Pan-Sharpening via Deep Locally Linear Embedding Residual Network

Jiaming Wang, Zhenfeng Shao, Xiao Huang, Tao Lu, *Member, IEEE*, Ruiqian Zhang, and Gui Cheng

***Abstract*— The goal of pan-sharpening tasks is to fuse panchromatic (PAN) images and low-spatial-resolution (LR) multispectral (MS) images for the purpose of aggregating texture and spectral information. Although traditional embedding-based pan-sharpening methods achieve competitive results, they are limited by the shallow network and not suitable for large-scale datasets. In this study, we design a novel multiscale locally linear embedding residual network (LLERN) that consists of two phases: the spectral preservation phase and the structural preservation phase. As the pretreatment of the structural preservation network, the spectral preservation network aims to upscale the LR MS image while retaining spectral information. The proposed locally linear embedding residual block (LLERB) in the structural preservation phase can search for similar sparse patches from the PAN image space and embed the corresponding local geometric relationship into the residual space to enhance the MS image. Extensive experiments suggest that the proposed LLERN outperforms state-of-the-art methods from visual and quantitative perspectives, and confirm the assumption that LR image patches and residual image patches in a local region share a similar manifold structure, which can be used to guide deep-learning modeling with improved interpretability. The source code is available at https://github.com/jiaming-wang/LLERN.**

***Index Terms*— Convolutional neural network (CNN), deep learning, locally linear embedding, pan-sharpening.**

Manuscript received February 5, 2022; revised March 26, 2022; accepted April 16, 2022. Date of publication April 18, 2022; date of current version May 5, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42090012 and Grant 62072350, in part by the Special Fund of the Hubei Luojia Laboratory under Grant 220100009, in part by the 03 Special Research and 5G Project of Jiangxi Province in China under Grant 20212ABC03A09, in part by the Zhuhai Industry University Research Cooperation Project of China under Grant ZH22017001210098PWC, in part by the Key Research and Development Project of the Sichuan Science and Technology Plan under Grant 2022YFN0031, in part by the Zhizhuo Research Fund on Spatial-Temporal Artificial Intelligence under Grant ZZJJ202202, and in part by the Opening Fund of the Hubei Key Laboratory of Intelligent Robot under Grant HBIR202103. *(Corresponding author: Zhenfeng Shao.)*

Jiaming Wang and Gui Cheng are with the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: wjmecho@whu.edu.cn; chenggui@whu.edu.cn).

Zhenfeng Shao is with the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing and the Hubei Luojia Laboratory, Wuhan University, Wuhan 430079, China (e-mail: shaozhenfeng@whu.edu.cn).

Xiao Huang is with the Department of Geosciences, University of Arkansas, Fayetteville, AR 72701 USA (e-mail: xh010@uark.edu).

Tao Lu is with the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China (e-mail: lutxyl@gmail.com).

Ruiqian Zhang is with the Institute of Photogrammetry and Remote Sensing, Chinese Academy of Surveying and Mapping, Beijing 100830, China (e-mail: zhangruiqian@whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2022.3168593

I. Introduction

With the development of satellite sensors, remote sensing technology has been widely adopted to facilitate human understanding of the ground, such as the extraction of urban impervious surfaces [1], [2], land cover classification [3], [4], the detection of interesting objects [5], [6], and change detection [7], [8]. However, there exists an inevitable tradeoff of satellite sensors between high spatial resolution and high spectral resolution. Given this intrinsic tradeoff, pan-sharpening, an important pretreatment workflow in the remote sensing field, has been developed, whose major goal is to fuse the high spatial resolution panchromatic (PAN) images and the low-spatial-resolution (LR) multispectral (MS) images for the purpose of generating visually appealing color images with sharp details [9].

Existing traditional pan-sharpening methods can be generally classified into three major categories [10]: component substitution (CS)-, multiresolution analysis (MRA)-, and variational optimization (VO)-based methods. The sparse coding-based image restoration methods have achieved great performance [11], [12], opening the door for the design of the traditional pan-sharpening algorithms. Traditional embedding-based methods explore the manifold relation between HR/LR MS and HR/LR PAN dictionaries for image patch embedding and obtain remarkably performance promotion. Zhu and Bamler [13] achieved satisfactory fusion results by proposing a novel method that searches patches from HR/LR PAN dictionaries and predicts the HR image patches using local geometric structure captured from the LR image space. Later, many efforts have been made to investigate efficient searching and sparse embedding for performance improvement. Chang *et al.* [14] introduced the k-nearest neighbor (KNN) strategy as a locality constraint and aggregated patches for reconstruction. Considering the structural correlation in the MS space, the joint sparse prior [15] can be captured to further enhance textural details. Wang *et al.* [16] introduced the N-way block [17] pursuit strategy to compute the weight coefficients of image patches. Despite the success of the aforementioned algorithms, the mismatch between the coding patterns and geometric structures often leads to distorted spatial structures. Zhang *et al.* [10] proposed a spatially weighted neighbor embedding framework to explore the similar manifold structures in the LR MS space and the HR MS space. However, traditional embedding-based methods are computationally demanding and insufficient in terms of searching for similar images from large-scale datasets.

Despite that these embedding-based methods embed the local geometric relationship in the PAN space, textural information gaps exist between MS and PAN images, even for images of the same size, leading to mismatches when a direct searching between MS and PAN images is implemented. Furthermore, limited by the expression ability of the shallow network [10], the performance of these methods is not satisfactory.

With the improvement in deep-learning theories, a number of convolutional neural networks (CNNs) [18]–[20] are designed for pan-sharpening tasks to retain spectral information and recover spatial information using image-level fusion strategies depending on the large-scale training dataset. These methods feed MS patches and corresponding PAN patches into an end-to-end network and consider the PAN patches as an additional dimension of corresponding MS patches. CNN was first introduced to solve pan-sharpening tasks by Masi *et al.* [18], who designed a three-layer CNN to learn nonlinear functions between LR and HR MS spaces. To improve the efficiency, Scarpa *et al.* [21] proposed a target-adaptive CNN (TACNN) with a residual learning strategy. While achieving better performance, TACNN also obtains a faster convergence rate. Yang *et al.* [19] suggested that the MS image reconstructing process should contain two major phases, i.e., spectral preservation and structural preservation. Yuan *et al.* [20] proposed a multiscale and multidepth CNN to simulate multiscale receptive fields and handle the feature information of different scale objects in a flexible manner. However, the above algorithms all take HR-sized MS images as input, relying on a great number of computing resources and limited by the receptive field [22]. He *et al.* [23] explored the detail injection mode in the pan-sharpening task and introduced two detail injection-based CNNs with residual learning to improve the interpretability and convergence speed of the models. Based on the injection network [23], Liu *et al.* [24] proposed a dual model with deep residual block and shallow CNN, which can inject the high-pass information from the PAN image into the MS image.

Numerous efforts [25], [26] have been made to upscale the LR-sized feature maps into the HR space. Although these methods are able to better preserve textural details, detail recovery remains to be a challenge [27]. To improve the quality of reconstructed MS images, some methods [28]–[30] adopt novel loss functions, such as spectral–spatial structure loss [28] and the perceptual loss [29], [30], aiming to further enhance the preservation of spectral and structures. Recently, Deng *et al.* [31] introduced spectral information via a residual network to reconstruct image features along the spectral direction. Xu *et al.* [32] proposed an iterative strategy to separately justify the generation of MS and PAN images. Wang *et al.* [33] proposed a spectral-to-spatial convolution with distortion-free property to aggregate spectral features into the spatial domain to perform the upsampling operation rather than the direct upsampling operation. Dang *et al.* [34] introduced channel and spatial attention to the hyperspectral pan-sharpening framework to generate the fused image with high spectral fidelity. Dang *et al.* [35] proposed a mature Gaussian–Laplacian pyramid network to simplify the hyperspectral pan-sharpening problem into several pyramid-level shallow learnings with reduced parameters. Dang *et al.* [36] decoupled the pan-sharpening problem into spatial and spectral reconstruction subtasks and built generative adversarial networks separately.

The above CNN-based supervised methods, however, are tailored to certain image degradation models. In real scenarios, how images are degraded remains unclear, and the degrading process usually fails to be described via established mathematical formulas. Thus, unsupervised CNNs are further proposed to improve the universality of practical scenarios by designing loss functions to preserve spatial structures (e.g., the regression linear weighting [37] and channel replica [38]). Ma *et al.* [39] proposed two independent generative adversarial networks that generate abundant spectral information with rich spatial details in an unsupervised manner.

In traditional deep learning-based pan-sharpening frameworks, the PAN image is usually fused into the MS space at the global level. Despite the performance improvement of the above deep-learning-based pan-sharpening methods, challenges still remain.

1) Although the PAN image contributes to the fusion model, the introduction of excess low-frequency components (image content) of the PAN image can lead to notable spectral distortion in the reconstructed images.
2) By contrast, the lack of high-frequency components (image textural and edge information) captured from the PAN image can result in blurry details of the reconstructed image.

