Pansharpening via Neighbor Embedding of Spatial Details

Junmin Liu, Changsheng Zhou, Rongrong Fei, Chunxia Zhang, and Jiangshe Zhang

***Abstract*—The *spatial details* injection model has been considered as a general framework in the literature of *pansharpening*, and recently there have been significant advances in this framework based on *sparse representation* (SR) of spatial details. However, the SR-based methods have greater computational burden in estimating the sparse vectors and limited ability in detail edge preservation. In this paper, we introduce the *neighbor embedding* (NE) instead of SR-based model and the edge-preserving filter into the spatial detail injection framework to address the above two drawbacks. By utilizing the best quality of NE, we propose the *detail injection via neighbor embedding* (DINE) algorithm for pansharpening, and DINE+, an improved variant of DINE by using the edge-preserving filter to enhance the spatial details. Experiments carried on three data sets captured by different satellite sensors and compared with current state-of-the-art methods validate the effectiveness of the proposed methods.**

***Index Terms*—Image fusion, pansharpening, neighbor embedding, sparse representation**

I. Introduction

THE demand of high resolution images, especially the *multi-spectral* (MS) images of *high spatial resolution* (HSR), is continuously growing with their commercial use (e.g., Google Earth and Bing Maps) [1]. However, due to the physical constraints of sensors and other technical factors [2] [3], current optical Earth observation satellites, instead of acquiring HSR MS images, can only provide two *different* yet *complementary* types of images, i.e., *low spatial resolution* (LSR) MS image with fine *spectral diversity* and HSR *panchromatic* (PAN) image with rich *spatial details*. To meet this demand, *pansharpening* [4] [5], which aims at generating a HSR MS image with the spectral characteristics of the former and the spatial features of the latter by merging them together, has been proposed and become a hot topic in the literature of remote sensing.

Pansharpening can be seen as a particular branch of the general image *fusion* or *superresolution* [6] [7], and often performs as a *preliminary* step for other remote sensing tasks, such as *environment monitoring* [8], *change detection* [9], *object recognition* [10], and so on. Over the past decades, many methods have been developed for this problem, and following the recent critical surveys [11] [12] they can be broadly classified into three main classes: *component substitution* (CS)-based methods, *multiresolution analysis* (MRA)-based methods, and *variational optimization* (VO)-based methods. The three main categories of methods are based on different fusion process and thus may have *different* even *complementary* performances.

The CS-based methods first extract one component (usually called *intensity component*) from the LSR MS image through some transformations, such as *intensity-hue-saturation* (IHS) [14], *principal component analysis* (PCA) [13], *Gram-Schmidt* (GS) [15], and then totally or partially substitutes the intensity component with the histogram-matched PAN image, and finally applies the inverse transform to obtain the fused HSR MS images. Representative CS-based methods are *adaptive* GS (GSA) [16], *band-dependent spatial detail* (BDSD) [17], *partial replacement adaptive CS* (PRACS) [18]. These methods are characteristic of easy implementation and good performance in spatial detail enhancement. However, they may suffer from severe spectral distortion due to the mismatch between the intensity component and the PAN image in some local areas [19] [20]. Therefore, recent studies for improving the CS-based methods mainly focus on context-adaptive approaches [21]–[23].

The MRA-based methods are to inject spatial details extracted from the PAN image using multi-resolution decomposition, such as *undecimated wavelet transform* (UDWT) [24], *Laplacian pyramid* (LP) [25], or spatial filter such as *high-pass filter* (HPF), *morphological filter* (MF) [26], *guided filter* (GF) [27] of the PAN image, into the upscaled MS image. Some popular methods belonging to this class are the *smoothing filter-based intensity modulation* (SFIM) [28], *generalized Laplacian pyramid with context-based decision* (GLP-CBD) [29], etc. Even though the MRA-based methods have the advantage of preserving spectral consistency with the MS image, they may result in spatial artifacts, e.g. ringing and aliasing in the pansharpened MS images.

The VO-based methods treat the reconstruction of high resolution MS images as an optimization problem of a variational model [12]. Generally, they have two major parts: construction of the energy functional and the optimization solution. First, the energy functional is constructed based on the observation models of the LSR MS images [30], [31], [33]–[36], or based on the *sparse representation* (SR) theory [37]–[39]. The energy functional is generally ill-posed, thus, to deal with this problem, some priors in spatial and/or spectral spaces, e.g. *nonlocal* prior [34], low-rank prior [35], *total*

This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0105601 and by the National Natural Science Foundation of China under Grants 61877049 and 61976174. *(Corresponding author: junmin Liu.)*

J. Liu, C. Zhang, and J. Zhang are with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China.

C. Zhou is with the School of Mathematics and Information Science, Guangzhou University, Guangzhou 510006, China.

R. Fei is with the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China.

variation prior [36], *sparse* prior [37]–[40] are often added to regularize the energy functional. And then an optimization algorithm is designed to minimize the regularized energy functional. Representative works in this class are P+XS [30], *bayesian data fusion* (BDF) [31] [32], *sparse fusion of images* (SparseFI) [38], *SR-based details injection* (SR-D) [39], etc. Although presenting comparable performances with the CS-based and/or MRA-based methods, the VO-based methods are of low efficiency due to the iterative optimization procedure, which seriously hinders their application in practice.

In recent years, inspired by a dramatic performance jump of *deep learning* (DL) methods in the field of machine learning and computer vision [41], some DL-based methods are proposed in the literature of remote sensing image fusion [5] [42], shadow detection [43], and cloud detection [44]. In the literature of pansharpening, e.g. the *pansharpening by neural networks* (PNN) [45], *deep residual pan-sharpening neural network* (DRPNN) [46], *pan-sharpening network with residual blocks* (PanNet) [47], or others [48]–[53]. Similar to the VO-based methods, the DL-based methods have the drawback of requiring *copious amounts of data* to train the model, even though they have shown comparable performance to the state-of-the-art methods. We here focus our study on reducing the hue computational complexities of VO-based or DL-based methods, typically to improve the efficiency and performances of SR-based methods.

In this paper, motivated by the impressive performance of SR in the fields of remote sensing [37]–[39], and also aimed at reducing its time complexity significantly without sacrificing the quality of the fused products, a new pansharpening method is proposed based on *neighbor embedding* (NE) [54] instead of SR. Similar to the SR theory, NE is based on the assumptions that the LSR image patches have the same *manifold representations* [55] as the corresponding HSR image patches, while different from the SR-based methods that require to iteratively solve an ℓ_1-minimization problem [56] to seek the sparse representations and thus resulting in time inefficiency, the NE looks for a compact but efficient representation by expressing each LSR image patch as a weighted combination of its K-*nearest neighbors* (K-NNs) in a dictionary and leading to low time complexity [57]. By taking this advantage of NE, we introduce it into the pansharpening problem and propose the *detail injection via neighbor embedding* (DINE) method for pansharpening. In addition, to further enhance the performance of DINE and obtain realistic fused MS images, we propose DINE+, an improved version of DINE, which uses the edge-preserving filter to enhance the spatial details. Extensive experiments on three different satellite data sets have shown that the proposed DINE method makes on sacrifices on quality of results and achieves an improved computational efficiency. It should be pointed out that, different from three similar works [58] [59] [60] which mainly take NE to directly generate the whole pansharpened MS image, our proposed methods only exploit the NE framework in the spatial detail space to reconstruct the missing spatial details, which would be then injected into the available LSR MS images.

In summary, the motivation for the design of the proposed model is to exploit the neighbor embedding of spatial details and the edge-preserving filter to enhance the spatial details of the pansharpened images. The main contributions of our work are as follows.

- A novel detail injection model, referred to *detail injection via neighbor embedding* (DINE), has been proposed in this paper by exploiting the neighbor embedding of spatial details patches. Empirical results show that it is effective in reducing the spectral distortion.
- We use an edge-preserving filter to extend the DINE and have proposed an improved version of DINE for further rendering the spatial details of the fused results.

This paper is organized as follows. Some background, including the spatial detail injection model and the NE-based method to superresolution, are described in Section II. We introduce the proposed DINE and DINE+ methods in Section III. Experimental results and comparison are presented in Section IV and finally, Section V concludes the paper.

II. Background

A. Spatial Detail Injection Model

Let $\boldsymbol{Y}^{\text{pan}}$ and $\boldsymbol{Z}_i^{\text{ms}}$ denote the observed HSR PAN image and the ith band of observed LSR MS images, respectively. In general, the ideal HSR MS band image $\boldsymbol{Y}_i^{\text{ms}}$ can be modeled by its low-frequency components $f_{\text{low}}(\boldsymbol{Y}_i^{\text{ms}})$ plus high frequency components $f_{\text{high}}(\boldsymbol{Y}_i^{\text{ms}})$ from an image frequency perspective, i.e.

$$\boldsymbol{Y}_i^{\text{ms}} = f_{\text{low}}(\boldsymbol{Y}_i^{\text{ms}}) + f_{\text{high}}(\boldsymbol{Y}_i^{\text{ms}}), \quad i = 1, 2, \cdots, B, \tag{1}$$

where B is the number of bands. The low frequency components of an image correspond to the basic characteristics of the image such as the background, the flat areas, whereas its high frequency components represent the *details* of an image such as the edges or boundaries of objects. Therefore, a LSR image can be seen as degradation from a HSR image by missing high frequency components through a low-pass filtering and down-sampling operators, i.e.

$$\boldsymbol{Z}_i^{\text{ms}} = (\boldsymbol{Y}_i^{\text{ms}} * \mathcal{G}_i) \downarrow r, \tag{2}$$

where $\mathcal{G}_i$ is a low-pass filter caused by the optical imaging systems of the ith MS band, $*$ is a convolution operator, and $\downarrow r$ is a down-sampling operator by ratio r. Let the LSR MS band with the same size of HSR PAN image $\boldsymbol{Y}^{\text{pan}}$ (i.e. LSR MS band without decimation) be $\boldsymbol{X}_i^{\text{ms}}$, according to the above Eq. (2), it can be further represented by

$$\boldsymbol{X}_i^{\text{ms}} = \big((\boldsymbol{Y}_i^{\text{ms}} * \mathcal{G}_i) \downarrow r\big) \uparrow r \approx \boldsymbol{Y}_i^{\text{ms}} * \mathcal{G}_i, \tag{3}$$

where $\uparrow r$ is an up-sampling operator by ratio r. It is reasonable to assume that

$$f_{\text{low}}(\boldsymbol{Y}_i^{\text{ms}}) = \boldsymbol{X}_i^{\text{ms}}, \tag{4}$$

and then the lost high frequency component of $\boldsymbol{Y}_i^{\text{ms}}$ can be given by

$$f_{\text{high}}(\boldsymbol{Y}_i^{\text{ms}}) = \boldsymbol{Y}_i^{\text{ms}} - \boldsymbol{X}_i^{\text{ms}}. \tag{5}$$

Therefore, using Eqs. (3), (4), and (5) in Eq. (1) yields [77] [76]

$$\boldsymbol{Y}_i^{\text{ms}} = \boldsymbol{X}_i^{\text{ms}} + \boldsymbol{D}_{h,i}^{\text{ms}}, \tag{6}$$

where $\boldsymbol{D}_{h,i}^{\text{ms}} \approx \boldsymbol{Y}_i^{\text{ms}} - \boldsymbol{Y}_i^{\text{ms}} * \mathcal{G}_i$ are the missing high frequency components of $\boldsymbol{X}_i^{\text{ms}}$. The loss of high frequency components in an image appears as the missing of spatial details. According to above equation, to reconstruct the target HSR MS band image $\boldsymbol{Y}_i^{\text{ms}}$, one has to estimate the missing high resolution spatial details $\boldsymbol{D}_{h,i}^{\text{ms}}$ and then inject them into the corresponding LSR MS band image $\boldsymbol{X}_i^{\text{ms}}$. Therefore, the Eq. (6) has been coined the name of *spatial detail injection model* [76] [11]. Different ways of inferring the spatial details $\boldsymbol{D}_{h,i}^{\text{ms}}$ yield different pan-sharpening methods. For example, the CS-based methods infer the spatial details $\boldsymbol{D}_{h,i}^{\text{ms}}$ by

$$\boldsymbol{D}_{h,i}^{\text{ms}} = g_i \cdot (\boldsymbol{Y}_i^{\text{pan}} - \boldsymbol{I}), \tag{7}$$

while the MRA-based methods estimate $\boldsymbol{D}_{h,i}^{\text{ms}}$ by

$$\boldsymbol{D}_{h,i}^{\text{ms}} = g_i \cdot (\boldsymbol{Y}_i^{\text{pan}} - \boldsymbol{X}_i^{\text{pan}}), \tag{8}$$

where g_i is the injection gain, $\boldsymbol{I}$ is defined as a weighted average of the MS bands, $\boldsymbol{Y}_i^{\text{pan}}$ and $\boldsymbol{X}_i^{\text{pan}}$ respectively are the histogram-matched version of $\boldsymbol{Y}^{\text{pan}}$ and a low resolution version of the PAN image that is usually dependent on the ith MS band and derived from the available PAN image through a low-pass filter or a multi-resolution decomposition.

