SSconv: Explicit Spectral-to-Spatial Convolution for Pansharpening

Yudong Wang
yudongwang1121@126.com
Yingcai Honors college, University of Electronic Science and Technology of China

Liang-Jian Deng*
liangjian.deng@uestc.edu.cn
School of Mathematical Sciences, University of Electronic Science and Technology of China

Tian-Jing Zhang
zhangtianjinguestc@163.com
Yingcai Honors College, University of Electronic Science and Technology of China

Xiao Wu
wxwsx1997@gmail.com
School of Mathematical Sciences, University of Electronic Science and Technology of China

ABSTRACT

Pansharpening aims to fuse a high spatial resolution panchromatic (PAN) image and a low resolution multispectral (LR-MS) image to obtain a multispectral image with the same spatial resolution as the PAN image. Thanks to the flexible structure of convolution neural networks (CNNs), they have been successfully applied to the problem of pansharpening. However, most of the existing methods only simply feed the up-sampled LR-MS into the CNNs and ignore the spatial distortion caused by direct up-sampling. In this paper, we propose an explicit spectral-to-spatial convolution (SSconv) that aggregates spectral features into the spatial domain to perform the up-sampling operation, which can get better performance than the direct up-sampling. Furthermore, SSconv is embedded into a multiscale U-shaped convolution neural network (MUCNN) for fully utilizing the multispectral information of involved images. In particular, multiscale injection branch and mixed loss on cross-scale levels are employed to fuse pixel-wise image information. Benefiting from the distortion-free property of SSconv, the proposed MUCNN can generate state-of-the-art performance with a simple structure, both on reduced-resolution and full-resolution datasets acquired from WorldView-3 and GaoFen-2.

CCS CONCEPTS

• **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Spectral-to-Spatial, multiscale, convolution neural networks, pansharpening

ACM Reference Format:
Yudong Wang, Liang-Jian Deng, Tian-Jing Zhang, and Xiao Wu. 2021. SSconv: Explicit Spectral-to-Spatial Convolution for Pansharpening. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475600

1 INTRODUCTION

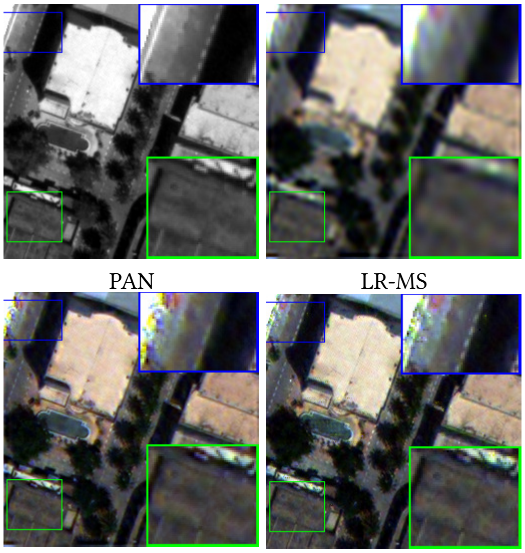
With the development of spectral imaging technology, the application of multispectral (MS) images in medicine, geology, agriculture, and other fields have become more and more important. MS images are usually acquired by sensors that are deployed on satellites. However, due to hardware limitations, the sensor cannot guarantee the spectrum and spatial resolution of the captured image at the same time [31]. Sensors usually acquire either a high-resolution (HR) PAN image or a low-resolution (LR) MS image. The popularity of pansharpening is proved by the contest in 2006 [3, 7] and many recently review papers [27, 29]. In order to make full use of the rich spectral information in the LR-MS image and the spatial information in the HR-PAN image, researchers come up with the idea of pansharpening, which attempts to fuse an HR-PAN image and an LR-MS image to obtain an HR-MS image. The main challenge of pansharpening is to achieve a balance between spectral and spatial information on the basis of avoiding distortion. Therefore, it is necessary to fully master the feature of the HR-PAN image and LR-MS image, and their potential relationship, especially the margin between their spectral and spatial resolution.