Inspired by the spatially weighted neighbor embedding pan-sharpening strategy [10] and the residual learning-based embedding super-resolution method [40], we address these above issues by designing a novel multiscale locally linear embedding residual network (LLERN) for image pan-sharpening tasks. The proposed LLERN contains two important stages: 1) spectral preservation and 2) structural preservation. The goal of the spectral preservation network is to upscale the LR MS image while retaining spectral information. Note that PAN images supply textural information in the locally linear embedding residual block (LLERB) of the structural preservation stage but are not involved in the spectral preservation stage. Similar to the traditional embedding-based pan-sharpening methods, we first cast an assumption that LR MS image patches share a similar manifold structure with their residual image patches in a local region, as shown in Fig. 1. The proposed LLERB is a refined strategy to fuse the PAN image into the MS space. Specifically, the proposed LLERN learns a similar local geometric relationship between the MS space and the PAN space rather than the relationship between the LR space and the HR space in the deep-learning framework. On the other hand, similar to residual learning, the proposed LLERN searches the local geometric relationship in the high-pass domain for stability and fast convergence. Furthermore, the corresponding high-pass patches are introduced to estimate target residual patches from a neighborhood region. That is to say, each LR MS image patch can be represented using a linear weight of its nearest neighbors via the LR MS image and high-pass version HR PAN image.

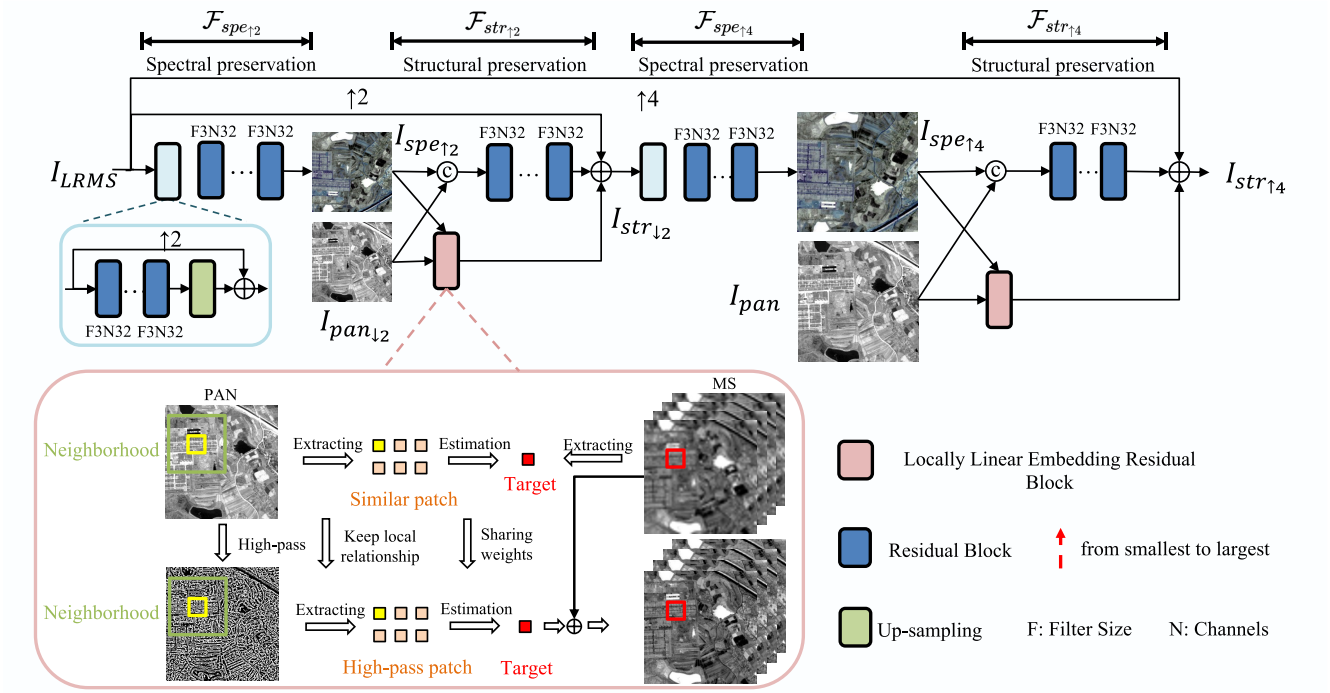


Fig. 1. Overall structure of the proposed LLERN. The proposed LLERN consists of two major phases: spectral preservation and structural preservation. Each LR MS patch can be represented using a linear weight of its nearest neighbors. Then, the corresponding high-pass patches are introduced to estimate the target residual patches from a neighboring region.

The main contributions of this work are given as follows.

1) We propose an end-to-end multiscale LLERN for satellite image pan-sharpening tasks. The proposed method approach mitigates the problems of spectral distortions and blurry textures. Both qualitative and quantitative results confirm that the proposed method outperforms state-of-the-art pan-sharpening methods.
2) The proposed LLERN consists of two major phases: spectral preservation and structural preservation. To improve the precision in image patch matching of the proposed LLERB in the structural preservation phase, the proposed spectral preservation network aims to upscale MS images and reduce the information gap between MS images and corresponding PAN images.
3) In the proposed LLERB, the high-frequency information from the PAN image is embedded into the MS image (structural preservation network) at the patch level to mitigate the issues of spectral distortion or the lack of texture. The extensive experiments confirm the existence of a similar local geometric relationship between the MS space and high-pass PAN space, which can be used to guide deep-learning modeling with improved interpretability.

II. Proposed Method

In this section, we describe the proposed LLERN in detail from three perspectives, i.e., problem formulation, LLERB, and the loss function.

A. Problem Formulation

Fig. 1 presents the flowchart of the proposed multiphase framework. We denote the LR MS image as $I_{\text{LRMS}} \in \mathbb{R}^{mn \times C}$, where m and n are the height and the width of the LR MS image, respectively, with C being the number of image bands ($C = 4$). $I_{\text{HRMS}} \in \mathbb{R}^{(sm)(sn) \times C}$ denotes an HR MS image with the spatial resolution scale s between I_{LRMS} and I_{HRMS}. $I_{\text{PAN}} \in \mathbb{R}^{(sm)(sn) \times c}$ denotes a PAN image with the number of image bands of c ($c = 1$).

Selected matching results of the five most similar PAN patches of an MS patch are shown in Fig. 2. The numbers below images indicate the indexes of PAN image patches. We take the PAN matching results of the HR MS as the standard and mark the exact match in red. We observe that directly matching similar PAN patches of an LR MS patch can bring ineluctable mismatches. Therefore, we design a progressive framework that consists of the spectral preservation phase and the structural preservation phase. For the spectral preservation phase, we first extract features and feed them to an upsampling layer. We denote I_{spe} as the spectral preservation version LR MS image. As shown in Fig. 2, the matching accuracy of I_{spe} is considerably improved than LR MS images. The structural preservation phase aims to embed the high-pass structural information into the MS image.

In the following, we introduce the spectral preservation phase and the structural preservation phase in a detailed manner.

1) Spectral Preservation Phase: Specifically, we feed the LR MS image I_{LRMS} into the spectral preservation block (an upsampling block with an upsampling scale of 2) to obtain its shallow features $f'_{\text{spe}_{\uparrow 2}}$. The upscaled map $f'_{\text{spe}_{\uparrow 2}}$ is then used for image reconstruction with deep reconstruction networks $H_r(.)$ with l residual blocks as

$$I_{\text{spe}_{\uparrow 2}} = H_r\left(f'_{\text{spe}_{\uparrow 2}}\right) \tag{1}$$

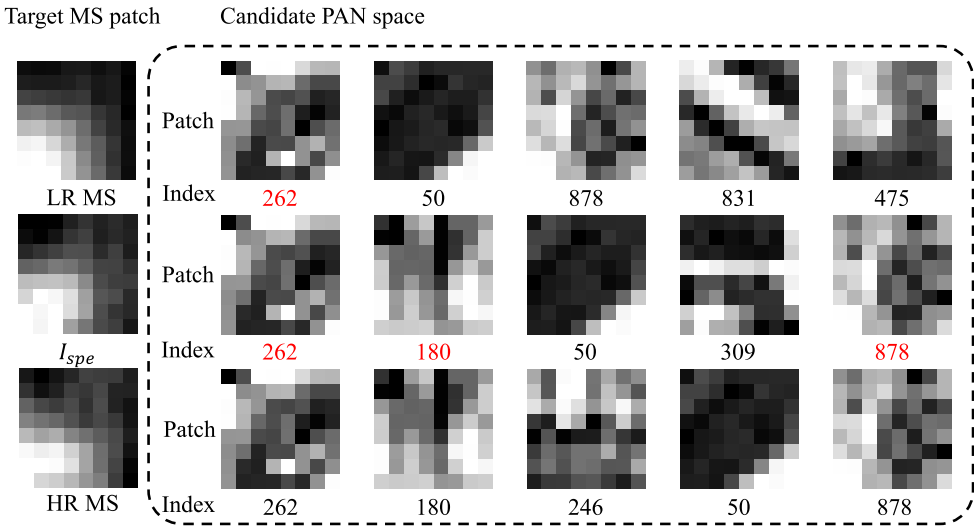


Fig. 2. Selected examples of the matching results with different patterns. We list the five most similar PAN patches of an MS patch. The numbers below images indicate the indexes of PAN image patches. We consider the matched PAN results of the HR MS as the standard and mark the exact matching in red. The neighbors found for LR MS patches are not consistent with the neighbors of HR patches.

where $I_{\text{spe}_{\uparrow 2}}$ denotes the upscaled version MS image with an upsampling scale of 2.