B. Neighbor Embedding

Inspired by the manifold learning methods, particularly the *locally linear embedding* (LLE) [55], Chang *et al.* [54] proposed the *neighbor embedding* (NE) approach for *single image super-resolution problem*. The basic idea of NE is that each LSR input patch can be represented as a weighted combination of its K-NNs selected from a LSR dictionary $\boldsymbol{D}^l$, which is a collection of atoms formulated as column vectors, and then apply the same weights to the corresponding HSR patches in a HSR dictionary $\boldsymbol{D}^h$, which is coupled with the LSR dictionary $\boldsymbol{D}^l$, to reconstruct the HSR output *patch*. Note that image as a whole is generally complicated and non-stationary, while local small image patch is simple and tend to redundantly recur many times inside the image resulting in consistent structure [37]. Therefore, the reconstruction of high resolution image is based on overlapping patches, and thus the first step is to divide the LSR input image $\boldsymbol{X}$ into $\sqrt{d} \times \sqrt{d}$ patches (with a s-pixel overlap) and ordered them lexicographically as d-dimensional column vectors, which are of the same size of that in the LSR dictionary $\boldsymbol{D}^l$, so obtaining the set of LSR input patch vectors $\mathcal{X} = \{\mathbf{x}_i | \mathbf{x}_i \in \mathcal{R}^d\}_{i=1}^N$. The detail procedures of NE is as follows.

Given the coupled dictionary set $\mathcal{D} = \{\boldsymbol{D}^l, \boldsymbol{D}^h\}$, where $\boldsymbol{D}^l = [\mathbf{d}_1^l, \mathbf{d}_2^l, \cdots, \mathbf{d}_n^l]$ is composed by a set of LSR patch vectors $\mathbf{d}_i^l \in \mathcal{R}^d, i = 1, 2, \cdots, n$ and $\boldsymbol{D}^h = [\mathbf{d}_1^h, \mathbf{d}_2^h, \cdots, \mathbf{d}_n^h]$ is composed by a set of HSR patch vectors $\mathbf{d}_i^h \in \mathcal{R}^{r^2 d}, i = 1, 2, \cdots, n$. There are three basic steps in NE for reconstructing the HSR patch vectors $\mathcal{Y} = \{\mathbf{y}_i | \mathbf{y}_i \in \mathcal{R}^{r^2 d}\}_{i=1}^N$ corresponding to the LSR input patch vectors $\mathcal{X} = \{\mathbf{x}_i | \mathbf{x}_i \in \mathcal{R}^d\}_{i=1}^N$. The first is to find K-NN set $\mathcal{N}_i = \{\mathbf{d}_{i_1}^l, \mathbf{d}_{i_2}^l, \cdots, \mathbf{d}_{i_K}^l\}$ of each input LSR patch vector $\mathbf{x}_i \in \mathcal{R}^d$ among all atoms from dictionary $\boldsymbol{D}^l$. Based on the assumption that a LSR patch and the corresponding HSR unknown patch share similar neighborhood structures, the next step is to calculate the optimal combination weights $\{w_{i_j}, j = 1, 2, \cdots, K\}$ by minimizing the linear reconstruction errors with the selected K-NNs, i.e.

$$\min_{w_{ij}} \|\mathbf{x}_i - \sum_{\mathbf{d}_{i_j}^l \in \mathcal{N}_i} w_{i_j} \mathbf{d}_{i_j}^l\|, \quad s.t. \quad \sum_{j=1}^{K} w_{i_j} = 1. \tag{9}$$

And then, the optimal weights $\{\hat{w}_{i_j}\}$ based on (9) are applied to the corresponding HSR patch vectors in the dictionary $\boldsymbol{D}^h$ to reconstruct the HSR output patch vector $\mathbf{y}_i \in \mathcal{R}^{r^2 d}$, i.e.

$$\mathbf{y}_i = \sum_{\mathbf{d}_{i_j}^h \in \mathcal{N}_i} \hat{w}_{i_j} \mathbf{d}_{i_j}^h. \tag{10}$$

The final HSR image is obtained by averaging all the reconstructed patch vectors $\mathbf{y}_i$.

C. Sparse Representation-based Details Injection

By exploiting the SR theory to construct the spatial detail image $\boldsymbol{D}_{h,i}^{\text{ms}}$, Vicinanza *et al.* [39] proposed the *SR-based details injection* (SR-D) model. Similar to NE for super-resolution, SR-D implemented the model by dividing the LSR details $\boldsymbol{D}_{h,i}^{\text{ms}}$ into overlapped patch $\mathbf{p}$, which is concatenated as column vector and assumed to be approximated as a linear combination of the atoms of a LSR dictionary $\boldsymbol{D}^l$, i.e.

$$\mathbf{p} = \boldsymbol{D}^l \boldsymbol{\alpha}, \tag{11}$$

where the dictionary $\boldsymbol{D}^l$ is constructed at the reduced scale based on the *scale invariance assumption* [17] [61], and with each column being the patch vector of the the LSR details of the PAN image. Based on the sparsity prior, namely, the coefficient vector $\boldsymbol{\alpha}$ is sparse (i.e. with few nonzero elements), SR-D tries to find the sparse coefficient vector $\hat{\boldsymbol{\alpha}}$ by solving the following *nondeterministic polynomial-time hard* (NP-hard) problem

$$\hat{\boldsymbol{\alpha}} = \arg_{\boldsymbol{\alpha}} \min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_0, \quad s.t. \quad \mathbf{p} = \boldsymbol{D}^l \boldsymbol{\alpha}, \tag{12}$$

where $\|\cdot\|_0$ is the well-known ℓ_0-norm defined as the number of nonzero elements of a vector. After getting a solution $\hat{\boldsymbol{\alpha}}$ of the problem (12), the HSR patch vector $\mathbf{q}$ corresponding to the LSR patch vector $\mathbf{p}$ can be reconstructed by

$$\mathbf{q} = \boldsymbol{D}^h \hat{\boldsymbol{\alpha}}, \tag{13}$$

where the dictionary $\boldsymbol{D}^h$ is the full scale counterpart of $\boldsymbol{D}^l$. Finally, the whole spatial detail of each band, $\boldsymbol{D}_{h,i}^{\text{ms}}$, is obtained by averaging the overlapped patch vectors $\mathbf{q}$.

Although the SR-D algorithm has shown encouraging performances in pansharpening, it suffers from high computational cost for solving the NP-hard problem (12), which usually use greedy strategies to obtain a local solution (e.g. orthogonal matching pursuit [62]) or relax the ℓ_0-norm by the ℓ_1-norm to obtain an approximate solution [63].

III. Neighbor Embedding of Spatial Details

A. DINE: detail injection by neighbor embedding

The detail injection model (6) has become a standard model in pansharpening, and most of the classical methods are developed based on it. Among them, the SR-based methods have shown significant performances in preserving the spectral information of MS images. However, the regularization term of SR-based method using the ℓ_0-norm (Eq. (12)) or its relaxed version ℓ_1-norm [38] of the coefficients make them more challenging to use due to high computational cost. In fact, the NE model is very similar to the SR-based model with the difference that, instead of solving a ℓ_0-minimization problem (Eq. (12)) for the SR theory, the NE formulates the reconstruction problem as a constrained least squares regression (Eq. (9)), which has a closed-form solution. Let $\boldsymbol{D}^l_{\mathcal{N}_i} = [\mathbf{d}^l_{i_1}, \mathbf{d}^l_{i_2}, \cdots, \mathbf{d}^l_{i_K}] \in \mathcal{R}^{d\times K}$ and $\mathbf{w}_i = (w_{i1}, w_{i2}, \cdots, w_{iK})^T \in \mathcal{R}^K$, the algebraic solution of problem (9) can be given by

$$\hat{\mathbf{w}} = \frac{\boldsymbol{H}_i^{-1}\mathbf{1}}{\mathbf{1}^T\boldsymbol{H}_i^{-1}\mathbf{1}}, \tag{14}$$

where $\mathbf{1}$ is a column vector of ones and

$$\boldsymbol{H}_i = (\mathbf{x}_i\mathbf{1}^T - \boldsymbol{D}^l_{\mathcal{N}_i})^T(\mathbf{x}_i\mathbf{1}^T - \boldsymbol{D}^l_{\mathcal{N}_i}). \tag{15}$$

Thus, the NE model can boost performance speed but shares more properties with sparse representation model.

Motivated by the low computational complexity of NE and its simplicity to be understood, we here use the NE model instead of the SR model to perform linear reconstruction of the corresponding HSR details patches, and propose the DINE method. The proposed DINE method, similar to [39], exploits the *self-similarity* of spatial details based on the *scale invariance assumption* [64], which has already been proven profitable for classical methods and recently developed DL-based methods, to discover the representation coefficients. The procedure of DINE is described in Algorithm 1. The inputs of DINE are the LSR MS images $\boldsymbol{Z}^{\text{ms}}_i, i = 1, 2, \cdots, N$, and the HSR PAN image $\boldsymbol{Y}^{\text{pan}}$. The first step is to upsample the observed LSR MS images into the size of PAN and then obtain the histogram-matched PAN images based on the upsampled MS images, which could reduce the the spectral mismatch between the PAN and each MS band [16], [18], [65]. The second step is to extract the HSR details from the histogram matched image at full scale by simulating the image generation. While steps 3 and 4 extract the LSR details from the LSR PAN and the LSR MS images at reduced scale, respectively. Note that the *modulation transfer functions* (MTFs) of the satellite sensors are taken into account to make the results obey the *spatial consistency* property [64]. Step 5 is to divide the spatial detail images into overlapping patches and transform into vectors to generate the coupled HSR and LSR dictionaries. And steps 6-10 is to estimate the missing HSR spatial details by the NE method. Finally, the missing HSR detail for each MS band is estimated based on step 11 and the reconstructed HSR MS images are obtained by step 12.

Algorithm 1 DINE(+) algorithm

Input: LSR MS images $\boldsymbol{Z}^{\text{ms}}_i, i = 1, 2, \cdots, B$, HSR PAN image $\boldsymbol{Y}^{\text{pan}}$, the number of neighbors K.

Output: HSR MS images $\boldsymbol{Y}^{\text{ms}}_i, i = 1, \cdots, B$.