The up to date strategies of pansharpening can be divided into four categories [31]: (1) component substitution (CS)-based methods; (2) multi-resolution analysis (MRA)-based methods; (3) variational model-based methods; (4) deep learning (DL)-based methods. The first three categories can be classified as traditional methods, while the deep learning that based on convolutional neural networks (CNNs) recently has achieved great success in a wide range of vision tasks, such as image recognition, target detection [26], and single image super-resolution [14, 39, 41]. Driven by the mapping requirement of relationship among LR-MS image, HR-PAN image, and the desired HR-MS image, various DL-based methods have been proposed to improve the fusion results of pansharpening since they can generate more details after training on a large number of existing datasets. The reason why the DL-based methods can

[*] Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10...$15.00
https://doi.org/10.1145/3474085.3475600



PAN LR-MS

DMDNet [9] MUCNN

Figure 1: The PAN image and LR-MS image and the pansharpened images by DMDNet [9], and our model MUCNN.

achieve advanced results lies in the powerful non-linear fitting and feature extraction capabilities of CNNs.

An inevitable problem for pansharpening is to mitigate the gap between the spatial resolution of HR-PAN images and LR-MS images. Up-sampling, as an operation to expand spatial resolution, is important in the process of information fusion. Existing up-sampling methods include linear interpolation, deconvolution, and unpooling [38]. Direct linear interpolation up-sampling is the most common one, which roughly supplements the image based on the average value of adjacent pixels. Although it is simple and fast, its results often appear unexpectedly smooth. Different from the pre-defined interpolation method, deconvolution with parameters that can be learned, has been widely used in segmentation tasks [20] and has achieved good results. However, the feature maps need to be padded with zero before the convolution operation, thus a large amount of information is fairly useless, and its calculation process is computationally expensive. Another method, unpooling, which upgrades the resolution of the feature maps through direct zero paddings, thus fails to explore the underlying information between the pixel and its neighbors. It is worth mentioning that up-sampling is equally critical for the single image super-resolution task. In [25], an efficient and effective up-sampling method for a single-channel feature map is proposed. In their work, the original LR image is reconstructed into an HR image through convolutions and periodic shuffling. Inspired by this, we believe that through similar pixel rearrangement operations, a spectral to spatial feature mapping can be learned with the help of convolution, which is reasonably suitable for processing multispectral images like pansharpening.

In addition to the specific operation of the upsampling method, we also consider the ratio of upsampling to LR-MS. Most prior DL-based methods upsample the original LR-MS image directly to the same resolution as the HR-PAN image, which may lead to spectral distortion and loss of information. In this paper, we propose a new U-shaped network with a multiscale injection branch to fully explore and utilize the information provided by the original LR-MS image and HR-PAN image. Particularly, we design a Spectral-to-Spatial convolution (SSconv) for the up-sampling in pansharpening to avoid the distortion caused by the conventional up-sampling methods. Following the U-shaped network and multiscale injection, feature maps with different scales are produced in the process of our network. To supervise the intermediate products of the network learning process, a mixed loss strategy is proposed. Finally, the proposed approach is validated on several datasets acquired from two satellites, i.e., WorldView-3 and GaoFen-2. Through the experimental analysis performed on both reduced and full resolutions, the proposed multiscale U-shaped convolutional neural network (MUCNN) is confirmed to be able to outperform a wide range of competitive methods.

The main contributions of our work are as follows:

(1) We design a Spectral-to-Spacial convolution to aggregate the spectral feature to the spacial domain. In addition to increasing the spatial resolution of the feature maps by making full use of the spectral information, SSconv also helps with the construction of the feature maps in the MUCNN.
(2) We propose a U-shaped convolution neural network with a multiscale injection branch to fuse the information both in spatial and spectral domains.
(3) A mixed loss strategy is adopted to supervise the output MS images with three different scales and train via backpropagation, which could capitalize on rich feature hierarchies. In addition, our method significantly exceeds the existing state-of-the-art methods with a simple structure.

The remaining of this paper is organized as follows. The notations and related works are introduced in Section 2. The proposed network architectures will be detailed in Section 3. Section 4 is devoted to the description of the experimental results and the related discussions. Finally, conclusions are drawn in Section 5.