2) Structural Preservation Phase: $I_{\text{spe}_{\uparrow 2}}$ and $I_{\text{pan}_{\downarrow 2}}$ (the downsampled version of I_{pan} with a downsampling scale of 2) are fed into the structural preservation phase. We enhance textural information of the MS image at the image level and the patch level. This process can be formulated as

$$I_{\text{str}_{\uparrow 2}} = H_{\text{fus}}\left(I_{\text{spe}_{\uparrow 2}}, I_{\text{pan}_{\downarrow 2}}\right) + \text{up}_{\uparrow 2}(I_{\text{LRMS}}) + H_{\text{llerb}}\left(I_{\text{spe}}, I_{\text{pan}_{\downarrow 2}}\right) \quad (2)$$

where $H_{\text{fus}}(.)$ refers to the image-level fusion operation with deep networks and l residual blocks, and $\text{up}_{\uparrow 2}(.)$ is the PixelShuffle [41] operator with an upsampling scale of 2. $I_{\text{str}_{\uparrow 2}}$ denotes the pan-sharpened version of an MS image. $H_{\text{llerb}}(.)$ refers to the proposed LLERB that fuses images at patch level.

By cascading these two phases, we are able to super-resolve the input LR MS image with scale $\times 4$ as

$$I_{\text{spe}_{\uparrow 4}} = H_r\left(H_{\text{up}}\left(I_{\text{str}_{\uparrow 2}}\right)\right) \quad (3)$$

$$I_{\text{str}_{\uparrow 4}} = H_{\text{fus}}\left(I_{\text{spe}_{\uparrow 4}}, I_{\text{pan}}\right) + \text{up}_{\uparrow 4}(I_{\text{LRMS}}) + H_{\text{llerb}}\left(I_{\text{spe}_{\uparrow 4}}, I_{\text{pan}}\right) \quad (4)$$

where $I_{\text{spe}_{\uparrow 4}}$ denotes the output image of the spectral preservation network. $\text{up}_{\uparrow 4}$ denotes the PixelShuffle operator with an upsampling scale of 4. $I_{\text{str}_{\uparrow 4}}$ denotes the final reconstructed image.

B. Locally Linear Embedding Residual Block

From an existing study [10], LLE-based processing can be, respectively, generalized as follows: the generation of LR version PAN image, image block-dividing, neighbors' searching, the estimate of the weight, and representation. Considering the differences between deep-learning and traditional algorithms, the proposed LLERB contains five major components: 1) high-pass extraction; 2) image block-dividing; 3) weight calculation; 4) representation; and 5) embedding. Here, we describe these components in detail and take the subnetworks in the phase with scale factor $\times 2$ as an example.

1) High-Pass Extraction: First, we feed the PAN image $I_{\text{pan}_{\downarrow 2}}$ to an extractor (e.g., the Sobel operator, the Candy operator, difference convolution, and the difference between the HR PAN image and the LR PAN image) for the high-frequency extraction. Please refer to Section III for more discussion on the high-pass extractions.

2) Image Block-Dividing: After the high-pass extraction, the high-pass PAN images $I_{\text{hp}_{\downarrow 2}}$, $I_{\text{pan}_{\downarrow 2}}$, and $I^{q}_{\text{spe}_{\uparrow 2}}(q = R, G, B, \text{NIR})$ are divided into small image patches $k \times k$ without overlapping, as $p^{m}_{\text{hp}_{\downarrow 2}} \in \mathbb{R}^{(m/k)(n/k)}$, $p^{m}_{\text{pan}_{\downarrow 2}} \in \mathbb{R}^{(m/k)(n/k)}$, and $p^{m,q}_{\text{spe}_{\uparrow 2}} \in \mathbb{R}^{(m/k)(n/k)}$ via the unary function layers (convolution layers $\theta(.)$, $\delta(.)$, and $\zeta(.)$ in Fig. 3), respectively, where k^2 is the number of pixels in an image patch, and M is the number of partitioned patches. R, G, B, and NIR denote red, green, blue, and near-infrared bands of an MS image. Note that the above blocking operations are realized by $k \times k$ convolution layers.

3) Weight Calculation: We further calculate patch correlation between $p^{m,q}_{\text{spe}_{\uparrow 2}} \in \mathbb{R}^{(m/k)(n/k)}$ and $p^{m}_{\text{pan}_{\downarrow 2}} \in \mathbb{R}^{(m/k)(n/k)}$ as the geometric relationship between the MS and PAN space. This process can be expressed as

$$w = \text{up}_{\text{near}}\left(e^{p_{\text{spe}_{\uparrow 2}}\left(p_{\text{pan}_{\downarrow 2}}\right)^{T}}\right) \quad (5)$$

where w denotes the weight map for each query window and $\text{up}_{\text{near}}(.)$ denotes the near interpolation function.

4) Representation: We introduce KNN to constraint the boundary of the locally linear representation space, i.e., only the values ordered to the top k are considered via the

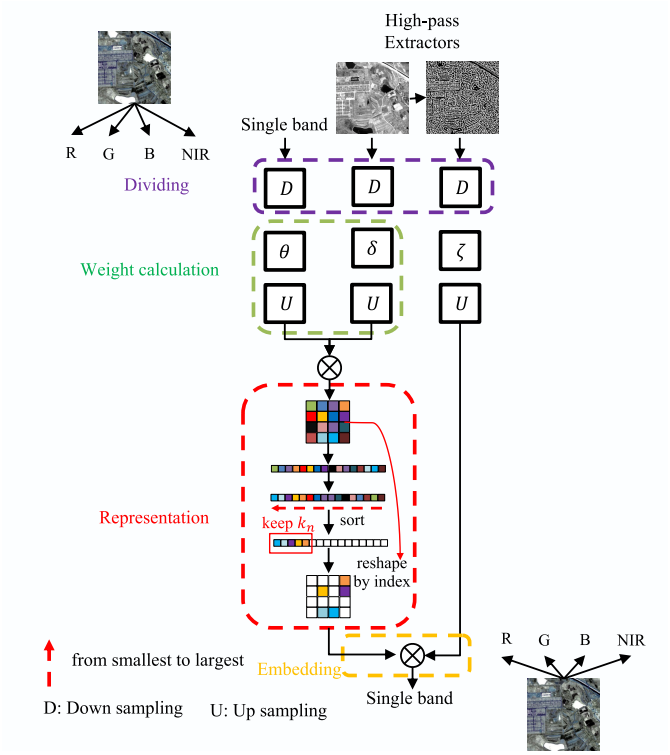


Fig. 3. Architecture of the proposed LLERB contains five major components: 1) high-pass extraction; 2) image block-dividing; 3) weight calculation; 4) representation; and 5) embedding.

following formula:

$$w_{k_n} = \begin{cases} \text{softmax}(w_{\text{idx}}), & \text{idx} \in \text{top}K\{w\} \\ 0, & \text{other} \end{cases} \tag{6}$$

where top$K\{w\}$ denotes the k_nth largest values of the weight map w. softmax(.) denotes the softmax function.

Then, we represent $p_{r_{\downarrow 2}}$ by k_n image patches as

$$p_{r_{\downarrow 2}} \approx p_{\text{spe}_{\uparrow 2}} w_{k_n}. \tag{7}$$

5) Embedding: In the training phase, we optimize the following loss function of mean absolute errors between all reconstructed images and ground-truth images to avoid generating oversmoothed results

$$p_{\text{res}_{\downarrow 2}} = p_{\text{hp}_{\downarrow 2}} w_{k_n} \tag{8}$$

where $p_{\text{res}_{\downarrow 2}}$ denotes the reconstructed residual image.

C. Loss Function

Mean square error (mse) loss has been widely used as a loss function. However, the mse loss can lead to the generation of oversmoothing details and textures. In the training phase, we optimize the mse loss function between all reconstructed images and ground truth in an adaptive manner, thus leading to more realistic results.

