0.00	**0.26**
Harvard					
Method	PSNR	SAM	ERGAS	SSIM	Time(s)
RB	37.45±7.4	5.04±2.7	12.7±28.8	0.953±0.05	0.72
ERB	**37.96**±6.8	**4.48**±2.0	**3.52**±2.5	**0.961**±0.03	**0.50**

2) Efficient Residual Block

This section compares the proposed GuidedNet with the same network using a general residual block. We only replace the ERB structure with the general residual block, re-training it on the same training dataset and with the same settings. Tab. V shows the average running times and quality indexes on 11 testing CAVE images and 10 Harvard testing images. It is clear that the ERB can significantly reduce the computational burden and improve the performance.

3) Multi-scale Loss

In this section, we investigate the role of the weights, α_k, in the loss function. We set some weights for the loss function, then re-training the network and obtaining the results on the CAVE dataset. A set of weights is tested on the same training set. We will use the following notation: (w_1, w_2, w_3), where we have three layers with three different weights, *i.e.*, w_1, w_2, and w_3. w_1 is related to the first (initial) layer, w_2 is about the second (intermediate) layer, and w_3 refers to the final layer. For instance, if the $\{\alpha_k\}_{k=1,\dots,3}$ are set to $(0, 0, 1)$, no initial and intermediate multi-scale losses are considered in the loss function. The average quality indexes are in Tab. VI. We can note that when we have that $\{\alpha_k\}_{k=1,\dots,3} = (4, 2, 1)$, the proposed method produces the best results avoiding instability caused by too low or too high weights.

4) PMS and DDS

We modify the network to verify the validity of the FRB's progressive multi-scale structure (PMS) and the dual data

stream (DDS). More specifically, the network uses only one DFM with a scaling factor of 8 and does not downsample multi-scale MSIs to obtain a network without PMS (w/o PMS). Note that the GuidedNet (w/o PMS) requires an 8x upsampling, and we expanded the final convolution kernel size to 7 to avoid significant degradation in performance. The DDS in the FRB is set to a single data stream by modifying the DFM to make the input and output only having one HSI. The experimental results using the same hyperparameters are shown in Tab. VII. The results in the table indicate that the performance significantly drops when we remove the PMS. Moreover, removing the DDS structure results in a performance reduction. The complete GuidedNet, holding both the structures, yields the best outcome demonstrating the importance of the PMS and the DDS structure in the proposed GuidedNet.

TABLE VI
AVERAGE QUALITY INDEXES WITH THE RELATED STANDARD DEVIATIONS OF THE RESULTS ON THE CAVE DATASET USING DIFFERENT WEIGHTS CONFIGURATIONS. THE BEST RESULTS ARE HIGHLIGHTED.

$\{\alpha_k\}_{k=1,\dots,3}$	PSNR	SAM	ERGAS	SSIM
$(0,0,1)$	44.55±8.9	3.97±1.2	1.06±0.8	**0.991**±0.00
$(1,2,4)$	43.55±4.5	3.96±1.1	1.25±1.2	0.990±0.01
$(1,1,1)$	45.09±3.7	**3.82**±1.2	0.984±0.7	**0.991**±0.00
$(16,4,1)$	44.93±3.5	4.31±1.5	1.01±0.7	0.990±0.00
$(4,2,1)$	**45.41**±3.6	4.03±1.3	**0.969**±0.7	**0.991**±0.00

TABLE VII
THE EFFECTS OF THE PMS AND THE DDS IN THE PROPOSED GUIDEDNET ON THE CAVE DATASET. THE BEST RESULTS ARE HIGHLIGHTED.

Method	PSNR	SAM	ERGAS	SSIM
w/o PMS	43.57±4.9	4.43±1.3	1.38±1.2	0.989±0.01
w/o DDS	44.38±3.9	4.26±1.5	1.22±0.9	0.990±0.00
GuidedNet	**45.41±3.6**	**4.03±1.3**	**0.97±0.7**	**0.991±0.00**

D. Comparison with DL-based methods

In this section, the two DL-based HISR methods are compared in more detail giving information about some aspects, such as network generalization and complexity.

1) Network Generalization

Network generalization is crucial to demonstrate the effectiveness of data-driven approaches. Thus, this section investigates the network generalization ability of the MHF-net, the HSRnet, and the GuidedNet. All the approaches are trained on the CAVE training set and then tested on the Harvard testing set. Tab. VIII reports the average quality indexes and standard deviations. The proposed method outperforms the other methods considering the PSNR and SSIM metrics, while HSRnet shows advantages referring to the SAM and ERGAS metrics.