2 NOTATIONS AND RELATED WORKS

2.1 Notations

For clearness and convenience, it is necessary to introduce the notations used in this paper. $\mathbf{MS} \in \mathbb{R}^{w \times h \times b}$ denotes the observed LR-MS image, where w, h, and b represent the width, height, and spectral band of the image, respectively. $\mathbf{P} \in \mathbb{R}^{W \times H \times 1}$ denotes the observed PAN image, where $H = 4h$, $W = 4w$, and $\mathbf{GT} \in \mathbb{R}^{W \times H \times b}$ is the ground-truth image. The desired HR-MS image is defined as $\widehat{\mathbf{MS}}_{4\times} \in \mathbb{R}^{W \times H \times b}$. Apart from that, we upsample the $\mathbf{MS}$ through SSconv to obtain the $2\uparrow$ and $4\uparrow$ $\mathbf{MS}$ images, defined as $\mathbf{MS}_{2\uparrow} \in \mathbb{R}^{2w \times 2h \times b}$ and $\mathbf{MS}_{4\uparrow} \in \mathbb{R}^{W \times H \times b}$. And we use 2×2 convolution with the stride of 2 and 4×4 convolution with the stride of 4 to downsample the $\mathbf{P}$ thus obtain the $2\downarrow$ and $4\downarrow$ $\mathbf{P}$ images, defined as $\mathbf{P}_{2\downarrow} \in \mathbb{R}^{2w \times 2h \times 1}$ and $\mathbf{P}_{4\downarrow} \in \mathbb{R}^{w \times h \times 1}$.

2.2 CNNs for pansharpening

As mentioned in the introduction, most of the DL-based methods that have emerged in the field of pansharpening in recent years are based on CNNs. The first DL-based method for pansharpening is proposed by Masi *et. al* in [21] named as PNN, which just simply stacks three convolutional layers and achieves remarkable results. Since then, more and more DL-based approaches have been proposed. A noteworthy work called PanNet [36] proposes a simple structure with a certain degree of physical interpretability focusing on spectral and spatial preservation. Subsequent works, *e.g.*, DMDNet [9], and FusionNet [8] further explore the potential of CNNs and achieve the promising results. Overall, the main framework of the application of CNNs in pansharpening can be described as a non-linear mapping $f_{\Theta_{FS}}$, where Θ_{FS} denote the parameters of CNNs. And their loss function can be unified as follows:

$$Loss(\Theta_{FS}) = \|f_{\Theta_{FS}}(\mathbf{P}, \mathbf{MS}) - \mathbf{GT}\|$$

where $\|\cdot\|$ represents a kind of norm which can be seen as a measurement of vector or matrix, e.g., L1 Loss or L2 Loss.

However, existing methods may fail to capture complex features caused by variations of scales and resolution ratios. Most of their network structure extract and learn the features of upsampled LR-MS images of the same size as HR-PAN images. And only focus on the final output without considering the products of the intermediate convolutional layers.

2.3 U-Net

U-Net [24], a classic network architecture designed for pixel-wise segmentation, has been proven to perform promisingly [6, 23, 42]. In particular, it learns different levels of semantic features and reduces the size of the feature maps through several down-sampling steps. Then the size of the feature map is gradually restored through the up-sampling steps, and the extracted semantic features are successfully used to complete the final segmentation task.

Concurrent with our work, there are several reasons that motivated us to chose it as our backbone. First, pansharpening is also a pixel-wise task, which needs to be refined to the characteristics of each pixel and the relationship with its neighborhood. Therefore, we believe that the powerful targeting and depicting the ability of a U-shaped network can be applied to the pansharpening task. Second, pyramid features meet our expectations for overcoming the spatial resolution gap between LR-MS and HR-PAN images. A U-shaped network provides a possibility to fuse images across scales by stages. Third, in the structure of U-Net, the feature maps are propagated progressively, which is consistent with the aim of the pansharpening task, since more detailed information can be restored in the feature maps.

3 THE PROPOSED METHOD

Our proposed model adopts multiscale input and U-shaped CNN to explore the feature of spatial, spectral, and their relationship. The proposed MUCNN consists of four parts, which are: (1) SSconv for upsampling operation of multi-spectral images, (2) multiscale injection branch which feeds MS images and PAN images progressively to the network, (3) U-shaped overall network structure which performs excellently on the pixel-wise problem, (4) mixed multiscale loss, which plays a role in accelerating the backpropagation of network and examine the fusion results by stages.

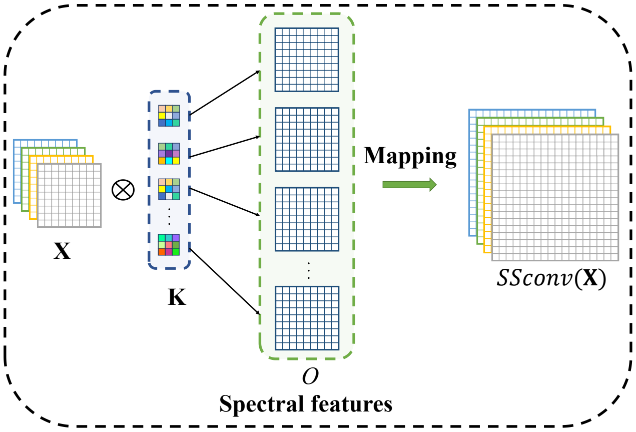


Figure 2: Schematic diagram of SSconv.