1: Obtain the expanded LSR MS images $\boldsymbol{X}^{\text{ms}}_i$ by **upsamling** $\boldsymbol{Z}^{\text{ms}}_i$ to the size of PAN $\boldsymbol{Y}^{\text{pan}}$, and the **histogram-matched** PAN image $\boldsymbol{Y}^{\text{pan}}_i$ by

$$\boldsymbol{Y}^{\text{pan}}_i = (\boldsymbol{Y}^{\text{pan}} - \mu_{\boldsymbol{Y}^{\text{pan}}}) \cdot \frac{\sigma_{\boldsymbol{X}^{\text{ms}}_i}}{\sigma_{\boldsymbol{Y}^{\text{pan}}}} + \mu_{\boldsymbol{X}^{\text{ms}}_i}, \tag{16}$$

where $\mu_{(\cdot)}$ and $\sigma_{(\cdot)}$ denote the mean and standard deviation of an image.

2: **Extract** the HSR details $\boldsymbol{D}^{\text{pan}}_{h,i}$ from **histogram-matched** PAN image $\boldsymbol{Y}^{\text{pan}}_i$ **at full scale** by

$$\boldsymbol{D}^{\text{pan}}_{h,i} = \boldsymbol{Y}^{\text{pan}}_i - \big((\boldsymbol{Y}^{\text{pan}}_i * \mathcal{G}^{\text{MTF}}_i) \downarrow r\big) \uparrow r, \tag{17}$$

where $\mathcal{G}^{\text{MTF}}_i$ is a MTF-matched filter [29].

3: **Extract** the LSR details $\boldsymbol{D}^{\text{pan}}_{l,i}$ from the LSR PAN image $\boldsymbol{Z}^{\text{pan}}_i$ **at reduced scale** by

$$\boldsymbol{D}^{\text{pan}}_{l,i} = \boldsymbol{Z}^{\text{pan}}_i - \big((\boldsymbol{Z}^{\text{pan}}_i * \mathcal{G}^{\text{MTF}}_i) \downarrow r\big) \uparrow r, \tag{18}$$

where $\boldsymbol{Z}^{\text{pan}}_i = (\boldsymbol{Y}^{\text{pan}}_i * \mathcal{G}^{\text{MTF}}) \downarrow r$ and $\mathcal{G}^{\text{MTF}}$ is a MTF-matched filter [29] corresponding to the PAN sensor.

4: **Extract** the LSR details $\boldsymbol{D}^{\text{ms}}_{l,i}$ from each LSR MS band $\boldsymbol{Z}^{\text{ms}}_i$ at **reduced scale** by

$$\boldsymbol{D}^{\text{ms}}_{l,i} = \boldsymbol{Z}^{\text{ms}}_i - \big((\boldsymbol{Z}^{\text{ms}}_i * \mathcal{G}^{\text{MTF}}_i) \downarrow r\big) \uparrow r. \tag{19}$$

Note that $\boldsymbol{D}^{\text{ms}}_{l,i}$ and $\boldsymbol{D}^{\text{pan}}_{l,i}$ have the same image size.

5: **Generate** the HSR and LSR dictionaries $\boldsymbol{D}_h$ and $\boldsymbol{D}_l$ with the overlapping patches (transformed into vectors) of $\boldsymbol{D}^{\text{pan}}_{h,i}$ and $\boldsymbol{D}^{\text{pan}}_{l,i}$ as shown in Fig. 1, which is similar to [39], and the LSR detail patch vectors $\mathcal{P} = \{\mathbf{p}_i\}^M_{i=1}$ from $\boldsymbol{D}^{\text{ms}}_{l,i}, i = 1, \cdots, B$.

6: **For** each LSR patch vector $\mathbf{p}_i \in \mathcal{P}$ **do**

7: Find its K-NNs $\mathcal{N}_i = \{\mathbf{d}^l_{i_1}, \cdots, \mathbf{d}^l_{i_K}\}$ among all atoms from $\boldsymbol{D}_l$.

8: Compute the optimal weights $\{\hat{w}_{ij}\}^K_{j=1}$ which best approximate $\mathbf{x}_i$ with a linear combination of $\mathcal{N}_i$ by Eqs. (14) and (15).

9: Apply the same weights to estimate the corresponding HSR detail patch vector $\mathbf{q}_i$ with the corresponding neighbors in $\boldsymbol{D}_h$ as follows

$$\hat{\mathbf{q}}_i = \sum_{\mathbf{d}^h_{i_j} \in \mathcal{N}_i} \hat{w}_{ij}\mathbf{d}^h_{i_j}. \tag{20}$$

10: **EndFor**

11: Combine the estimated HSR detail patch vectors $\mathcal{Q} = \{\hat{\mathbf{q}}_i\}^M_{i=1}$ to form the HSR details $\widehat{\boldsymbol{D}}^{\text{ms}}_{h,i}, i = 1, \cdots, B$.

12: Obtain the HSR MS images by

$$\boldsymbol{Y}^{\text{ms}}_i = \boldsymbol{X}^{\text{ms}}_i + \widehat{\boldsymbol{D}}^{\text{ms}}_{h,i} \quad \text{for DINE}, \tag{21}$$

or by

$$\boldsymbol{Y}^{\text{ms}}_i = \boldsymbol{X}^{\text{ms}}_i + \widehat{\boldsymbol{G}}_i \odot \widehat{\boldsymbol{D}}^{\text{ms}}_{h,i} \quad \text{for DINE+}, \tag{22}$$

where $\odot$ is the Hadamard product and $i = 1, 2, \cdots, B$.

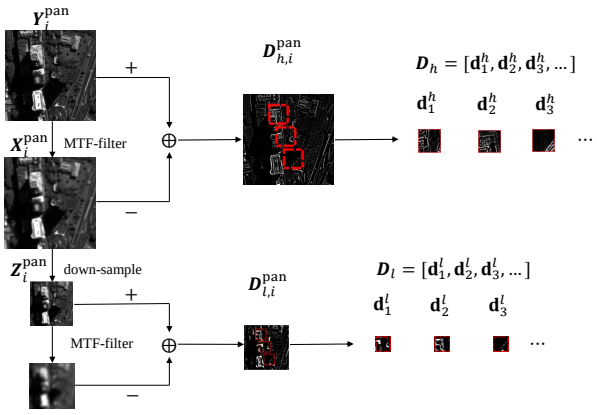

Fig. 1. The generation of HSR and LSR dictionaries from the PAN image. Note that the same procedure is used for the MS image to generate the LSR MS image patches.

Remark 1. Currently, several works [58] [59] [60] share the similar philosophy of the proposed DINE method. Although they are all inspired by the manifold learning and based on NE for the fusion of PAN and MS images, they are different from our proposed DINE method in that they consider generating the whole image patches in the pixel or feature space, while the proposed DINE method apply the NE to directly reconstruct the missing HSR detail patches (or image residuals) instead of the pixel patches. It has been shown empirically [39] [48] [49] that working on spatial details (or image residuals) may boost fusion accuracy and efficiency.

Remark 2. The running time of the recently developed pan-sharpening methods, especially the SR-based methods, is often ignored despite the fact that it is a critical factor for real applications. The proposed DINE method is very simple and aims at fast running while making no compromise on quality compared to the SR-based methods [37] [38] [39]. Since NE has a closed-form solution for the representation inference, therefore the proposed DINE based on NE is very fast in the HSR detail patch reconstruction.

Remark 3. It should be noted that the proposed DINE algorithm (*i.e.*, **Algorithm 1**) involves a procedure of selecting the K-NNs set $\mathcal{N}_i$ for each LSR patch. Usually, we find the K-NNs set $\mathcal{N}_i$ of the LSR patch $\mathbf{p}_i$ in $\boldsymbol{D}_l$ in terms of Euclidean distance as

$$\mathcal{N}_i = \underset{\{\mathbf{d}_{i_j}^l\}_{j=1}^K \in \boldsymbol{D}_l}{\arg\min} \sum_{j=1}^{K} \|\mathbf{p}_i - \mathbf{d}_{i_j}^l\| \tag{23}$$

The distance metric of a data point to all other data points can be any type of distance. And the size of the neighborhood K that is used by the DINE method is a critical parameter. In the following, we will show how sensitive this parameter is for our proposed methods. Additionally, the computation cost of our proposed DINE mainly relies on that of the nearest neighbor search, it will be high as the number of dimension and data points grows because of calculating the distance between the data points to all other data points.

B. DINE+: Spatially Improved DINE by Edge-Preserving Model

Although having an excellent time complexity, the DINE suffers from limited capacity in detail enhancement, as will be shown in the following experiments, possibly due to two reasons. First, the reconstruction based on the linear combination of the K-NNs may lead to the under-fitting problem. Second, according to Eq. (20), the reconstructed HSR details $\widehat{\boldsymbol{D}}_{h,i}^{\text{ms}}, i = 1, \cdots, B$ of the HSR MS images can be seen as a weighted average of the HSR details of the HSR PAN image, leading to smooth fusion results. To overcome these drawbacks, we propose DINE+, an improved variant of DINE, by introducing the edge-preserving filter to enhance the contrast of the fine details in each MS band.

With the above in mind, we would like to weight the estimated HSR details $\widehat{\boldsymbol{D}}_{h,i}^{\text{ms}}, i = 1, \cdots, B$ by exploring the local linear relationships between each ideal HSR MS details $\boldsymbol{D}_{h,i}^{\text{ms}}$ and HSR PAN details $\boldsymbol{D}_{h,i}^{\text{pan}}$ in a window ω_k centered at the pixel k, i.e. for each pixel $j \in \omega_k$, we have

$$\boldsymbol{D}_{h,i}^{\text{ms}}(j) = \boldsymbol{G}_i(k)\boldsymbol{D}_{h,i}^{\text{pan}}(j) + \boldsymbol{B}_i(k), \tag{24}$$

where $\boldsymbol{D}_{h,i}^{\text{ms}}(j)$ and $\boldsymbol{D}_{h,i}^{\text{pan}}(j)$ are the gray values of the spatial detail images $\boldsymbol{D}_{h,i}^{\text{ms}}$ and $\boldsymbol{D}_{h,i}^{\text{pan}}$ located at the pixel j, $\boldsymbol{G}_i(k)$ and $\boldsymbol{B}_i(k)$ are some linear coefficients corresponding the pixel k, with constant value assumption in the window ω_k for the ith MS band. By also resorting to the scale invariance assumption [17] [39] [64] of the above model, we have

$$\boldsymbol{D}_{l,i}^{\text{ms}}(j) = \boldsymbol{G}_i(k)\boldsymbol{D}_{l,i}^{\text{pan}}(j) + \boldsymbol{B}_i(k), \tag{25}$$

where $\boldsymbol{D}_{l,i}^{\text{ms}}$ and $\boldsymbol{D}_{l,i}^{\text{pan}}$ are the LSR spatial details of the MS and PAN images and obtained by Eqs. (18) and (19), respectively. Since $\boldsymbol{D}_{l,i}^{\text{ms}}$ and $\boldsymbol{D}_{l,i}^{\text{pan}}$ are known, we can determine the linear coefficients by minimizing the following objective function with respect to $\boldsymbol{G}_i(k)$ and $\boldsymbol{B}_i(k)$:

$$\mathcal{L} = \sum_{j \in \omega_k} \left[\boldsymbol{G}_i(k)\boldsymbol{D}_{l,i}^{\text{pan}}(j) + \boldsymbol{B}_i(k) - \boldsymbol{D}_{l,i}^{\text{ms}}(j)\right]^2. \tag{26}$$