2) Network Complexity

Tab. IX shows the network parameters number, the floating point operations (FLOPs), and the training times of the three compared approaches. It is easily remarked that the GuidedNet has fewer parameters and computations with respect to the MHF-net and HSRnet. Furthermore, the GuidedNet takes less training time, and, as discussed earlier, the average testing times of the GuidedNet on both the CAVE and the Harvard datasets are shorter. Moreover, we also compare the hardware consumption and training times of the GuidedNet without PMS and DDS. From Tab. IX, it can be seen that PMS and DDS can improve performance without significantly increasing hardware consumption.

TABLE VIII
AVERAGE QUALITY INDEXES WITH THE RELATED STANDARD DEVIATIONS OF THE PROPOSED GUIDENET AND THE MHF-NET TRAINED ON THE CAVE TRAINING SET AND TESTED ON 10 TESTING IMAGES FROM THE HARVARD DATASET. THE BEST RESULTS ARE HIGHLIGHTED.

Method	PSNR	SAM	ERGAS	SSIM
MHF-net [6]	37.24±7.5	6.21±3.9	17.27±39.84	0.943±0.06
HSRnet [37]	37.85±7.2	**4.35**±1.7	**3.48**±1.5	0.958±0.05
GuidedNet	**37.96**±6.8	4.48±2.0	3.52±2.5	**0.961**±0.03

TABLE IX
PARAMETER AMOUNT AND FLOATING-POINT OPERATIONS PER SECOND (FLOPS) OF THE GUIDEDNET, THE MHF-NET AND THE HSRNET.

Method	# Params.	FLOPs	Training time
MHF-net [6]	2.03M	53.27G	$14.9 \times 10^4 s$
HSRnet [37]	1.98M	45.33G	$2.6 \times 10^4 s$
GuidedNet (w/o PMS)	0.81M	34.68G	$2.2 \times 10^4 s$
GuidedNet (w/o DDS)	0.69M	32.79G	$2.1 \times 10^4 s$
GuidedNet	0.70M	35.31G	$2.1 \times 10^4 s$

TABLE X
AVERAGE QUALITY INDEXES WITH THE RELATED STANDARD DEVIATIONS OF THE RESULTS PROVIDED BY THE ALL METHODS ON 11 TESTING IMAGES FROM THE CAVE DATASET WITH A SCALING FACTOR OF 4×. THE BEST RESULTS ARE HIGHLIGHTED.

Method	PSNR	SAM	ERGAS	SSIM
CNMF [18]	41.59±2.9	8.10±3.4	3.99±3.2	0.972±0.02
FUSE [57]	39.71±3.5	5.83±2.0	4.19±3.1	0.975±0.02
GLP-HS [58]	37.81±3.1	5.36±1.8	4.66±2.7	0.972±0.01
LTTR [19]	36.76±2.8	6.60±2.5	5.65±2.8	0.957±0.03
LTMR [5]	36.19±2.7	7.66±2.8	5.70±2.7	0.949±0.03
IR-TenSR [59]	36.38±2.6	8.7±3.0	5.52±2.6	0.948±0.03
MHF-net [6]	46.27±2.7	4.33±1.8	1.74±1.2	0.992±0.00
HSRnet [37]	**47.71**±2.7	**2.95**±1.0	**1.39**±0.8	0.994±0.00
GuidedNet	47.64±3.2	3.29±1.2	1.47±1.0	**0.994**±0.00

3) Results on Different Scaling Factors

By changing the number of the recursive DFMs, the proposed GuidedNet can easily reach any super-resolution scaling factor power of 2. In the previous experiments, we tested the performance of HISR on a scaling factor equal to 8. This section investigates fusion performance varying the scaling factors (*e.g.*, 4, 8, 16, and 32). Concerning the data simulation, we only need to change the scaling factor of the downsampling while keeping the other network settings unchanged. In Tab. X, we compare the performance of the used benchmark exploiting a scaling factor of 4 and measuring an average and the corresponding standard deviations for all the quality indexes.