3.1 Spectral-to-Spatial Convolution

As the LR-MS image contains different spectral bands of the same scene, which can be treated as a multiple image super-resolution problem. We believe that information aggregation from different spectral bands will be helpful for the reconstruction of spatial information. Therefore, we propose a novel Spectral-to-Spatial convolution to perform upsampling operations. The operation process is shown in Figure 2, we use r^2b convolutions (3×3) to integrate spatial and spectral features, where r is the ratio of upsampling. Through the pixel-mapping for each r^2b band, the upsampled LR-MS image is generated. For example, the i-th feature map $O_i \in \mathbb{R}^{w\times h}$ is obtained from the following operation conducting on $X \in \mathbb{R}^{w\times h\times b}$:

$$O_i = \mathbf{X} \otimes \mathbf{K}_i \qquad (i = 0, 1, \ldots, r^2b - 1) \tag{1}$$

where the $\mathbf{K}_i \in \mathbb{R}^{1\times 3\times 3\times b}$ denotes i-th convolution kernel and $\otimes$ denotes the convolution operation in conventional CNNs. Then, we can obtain upsampled $\mathbf{X}$ through the mapping (mentioned in Figure 2):

$$\begin{aligned} &SSconv(\mathbf{X})_{ri+c_1, rj+c_2, k} = O_{i,j,kr^2+c_1r+c_2} \\ &(i = 0, 1, \ldots, w-1; j = 0, 1, \ldots, h-1; \\ &c_1 = 0, 1, \ldots, r-1; c_2 = 0, 1, \ldots, r-1; \\ &k = 0, 1, \ldots, b-1) \end{aligned} \tag{2}$$

where the $SSconv(\mathbf{X})_{ri+c_1,rj+c_2,k}$ denotes the pixel of upsampled image, $SSconv(\cdot)$ is the SSconv operation. When the number of spectral bands b equals to 1, the problem degenerates into the single-image super-resolution. Besides, the SSconv also degenerates into the pixel shuffle [25]. More details please refer to Figure 2.

3.2 Multiscale injection branch

The ratio of the spatial resolution between the **P** and the **MS** is four. In order to fully explore the potential information of the images and model the relationship among the **P**, **MS**, and $\widehat{\mathbf{MS}}_{4\times}$. We intend to take the known image, i.e., **P**, **MS** as input in multiscale pyramid form. As shown in Figure 3, the **MS** is upsampled by SSconv twice, obtain $\mathbf{MS}_{2\uparrow} \in R^{2w\times 2h\times b}$ and $\mathbf{MS}_{4\uparrow} \in R^{4w\times 4h\times b}$ as follows:

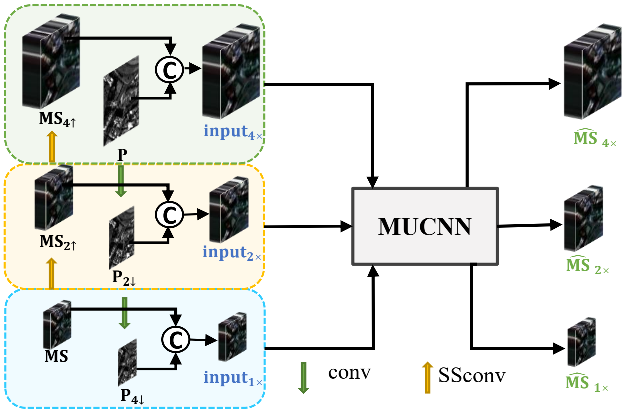

$\mathbf{MS}_{4\uparrow}$
P
$\mathbf{input}_{4\times}$
$\widehat{\mathbf{MS}}_{4\times}$
$\mathbf{MS}_{2\uparrow}$
$\mathbf{P}_{2\downarrow}$
$\mathbf{input}_{2\times}$
MUCNN
$\widehat{\mathbf{MS}}_{2\times}$
MS
$\mathbf{P}_{4\downarrow}$
$\mathbf{input}_{1\times}$
conv
SSconv
$\widehat{\mathbf{MS}}_{1\times}$

Figure 3: Schematic diagram of Multiscale injection branch. $\mathbf{MS}_{4\uparrow}$ is directly generated from the MS and the $\mathbf{P}_{4\downarrow}$ is directly generated from the P.