III. Experiments

A. Datasets and Implementation Details

The experiments are conducted on the WorldView II (WV), Gaofen-2 (GF), and Quick Bird (QB). The WV, GF, and QB datasets used in this study provide PAN images at 0.5-, 0.8-, and 0.61-m spatial resolutions, respectively, with the spatial resolution ratio of $\times 4$. The image sizes of original MS images (red, green, blue, and near-infrared four bands) of the WV (200 train, 20 validation, and 20 test), GF (730 train, 78 validation, and 78 test), and QB (450 train, 50 validation, and 50 test) are $3848 \times 4096 \times 4$ pixels, $7260 \times 6864 \times 4$ pixels, and $115\,200 \times 115\,200 \times 4$ pixels, respectively. The spatial resolution of original PAN images is four times bigger than the original MS images. We downsample the original PAN images via the bicubic function with a factor $\times 4$, which are regarded as the HR PAN images. The same downsampling version MS images are treated as the LR MS images. The size of cropped downsampled LR-MS images is $64 \times 64 \times 4$ pixels, and the size of the corresponding PAN images is $256 \times 256 \times 1$ pixels. Seven widely used image quality assessment (IQA) metrics are employed as evaluation metrics, including four supervised IQAs, i.e., the erreur relative globale adimensionnelle de synthese (ERGAS), the peak signal-to-noise ratio (PSNR), the spectral angle mapper (SAM), and the universal image quality index (UIQI), and three unsupervised IQAs, i.e., the spectral distortion index (D_λ), the spatial distortion index D_S, and the hybrid quality with no reference (HQNR). The best values for these metrics are 0, $+\infty$, 0, 1, 0, 0, and 1, respectively.

All models are trained on the desktop with Ubuntu 18.04, CUDA 10.2, and CUDNN 7.5 with two Nvidia TITAN GPUs (24 GB). We use the Adam optimizer [42] for optimization with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1e-8$, and a minibatch size of 8. The learning rate is initialized to 1×10^{-4} and decays by a factor of 0.1 every 500 epochs. We train the network for 1000 epochs with the above settings. The patch size of the LR MS image and the PAN image is set to 40×40 pixels and 160×160 pixels, respectively. For the configuration parameter, we set $l = 11$ (more details in Section III-C).

B. Comparison With State-of-the-Art Methods

To verify the effectiveness of the proposed method, we compare our method against ten state-of-the-art pan-sharpening algorithms that include PRACS [43], BDSD [44], Brovey [45], AWLP-H [46], PNN [18], PANNet[1] [19], MSDCNN[2] [20], DARN[3] [47], GPPNN[4] [32], and MUCNN[5] [33]. We summarize the optimal parameters of all compared deep-learning-based methods in Table I, where all key parameters are derived from the corresponding articles. For a fair comparison, we adjust the parameters for the highest performance.

[1]https://github.com/codegaj/py_pansharpening/tree/master/methods
[2]https://github.com/Decri/Multi-Scale-and-Depth-CNN-for-Pan-sharpening
[3]https://githubmemory.com/index.php/repo/yxzheng24/IEEE_TGRS_DHP-DARN
[4]https://github.com/xsxjtu/GPPNN
[5]https://liangjiandeng.github.io/

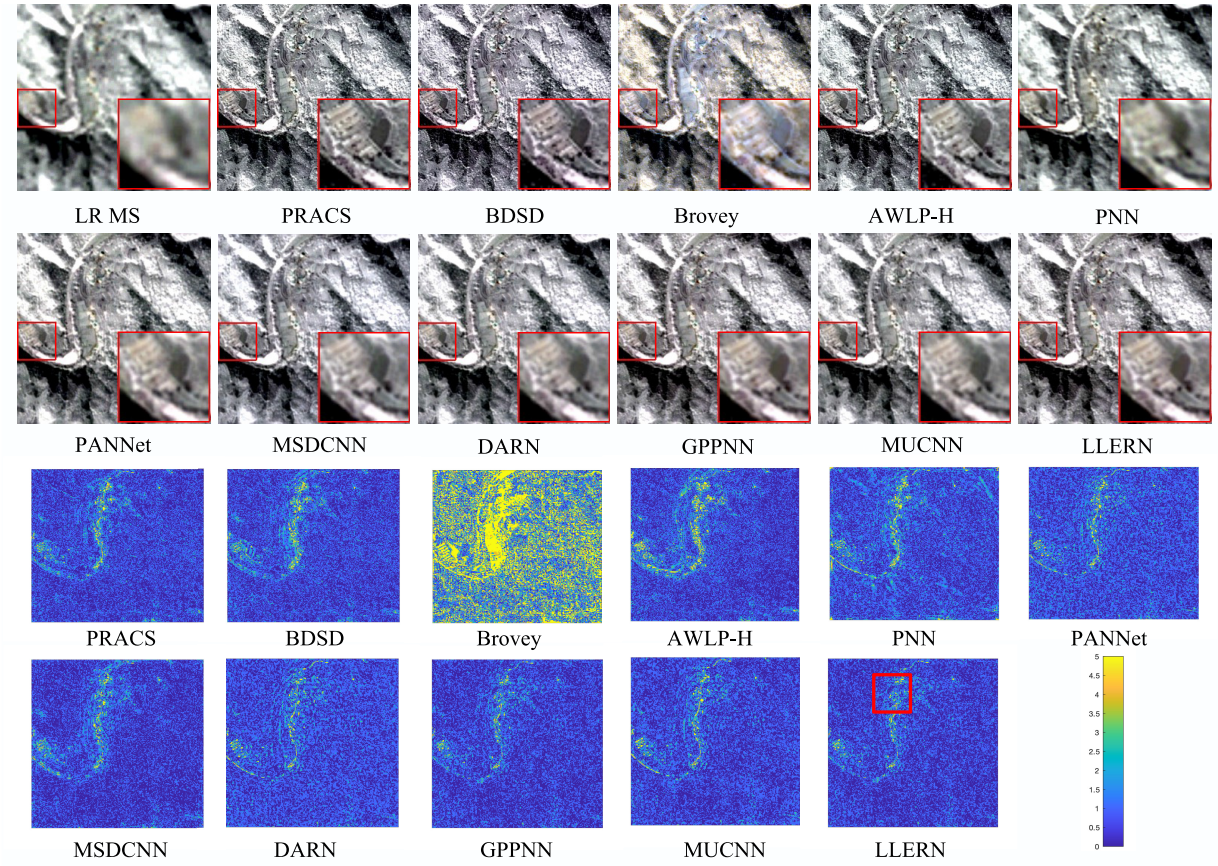


Fig. 4. Qualitative comparison of LLERN with ten counterparts on a typical satellite image pair from the GF dataset. Images in the last two rows visualize the mse between the pan-sharpened results and the ground truth. (Please zoom in to see more details.)

TABLE I

OPTIMAL PARAMETERS FOR THE COMPARED DEEP-LEARNING-BASED METHODS. NOTATIONS: **ITER** (ITERATION EPOCH), **BS** (MINIBATCH SIZE), **ALGO** (OPTIMIZATION ALGORITHM), **LR** (LEARNING RATE), **FS** (FILTER SIZE), AND **PS** (PATCH SIZE)

	PNN	PANNet	MSDCNN	DARN	GPPNN	MUCNN
Iter.	2000	300	3000	600	1000	1000
Bs	128	128	64	16	16	16
Algo	SGD	SGD	SGD	Adam	Adam	Adam
Lr	1e-1	1e-1	1e-1	1e-3	1e-4	1e-3
PS	33	33	41	32	33	32

The performance of selected models on the WV and GF datasets is shown in Table II. We notice that the proposed method obtains the best average quantitative performance for all supervised IQA metrics. The proposed LLERN outperforms the other two competitive deep-learning-based algorithms (DARN [47] and MUCNN [33]) by 0.62/0.51 and 1.19/1.11 dB in PSNR, respectively. However, traditional methods perform better than deep-learning-based methods for unsupervised IQA metrics on the GF dataset. This can be explained by the fact that traditional methods fail to directly recover the missing high-frequency information and spectral information from the prior information but replace the LR MS images with the PAN images through a certain strategy. Despite that these methods obtain better unsupervised IQA results, spectral distortion is inevitably introduced.

Selected fusion results are presented in Figs. 4 and 5, where the last two rows present the visualization of error maps, i.e., the mse between the pan-sharpened results and the ground truth. The error maps of the proposed method present notable advantages, evidenced by the region highlighted by the red box. The images on the WV dataset cover typical urban areas with richer textural information, while the images are mainly concentrated in mountain areas on the GF dataset. The fusion images of PRACS [43] and BDSD [44] show notable color distortion, e.g., the road surface on the WV image and lake water on the GF image. The proposed method effectively preserves and enhances the spectrum and texture information compared to other deep-learning-based methods.