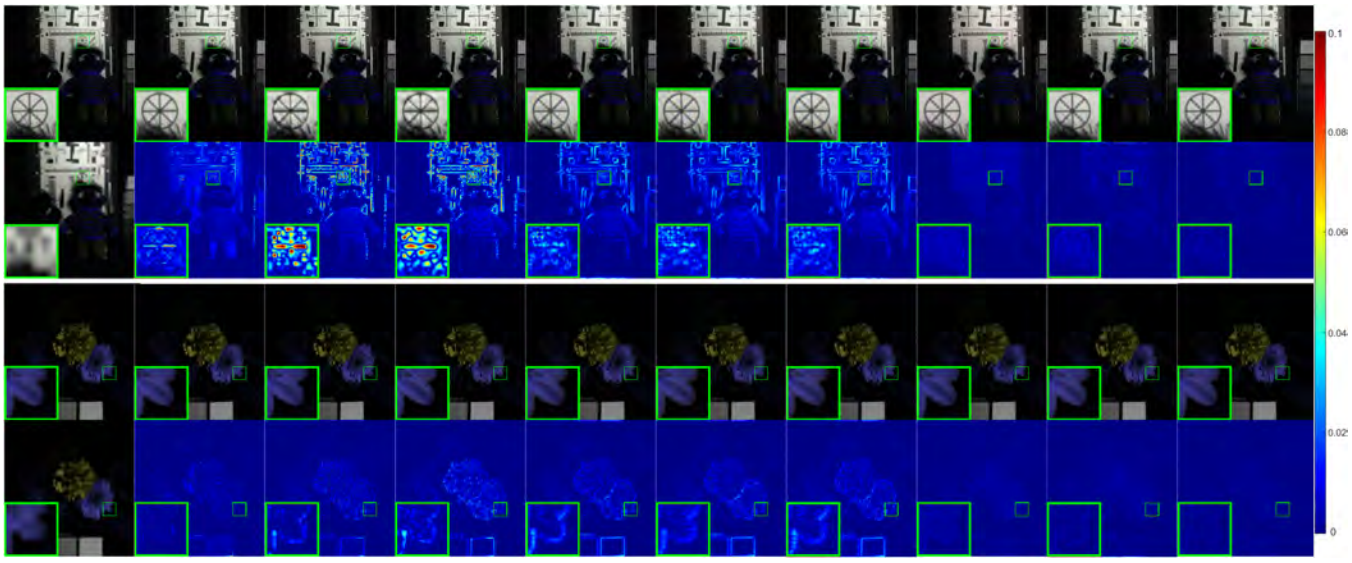


Fig. 10. HISR with a scaling factor 4×. The first column: the ground-truths and the corresponding LR-HSI images (in pseudo-colors) for the *chart and stuffed toy* (R-2, G-11, B-5) (1st-2nd rows) and the *flowers* (R-27, G-21, B-26) (3rd-4th rows) test cases from the CAVE dataset. The 2nd-8th columns: the visual results and the related error maps for all the compared approaches. A zoomed area has been added to aid the visual inspection.

This table shows that the HSRnet obtains the best results on PSNR, SAM, and ERGAS, and our GuidedNet gets the best SSIM. All the traditional approaches show a significant gap comparing them with the two DL-based methods (*i.e.*, the MHF-net and the HSRnet) and the GuidedNet.

TABLE XI
AVERAGE QUALITY INDEXES WITH THE RELATED STANDARD DEVIATIONS OF THE RESULTS PROVIDED BY THE MHF-NET, THE HSRNET, AND THE GUIDEDNET FOR 11 TESTING IMAGES ON THE CAVE DATASET CONSIDERING AS SCALING FACTORS 16× AND 32×.

16×				
Method	PSNR	SAM	ERGAS	SSIM
MHF-net [6]	43.33±3.5	5.58±1.9	0.627±0.45	0.987±0.01
HSRnet [37]	42.82±3.8	5.05±1.8	0.691±0.44	0.984±0.01
GuidedNet	**43.39**±3.9	**4.70**±1.5	**0.605**±0.47	**0.989**±0.01
32×				
Method	PSNR	SAM	ERGAS	SSIM
MHF-net [6]	41.91±4.0	**6.23**±2.1	0.371±0.30	**0.985**±0.01
HSRnet [37]	39.64±2.8	6.85±2.4	0.507±0.34	0.969±0.02
GuidedNet	**41.98**±3.5	6.66±2.3	**0.336**±0.21	0.984±0.01

Moreover, we also depicted the corresponding pseudo-color images of the HISR outputs in Fig.10. We can see that the proposed GuidedNet obtains fewer residuals than the other approaches demonstrating its effectiveness. Finally, Tab. XI also reports the quantitative outcomes of the three DL-based methods on larger scaling factors, *i.e.,* 16 and 32. Some other traditional methods cannot achieve the task of 32×, or codes are not runnable on larger scale factors. Thus, we here only add a comparison with MHF-net and HSRnet since they can be run on larger scale factors and are also DL-based methods. From Tab. XI, it is clear that the GuidedNet approach shows competitive performance in these other configurations demonstrating a good adaptation for addressing diverse scale fusion problems. Compared to the scale factor of 4, the performance of HSRnet decreases significantly as the scale factor increases because of the used upsampling strategy.