By taking the partial gradient to be zero, the coefficients can be explicitly expressed by

$$\boldsymbol{G}_i(k) = \frac{\frac{1}{|\omega_k|}\sum_{j \in \omega_k} \xi_{l,i}(j)}{\sigma_{l,i}^{\text{pan}}(k)}, \tag{27}$$

$$\boldsymbol{B}_i(k) = \boldsymbol{\mu}_{l,i}^{\text{pan}}(k) - \boldsymbol{G}_i(k)\boldsymbol{\mu}_{l,i}^{\text{ms}}(k), \tag{28}$$

where

$$\xi_{l,i}(j) = \boldsymbol{D}_{l,i}^{\text{pan}}(j)\boldsymbol{D}_{l,i}^{\text{ms}}(j) - \boldsymbol{\mu}_{l,i}^{\text{pan}}(k)\boldsymbol{\mu}_{l,i}^{\text{ms}}(k), \tag{29}$$

$\boldsymbol{\mu}_{l,i}^{\text{pan}}(k)$ and $\sigma_{l,i}^{\text{pan}}(k)$ are the mean and variance of $\boldsymbol{D}_{l,i}^{\text{pan}}$ in window ω_k, $\boldsymbol{\mu}_{l,i}^{\text{ms}}(k)$ is the mean of $\boldsymbol{D}_{l,i}^{\text{ms}}$ in ω_k, and $|\omega_i|$ is the number of pixels in window ω_k. The linear model (24) is applied to all local windows in the whole image. However, the value of $\boldsymbol{D}_{l,i}^{\text{ms}}(j)$ obtained by Eq. (24) is not identical in all the overlapping windows ω_k that cover the jth pixel. To deal with this, we average all the possible values of $\boldsymbol{G}_i(k)$ and $\boldsymbol{B}_i(k)$ and then obtain the local linear expression

$$\boldsymbol{D}_{l,i}^{\text{ms}}(j) = \widehat{\boldsymbol{G}}_i(j)\boldsymbol{D}_{l,i}^{\text{pan}}(j) + \widehat{\boldsymbol{B}}_i(j), \tag{30}$$

where

$$\widehat{\boldsymbol{G}}_i(j) = \frac{1}{|\omega_j|}\sum_{k\in\omega_j}\boldsymbol{G}_i(k), \widehat{\boldsymbol{B}}_i(j) = \frac{1}{|\omega_j|}\sum_{k\in\omega_j}\boldsymbol{B}_i(k). \quad (31)$$

This local linear model (30) ensures that the edges (gradient information) in $\boldsymbol{D}_{l,i}^{\text{ms}}$ (or $\boldsymbol{D}_{h,i}^{\text{ms}}$) can be preserved, because we have $\nabla\boldsymbol{D}_{l,i}^{\text{ms}} = \widehat{\boldsymbol{G}}_i \odot \nabla\boldsymbol{D}_{l,i}^{\text{pan}}$ (or $\nabla\boldsymbol{D}_{h,i}^{\text{ms}} = \widehat{\boldsymbol{G}}_i \odot \nabla\boldsymbol{D}_{h,i}^{\text{pan}}$), where $\widehat{\boldsymbol{G}}_i$ is the coefficient matrix composed of $\widehat{\boldsymbol{G}}_i(j)$, $\nabla(\cdot)$ denotes the gradients of an image, and $\odot$ is the Hadamard product. This edge-preserving property of the model (refer to [27] [66] [67] for more details) motivates us to apply the coefficient matrix $\widehat{\boldsymbol{G}}_i$ as injection gain to weight the estimated HSR details $\widehat{\boldsymbol{D}}_{h,i}^{\text{ms}}$ instead of directly using it to reconstruct the HSR MS images, thus one has

$$\boldsymbol{Y}_i^{\text{ms}} = \boldsymbol{X}_i^{\text{ms}} + \widehat{\boldsymbol{G}}_i \odot \widehat{\boldsymbol{D}}_{h,i}^{\text{ms}}, i = 1, \cdots, B. \quad (32)$$

Here we refer to the pansharpening algorithm based on Eq. (32) as the DINE+ algorithm.

IV. Experimental Results and Analysis

In this section, we apply the proposed DINE and DINE+ algorithms to three remote sensing image data sets and compare them with the baselines (including the traditional methods and the recently developed DL-based methods) to analysis their performances.

A. Experimental settings

Baselines. For comparison, we select the classical methods (i.e. the CS- and MRA-based ones) and the recently developed methods (i.e. the DL- and SR-based ones) as the baselines, including:

- EXP, which is the expanded MS image, *i.e.*, the original MS image upsampled to the panchromatic scale by using a polynomial with 23 coefficients interpolator [25] [78].
- GSA [16], which is one of the typical CS-based methods and can be seen as an improved version of the well-known GS method [15].
- PRACS [18], which is based on the CS framework by partial replacement of the PAN image and statistical ratio-based injection gains.
- *Nonlinear IHS* (NLIHS) [22], which is a nonlinear version of the IHS method by utilizing nonlinear synthesis approach instead of the common linear one to approximate the intensity component.
- MF [26], which is a recently proposed method belonging to the MRA-based class.
- GFPCA [68], which is a hybrid method of the CS- and MRA-based classes by applying the guided filter in the PCA domain.
- SR-D [39], which is a representative method of VO-based and/or SR-based methods and has shown a clear superiority against the classical ones.
- APNN [48], which is an advanced version of the DL-based pansharpening method PNN [45].
- SFIM [28], which aims at using ratios between the high resolution PAN image pixels and its low resolution ones.
- MTF-GLP-CBD [29], which employs the *generalized Laplacian Pyramid* (GLP) with filters matching to sensor MTF.
- ATWT [76], which is based on the *"à trous" wavelet transform* (ATWT) for extracting the spatial details.

The proposed methods and baselines were implemented with MATLAB in a machine equipped with a single Intel Core i3 CPU, 16GB of memory. To make fair comparisons, all parameter settings of baselines are the same as that depicted in their papers [16], [18], [22], [26], [28], [39], [68]. For the APNN [48], we use its fine-tune version with 50 iterations. It should be pointed out that the way that the low-resolution MS images expanded into the size of the high-resolution PAN images have an important effect on the fusion results, we have used the polynomial with 23 coefficients interpolator [25] to get the expanded MS images for our proposed methods and all of the baselines.

Data Sets. To evaluate the methods, we conducted experiments on three data sets, which were collected by three different satellites, namely, ***QuickBird***, ***WorldView-2***, ***GeoEye-1***. As shown in Fig. 2, each data set consists of 16 pairs of MS and PAN images. These data sets capture image with different scenes such as rural or urban area and with different objects such as ship, buildings, cars, roads, and so on. We believe that these images under different conditions provide a comprehensive validation for performances. Additionally, Table I provides the detail parameters for these three data sets and satellites. For simplicity, we use QB to denote QuickBird, GE1 to denote GeoEye-1, and WV2 to denote WorldView-2 throughout the figures, tables, and texts in the following.

Evaluation Indexes. The quantitative evaluation of the pansharpening results is a challenge problem, since the ideal HSR MS images are not available. Generally, quantitative comparison can be performed in the following two ways: evaluation at a *reduced scale* and evaluation at a *full scale*. The first way follows a Wald's protocol [64], which is based on the scale invariance assumption. The Wald's protocol performs evaluation by first reducing the resolutions of the original LSR MS and HR PAN images, then fusing the MS and PAN images at a reduced scale, and finally measuring the quality with the original LSR MS image as a reference. By this way, the following indexes can be used to quantitatively measure the performance of the different pansharpening methods:

- *Root Mean Square Error* (RMSE) [69] is defined as

$$\text{RMSE}(k) \triangleq \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{Y}_k^{\text{ms}}(j) - \widehat{\boldsymbol{Y}}_k^{\text{ms}}(j))^2}, \quad (33)$$

  where N is the number of pixels for the kth band HSR MS reference $\boldsymbol{Y}_k^{\text{ms}}$ and pansharpened MS image $\widehat{\boldsymbol{Y}}_k^{\text{ms}}$.
- *Erreur Relative Globale Adimensionnelle de Synthése* (ERGAS, or relative dimensionless global error in synthesis) [70] is defined as

$$\text{ERGAS} \triangleq \frac{100}{\beta}\sqrt{\frac{1}{B}\sum_{k=1}^{B}\left(\frac{\text{RMSE}(k)}{\mu_{\boldsymbol{Y}_k^{\text{ms}}}}\right)^2}, \quad (34)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3067877, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

(a) QuickBird Data Set

(b) GeoEye-1 Data Set

(c) Worldview-2 Data Set

Fig. 2. The satellite image data sets used in our experiments. Note that only the MS images are presented for the purpose of visualization.

TABLE I
PARAMETERS OF DIFFERENT SATELLITES.

Satellites		QuickBird	WorldView-2	GeoEye-1
Launch date		18 October 2001	08 October 2009	06 September 2008
Temporal resolution		1-5 days	1.1 day	<3 days
Radiometric resolution (bits)		11	11	11
Spatial resolution	PAN	0.6 meter	0.46 meter	0.41 meter
	MS	2.4 meter	1.84 meter	1.64 meter
Spectral range (MTF gain)	PAN	450-900 nm (0.15)	450-800 nm (0.11)	450-800 nm (0.16)
	Blue	450-520 nm (0.34)	450-510 nm (0.35)	450-510 nm (0.23)
	Green	520-600 nm (0.32)	510-580 nm (0.35)	510-580 nm (0.23)
	Red	630-690 nm (0.30)	630-690 nm (0.35)	655-690 nm (0.23)
	NIR1	760-900 nm (0.22)	770-895 nm (0.35)	780-920 nm (0.23)
	Red edge		705-745 nm (0.35)	
	Coastal		400-450 nm (0.35)	
	Yellow		585-625 nm (0.27)	
	NIR2		860-1040 nm (0.35)	
Image size	PAN	1024×1024	1024×1024	1024×1024
	MS	256×256	256×256	256×256

where β is the scale ratio between the PAN and the original MS images, B is the number of bands, and $\mu_{\boldsymbol{Y}_k^{\text{ms}}}$ is the mean of the kth reference MS band $\boldsymbol{Y}_k^{\text{ms}}$.

- Q4/Q8 [71] [72] is an extension of the *universal image quality index* (UIQI) [73]. Q4 is defined as

$$\text{Q4} = \frac{\sigma_{\boldsymbol{\omega}_1\boldsymbol{\omega}_2}}{\sigma_{\boldsymbol{\omega}_1}\sigma_{\boldsymbol{\omega}_2}} \cdot \frac{2|\mu_{\boldsymbol{\omega}_1}||\mu_{\boldsymbol{\omega}_1}|}{|\mu_{\boldsymbol{\omega}_1}|^2 + |\mu_{\boldsymbol{\omega}_2}|} \cdot \frac{2\sigma_{\boldsymbol{\omega}_1}\sigma_{\boldsymbol{\omega}_2}}{\sigma_{\boldsymbol{\omega}_1}^2 + \sigma_{\boldsymbol{\omega}_2}^2}, \quad (35)$$

where $\boldsymbol{\omega}_1 = \boldsymbol{Y}_1^{\text{ms}} + \text{i}\boldsymbol{Y}_2^{\text{ms}} + \text{j}\boldsymbol{Y}_3^{\text{ms}} + \text{k}\boldsymbol{Y}_4^{\text{ms}}$, $\boldsymbol{\omega}_2 = \widehat{\boldsymbol{Y}}_1^{\text{ms}} + \text{i}\widehat{\boldsymbol{Y}}_2^{\text{ms}} + \text{j}\widehat{\boldsymbol{Y}}_3^{\text{ms}} + \text{k}\widehat{\boldsymbol{Y}}_4^{\text{ms}}$, $\boldsymbol{Y}_k^{\text{ms}}$ and $\widehat{\boldsymbol{Y}}_k^{\text{ms}}$ are the $k$the band of the reference and fused MS images, respectively. Here, i, j and k are imaginary units, $\mu_{\boldsymbol{\omega}}$ and $\sigma_{\boldsymbol{\omega}}$ are the mean and variance of variable $\boldsymbol{\omega}$, and $\sigma_{\boldsymbol{\omega}_1\boldsymbol{\omega}_2}$ is the covariance between $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$. Q4 is usually calculated using a sliding window (typically 32×32 in our experiments) and averaged on the entire image. Q4 has been extended to Q8 index such that it is suitable for images whose number of bands is any power of two, refer to [72] for

TABLE II
THE PERFORMANCES IN TERMS OF SAM (AT REDUCED SCALE) AND QNR (AT FULL SCALE) FOR THE PROPOSED DINE AND DINE+ ALGORITHMS UNDER DIFFERENT NUMBER OF NEIGHBORS K ON THE QUICKBIRD AND GEOEYE-1 DATA SETS.