We compare the proposed method with state-of-the-art deep-learning-based and traditional-based pan-sharpening algorithms in Table III. We notice that the proposed method also obtains the best supervision performance and the best HQNR than DARN [47] and MUCNN [33]. The visual analysis supports these quantitative evaluations, as shown in Fig. 6, where the last two rows present the visualization of the error maps, i.e., the mse between the pan-sharpened results and the ground truth. The error map of the proposed

TABLE II

QUANTITATIVE COMPARISON OF ELEVEN METHODS. BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED BY RED AND BLUE, RESPECTIVELY. ↑ INDICATES THAT THE LARGER THE VALUE, THE BETTER THE PERFORMANCE, AND ↓ INDICATES THAT THE SMALLER THE VALUE, THE BETTER THE PERFORMANCE

Dataset	Methods	ERGAS ↓	PSNR ↑	SAM ↓	UIQI ↑	D_λ ↓	D_S ↓	HQNR ↑
WV	PRACS	1.5459	36.03	2.0222	0.6440	0.0206	0.0917	0.8931
	BDSD	1.7586	35.45	2.3177	0.6447	0.0551	0.1063	0.8506
	Brovey	1.8615	35.10	2.2997	0.6017	0.0629	0.1706	0.7141
	AWLP-H	1.4617	36.39	1.8656	0.5897	0.0683	0.1109	0.8671
	PNN	1.9399	35.21	2.5682	0.5875	0.0321	0.0790	0.8398
	PANNet	1.3243	37.99	1.9270	0.6700	0.0396	0.1072	0.8522
	MSDCNN	1.3847	37.88	1.8307	0.6903	0.0407	0.0919	0.8212
	DARN	0.9819	40.15	1.3338	0.7558	0.0204	0.0695	0.9045
	GPPNN	1.2594	38.31	1.7385	0.6964	0.0372	0.0779	0.8783
	MUCNN	0.9965	40.04	1.3808	0.7565	0.0271	0.0703	0.8945
	LLERN (our)	0.9331	40.66	1.2963	0.7616	0.0193	0.0679	0.9154
GF	PRACS	1.3730	38.15	1.5389	0.5915	0.0385	0.0686	0.8669
	BDSD	1.7637	36.32	1.8615	0.5171	0.0484	0.0600	0.8269
	Brovey	1.1886	38.07	1.2460	0.4729	0.0784	0.0937	0.6209
	AWLP-H	1.4732	37.63	1.5844	0.5758	0.0596	0.0653	0.8553
	PNN	0.7924	42.46	0.9273	0.6492	0.0207	0.1069	0.8350
	PANNet	0.6697	43.87	0.7446	0.7321	0.0105	0.1014	0.8419
	MSDCNN	0.5876	44.36	0.6859	0.7482	0.0098	0.1082	0.8405
	DARN	0.5986	44.67	0.6434	0.7600	0.0089	0.1102	0.7812
	GPPNN	0.6467	44.07	0.7263	0.7450	0.0150	0.1100	0.8069
	MUCNN	0.6140	44.57	0.6822	0.7588	0.0143	0.1085	0.7889
	LLERN (our)	0.5059	45.76	0.5638	0.7845	0.0088	0.1063	0.8480
	Desired value	0	$+\infty$	0	1	0	0	1

TABLE III

QUANTITATIVE COMPARISON OF ELEVEN METHODS ON THE QB DATASET. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY. ↑ INDICATES THAT THE LARGER THE VALUE, THE BETTER THE PERFORMANCE, AND ↓ INDICATES THAT THE SMALLER THE VALUE, THE BETTER THE PERFORMANCE

Methods	ERGAS ↓	PSNR ↑	SAM ↓	UIQI ↑	D_λ ↓	D_S ↓	HQNR ↑
PRACS	0.9188	29.80	0.6590	0.8348	0.0065	0.0386	0.8062
AWLP-H	0.8496	30.41	0.6336	0.6845	0.0140	0.0247	0.8002
PNN	0.6761	32.16	0.6416	0.8247	0.0114	0.0129	0.9639
PANNet	0.6224	32.85	0.5715	0.7838	0.0078	0.0164	0.9635
MSDCNN	0.6096	33.07	0.5790	0.8586	0.0042	0.0207	0.9492
DARN	0.5418	34.03	0.5418	0.8654	0.0041	0.0194	0.9673
GPPNN	0.5887	33.33	0.6185	0.8569	0.0105	0.1091	0.9462
MUCNN	0.4461	35.33	0.5212	0.9013	0.0062	0.0203	0.9616
LLERN (our)	0.4370	35.82	0.5120	0.9091	0.0049	0.0200	0.9698
Desired value	0	$+\infty$	0	1	0	0	1

method is more blue-dominant, indicating that the fused results from the proposed method are closer to the ground truth, where traditional methods exhibit significant spectral distortion.

C. Ablation Experiments

1) Effect of the Downgrading Process: Many studies [9], [48] revealed that the modulation transfer function (MTF) might be more appropriate for MS and PAN images. Therefore,

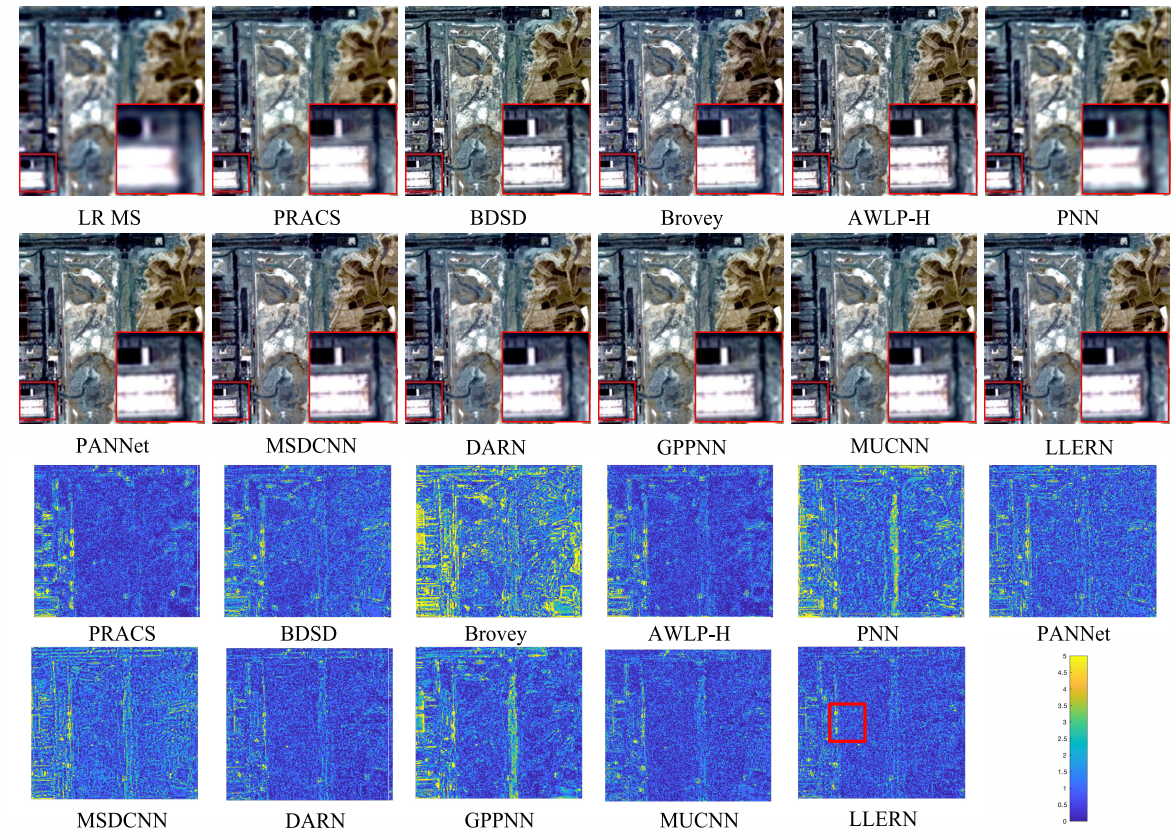


Fig. 5. Qualitative comparison of LLERN with ten counterparts on a typical satellite image pair from the WV dataset. Images in the last two rows visualize the mse between the pan-sharpened results and the ground truth. (Please zoom in to see more details.)