TABLE XII
AVERAGE QUALITY INDEXES WITH THE RELATED STANDARD DEVIATIONS OF THE PANSHARPENING RESULTS PROVIDED BY DIFFERENT METHODS FOR 1258 TESTING IMAGES ON THE WORLDVIEW-3 DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLDFACE.

Method	SAM	ERGAS	SCC	Q8
PNN [48]	4.00±1.3	2.72±1.0	0.962±0.05	0.908±0.11
PanNet [47]	4.09±1.3	2.95±1.0	0.949±0.05	0.894±0.11
DMDnet [49]	3.97±1.2	2.86±1.0	0.953±0.04	0.900±0.11
FusionNet [52]	3.74±1.2	2.57±0.9	0.958±0.05	0.914±0.11
GuidedNet	**3.50**±1.2	**2.39**±0.9	**0.963**±0.04	**0.922**±0.10

E. Extension to Other Applications

As mentioned before, the GuidedNet is a general fusion framework that can effectively fuse an LR input with HR guidance to reach a higher resolution. Thanks to the proposed general paradigm, we can extend the GuidedNet to other resolution enhancement tasks when there is a HR guidance. In what follows, we apply GuidedNet to two image resolution enhancement problems, *i.e.*, remote sensing pansharpening and single image super-resolution (SISR).

1) Pansharpening

Pansharpening is about fusing a low-resolution multispectral image (LR-MSI) and a panchromatic (PAN) image with high spatial resolution, aiming to obtain an HR-MSI with the exact spatial resolution as the PAN image. More details about pansharpening can be found in a recent review literature [66]. The pansharpening task shares some similarities with the MSI/HSI fusion task. Therefore, following the MSI/HSI fusion framework of the GuidedNet, we only need to replace the HR-MSI in Fig. 2 with the PAN image and substitute the

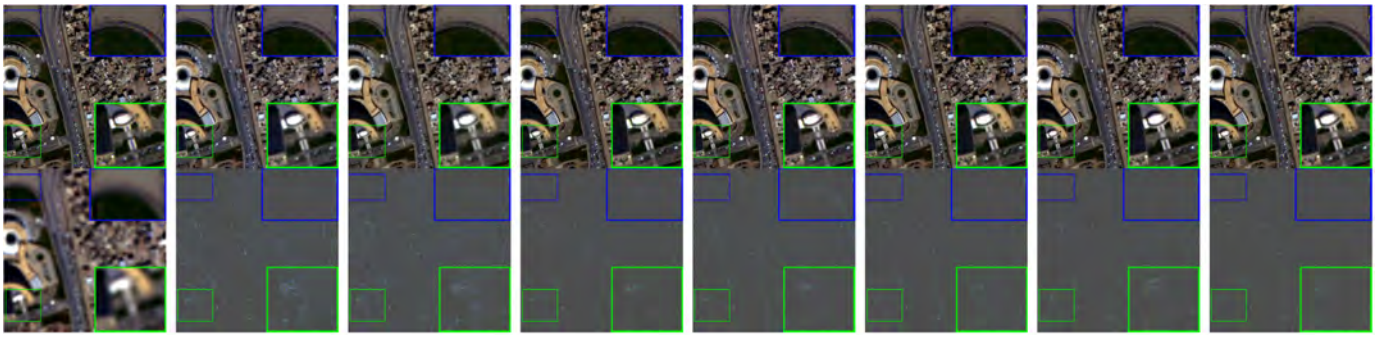


Fig. 11. Visual comparison of pansharpening products on the *Rio* datasets acquired by the WorldView-3 sensor. Two zoomed areas have been added to aid the visual inspection.