Data sets	Methods	Indexes	Number of Neighbors (K)													
			3	4	5	6	7	8	9	10	11	12	13	14	15	16
QB	DINE	SAM	2.885	2.867	2.862	2.862	2.860	2.909	2.925	2.944	2.962	2.982	3.000	3.017	3.035	3.051
		QNR	0.933	0.935	0.937	0.937	0.937	0.937	0.937	0.937	0.936	0.936	0.936	0.936	0.936	0.936
	DINE+	SAM	3.267	3.282	3.271	3.252	3.222	3.295	3.261	3.227	3.296	3.266	3.236	3.227	3.994	3.976
		QNR	0.960	0.961	0.961	0.961	0.960	0.959	0.958	0.958	0.957	0.957	0.956	0.956	0.956	0.955
GE1	DINE	SAM	5.581	5.575	5.562	5.556	5.542	5.545	5.549	5.544	5.550	5.552	5.554	5.552	5.598	5.594
		QNR	0.949	0.952	0.953	0.955	0.956	0.956	0.950	0.950	0.952	0.953	0.950	0.948	0.946	0.945
	DINE+	SAM	6.157	6.160	6.138	6.120	6.086	6.081	6.083	6.085	6.164	6.169	6.162	6.188	6.164	6.142
		QNR	0.916	0.916	0.917	0.917	0.919	0.919	0.919	0.919	0.919	0.918	0.918	0.918	0.918	0.918

TABLE III
THE PERFORMANCES IN TERMS OF SAM (AT REDUCED SCALE) AND QNR (AT FULL SCALE) FOR THE PROPOSED DINE AND DINE+ ALGORITHMS UNDER DIFFERENT PATCH SIZES ON THE QUICKBIRD AND GEOEYE-1 DATA SETS.

Data sets	Methods	Indexes	Patch Size													
			3	4	5	6	7	8	9	10	11	12	13	14	15	16
QB	DINE	SAM	0.398	0.399	0.400	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401	0.401
		QNR	0.996	0.995	0.994	0.993	0.993	0.994	0.993	0.995	0.992	0.991	0.987	0.987	0.986	0.986
	DINE+	SAM	0.426	0.428	0.428	0.428	0.428	0.429	0.429	0.429	0.429	0.429	0.429	0.429	0.429	0.429
		QNR	0.988	0.987	0.987	0.986	0.986	0.986	0.986	0.986	0.986	0.985	0.985	0.985	0.985	0.984
GE1	DINE	SAM	4.567	4.784	4.899	4.956	5.024	5.075	5.080	5.047	4.963	4.987	5.013	4.984	4.973	4.968
		QNR	0.967	0.924	0.948	0.964	0.962	0.959	0.945	0.932	0.920	0.907	0.903	0.896	0.885	0.884
	DINE+	SAM	5.023	5.417	5.523	5.599	5.583	5.548	5.498	5.478	5.339	5.306	5.273	5.227	5.227	5.142
		QNR	0.957	0.956	0.952	0.945	0.939	0.929	0.920	0.913	0.903	0.896	0.893	0.889	0.882	0.881

more details.

- *Spectral Angle Mapper* (SAM) [74] between two image vectors $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$ is defined as

$$\mathrm{SAM}(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) \triangleq \arccos\left(\frac{\langle \boldsymbol{Y}, \widehat{\boldsymbol{Y}} \rangle}{\|\boldsymbol{Y}\|_2 \|\widehat{\boldsymbol{Y}}\|_2}\right), \tag{36}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product and $\| \cdot \|_2$ denotes the l_2-norm.

The ideal values of the full-reference quality indexes ERGAS, RMSE, SAM, and Q4 (or Q8) are 0, 0, 0, 1, respectively. The RMSE and SAM indexes are averaged over all of the MS bands to obtain an overall score. Among them, the RMSE, ERGAS, and $Q4$ (or $Q8$) mainly quantify the spatial distortions, while the SAM is used to quantify the spectral quality.

As for the full scale assessment, the *Quality with No Reference* (QNR) index [75] is often applied to measure the quality of the fused product. The QNR index is defined as

$$\mathrm{QNR} \triangleq (1 - D_\lambda)(1 - D_s), \tag{37}$$

where D_λ is a *spectral* quality index, defined as

$$D_\lambda = \frac{1}{B(B-1)} \sum_{k=1}^{B} \sum_{l=1}^{B} |Q(\boldsymbol{Y}_k^{\mathrm{ms}}, \boldsymbol{Y}_l^{\mathrm{ms}}) - Q(\widehat{\boldsymbol{Y}}_k^{\mathrm{ms}}, \widehat{\boldsymbol{Y}}_l^{\mathrm{ms}})|,$$

and D_s is a *spatial* quality index, defined as

$$D_s = \frac{1}{B} \sum_{k=1}^{B} |Q(\hat{\boldsymbol{Y}}_k^{\mathrm{ms}}, \boldsymbol{Y}^{\mathrm{pan}}) - Q(\boldsymbol{Y}_k^{\mathrm{ms}}, \boldsymbol{Y}^{\mathrm{pan}})|. \tag{38}$$

Here $Q(\cdot)$ is the *universal image quality index* (UIQI) [73]. The optimal values of D_λ and D_s are 0 and the optimum theoretical value of QNR is 1.

In addition, a ***visual analysis*** on a false color display of the fused product is often used to subjectively evaluate the quality of the pansharpening results. The visual analysis is to inspect whether the objects in the fused images are clear and/or the colors are similar to the original LSR MS ones,

B. Parameters Analysis

There are two key parameters: the number of neighbors K and the patch size, which may have an impact on the performances of our proposed DINE and DINE+ methods. Therefore, we here assess how the two parameters influence the results.

There is no particular rules to determine the best K, so we have to try some values to find the best setting. Table II reports the SAM values at reduced scale and QNR values at full scale obtained when DINE and DINE+ implement with different K. We observe that, a very low value for K such as $K = 3$ or a very high value such as $K \geq 8$, can be noisy and lead to a little worse results. When the number of neighbors is 7, both the DINE and DINE+ achieve the best results in most cases. Thus, We could conclude that setting the number of neighbors to be 7 is a good heuristic towards good performances.

Table III reports the SAM values at reduced scale and QNR values at full scale obtained by using DINE and DINE+ under different patch size when the number of neighbors fixed to be 7. As it can be found, the larger patch size the worse. This may due to that patch-based methods attempt to reconstruct pixels by using similar neighboring pixels, while larger patch size may lead to dissimilar matching of patches, resulting in redundant spatial details injection.

Therefore, we respectively set $K = 7$ and the patch size to 3 for DINE and DINE+ in the subsequent experiments.

C. Results Analysis

The quantitative assessment has been performed both at reduced and full scale experiments.

Reduced Scale Experiments. In this case, we follow Wald's protocol [64] to quantitatively assess the results, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3067877, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Fig. 3. Fusion results on a typical scene from the QB data set at the reduced scale experiments. From left to right and top to bottom are the observed MS image, the fusion results of MTF-GLP-CBD [29], GSA [16], PRACS [18], NLIHS [22], MF [26], GFPCA [68], SR-D [39], ATWT [76], SFIM [28], DINE and DINE+. Best viewed in color. Note that the model of APNN [48] is not available (n/a) for QB data set.

TABLE IV

THE MEAN VALUES IN TERM OF ERGAS, RMSE, SAM, AND Q4/Q8 OBTAINED BY OUR PROPOSED DINE AND DINE+, AND BY OTHER STATE-OF-THE-ART METHODS SUCH AS GSA [16], PRACS [18], NLIHS [22], MF [26], GFPCA [68], SR-D [39], APNN [48], SFIM [28], GLP-CBD [29], ATWT [76] AND THE EXP IMAGES EVALUATED ON THREE DATA SETS AT THE REDUCED SCALE EXPERIMENTS.

Dataset	Indexes	EXP	GSA [16]	PRACS [18]	NLIHS [22]	MF [26]	GFPCA [68]	SR-D [39]	APNN [48]	SFIM [28]	GLP-CBD [29]	ATWT [76]	DINE	DINE+
QB	ERGAS	2.078	1.777	1.651	1.885	1.608	2.134	1.574	n/a	1.885	1.661	1.598	1.504	1.548
	RMSE	21.75	18.71	17.43	19.97	16.82	22.34	16.59	n/a	19.55	16.93	16.37	15.99	16.36
	SAM	2.244	2.328	2.042	2.265	1.975	2.536	1.946	n/a	2.097	2.026	1.962	1.854	1.940
	Q4	0.607	0.715	0.771	0.667	0.758	0.577	0.773	n/a	0.759	0.706	0.720	0.820	0.811
GE1	ERGAS	5.506	4.400	4.310	4.856	4.071	5.704	3.827	3.821	4.894	4.057	3.816	3.524	3.811
	RMSE	50.34	41.27	40.59	44.65	37.84	53.04	33.94	33.01	45.31	35.16	33.09	32.97	36.50
	SAM	5.388	6.701	5.149	5.360	5.245	7.393	5.162	5.248	5.879	5.194	5.255	4.466	5.144
	Q4	0.685	0.813	0.810	0.752	0.839	0.618	0.884	0.885	0.857	0.853	0.876	0.888	0.919
WV2	ERGAS	7.866	5.535	5.263	6.210	4.867	7.175	4.966	4.736	5.513	5.197	4.734	4.618	4.743
	RMSE	62.49	42.70	40.03	48.45	38.37	57.11	37.70	37.65	43.44	40.01	36.35	36.13	37.77
	SAM	3.816	4.141	3.503	3.737	3.212	4.685	4.337	3.218	4.005	3.499	3.399	2.917	3.184
	Q8	0.545	0.761	0.784	0.739	0.795	0.667	0.788	0.813	0.814	0.687	0.703	0.827	0.847

TABLE V

THE MEAN VALUES IN TERM OF INDEXES D_λ, D_s, AND QNR, OBTAINED BY OUR PROPOSED DINE AND DINE+, AND BY OTHER STATE-OF-THE-ART METHODS SUCH AS GSA [16], PRACS [18], NLIHS [22], MF [26], GFPCA [68], SR-D [39], APNN [48], SFIM [28], GLP-CBD [29], ATWT [76] AND THE EXP IMAGES EVALUATED ON THREE DATA SETS AT THE FULL SCALE EXPERIMENTS.