$$\begin{aligned} \mathbf{MS}_{2\uparrow} &= SSconv(\mathbf{MS}) \qquad (r = 2) \\ \mathbf{MS}_{4\uparrow} &= SSconv(\mathbf{MS}) \qquad (r = 4) \end{aligned} \tag{3}$$

Even-sized convolutions, i.e., 2×2 and 4×4, are used to downsample the images, whose effectiveness has be verified in [12, 22, 35]. And we obtain $\mathbf{P}_{2\downarrow} \in \mathbb{R}^{2w\times2h\times1}$ and $\mathbf{P}_{4\downarrow} \in \mathbb{R}^{w\times h\times1}$:

$$\begin{aligned} \mathbf{P}_{2\downarrow} &= \mathbf{K}_1 \otimes \mathbf{P} \\ \mathbf{P}_{4\downarrow} &= \mathbf{K}_2 \otimes \mathbf{P} \end{aligned} \tag{4}$$

where $\mathbf{K}_1 \in \mathbb{R}^{1\times2\times2\times1}$ and $\mathbf{K}_2 \in \mathbb{R}^{1\times4\times4\times1}$ represent the convolution kernels.

Finally, we concatenate the images in the same resolution to get the following three inputs:

$$\begin{aligned} \mathbf{input}_{1\times} &= C[\mathbf{MS}; \mathbf{P}_{4\downarrow}] \\ \mathbf{input}_{2\times} &= C[\mathbf{MS}_{2\uparrow}; \mathbf{P}_{2\downarrow}] \\ \mathbf{input}_{4\times} &= C[\mathbf{MS}_{4\uparrow}; \mathbf{P}] \end{aligned} \tag{5}$$

where $\mathbf{input}_{1\times} \in \mathbb{R}^{w\times h\times(b+1)}$, $\mathbf{input}_{2\times} \in \mathbb{R}^{2w\times2h\times(b+1)}$, $\mathbf{input}_{4\times} \in \mathbb{R}^{4w\times4h\times(b+1)}$, and $C[\cdot]$ stands for the concatenate operation. All these inputs are fed into the following U-shaped network in their corresponding scales.

3.3 MUCNN

The network arichitecture is shown in Figure 4. It consists of a feature extraction path (left side) and a reconstruction path (right side). The extraction path has two steps, each step contains a 3 × 3 convolution, a rectified linear unit (ReLU) and a max pooling operation. Between each step, a new input is concatenated after max pooling. As for the reconstruction path, it is made up of three steps, each step contains a SSconv operation and a 3×3 convolution. The reconstruction path is connected with the extraction path by two skip connections and a 3 × 3 convolution in the bottom of the network, see Figure 4 for more details.

To accelerate the back propagation and promote the network to learn the rich feature hierarchies, we set three 3 × 3 convolutions to get three outputs $\widehat{\mathbf{MS}}_{1\times} \in \mathbb{R}^{w\times h\times b}$, $\widehat{\mathbf{MS}}_{2\times} \in \mathbb{R}^{2w\times2h\times b}$ and $\widehat{\mathbf{MS}}_{4\times} \in \mathbb{R}^{4w\times4h\times b}$, while the $\widehat{\mathbf{MS}}_{4\times}$ is the desired HR-MS image. Overall, the MUCNN can be summarized as follows:

$$[\widehat{\mathbf{MS}}_{1\times}, \widehat{\mathbf{MS}}_{2\times}, \widehat{\mathbf{MS}}_{4\times}] = f_{\Theta_{MUCNN}}(\mathbf{input}_{1\times}, \mathbf{input}_{2\times}, \mathbf{input}_{4\times}) \tag{6}$$

where $f_{\Theta_{MUCNN}}$ present the network and Θ_{MUCNN} denotes the parameters inside the network. $\widehat{\mathbf{MS}}_{1\times}, \widehat{\mathbf{MS}}_{2\times}, \widehat{\mathbf{MS}}_{4\times}$ denote the outputs of MUCNN of different scales.