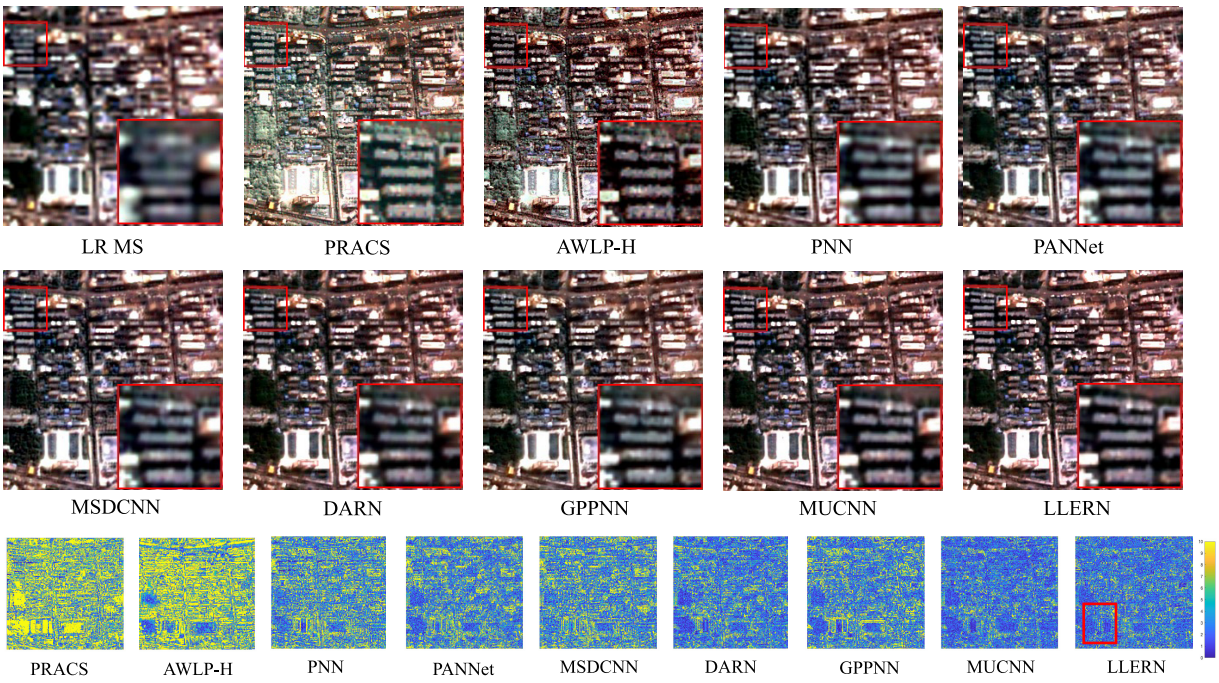


Fig. 6. Qualitative comparison of LLERN with five counterparts on a typical satellite image pair from the QB dataset. Images in the last row visualize the mse between the pan-sharpened results and the ground truth. (Please zoom in to see more details.)

TABLE IV

QUANTITATIVE COMPARISON ON THE WV DATASET. BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED BY RED AND BLUE, RESPECTIVELY. ↑ INDICATES THAT THE LARGER THE VALUE, THE BETTER THE PERFORMANCE, AND ↓ INDICATES THAT THE SMALLER THE VALUE, THE BETTER THE PERFORMANCE

Downgrade	Methods	ERGAS ↓	PSNR ↑	SAM ↓
Bicubic	PRACS	1.5459	36.03	2.0222
	AWLP-H	1.4617	36.39	1.8656
	DARN	0.9819	40.15	1.3338
	MUCNN	0.9965	40.04	1.3808
	LLERN	0.9331	40.66	1.2963
MTF	PRACS	1.8989	34.36	2.4564
	AWLP-H	1.8040	34.89	2.2586
	DARN	1.3397	37.87	1.8151
	MUCNN	1.3589	37.69	1.8124
	LLERN	1.2684	38.32	1.6541

we first conduct ablation experiments on the WV dataset using some methods with decent performances. The MTF of the MS sensor consists of a bank of Gaussian filters, representing a more complex degradation process. Traditional methods usually fail to regenerate lost high-frequency information in the MS image, while directly introducing the PAN images into MS images can lead to ghost images. With the powerful feature extraction capability of CNN, deep-learning-based algorithms have shown notable improvement in performance (more than 3 dB). Specifically, the proposed method is able to learn the missing MS high-frequency information in the spectral preservation phase and further fuse the PAN image in the structural preservation phase, thus demonstrating greater robustness in the complex data scenes than one-stage deep-learning-based methods.

2) Effect of Configurations:

a) Number of residual blocks l: We conduct the experiments on the QB dataset to test the impact of the number of residual blocks l (with values that range from 8 to 13) on model performance. As shown in Fig. 7, we report the performance and running time of the proposed method with different l's. With the increase in l, the performance increases gradually. When $l > 11$, the PSNR value increases slowly with an increased time overhead. The proposed structural preservation occurs in the high-resolution space; therefore, the running time is expected to increase significantly with the increase in the number of residual blocks. Considering this, we set $l = 11$, a tradeoff between the time cost and objective performance.

b) Multiscale preservation phases: To validate the effectiveness of the proposed two preservation phases, we conduct experiments with different combinations of preservation phases on the QB dataset, as shown in Table V. As PAN images only participated in the two structure preservation phases ($\mathcal{F}_{\text{str}_{\uparrow 2}}$ and $\mathcal{F}_{\text{str}_{\uparrow 4}}$), the spectral and texture performance of the fused image is degraded with the removal of $\mathcal{F}_{\text{str}}$. Especially, the PSNR value decreases by 5 dB without $\mathcal{F}_{\text{str}}$.

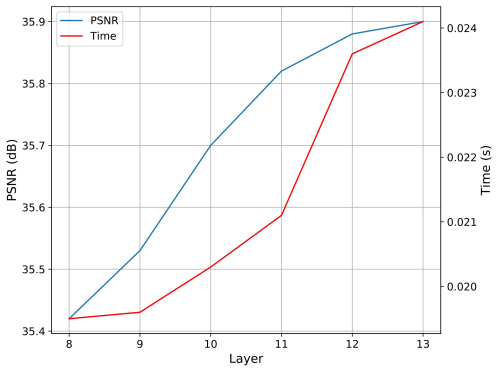


Fig. 7. Ablation experiments on residual blocks on the QB dataset.

TABLE V

ABLATION EXPERIMENTS ON DIFFERENT CONFIGURATIONS ON THE QB DATASET

$\mathcal{F}_{spe_{\uparrow 2}}$	$\mathcal{F}_{str_{\uparrow 2}}$	$\mathcal{F}_{spe_{\uparrow 4}}$	$\mathcal{F}_{str_{\uparrow 4}}$	ERGAS ↓	PSNR ↑
Bicubic				0.8834	29.85
✓	-	✓	-	0.7853	30.86
✓	✓	✓	-	0.5464	33.91
✓	-	✓	✓	0.5463	33.91
-	✓	-	✓	0.4546	35.18
✓	✓	-	✓	0.4439	35.45
-	✓	✓	✓	0.4431	35.49
One-scale preservation phases				0.4412	35.71
✓	✓	✓	✓	0.4370	35.82

f_{spe} denotes the function of spectral preservation phase.
f_{str} denotes the function of structural preservation phase.

When we remove $\mathcal{F}_{\text{spe}_{\uparrow 2}}$ and $\mathcal{F}_{\text{spe}_{\uparrow 4}}$, the performance degradation problem resulting from the mismatches can be found in the structural preservation phase with $\times 2$. We notice that the proposed LLERN maintains great performance when only one structural preservation phase is removed.

In addition, we employ one-scale preservation phases to investigate the impact of the number of phases on the fusion results. The multiscale version network based on a coarse-to-fine strategy has more parameters than the one-scale version network, thus demanding more computation power. Despite its computational demand, the multiscale network outperforms the one-scale version network by 0.11 dB due to its capability in obtaining multiscale image features leveraging LR MS images.

3) Impact of LLERN: Fig. 8 presents the curves of loss values per epoch on the WV dataset. LLERN* denotes the proposed methods without LLERB. These curves indicate that LLERB guarantees fast and stable convergence. To further investigate the choice of hyperparameter k_n and high-pass extractors in the proposed LLERN, we conduct a series of ablation experiments. Model performances with four different configurations, i.e., the Laplace operator, the Candy operator, difference convolution [49] (a universal and popular automatic edge extraction layer), and the difference between the HR PAN

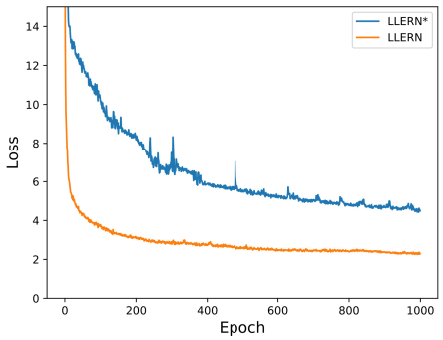


Fig. 8. Curves of the training procedure on the WV dataset. LLERN* denotes the proposed methods without LLERB. These curves show that LLERB is characterized by its fast and stable convergence.

image and the LR PAN image, i.e., $I_{\text{PAN}} - \text{bic}(I_{\text{PAN}})$ with different hyperparameters k_n, are listed in Table VI.

a) High-pass extractors: In a multiscale framework, the high-pass component of $I_{\text{PAN}} - \text{bic}(I_{\text{PAN}})$ can bring low-frequency information with decreased information gap. As the most widely used feature extractors, Laplace and Candy operations attach more importance to image edge information. Therefore, the average quantitative results with Laplace and Candy operations are similar. As shown in Table VI, Laplace and Candy operations reduce the values of SAM and ERGAS, suggesting that they are able to mitigate the spectral distortion problem. The difference convolution layer can be seen as an adaptive edge extraction operator and, therefore, also has greater robustness. Due to the limitation of feature extraction ability, the performance improvement brought by one difference convolution layer is not obvious. The increase in difference convolution layers [49] directs the feature map attention to edges. It is clear that the increase in this difference convolution blocks in number leads to performance improvement at the sacrifice of efficiency. To balance performance and computational overhead, we employ the Candy extractor as the extractor.

b) Hyperparameter k_n: To investigate the effectiveness of the hyperparameter k_n that controls the boundary of the locally linear representation space, we list the quantitative results with hyperparameter k_n ranging from 1 to 20. It is notable that model performance first increases and then decreases with the increase in k_n. This phenomenon can be explained by the fact that a small number of image patches fail to accurately describe HR MS patches, while an excessively large k_n tends to introduce image artifacts, resulting in performance degradation.