Dataset	Indexes	EXP	GSA [16]	PRACS [18]	NLIHS [22]	MF [26]	GFPCA [68]	SR-D [39]	APNN [48]	SFIM [28]	GLP-CBD [29]	ATWT [76]	DINE	DINE+
QB	D_λ	0.000	0.142	0.056	0.086	0.155	0.072	0.068	n/a	0.127	0.143	0.107	0.010	0.017
	D_s	0.088	0.206	0.128	0.042	0.179	0.075	0.067	n/a	0.132	0.163	0.138	0.064	0.020
	QNR	0.912	0.696	0.829	0.877	0.705	0.859	0.870	n/a	0.771	0.728	0.778	0.927	0.964
GE1	D_λ	0.000	0.094	0.022	0.049	0.131	0.075	0.041	0.014	0.096	0.117	0.109	0.012	0.003
	D_s	0.065	0.156	0.093	0.035	0.138	0.107	0.044	0.042	0.114	0.131	0.131	0.028	0.039
	QNR	0.935	0.765	0.888	0.918	0.749	0.826	0.916	0.945	0.801	0.768	0.774	0.960	0.958
WV2	D_λ	0.000	0.120	0.044	0.012	0.153	0.129	0.073	0.023	0.136	0.140	0.132	0.012	0.022
	D_s	0.052	0.220	0.149	0.031	0.219	0.044	0.065	0.026	0.202	0.205	0.208	0.025	0.022
	QNR	0.948	0.695	0.816	0.957	0.669	0.831	0.867	0.952	0.699	0.693	0.697	0.964	0.956

Fig. 4. Fusion results on a cropped region from the GE1 data set at the reduced scale experiments. From left to right and top to bottom are the observed MS image, the fusion results of MTF-GLP-CBD [29], GSA [16], PRACS [18], NLIHS [22], MF [26], GFPCA [68], SR-D [39], ATWT [76], APNN [48], DINE and DINE+. Best viewed in color.

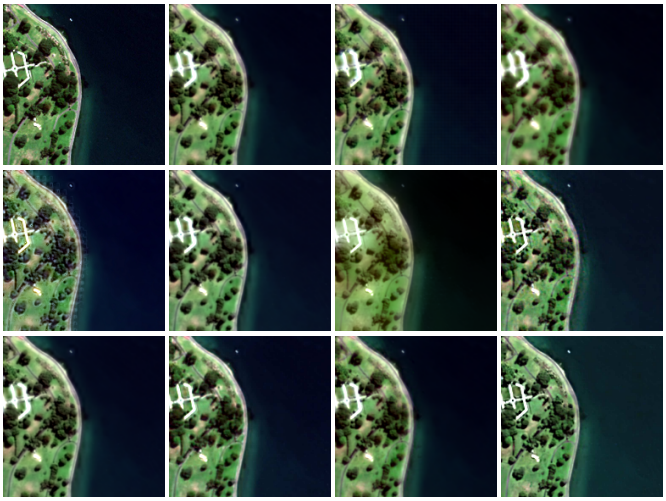

Fig. 5. Fusion results on a cropped region from the WV2 data set at the reduced scale experiments. From left to right and top to bottom are the observed MS image, the fusion results of MTF-GLP-CBD [29], GSA [16], PRACS [18], NLIHS [22], MF [26], GFPCA [68], SR-D [39], ATWT [76], APNN [48], DINE and DINE+. Best viewed in color.

available data sets (including both the original MS and PAN images) are first degraded to a reduced scale by using the MTF-matched filters [29], and then the original MS image are used as a reference for comparing it with the pansharpened result. All 48 pairs of MS and PAN images, as shown in Fig. 2, are used to carry out the experiments.

The mean values of ERGAS, RMSE, SAM, Q4/Q8 indexes of the baselines and the proposed methods for all the tested images are shown in Table IV, where the best results are highlighted in red, and the second best are indicated in blue. As we can see from the table that the proposed DINE method has the best values in terms of the RMSE, ERGAS, and SAM for most of the image pairs, while the proposed DINE+ performs the second best in most cases, and the best values in terms of Q4/Q8 for GE1 and WV2 data sets. The APNN and ATWT achieve a relatively higher scores among all the compared methods.

Moreover, three typical test images selected from QB, GE1, and WV2 data sets, are shown in Figs. 3-5. Due to limited spaces, we show only the cropped regions for GE1 and WV2 image scene. The fused images yielded by our proposed DINE and DINE+ look much more similar to the references, without significant spectral distortions or noticeable artifacts. Significant spectral distortions of the NLIHS for QB can be found in the forest areas, while the APNN, PRACS, MF, GFPCA, and GSA produce less color distortions. When comparing our DINE and DINE+ to SR-D and GFPCA, the visual improvements are evident, and our methods are able to recover finer details, whereas the fusion results of SR-D and GFPCA are insufficient in extracting high-frequency contents from the PAN image. We can find also that although our DINE+ has sharp results due to the edge-preserving filter and is capable of recovering structured spatial details which were missing in the expanded MS by cubic interpolation, it seems to have some spectral distortions compared to the DINE, GLP-CBD and APNN especially on the GE1 and WV2 data sets. This may be due to simulation procedure following the Wald's protocol [70], which results in an inconsistency between the spatial details of PAN and the reference MS, and thus extra spatial details will be extracted from the PAN by the DINE+. This conclusion is also in favor of the aforementioned observations in Table IV, where the DINE+ has better spatial quantitative index Q4/Q8 but relatively worse spectral quantitative indexes on GE1 and WV2 data sets.

Full Scale Experiments. In this case, assessment procedure is carried out by merging the PAN and MS images at the full scale (*i.e.* the original resolution), thus avoiding the simulation procedures required by the reduced scale experiments.

Table V reports the mean values of the D_s, D_λ and QNR indexes yielded by the baselines and the proposed DINE and DINE+ methods calculated on the 48 pairs of MS and PAN images. Typically, our proposed DINE and DINE+, and the state-of-the-art methods such as NLIHS and APNN methods achieve better scores than the other methods, while GSA, MF and SFIM methods perform a little poorer than the NLIHS, PRACS, SR-D and APNN methods. As it can also be seen that the spectral quality of our DINE+ are slightly worse than that of the DINE method for GE1 and WV2 data sets in terms of D_λ index, possibly because of that the DINE+ may introduce some extra spatial details. As is well-known that it is very challenging to find an appropriate trade-off between the spectral preservation and the spatial detail injections. On the other hand, because of the unavailable ground truth, the evaluation metrics (*i.e.*, D_s, D_λ and QNR) may not fully reflect the performance of the proposed methods.

Since quantitative indexes may not correlate well with visual perception, we present some typical scenes, as shown in Figs. 6-8, for the readers to examine the visual quality of the fusion results for better evaluation. From these figures, we can see that all methods can generate a plausible and visually pleasing fused MS images from the LSR MS images

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3067877, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

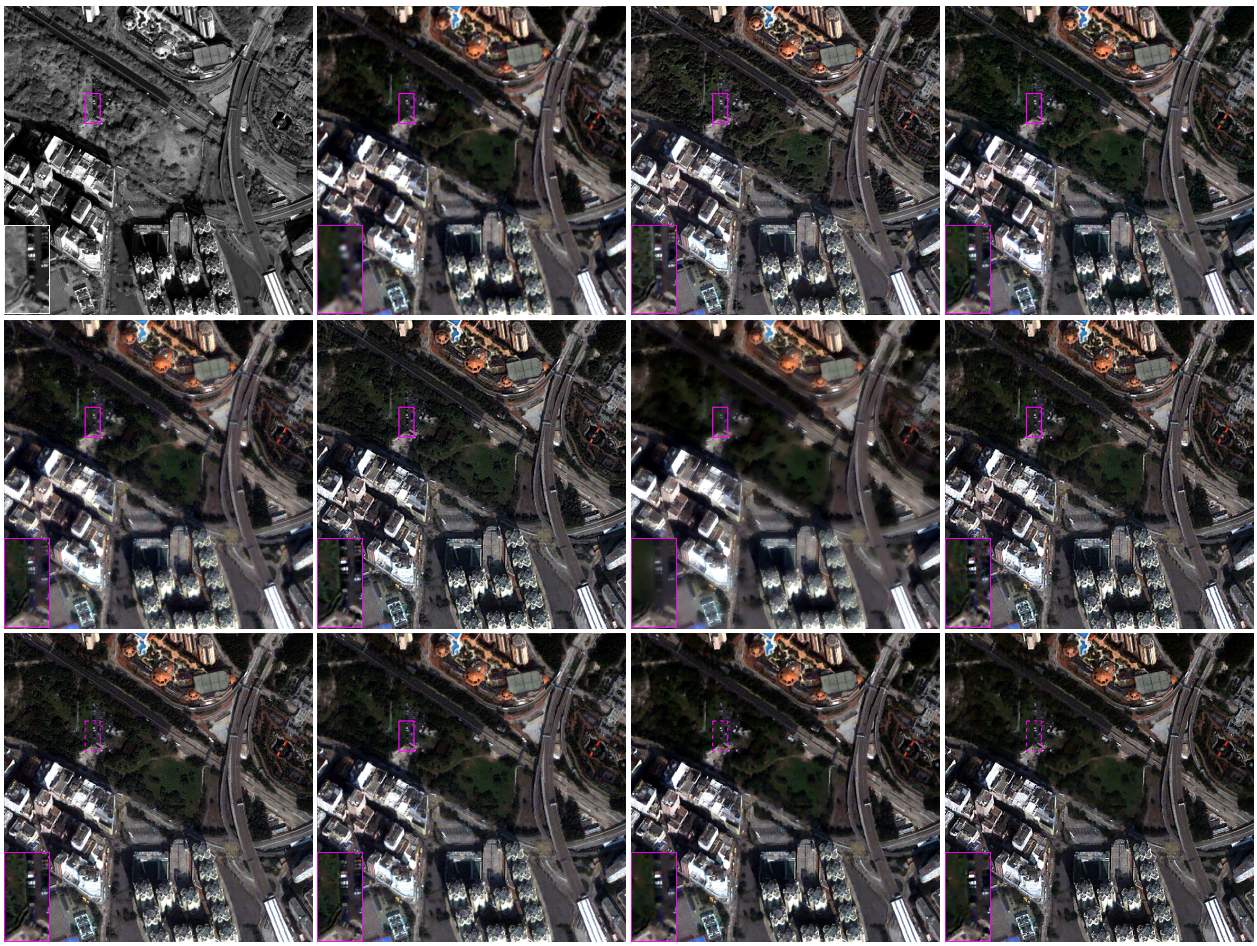

Fig. 6. Fusion results on one typical scene from the QuickBird data set at the full scale experiments. From left to right and top to bottom are the PAN, the EXP, the fusion results of GSA [16], PRACS [18], NLIHS [22], MF [26], GFPCA [68], SR-D [39], ATWT [76], SFIM [28], DINE and DINE+. Best viewed in color.

TABLE VI
THE AVERAGE RUNNING TIME (IN SECONDS) FOR OUR PROPOSED DINE, DINE+, AND OTHER STATE-OF-THE-ART METHODS SUCH AS GSA [16], PRACS [18], NLIHS [22], MF [26], GFPCA [68], SR-D [39], APNN [48], SFIM [28], GLP-CBD [29], AND ATWT [76] ON THREE DATA SETS AT THE *reduced scale* (RS) AND *full scale* (FS) EXPERIMENTS.

	Dataset	GSA [16]	PRACS [18]	NLIHS [22]	MF [26]	GFPCA [68]	SR-D [39]	APNN [48]	SFIM [28]	DINE	DINE+	ATWT [76]	GLP-CBD [29]
RS	QB	0.0269	0.0551	0.0272	0.0354	0.0234	1.1995	n/a	0.0116	0.3966	0.4047	0.0466	0.0362
	GE1	0.0269	0.0537	0.0266	0.0351	0.0226	1.1816	1.1172	0.0117	0.3928	0.3996	0.0459	0.0337
	WV2	0.0530	0.1426	0.0334	0.0712	0.0541	2.4668	1.7360	0.0224	0.8058	0.8306	0.0915	0.0688
FS	QB	0.4805	1.0544	0.3509	0.4862	0.8379	114.49	n/a	0.2000	107.74	107.79	1.4064	0.4460
	GE1	0.4491	1.0396	0.3482	0.4819	0.8299	111.93	14.481	0.1820	106.15	106.19	1.3963	0.4379
	WV2	0.8901	2.9484	0.4264	0.9978	0.9490	240.84	19.210	0.3673	219.12	219.21	2.8341	0.8983

and can improve the spatial details from the PAN image to some extent. Among all pansharpening methods, our proposed DINE and DINE+ methods, and the recently developed SR-based and DL-based methods such as SR-D and APNN have been generally acknowledged as the promising methods by producing clearer and visually pleasing results.