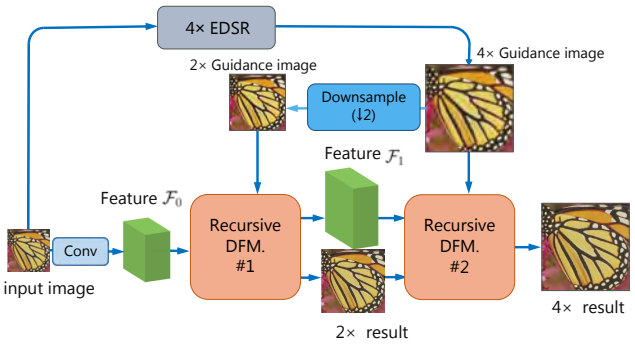


Fig. 12. The extended architecture of the GuidedNet for the 4× SISR task. EDSR has been pre-trained and its results are directly used as input.

TABLE XIII
QUALITY INDEXES OF THE RESULTS PROVIDED BY THE BENCHMARK ON THE SET5 DATASET FOR SISR WITH SCALING FACTOR OF 4×.

Set5			
Method	bicubic PSNR/SSIM	EDSR [40] PSNR/SSIM	GuidedNet PSNR/SSIM
baby	31.83/0.858	32.54/0.869	**33.64/0.892**
bird	30.05/0.870	31.34/0.898	**34.00/0.934**
butterfly	22.15/0.734	23.54/0.801	**27.30/0.901**
head	32.67/0.754	32.70/0.764	**32.91/0.794**
woman	26.44/0.831	27.56/0.861	**30.02/0.907**
average	28.43/0.810	29.54/0.839	**31.57/0.886**

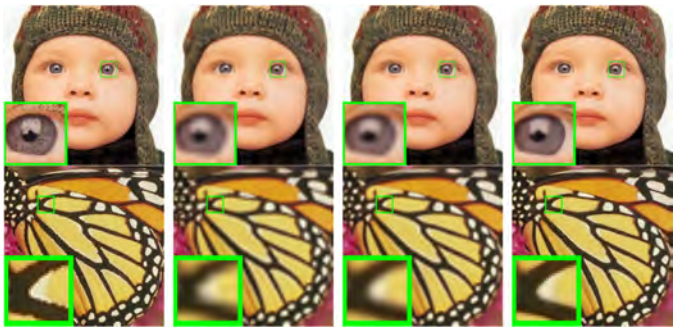


Fig. 13. Results provided by the benchmark with a scaling factor of 4 for the *baby* and the *butterfly* test cases from the Set5 dataset. A zoomed area has been added to aid the visual inspection.

LR-HSI in Fig. 2 with the LR-MSI. It is worth noting that the scaling factor for pansharpening is often 4 (at least for the primary adopted sensors). Thus, we reduced the number of recursive DFMs to 2. We employed an 8-band multispectral dataset acquired by the WorldView-3 (WV-3) sensor for the training. The process of building training and testing data is described in [52]. Thus, we have 8806 PAN (64×64), LR-MSI ($16\times16\times8$), and HR-MSI ($64\times64\times8$) image patch pairs as training set. For the sake of brevity, we do not introduce details about the data used. Readers can refer to [52] and [67] for more information. Moreover, the quality of the fusion results is evaluated using the spectral angle mapper (SAM) [63], the erreur relative globale adimensionnelle de synthèse (ERGAS) [64], the spatial correlation coefficient (SCC) [68], and the universal image quality index for eight-band images (Q8) [69].

For this application, we compare our approach with four state-of-the-art DL-based pansharpening methods, *i.e.*, PNN [48], PanNet [47], DMDnet [49], and FusionNet [52]. Tab. XII reports the outcomes for all the compared approaches on 1258 randomly selected training samples. From the average QIs and the related standard deviations shown in Tab. XII, it is clear that the GuidedNet yields the best quantitative performance on all the indicators, *i.e.*, SAM, ERGAS, SCC, and Q8. Besides, for qualitative comparison, we augment the benchmark even including a CS-based approach, *i.e.*, the BDSD-PC, and an MRA-based method, *i.e.*, the GLP-Reg. Fig. 11 depicts the results on WV-3, indicating that the image reconstructed by the proposed method is more precise than the comparison methods. The GuidedNet's satisfactory results on pansharpening demonstrate its ability to address different tasks.