D. Efficiency Analysis

We also investigate the efficiency of LLERN compared with other competing approaches on the desktop with two Nvidia TITAN GPUs (24 GB) and Intel[6] Xeon[6] E5-2680. Quantitative results in terms of PSNR, running time, and the number of parameters are tabulated in Table VII. As traditional pan-sharpening algorithms have no trainable parameters, their running time on the CPU is listed. In terms of objective results, the proposed method gains a significant improvement (more than 0.5 dB). The running time of the proposed algorithm is greater than that of DARN [47] but less than that of MUCNN [33]. We can conclude that the proposed method owns a better tradeoff between the computational overhead and performance.

[6]Registered trademark.

TABLE VI

AVERAGE QUANTITATIVE RESULTS OF THE PROPOSED METHOD WITH DIFFERENT HIGH-PASS EXTRACTORS ON THE WV DATASET

High pass extractor	k_n	ERGAS ↓	PSNR ↑	SAM ↓
$I_{PAN} - bic(I_{PAN})$	1	1.0071	40.03	1.3552
	5	**0.9839**	**40.53**	1.3453
	10	0.9950	40.23	**1.3440**
	15	1.0689	40.13	1.4278
	20	1.0432	40.05	1.4105
$f_{dc}(I_{PAN}) \times 1$	1	0.9931	40.23	1.3482
	5	**0.9788**	**40.57**	**1.3322**
	10	0.9878	40.43	1.3440
	15	0.9950	40.33	1.3478
	20	0.9915	40.25	1.3405
$f_{dc5}(I_{PAN}) \times 5$	1	0.9234	40.61	1.2831
	5	**0.9196**	**40.73**	**1.2798**
	10	0.9261	40.60	1.2833
	15	0.9349	40.55	1.2966
	20	0.9411	40.44	1.3148
Laplace	1	0.9390	40.48	1.3054
	5	0.9352	40.50	1.2851
	10	**0.9166**	**40.65**	**1.2753**
	15	0.9316	40.55	1.2954
	20	0.9275	40.54	1.2888
Candy	1	**0.9266**	40.59	**1.2845**
	5	0.9331	**40.66**	1.2963
	10	0.9285	40.56	1.2931
	15	0.9349	40.53	1.2966
	20	0.9428	40.52	1.3019

$bic(.)$ denotes the Bicubic operation.
$f_{dc}(.) \times n$ denotes the differential convolution and n denotes the number of the differential convolution.

E. Scalability Analysis

In order to explore the scalability of the proposed method, we develop a flexible collocation of the dataset and conduct additional ablation experiments. As illustrated in Table VIII, we list the value of PSNR of the proposed LLERN and the state-of-the-art deep-learning-based method (MUCNN) with different training configurations. The proposed LLERN produces better scalability than MUCNN with the same training and testing collocation with an improved performance of up to 2 dB. Due to the difference in the imaging characteristics and the variations of remote sensing platforms, the scalability of the deep-learning-based methods is limited. After fine-tuning, the performance of the proposed LLERN improves significantly (more than 7 dB), presumably due to the fact that the proposed LLERN is able to better leverage prior informa-

TABLE VII

EFFICIENCY ANALYSIS WITH PSNR AND RUNNING TIME ON THE WV DATASET AND THE NUMBER OF PARAMETERS

Methods	PSNR	Time (s)	Par. (M)
PRACS	36.03	0.1805	-
BDSD	35.45	0.0605	-
Brovey	35.10	0.0537	-
AWLP-H	36.39	0.0868	-
PNN	35.21	0.0036	0.07
PANNet	37.99	0.0038	0.07
MSDCNN	37.88	0.0108	0.24
DARN	40.15	0.0184	0.31
GPPNN	38.30	0.0147	0.11
MUCNN	40.04	0.0274	1.37
LLERN (our)	40.66	0.0211	0.50

Average time (in second) and parameters (million) for a 64×64 LR MS image with the corresponding 256×256 HR PAN image are compared.

TABLE VIII

SCALABILITY EXPERIMENTS

Training	Fine-tune	Testing	Method	PSNR ↑
WV	-	GF-test	MUCNN	34.68
WV	-	GF-test	LLERN	39.27
WV	-	GF-train	MUCNN	33.74
WV	-	GF-train	LLERN	38.34
WV	GF-test	GF-train	MUCNN	45.30
WV	GF-test	GF-train	LLERN	45.53
GF	-	QB-test	MUCNN	20.86
GF	-	QB-test	LLERN	23.51
GF	-	QB-train	MUCNN	22.77
GF	-	QB-train	LLERN	25.32
GF	QB-test	QB-train	MUCNN	32.71
GF	QB-test	QB-train	LLERN	35.33

GF-test denotes the testing data on the GF dataset.
GF-train denotes the training data on the GF dataset.
QB-test denotes the testing data on the QB dataset.
QB-train denotes the training data on the QB dataset.

tion of the PAN and MS images for embedded generation, which leads to its great generalization ability.

F. Performance in High-Level Vision Task

Pan-sharpening can serve as a preprocessing step for many high-level vision tasks, such as image segmentation and dynamic monitoring applications. Thus, the accuracy of these subsequent tasks reflects the effectiveness of pan-sharpening algorithms. In this section, we employ the classic K-Means as an unsupervised satellite image semantic segmentation method to evaluate the quality of reconstructed images from the proposed LLERN and other state-of-the-art pan-sharpening methods on the QB dataset in Fig. 9. The number of categories

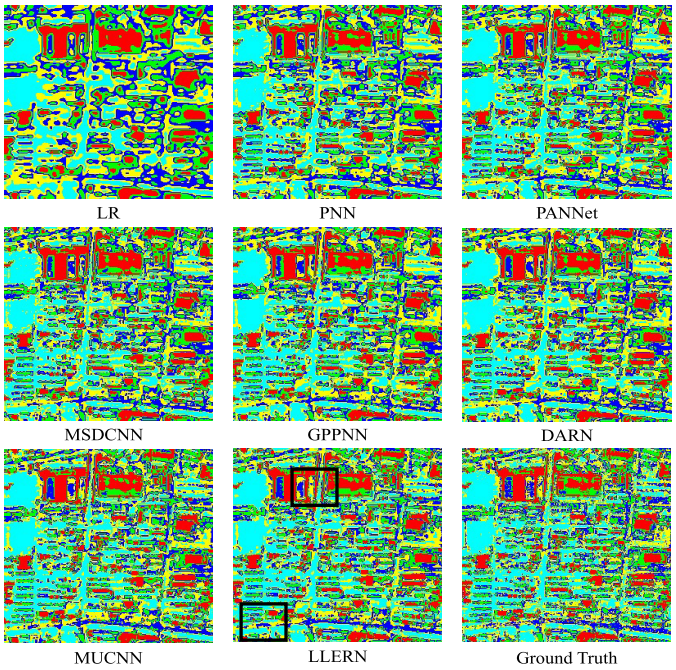


Fig. 9. Classification results via the unsupervised K-means classification method. The black boxes emphasize the areas where the proposed model outperforms other competing methods. (Please zoom in for more details.)

is set to five, with the change threshold set to 5%. Low-resolution textures are responsible for limiting the accurate segmentation process. The black boxes emphasize the areas where the proposed LLERN outperforms other competing models, as the proposed LLERN is able to recover more spatial details and textures.

IV. CONCLUSION AND FUTURE WORK

In this study, we propose a new deep LLERN for image pan-sharpening tasks. We design an LLERB to embed the high-frequency information achieved from the PAN image into the MS image in the residual space. To improve the search precision of LLERB, we divide the pan-sharpening task into two phases, i.e., the spectral preservation phase and the structural preservation phase. As the pretreatment of the structural preservation network, the spectral preservation network aims to upscale the LR MS image while retaining spectral information. The extensive experiments conducted on three satellite datasets, i.e., WV, GF, and QB, confirm the assumption that LR image patches and residual image patches in a local region share a similar manifold structure with LR MS images. Although the proposed method brings promising results, some notable issues remain and call for further research. Future works are expected to adaptive evaluate the sparse hyperparameter k_n. We also intend to build an embedding-based block in a more efficient way while maintaining the universality and expression ability of the proposed LLERN.

REFERENCES

[1] Z. Shao, H. Fu, D. Li, O. Altan, and T. Cheng, "Remote sensing monitoring of multi-scale watersheds impermeability for urban hydrological evaluation," *Remote Sens. Environ.*, vol. 232, Oct. 2019, Art. no. 111338.

[2] Z. Shao, Y. Zhang, C. Zhang, X. Huang, and T. Cheng, "Mapping impervious surfaces with a hierarchical spectral mixture analysis incorporating endmember spatial distribution," *Geo-Spatial Inf. Sci.*, pp. 1–18, Mar. 2022, doi: 10.1080/10095020.2022.2028535.