3.4 Mixed multiscale loss

The mixed loss strategy is proposed to make the most of rich feature hierarchies. On the premise that the reduced image is reliable, we compare the three outputs with the GT image of the corresponding scale to form the final loss function. Low-resolution GT ($\mathbf{GT}_{2\downarrow}$) and medium-resolution GT ($\mathbf{GT}_{4\downarrow}$) are obtained through linear interpolation. Finally, the mixed loss function of the MUCNN is defined as follows:

$$\begin{aligned} Loss(\Theta_{MUCNN}) = \frac{\lambda_1}{b}\sum_{i=1}^{b} \|\mathbf{GT} - \widehat{\mathbf{MS}}_{4\times}\|_F^2 \\ + \frac{\lambda_2}{b}\sum_{i=1}^{b} \|\mathbf{GT}_{2\downarrow} - \widehat{\mathbf{MS}}_{2\times}\|_F^2 \\ + \frac{\lambda_3}{b}\sum_{i=1}^{b} \|\mathbf{GT}_{4\downarrow} - \widehat{\mathbf{MS}}_{1\times}\|_F^2 \end{aligned} \tag{7}$$

where $\|\cdot\|_F$ is Frobenius norm, $\lambda_1, \lambda_2, \lambda_3$ in this work are three proportionality coefficients, which are set as $[0.5, 0.3, 0.2]$.

4 EXPERIMENTS

In this section, we conduct several comparative experiments using datasets acquired by WorldView-3 and GaoFen-2 sensors. The proposed MUCNN is compared with some state-of-the-art pansharpening methods belonging to the CS-based, MRA-based, and DL-based methods.

We set the kernel size in convolution in MUCNN as 3 × 3. As for the kernels in the multiscale injection branch, we have mentioned in Sections 3.1 and 3.2. The number of feature maps is shown in figure 4. The non-linear activation is ReLU [17]. We use Pytorch framework and the Adam [16] with a mini-batch size of 32 to train our network. We initialize the learning rate as 0.001 and divide it by 10 every 200 epoch, and terminate the training after 600 epochs.

4.1 Datasets

Datasets adopted in most of our work are downloaded on the public website[1]. For WorldView-3 (8-band) satellite whose resolutions are 0.5 m and 2 m for PAN and LR-MS, we simulate 12580 HR-PAN/LR-MS/GT image pairs with the size 64 × 64, 16 × 16 × 8, and 64 × 64 × 8, respectively. Besides, we divide them into 70%, 20%, and 10% for training (8806 examples), validation (2516 examples), and testing (1258 examples). The process simulating is as follows: 1) downsample the original HR-PAN and the original LR-MS image by a resolution ratio of 4 using modulation transfer function (MTF) based filters, 2) take the downsampled images as the HR-PAN and

[1] http://www.digitalglobe.com/samples?search=Imagery

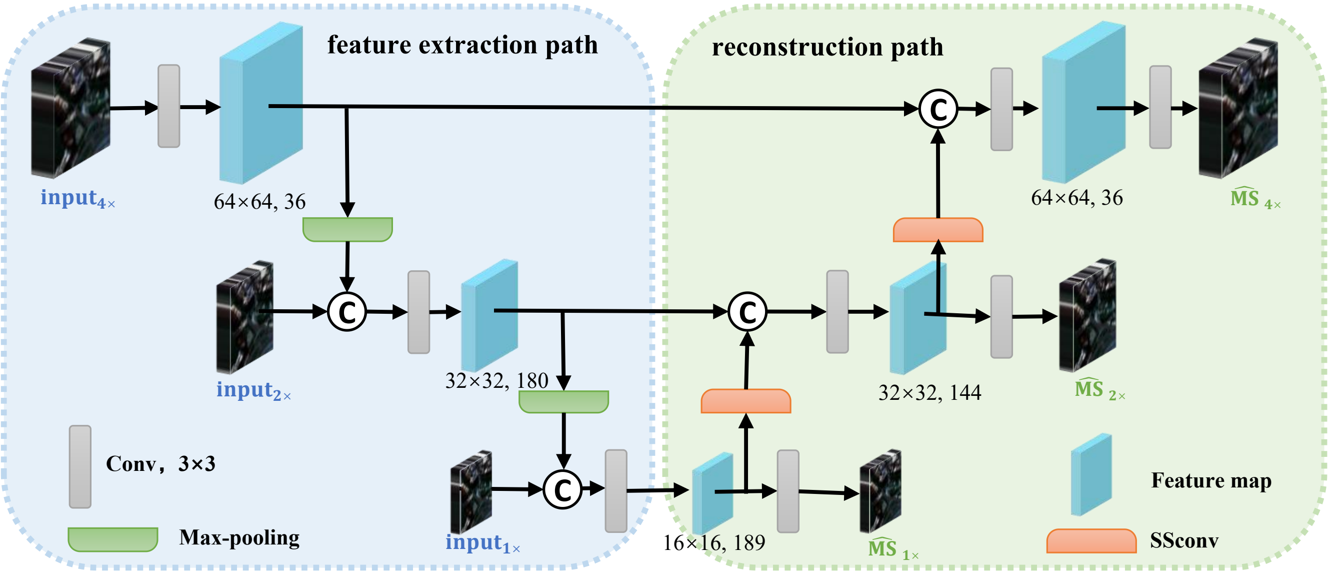


Figure 4: The architecture of the MUCNN. The sizes of feature maps are examples of the WorldView-3 training datasets mentioned in Section 4.1, while the numbers of the convolution filters are fixed.