2) Single Image Super-Resolution

The GuidedNet fusion framework can also be extended to the single image super-resolution (SISR) problem. However, the proposed fusion framework requires high-resolution guidance to enhance the resolution. Instead, the SISR has a unique input, the LR image. Therefore, we need to introduce high-resolution guidance into the framework. Here, we use the outcome of a competitive SISR method, *i.e.*, the DL-based SISR approach EDSR [40], to replace the HR-MSI in the HGB in Fig. 2. Fig. 12 depicts the GuidedNet structure for the

$4\times$ SISR task modified by the addition of EDSR. EDSR has been pre-trained using the DIV2K dataset by Adam optimizer, and we only utilize its testing outcomes as input in the HGB branch of our GudedNet. It is not necessary to re-train the EDSR again in GuidedNet. The training parameters are the default ones in [40], *i.e.*, the batch size is set to 16, and the learning rate is initialized to 0.0001 halved at every 1×10^6 batch updates. The scaling factor is 4. Thus, the GuidedNet for SISR requires two recursive DFMs. The proposed approach is again trained using the DIV2K dataset [70].

After ending the training of the GuidedNet, the trained network is evaluated on the Set5 testing dataset. Tab. XIII reports the average PSNR and SSIM values of the bicubic interpolation, the state-of-the-art EDSR, and the GuidedNet. The proposed method can significantly improve the results of EDSR, even outperforming the classical bicubic interpolator. By looking at the visual comparison shown in Fig. 13, the GuidedNet approach holds sharper details, especially comparing them with the ones of the baseline method, EDSR. In this case, we consider the outcome of the EDSR method into the HGB (viewed as a plug-in module). However, we can take any baseline SISR method into our GuidedNet framework to enhance the SR performance.

V. Conclusions

This paper proposed a general CNN fusion framework, GuidedNet, to deal with the HISR problem thanks to high-resolution guidance. Motivated by the specific problem (*i.e.*, the HISR), this framework has been formulated using two branches: the HGB and the FRB. Besides, by considering some strategies, such as the recursive mechanism and the progressive technique, the proposed GuidedNet can significantly reduce the network parameters getting high-quality outcomes. Extensive experiments on several hyperspectral image datasets demonstrate the superiority of the proposed GuidedNet, comparing it with recent state-of-the-art approaches. Furthermore, discussions about several aspects, such as network generalization, network complexity, robustness with respect to variations of scaling factors, and time comparison, have been provided to the readers. Finally, the proposed fusion framework has been easily extended to other resolution enhancement tasks, *i.e.*, remote sensing pansharpening and SISR.

References

[1] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Trans. Geosci. Remote Sens.*, vol. 1, no. 2, pp. 6–36, 2013.

[2] M. Bergeron, A. Hollinger, K. Staenz, M. Maszkiewicz, R. A. Neville, S. E. Qian, and D. G. Goodenough, "Hyperspectral environment and resource observer (HERO) mission," *Canadian Journal of Remote Sensing*, vol. 34, no. sup1, pp. S1–S11, 2008.

[3] H. Yuan and Y. Y. Tang, "Spectral–spatial shared linear regression for hyperspectral image classification," *IEEE Trans. Cyber.*, vol. 47, no. 4, pp. 934–945, 2017.

[4] M. Uzair, A. Mahmood, and A. Mian, "Hyperspectral face recognition using 3d-dct and partial least squares," in *British Machine Vis. Conf. (BMVC)*, 2013.

[5] 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, 2019.

[6] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "Mhf-net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1457–1473, 2022.

[7] T.-X. Jiang, T.-Z. Huang, X.-L. Zhao, L.-J. Deng, and Y. Wang, "A novel tensor-based video rain streaks removal approach via utilizing discriminatively intrinsic priors," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 2818–2827.

[8] J. Jiang, J. Ma, C. Chen, X. Jiang, and Z. Wang, "Noise robust face image super-resolution through smooth sparse representation," *IEEE Trans. Cyber.*, vol. 47, no. 11, pp. 3991–4002, 2017.

[9] T.-X. Jiang, T.-Z. Huang, X.-L. Zhao, L.-J. Deng, and Y. Wang, "Fastderain: A novel video rain streak removal method using directional gradient priors," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2089–2102, 2019.

[10] L.-J. Deng, W. Guo, and T.-Z. Huang, "Single-image super-resolution via an iterative reproducing kernel hilbert space method," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 26, no. 11, pp. 2001–2014, 2016.

[11] Y. Chang, L. Yan, X.-L. Zhao, H. Fang, Z. Zhang, and S. Zhong, "Weighted low-rank tensor recovery for hyperspectral image restoration," *IEEE Trans. Cyber.*, vol. 50, no. 11, pp. 4558–4572, 2020.

[12] L.-J. Deng, G. Vivone, W. Guo, M. Dalla 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, 2018.