[3] X. Lv, D. Ming, T. Lu, K. Zhou, M. Wang, and H. Bao, "A new method for region-based majority voting CNNs for very high resolution image classification," *Remote Sens.*, vol. 10, no. 12, p. 1946, Dec. 2018.

[4] R. Li, S. Zheng, C. Duan, L. Wang, and C. Zhang, "Land cover classification from remote sensing images based on multi-scale fully convolutional network," *Geo-Spatial Inf. Sci.*, pp. 1–17, Jan. 2022, doi: 10.1080/10095020.2021.2017237.

[5] R. Zhang, Z. Shao, X. Huang, J. Wang, and D. Li, "Object detection in UAV images via global density fused convolutional network," *Remote Sens.*, vol. 12, no. 19, p. 3140, Sep. 2020.

[6] R. Zhang, S. Newsam, Z. Shao, X. Huang, J. Wang, and D. Li, "Multi-scale adversarial network for vehicle detection in UAV imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 180, pp. 283–295, Oct. 2021.

[7] S. Haigang, L. Deren, G. Jianya, and Z. Qing, "Analysis and representation of changes in change detection," *Geo-Spatial Inf. Sci.*, vol. 5, no. 2, pp. 13–16, Jan. 2002.

[8] C. Wu, B. Du, X. Cui, and L. Zhang, "A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion," *Remote Sens. Environ.*, vol. 199, pp. 241–255, Sep. 2017.

[9] G. Vivone, L. Alparone, J. Chanussot, M. D. Mura, A. Garzelli, G. Licciardi, R. Restaino, and L. Wald, "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, Dec. 2014.

[10] K. Zhang, F. Zhang, and S. Yang, "Fusion of multispectral and panchromatic images via spatial weighted neighbor embedding," *Remote Sens.*, vol. 11, no. 5, p. 557, Mar. 2019.

[11] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[12] K. Zhang, M. Wang, S. Yang, and L. Jiao, "Convolution structure sparse coding for fusion of panchromatic and multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1117–1130, Feb. 2019.

[13] X. X. Zhu and R. Bamler, "A sparse image fusion algorithm with application to pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2827–2836, May 2013.

[14] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2004, p. 1.

[15] X. X. Zhu, C. Grohnfeld, and R. Bamler, "Exploiting joint sparsity for pansharpening: The J-SparseFI algorithm," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2664–2681, May 2016.

[16] M. Wang, K. Zhang, X. Pan, and S. Yang, "Sparse tensor neighbor embedding based pan-sharpening via N-way block pursuit," *Knowl.-Based Syst.*, vol. 149, pp. 18–33, Jun. 2018.

[17] C. F. Caiafa and A. Cichocki, "Block sparse representations of tensors using Kronecker bases," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 2709–2712.

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

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

[20] 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.

[21] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5443–5457, Sep. 2018.

[22] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3262–3271.

[23] L. He *et al.*, "Pansharpening via detail injection based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1188–1204, Apr. 2019.

[24] 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, 2020.

[25] 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. 2090–2104, May 2021.

[26] H. Xu, J. Ma, Z. Shao, H. Zhang, J. Jiang, and X. Guo, "SDPNet: A deep network for pan-sharpening with enhanced information representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4120–4134, May 2021.

[27] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, Dec. 2021.

[28] J.-S. Choi, Y. Kim, and M. Kim, "S3: A spectral-spatial structure loss for pan-sharpening networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 5, pp. 829–833, May 2020.

[29] J. Wang, Z. Shao, X. Huang, T. Lu, R. Zhang, and J. Ma, "Pan-sharpening via high-pass modification convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1714–1718.

[30] J. Wang, Z. Shao, X. Huang, T. Lu, and R. Zhang, "A dual-path fusion network for pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[31] L.-J. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6995–7010, Aug. 2021.

[32] S. Xu, J. Zhang, Z. Zhao, K. Sun, J. Liu, and C. Zhang, "Deep gradient projection networks for pan-sharpening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1366–1375.

[33] Y. Wang, L.-J. Deng, T.-J. Zhang, and X. Wu, "SSconv: Explicit spectral-to-spatial convolution for pansharpening," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4472–4480.

[34] W. Dong, Y. Yang, J. Qu, W. Xie, and Y. Li, "Fusion of hyperspectral and panchromatic images using generative adversarial network and image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[35] W. Dong, T. Zhang, J. Qu, S. Xiao, J. Liang, and Y. Li, "Laplacian pyramid dense network for hyperspectral pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[36] W. Dong, S. Hou, S. Xiao, J. Qu, Q. Du, and Y. Li, "Generative dual-adversarial network with spectral fidelity and spatial enhancement for hyperspectral pansharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 10, 2021, doi: 10.1109/TNNLS.2021.3084745.

[37] S. Luo, S. Zhou, Y. Feng, and J. Xie, "Pansharpening via unsupervised convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4295–4310, 2020.

[38] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12797–12804.

[39] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion," *Inf. Fusion*, vol. 62, pp. 110–120, Oct. 2020.

[40] T. Lu, L. Pan, J. Wang, Y. Zhang, Z. Wang, and Z. Xiong, "AWCR: Adaptive and weighted collaborative representations for face super-resolution with context residual-learning," in *Proc. Pacific Rim Conf. Multimedia*. Cham, Switzerland: Springer, 2017, pp. 107–116.

[41] 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.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[43] J. Choi, K. Yu, and Y. Kim, "A new adaptive component-substitution-based satellite image fusion by using partial replacement," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 295–309, Jan. 2011.

[44] 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, Jan. 2008.

[45] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. II. Channel ratio and 'chromaticity' transformation techniques," *Remote Sens. Environ.*, vol. 22, no. 3, pp. 343–365, 1987.

[46] G. Vivone, L. Alparone, A. Garzelli, and S. Lolli, "Fast reproducible pansharpening based on instrument and acquisition modeling: AWLP revisited," *Remote Sens.*, vol. 11, no. 19, p. 2315, Oct. 2019.

[47] Y. Zheng, J. Li, Y. Li, G. Jie, X. Wu, and J. Chanussot, "Hyperspectral pansharpening using deep prior and dual attention residual network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8059–8076, Nov. 2020.

[48] G. Vivone *et al.*, "A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 53–81, Mar. 2021.
[49] Z. Yu *et al.*, "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5295–5305.

Jiaming Wang received the B.S. degree from the College of Post and Telecommunication, Wuhan Institute of Technology, Wuhan, China, in 2016, and the master's degree from the Wuhan Institute of Technology in 2018. He is currently pursuing the Ph.D. degree with the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, under the supervision of Prof. Zhenfeng Shao.

His research fields include image/video processing and computer vision.

Zhenfeng Shao received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2004.

Since 2009, he has been a Full Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. He has authored or coauthored over 50 peer-reviewed articles in international journals. His research interests include high-resolution image processing, pattern recognition, and urban remote sensing applications.

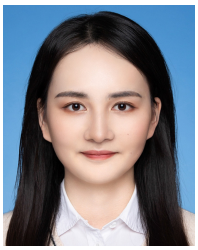
Dr. Shao was a recipient of the Talbert Abrams Award for the Best Paper in Image Matching from the American Society for Photogrammetry and Remote Sensing in 2014 and the New Century Excellent Talents in University from the Ministry of Education of China in 2012. Since 2019, he has been serving as an Associate Editor for the Photogrammetric Engineering and Remote Sensing (PE & RS) specializing in smart cities, photogrammetry, and change detection.

Xiao Huang received the B.S. degree from Wuhan University, Wuhan, China, in 2015, the master's degree in geographic information science and technology from the Georgia Institute of Technology, Atlanta, GA, USA, in 2016, and the Ph.D. degree in geography from the University of South Carolina, Columbia, SC, USA, in 2020.

He is currently an Assistant Professor with the Department of Geosciences, University of Arkansas, Fayetteville, AR, USA. His research interests cover Geospatial Artificial Intelligence (GeoAI), deep learning, and human–environmental interactions.

Tao Lu (Member, IEEE) received the B.S. and M.S. degrees from the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China, in 2003 and 2008, respectively, and the Ph.D. degree from the National Engineering Research Center For Multimedia Software, Wuhan University, Wuhan, in 2013.

He held a post-doctoral position at the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA, from 2015 to 2017. He is currently an Associate Professor with the School of Computer Science and Engineering, Wuhan Institute of Technology. He is also a Research Member with the Hubei Provincial Key Laboratory of Intelligent Robot, Wuhan Institute of Technology. His research interests include image/video processing, computer vision, and artificial intelligence.



Ruiqian Zhang received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2021.

She currently holds a post-doctoral position at the Institute of Photogrammetry and Remote Sensing, Chinese Academy of Surveying and Mapping, Beijing, China. Her research interests include image processing, pattern recognition, and remote sensing.

Gui Cheng received the B.S. degree in geographic information science from Chang'an University, Xi'an, China, in 2019. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China.

His research interests include image processing and computer vision.