However, it is difficult to judge which method is the best by visually comparison from the whole image. Thus, for each scene, a small region (*i.e.* the purple box) in each image is zoomed in to inspect and compare the spatial details. By considering the zoomed region for each image, it can be found that other approaches often fail to reconstruct some structured details, our proposed DINE+ method can recover the missing high-frequency contents by properly exploiting the edge preserving filter. We can also find that the fusion result by our DINE+ have much sharper edges to a certain extent. The objects, for example, the cars as shown in Fig. 6, the buildings in Fig. 7, and the coastline in Fig. 8, in the fused image of NLIHS and GFPCA are not such obvious, which means that the spatial information in the PAN images is not well transferred into the fused MS images. The SFIM, GSA and MF contain slightly sharper edges and fewer artifacts such as ringing, while the APNN, SR-D and GLP-CBD have better spectral fidelity but with a more serious spatial distortion in all the cases. The fusion results obtained by our proposed DINE and DINE+ are comparable to those obtained by APNN, SR-D and PRACS, and are without noticeable artifacts and often visually more pleasing than others. The DINE method

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3067877, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Fig. 7. Results on one scene from the GE1 data set at the full scale experiments. From left to right and top to bottom are the PAN, the EXP, the fusion results of GSA [16], PRACS [18], NLIHS [22], MF [26], GFPCA [68], SR-D [39], APNN [48], MTF-GLP-CBD [29], DINE and DINE+.

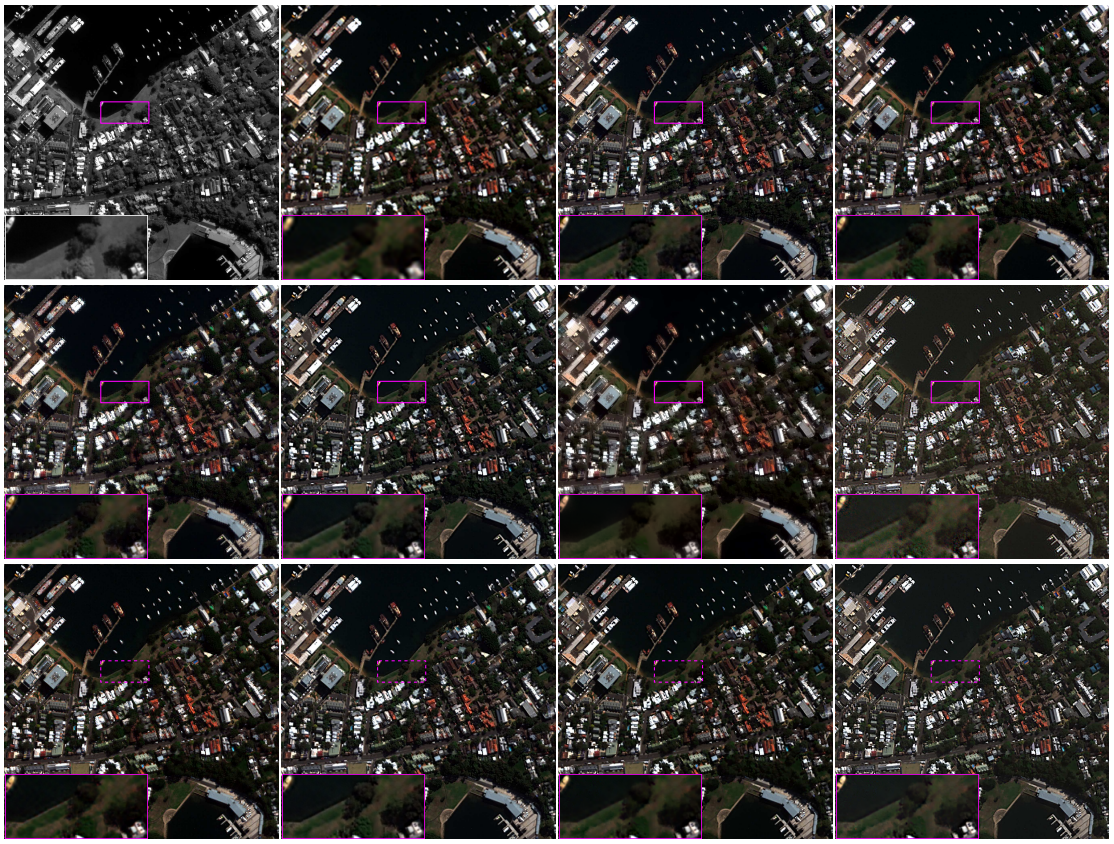

Fig. 8. Results on one scene from the WV2 data set at the full scale experiments. From left to right and top to bottom are the PAN, the EXP, the fusion results of GSA [16], PRACS [18], NLIHS [22], MF [26], GFPCA [68], SR-D [39], APNN [48], MTF-GLP-CBD [29], DINE and DINE+.

alleviates the artifacts yet the overall result is still blurry, as shown in the result of QB and GE1 data set. In this case, the improvement of the DINE+ is clear and on obvious spatial and spectral distortions are evident. The DINE+ generates sharp edges with rare artifacts in terms of both subjective visual evaluation and the objective quantitative evaluation, but at a cost of slight spectral distortion. This conclusion is consistent with the slightly worse spectral quantitative index D_λ as shown in Table V. In a word, the proposed methods, especially the DINE+, successfully add the detailed textures and edges extracted from the PAN images into the fused MS images and exhibit better-looking outputs compared with the-state-of-the-art works.

Time Complexity Analysis. Table VI reports the average running times of our proposed DINE, DINE+, and other state-of-the-art methods on three data sets (including 48 image pairs) at the reduced and full scale experiments, respectively. Note that the average running times of DINE and DINE+ reported in this table include the computational time for selecting the K-NN neighbors of each patch $\mathbf{q}$. From this table, We can observe that the proposed DINE+ and DINE methods clearly requires less time than the SR-D method. This maybe thank to the low complexity choice of the optimal weights. In addition, we can also find that the time cost of the proposed methods is not significantly improved at full scale experiments. This is due to that the computation cost of our methods mainly relies on the procedure of selecting K-NN set for each patch, it will be high as the number of divided patches in an image grows. At the full scale experiments, we set the size of patches as the same as that of the reduced scale experiments, this will lead to a large number of patches due to the large size of images at the full scale experiments. To reduce the running time, we can set a relative large patch size with a little sacrifice of the quality of the fusion results. Additionally, we can also resort to some more efficient nearest neighbor searching strategies such as *kd-trees*, *box-decomposition trees*, or *k-d generalized randomized forests* [80] to reduce the computational complexity of DINE and DINE+.

Above all, the DINE method can well keep the spectral information in the MS images, while the DINE+ can highlight the spatial details in the PAN image with a little sacrifices of spectral quality. This is concluded from the quantitative comparisons in tables IV and V at both reduced and full scale experiments.

V. Conclusion and Future Works

In this paper, we proposed a simple but effective *detail injection via neighbor embedding* (DINE) method for pansharpening, aiming at enhancing the performances of sharpened images and reducing the time-complexity resulting by the SR-based and DL-based methods. The DINE was unified with the spatial details injection framework, and was deployed as a locally linear neighbor embedding of overlapped image patches. Additionally, to further improve the spatial details of the fused results, an improved version of DINE, called DINE+, was developed by exploiting the edge-preserving model. Extensive experimental results on three kinds of satellites data sets and on comparing with the other state-of-the-art methods have demonstrated that our proposed DINE and DINE+ methods are successful in well keeping the spectral content of the original MS images and improving the structured spatial details. However, all of the conclusions with our proposed methods are obtained on the images with high qualities, we would like to validate their effectiveness when some outliers or noises (e.g., shadow or cloud appeared in the scene [43] [79]) as our future work. There is still a room for the improvement of the proposed DINE and DINE+ methods, for example, how to efficiently find K-nearest neighbors if the data is large and in high dimensional. Approximate nearest neighbor searching in high dimensions [80] may be a good approach, which offers a fair trade-off between accuracy and speed.

Acknowledgment

The authors would like to thank the authors of [11], [22], [26], [68], [39], [45], [48] for kindly sharing their source codes, and thank the Editors and three reviewers for their valuable comments and suggestions, which lead to a substantial improvement of this paper.

References

[1] G. Vivone, P. Addesso, R. Restaino, M. D. Mura, and J. Chanussot, "Pansharpening based on deconvolution for multiband filter estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 540–553, Jan. 2019.

[2] Y. Zhang, "Understanding image fusion," *Photogramm. Eng. & Remote Sens.*, vol. 70, no. 6, pp. 657–661, 2004.

[3] G. A. Shaw and H. K. Burke, "Spectral imaging for remote sensing," *Lincoln Laboratory Journal*, vol. 14, no. 1, pp. 3–28, 2003.

[4] L. Alparone, *et al.*, "Comparison of Pansharpening Algorithms: Outcome of the 2006 GRS-S Data-Fusion Contest," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3012–3021, Oct. 2007.

[5] G. Vivone, M. D. Mura, A. Garzelli, R. Restaino, G. Scarpa, M. O. Ulfarsson, L. Alparone, J. Chanussot, "A new benchmark based on recent advances in multispectral pansharpening," *IEEE Geoscience and Remote Sensing Magazine*, DOI: 10.1109/MGRS.2020.3019315, 2020.

[6] L. Loncan, *et al.*, "Hyperspectral pansharpening: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, no. 3, pp. 27–46, Sept. 2015.

[7] A. Garzelli, "A review of image fusion algorithms based on the super-resolution paradigm," *Remote Sens.*, vol. 8, pp. 797.1-797.20, 2016.

[8] P. Sirguey, R. Mathieu, Y. Arnaud, M. Khan, and J. Chanussot, "Improving MODIS spatial resolution for snow mapping using wavelet fusion and ARSIS concept," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 1, pp. 78–82, Jan. 2008.

[9] C. Souza, J. Firestone, L. Silva, and D. Roberts, "Mapping forest degradation in the Eastern Amazon from SPOT 4 through spectral mixture models," *Remote Sens. Environ.*, vol. 87, no. 4, pp. 494–506, Nov. 2003.

[10] A. Mohammadzadeh, A. Tavakoli, and M. Zoej, "Road extraction based on fuzzy logic and mathematical morphology from pan-sharpened IKONOS images," *Photogramm. Rec.*, vol. 21, no. 113, pp. 44–60, Mar. 2006.

[11] G. Vivone, *et al.*, "A Critical Comparison Among Pansharpening Algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

[12] X. Meng, H. Shen, H. Li, L. Zhang, and R. Fu, "Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and chanllenges," *Information Fusion*, vol. 46, pp. 102–113, 2019.

[13] P. Chavez and A. Kwateng, "Extracting Spectral Contrast in Landsat Thematic Mapper Image Data Using Selective Principal Component Analysis," *Photogramm. Eng. Remote Sens.*, vol. 55, no. 3, pp. 339–348, 1989.

[14] W. J. Carper, T. M. Lillesand, and R. W. Kiefer, "The Use of Intensity-Hue-Saturation Transformations for Merging SPOT Panchromatic and Multispectral Image Data," *Photogramm. Eng. Remote Sens.*, vol. 56, no. 4, pp. 459–467, Apr. 1990.