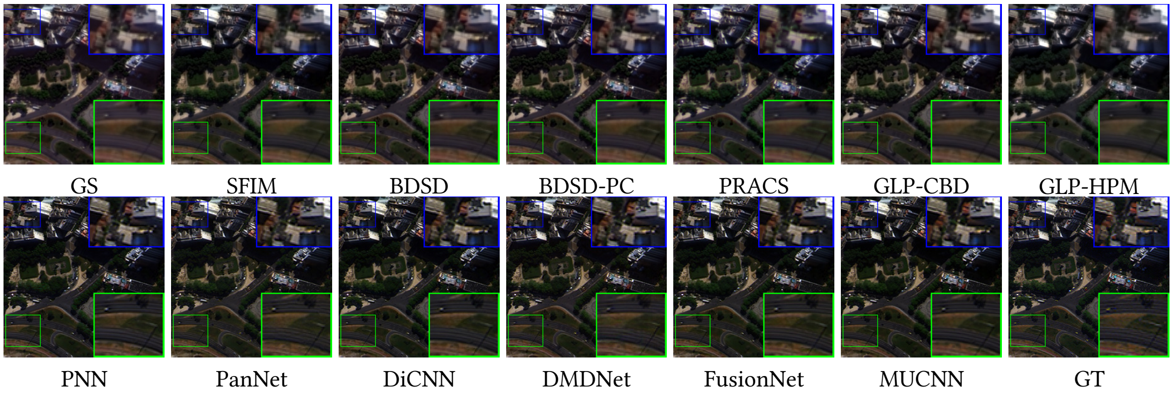


Figure 5: Visual comparisons in natural colors of the most representive approaches on reduced-resolution datasets (WorldView-3).

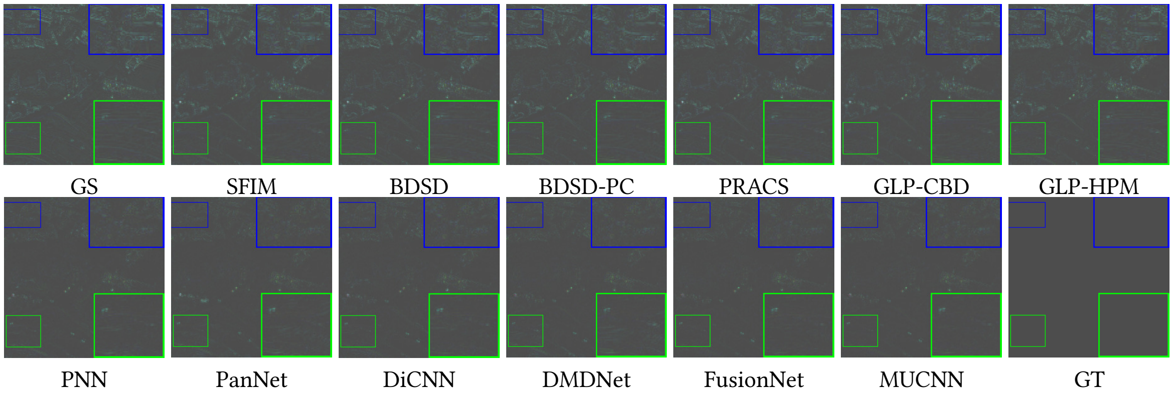
LR-MS images for training, validation, and testing, 3) the original LR-MS images are used as the GT images.

Apart from the 8-band WorldView-3 images, we also evaluate the performances of the proposed method on 4-band (red, green, blue, and near-infrared) datasets. In particular, we use the images acquired by GaoFen-2 whose resolutions are 3.2 m for the LR-MS images, 0.8 m for the PAN images, and 10 bits for radiometric. For GaoFen-2 (4-band) satellite, we generate the training and testing data using the same way as the WorldView-3 datasets. For training, we download a large dataset (6907 × 7300 × 4) over the city of Beijing from the website[2] to simulate 21607 examples. As for the testing, we use a huge image acquired by GaoFen-2 over the city of Guangzhou to simulate 81 testing data (size: 256 × 256 × 4).