[13] Z.-C. Wu, T.-Z. Huang, L.-J. Deng, G. Vivone, J.-Q. Miao, J.-F. Hu, and X.-L. Zhao, "A new variational approach based on proximal deep injection and gradient intensity similarity for spatio-spectral image fusion," *IEEE Jour. Selec. Topics Applied Earth Obser. & Remote Sens.*, vol. 13, pp. 6277–6290, 2020.

[14] P. Guo, P. Zhuang, and Y. Guo, "Bayesian pan-sharpening with multi-order gradient-based deep network constraints," *IEEE Jour. Selec. Topics Applied Earth Obser. & Remote Sens.*, vol. 13, pp. 950–962, 2020.

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

[16] 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. Cyber.*, vol. 50, no. 10, pp. 4469–4480, 2020.

[17] Z. H. Nezhad, A. Karami, R. Heylen, and P. Scheunders, "Fusion of hyperspectral and multispectral images using spectral unmixing and sparse coding," *IEEE Jour. Selec. Topics Applied Earth Obser. & Remote Sens.*, vol. 9, no. 6, pp. 2377–2389, 2016.

[18] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, 2012.

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

[20] 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.

[21] R. Dian, S. Li, and X. Kang, "Regularizing hyperspectral and multispectral image fusion by cnn denoiser," *IEEE Trans. Neural Net. Learn. Syst.*, vol. 32, no. 3, pp. 1124–1135, 2021.

[22] 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 Jour. Selec. Topics Applied Earth Obser. & Remote Sens.*, vol. 13, pp. 4381–4398, 2020.

[23] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2599–2613, 2019.

[24] R. Lan, L. Sun, Z. Liu, H. Lu, Z. Su, C. Pang, and X. Luo, "Cascading and enhanced residual networks for accurate single-image super-resolution," *IEEE Trans. Cyber.*, vol. 51, no. 1, pp. 115–125, 2021.

[25] Y. Zhou, X. Du, M. Wang, S. Huo, Y. Zhang, and S.-Y. Kung, "Cross-scale residual network: A general framework for image super-resolution, denoising, and deblocking," *IEEE Trans. Cyber.*, pp. 1–13, 2021.

[26] C. Ren, X. He, Y. Pu, and T. Q. Nguyen, "Learning image profile enhancement and denoising statistics priors for single-image super-resolution," *IEEE Trans. Cyber.*, vol. 51, no. 7, pp. 3535–3548, 2021.

[27] X. Liu, L. Li, F. Liu, B. Hou, S. Yang, and L. Jiao, "Gafnet: Group attention fusion network for pan and ms image high-resolution classification," *IEEE Trans. Cyber.*, pp. 1–14, 2021.

[28] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering,"

IEEE transactions on neural networks and learning systems, vol. 29, no. 12, pp. 5947–5959, 2018.

[29] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE Trans. Neural Net. Learn. Syst.*, pp. 1–11, 2018.

[30] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-D-convolutional neural network," *IEEE Geosci. and Remote Sens. Letters*, vol. 14, no. 5, pp. 639–643, 2017.

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

[32] X.-H. Han, Y. Zheng, and Y.-W. Chen, "Multi-level and multi-scale spatial and spectral fusion CNN for hyperspectral image super-resolution," in *Int. Conf. Comput. Vis. worksh. (ICCVW)*, 2019, pp. 4330–4339.

[33] Z. Zhu, J. Hou, J. Chen, H. Zeng, and J. Zhou, "Hyperspectral image super-resolution via deep progressive zero-centric residual learning," *IEEE Trans. Image Process.*, vol. PP, pp. 1–1, 12 2020.

[34] F. Zhou, R. Hang, Q. Liu, and X. Yuan, "Pyramid fully convolutional network for hyperspectral and multispectral image fusion," *IEEE Jour. Selec. Topics Applied Earth Obser. & Remote Sens.*, vol. 12, no. 5, pp. 1549–1558, 2019.

[35] S. Xu, O. Amira, J. Liu, C.-X. Zhang, J. Zhang, and G. Li, "Ham-mfn: Hyperspectral and multispectral image multiscale fusion network with rap loss," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4618–4628, 2020.

[36] K. Zheng, L. Gao, W. Liao, D. Hong, B. Zhang, X. Cui, and J. Chanussot, "Coupled convolutional neural network with adaptive response function learning for unsupervised hyperspectral super resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2487–2502, 2021.

[37] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 7251–7265, 2022.

[38] G. Vivone, "Multispectral and hyperspectral image fusion in remote sensing: A survey," *Information Fusion*, vol. 89, pp. 405–417, 2023.