[15] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using Pan-sharpening," U.S. Patent 6 011 875, Jan. 4, 2000. Tech. Rep., Eastman Kodak Company.

[16] B. Aiazzi, S. Baronti, and M. Selva, "Improving Component Substitution Pansharpening Through Multivariate Regression of MS+Pan Data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Oct. 2007.

[17] 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.

[18] 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.

[19] C. Thomas, T. Ranchin, L. Wald, and J. Chanussot, "Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1301–1312, May 2008.

[20] Q. Xu, Y. Zhang, B. Li, and L. Ding, "Pansharpening Using Regression of Classified MS and Pan Images to Reduce Color Distortion," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 28–32, Jan. 2015.

[21] A. Garzelli, "Pansharpening of multispectral images based on nonlocal parameter optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2096–2107, Apr. 2015.

[22] M. Ghahremani and H. Ghassemian, "Nonlinear IHS: A Promising Method for Pan-Sharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 11, pp. 1606–1610, Nov. 2016.

[23] R. Restaino, M. D. Mura, G. Vivone, and J. Chanussot, "Context-Adaptive Pansharpening Based on Image Segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 753–766, Feb. 2017.

[24] L. Alparone, S. Baronti, B. Aiazzi, and A. Garzelli, "Spatial methods for multispectral pansharpening: multiresolution analysis demystified," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2563–2576, May 2016.

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

[26] R. Restaino, G. Vivone, M. Dalla Mura, and J. Chanussot, "Fusion of multispectral and panchromatic images based on morphological operators," *IEEE Trans. Image Process.*, vol. 25, no. , pp. 2882–2895, 2016.

[27] J. Liu and S. Liang, "Pan-sharpening using a guided filter," *International Journal of Remote Sensing*, vol. 37, no. 8, pp. 1777-1800, 2016.

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

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

[30] C. Ballester, V. Caselles, L. Igual, J. Verdera, and B. Rougé, "A variational model for P+XS image fusion," *Int. J. Comput. Vis.*, vol. 69, no. 1, pp. 43–58, 2006.

[31] D. Fasbender, J. Radoux, and P. Bogaert, "Bayesian Data Fusion for Adaptable Image Pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1847–1857, Jun. 2008.

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

[33] L. 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. Geosci. Remote Sens.*, vol. 27, no. 9, pp. 4330–4344, Sep. 2018.

[34] J. Duran, A. Buades, B. Coll, and C. Sbert, "A Nonlocal Variational Model for Pansharpening Image Fusion," *SIAM J. Imaging Sciences*, vol. 7, no. 2, pp. 761–796, 2014.

[35] S. Yang, K. Zhang, and M. Wang, "Learning Low-Rank Decompostion for Pan-sharpening With Spatial-Spectral Offsets," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3647–3657, Aug. 2018.

[36] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "A New Pansharpening Algorithm Based on Total Variation," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 318–322, Jan. 2014.

[37] S. Li, and B. Yang, "A new pan-sharpening method usinga compressed sensing technique," *IEEE Trans. Geosci. Remot Sens.*, vol. 49, no. pp. 738–746, 2011.

[38] 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.

[39] M. R. Vicinanza, R. Restaino, G. Vivone, M. D. Mura, J. Chanussot, "A Pansharpeing Method Based on the Sparse Representation of Injected Details," *IEEE Geosci. Remote Sens. Lette.* , vol. 12, no. 1, pp. 180–184, Jan. 2015.

[40] H. Yin, "PAN-Guided Cross-resolution Projection for Local Adaptive Sparse Representation-Based Pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4938–4950, July 2019.

[41] Y. LeCun, Y. Bengio, G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436–444, 2015.

[42] Z. Shao, J. Cai, P. Fu, L. Hu, "Deep learning-based fusion of Landsat-8 and Sentinel-2 images for a harmonized surface reflectance product," *Remote Sensing of Environment*, vol. 235, pp. 111425, Dec. 2019.

[43] H. Luo, L. Wang, Z. Shao, and D. Li, "Development of a multi-scale object-based shadow detection method for high spatial resolution image ," *Remote Sensing Letters*, vol. 6, no. 1, pp. 59–68, 2015.

[44] Z. Shao, Y. Pan, C. Diao, and J. Cai, "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 4062–4076, June 2019.

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

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

[47] J. Yang, C. 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.*, Oct. 2017, pp. 1753–1761.

[48] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-Adaptive CNN-Based Pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vo. 56, no. 9, pp. 5443–5457, Sept. 2018.

[49] S. Vitale, G. Scarpa, "A detail-preserving cross-scale learning strategy for CNN-based pansharpening," *Remote Sensing*, vol. 12, no. 3, pp. 348, 2020.

[50] Z. Shao and J. Cai, "Remote sensing image fusion with deep convolutional neural network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 5, pp. 1656–1669, May 2018.

[51] D. Wang, Y. Li, L. Ma, Z. Bai, and J. Chan. "Going deeper with densely connected convolutional neural networks for multispectral pansharpening," *Remote. Sens.*, vol. 11, no. 22, pp.–2608, 2019.

[52] L. Liu, J. Wang, E. Zhang, B. Li, X. Zhu, Y. Zhang, and J. Peng, "Shallow-deep convolutional network and spectral-discrimination-based detail injection for multispectral imagery pan-sharpening," *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, vol. 13, pp. 1772–1783, 2020.

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

[54] H. Chang, D. Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. CVPR*, vol. 1, Washington, DC, USA, 2004, pp. 275–282.

[55] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[56] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197-2202, 2003.

[57] M. Bevilacqua, A. Roumy, C. Guillemot, and M. Alberi Morel, "Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding", in *Proc. of the British Machine Vision Conference*, Guildford, Britain, Sep. 3–7, 2012, pp. 135.1–135.10.

[58] Q. Liu, Y. Wang, Z. Zhang, "Pan-sharpening based on geometric clustered neighbor embedding," *Optical Engineering*, vol. 53, no. 9, pp. 093109.1–093109.16, 2014.

[59] M. Wang, K. Zhang, X. Pan, S. Yang, "Sparse tensor neighbor embedding based pan-sharpening via N-way block pursuit," *Knowle dge-Base d Systems*, vol. 149, pp. 18–33, 2018.

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

[61] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *International Conference on Computer Vision (ICCV)*, Sept. 2009, pp. 349–356.
[62] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit, " *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655-4666, Dec. 2007.
[63] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic Decompostion by Basisi Pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.
[64] L. Wald, T. Ranchin, M. Mangolini, "Fusion of satellite images of different spatail resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, pp. 691–699, 1997.
[65] J. Liu, J. Ma, R. Fei, H. Li, J. Zhang, "Enhanced back-projection as postprocessing for pansharpening," *Remote Sensing*, vol. 11, no. 6, pp. 712.
[66] J. Liu, Y. Hui, P. Zan, "Locally Linear Detail Injection for Pansharpening," *IEEE Access*, vol. 5, pp. 9728–9738, 2017.
[67] K. He, J. Sun, X. Tang, "Guided Image Filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.
[68] W. Liao, *et al.*, "Processing of multiresolution thermal hyperspectral and digital color data: Outcome of the 2014 IEEE GRSS Data Fusion Contest," *IEEE Journal of Seleted Topics in Applied Earth Observation and Remote Sensing*, vol. 8, no. 6, pp. 2984–2996, Jun. 2015.
[69] D. Yocky, "Multiresolution wavelet decomposition image merger of Landsat Thematic Mapper and SPOT panchromatic data," *Photogrammetric Engineering and Remote Sensing*, vol. 62, pp. 1067-1074, 1996.
[70] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?." In *Proceedings of the third conferences on Fusion of Earth data: merging point measurements, raster maps and remotely sensed images*, Sophia Antipolis, France, pp. 99–103, 2000.
[71] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini, "A global quality measurement of pan-sharpened multispectral imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 1, no. 4, pp. 313–317, Oct. 2004.
[72] A. Garzelli and F. Nencini, "Hypercomplex quality assessment of multi/hyperspectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 4, pp. 662-665, Oct. 2009.
[73] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81-84, Mar. 2002.
[74] R. Yuhas and J. Boardman, "Discrimination among semi-arid landscape endmembers using the Spectral Angle Mapper (sam) algorithm," In *Proceedings of Summar. 3rd Annual JPL Airborne Geoscience Workshop*, JPL Publication, Pasadena, CA, pp. 147–149, 1992.
[75] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogrammetric Engineering and Remote Sensing*, vol. 74, no. 2, pp. 193-200, 2008.
[76] G. Vivone, R. Restaino, M. D. Mura, G. Licciardi, and J. Chanussot, "Contrast and error-based fusion schemes for multispectral image pansharpening," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 5, pp. 930–934, May 2014.
[77] Z. Wang, D. Ziou, C. Armenakis, D. Li, and Q. Li, "A comparative analysis of image fusion methods," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1391–1402, Jun. 2005.
[78] B. Aiazzi, S. Baronti, M. Selva, and L. Alparone, "Bi-cubic interpolation for shift-free pan-sharpening," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 86, no. 6, pp. 65–76, Dec. 2013.
[79] Z. Shao, J. Deng, L. Wang, Y. Fan, N. S. Sumari, and Q. Cheng, "Fuzzy autoEncode based cloud detection for remote sensing imagery," *Remote Sensing*, vol. 9, no. 4, pp. 311, 2017.
[80] Y. Avrithis, I.Z. Emiris, I.Z. Samaras, "High-dimensional approximate nearest neighbor: k-d generalized randomized forests," *arXiv:1603.09596*, 2016.

Junmin Liu received the Ph.D. in applied mathematics from Xi'an Jiaotong University, Xi'an, China in 2013. He is currently an Associate Professor in the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China. From 2011 to 2012, he was a Research Assistant in the Department of Geography and Resource Management, The Chinese University of Hong Kong, and he was a visiting scholar in the Department of Geographical Sciences, University of Maryland, College Park, USA, from 2014 to 2015. His research interests are focused on remotely sensed image fusion, hyperspectral unmixing, object detection, deep learning, and so on.

Changsheng Zhou received the B.S. degree in applied mathematics, and Ph.D. degree in statistics from Xian Jiaotong University, Xian, China, in 2014 and 2020, respectively. He is currently a postdoc fellow in the school of mathematics and information science from Guangzhou University, Guangzhou, China. His research interests include generative model, deep learning, and Bayes modelling and computer vision.

Rongrong Fei was born in Shaanxi, China, in 1991. She received the Ph.D. degree in School of Mathematics and Statistics from Xian Jiaotong University, Xian, Shaanxi, China, in 2020. She is currently a Lecturer in School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xian, Shaanxi, China. Her research interests include sparse machine learning method, image processing and remote sensing.

Chunxia Zhang is currently a Professor in the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China. Her research interests are focused on high-dimensional data analysis and statistical learning.

Jiangshe Zhang received the B.S., M.S., and Ph.D. degrees in computational mathematics from Xian Jiaotong University, Xian, China, in 1984, 1987, and 1993, respectively. He is currently the Director of the Institute of Machine Learning and Statistical Decision Making, Xian Jiaotong University, where he is a Professor with the Department of Statistics. He is also the Vice-President of the Xian International Academy for Mathematics and Mathematical Technology, Xian, China. He has authored and co-authored one monograph and over 100 journal papers. His research interests include statistical computing, deep learning, cognitive representation, and statistical decision making.

Prof. Zhang received the National Natural Science Award of China (Third Place) and the First Prize in Natural Science from the Ministry of Education of China in 2007. He served as the President of the Shaanxi Mathematical Society and the Executive Director of the China Mathematical Society.