[39] C. I. Kanatsoulis, X. Fu, N. D. Sidiropoulos, and W.-K. Ma, "Hyperspectral super-resolution via coupled tensor factorization: Identifiability and algorithms," in *IEEE Int. Conf. Acoustics, Speech & Sign. Process. (ICASSP)*, 2018, pp. 3191–3195.

[40] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *IEEE Conf. Comput. Vis. Pattern Recog. Worksh. (CVPRW)*, 2017, pp. 136–144.

[41] K. Zeng, J. Yu, R. Wang, C. Li, and D. Tao, "Coupled deep autoencoder for single image super-resolution," *IEEE Trans. Cyber.*, vol. 47, no. 1, pp. 27–37, 2017.

[42] T. Wang, F. Fang, H. Zheng, and G. Zhang, "Frmlnet: Framelet-based multilevel network for pansharpening," *IEEE Trans. Cyber.*, pp. 1–12, 2021.

[43] A. Garzelli, F. Nencini, and L. Capobianco, "Optimal mmse pan sharpening of very high resolution multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 228–236, 2008.

[44] G. Vivone, "Robust band-dependent spatial-detail approaches for panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6421–6433, 2019.

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

[46] 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, 2018.

[47] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 5449–5457.

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

[49] X. Fu, W. Wang, Y. Huang, X. Ding, and J. Paisley, "Deep multiscale detail networks for multiband spectral image sharpening," *IEEE Trans. Neural Net. Learn. Syst.*, 2020.

[50] L. He, Y. Rao, J. Li, J. Chanussot, A. Plaza, J. Zhu, and B. Li, "Pansharpening via detail injection based convolutional neural networks," *IEEE Jour. Selec. Topics Applied Earth Obser. & Remote Sens.*, vol. 12, no. 4, pp. 1188–1204, 2019.

[51] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multispectral image pansharpening by learning a deep residual network," *IEEE Geosci. and Remote Sens. Letters*, vol. 14, no. 10, pp. 1795–1799, 2017.

[52] L.-J. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–16, 2020.

[53] Z.-C. Wu, T.-Z. Huang, L.-J. Deng, J.-F. Hu, and G. Vivone, "VO+Net: An adaptive approach using variational optimization and deep learning for panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–16, 2021.

[54] 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, 2019.

[55] T.-W. Hui, C. C. Loy, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 353–369.

[56] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 4700–4708.

[57] 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, 2015.

[58] M. Selva, B. Aiazzi, F. Butera, L. Chiarantini, and S. Baronti, "Hyper-sharpening: A first approach on SIM-GA data," *IEEE Jour. Selec. Topics Applied Earth Obser. & Remote Sens.*, vol. 8, no. 6, pp. 3008–3024, 2015.

[59] T. Xu, T.-Z. Huang, L.-J. Deng, and N. Yokoya, "An iterative regularization method based on tensor subspace representation for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[60] 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, p. 2241, 2010.

[61] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2011.

[62] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Int. Conf. on Learning Representations (ICLR)*, 12 2014.

[63] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *JPL Airborne Geosci. Worksh.*, vol. 1, 1992, pp. 147–149.

[64] L. Wald, *Data Fusion. Definitions and Architectures - Fusion of Images of Different Spatial Resolutions*. Presses des MINES, 2002.

[65] 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, 2004.

[66] G. Vivone, M. Dalla Mura, A. Garzelli, R. Restaino, G. Scarpa, M. O. Ulfarsson, L. Alparone, and J. Chanussot, "A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods," *IEEE Geosci. and Remote Sensing Magazine*, vol. 9, no. 1, pp. 53–81, 2021.

[67] L.-J. Deng, G. Vivone, M. E. Paoletti, G. Scarpa, J. He, Y. Zhang, J. Chanussot, and A. Plaza, "Machine learning in pansharpening: A benchmark, from shallow to deep networks," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 3, pp. 279–315, 2022.

[68] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge landsat TM and SPOT panchromatic data," *Int. Journal of Remote Sens.*, vol. 19, no. 4, pp. 743–757, 1998.

[69] A. Garzelli and F. Nencini, "Hypercomplex quality assessment of multi/hyperspectral images," *IEEE Geosci. and Remote Sens. Letters*, vol. 6, no. 4, pp. 662–665, 2009.

[70] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *IEEE Conf. Comput. Vis. Pattern Recog. Worksh. (CVPRW)*, July 2017.