4.2 Baseline Methods

We compare our network with four CS-based methods: the Gram-Schmidt sharpening approach (GS) [18], the band-dependent spatial-detail method (BDSD) [11], the robust band-dependent spatial-detail approach (BDSD-PC) [28], the partial replacement adaptive component substitution approach (PRACS) [5], there MRA-based methods: the smoothing filter-based intensity modulation (SFIM) [19], the GLP with MTF-matched filter [2] and multiplicative injection model [32] (GLP-HPM), the GLP with MTF-matched filter [2] and

[2] http://www.rscloudmart.com/dataProduct/sample


GS SFIM BDSD BDSD-PC PRACS GLP-CBD GLP-HPM

PNN PanNet DiCNN DMDNet FusionNet MUCNN GT

Figure 6: The AEMs of Figure 5.

regression-based injection model (GLP-CBD) [4], [1], and five DL-based methods, such as PNN [21], PanNet [36], DiCNN [13], DMDNet [9] and the FushionNet [8].

To be fair in comparison, all the CNNs are trained and tested on the same datasets, the same hardware, and the same software environments.

4.3 Evaluation Metrics

The performance assessment is implemented at both reduced and full resolution. For the reduced resolution, the five measures are chosen: the spectral angle mapper (SAM) [37], the relative dimensionless global error in synthesis (ERGAS) [33], the spatial correlation coefficient (SCC) [40], universal image quality index [34] averaged over the bands (QAVE), and the universal image quality index for 4-band image (Q4) and 8-band images (Q8) [10]. For Q4, Q8, QAVE, and SCC, the desired value is 1, while SAM and ERGAS are both 0. As for the performance assessment at full resolution, we use the QNR, the D_λ, and the D_s indexes [30]. The QNR's desired value is 1, instead of 0 for D_λ and D_s.

4.4 Reduced-Resolution Experimental Results

We first train our network on the training datasets of WorldView-3. Then, we test our model on 1258 reduced-resolution images from WorldView-3, whose number of spectral bands is 8. For color visualization, we only display the three selected bands, while all spectral bands are used for quality assessments.

In Table 1, we list the average quantitative results of those methods on the testing datasets. For each pair in testing datasets, PAN, LR-MS, and GT images are of the same scale as the training datasets. In the table, the mean and standard deviation of quantitative scores are shown. Among them, the model we proposed performs the best. What's more, we present natural color maps and the absolute error maps (AEM) with GT in Figure 5 and Figure 6 respectively. The better result, which is less different from GT, has a darker AEM. It is clear that MUCNN performs better.

4.5 Evaluation at the Full-Resolution

We evaluate the different methods at the full-resolution of the WorldView-3 satellite on 200 test images with the size of 256×256×8.

Table 1: Average quantitative comparisons of the most representive approaches on 1258 reduced-resolution WorldView-3 samples. Best results in boldface.

Method	*SAM*	*ERGAS*	*SCC*	*Q8*	*QAVE*
GS [18]	5.698 ± 2.008	5.282 ± 2.187	0.873 ± 0.071	0.766 ± 0.139	0.768 ± 0.146
SFIM [19]	5.452 ± 1.903	5.200 ± 6.574	0.866 ± 0.067	0.798 ± 0.122	0.811 ± 0.130
BDSD [11]	7.000 ± 2.853	5.167 ± 2.248	0.871 ± 0.080	0.813 ± 0.123	0.817 ± 0.126
BDSD-PC [28]	5.425 ± 1.972	4.246 ± 1.860	0.891 ± 0.069	0.853 ± 0.116	0.852 ± 0.124
PRACS [5]	5.286 ± 1.958	4.163 ± 1.775	0.890 ± 0.070	0.854 ± 0.114	0.849 ± 0.123
GLP-CBD [4]	5.286 ± 1.958	4.163 ± 1.775	0.890 ± 0.070	0.854 ± 0.114	0.849 ± 0.123
GLP-HPM [32]	5.604 ± 1.974	4.764 ± 1.935	0.873 ± 0.065	0.817 ± 0.128	0.810 ± 0.135
PNN [21]	4.002 ± 1.329	2.728 ± 1.004	0.952 ± 0.046	0.908 ± 0.112	0.911 ± 0.114
PanNet [36]	4.092 ± 1.273	2.952 ± 0.978	0.949 ± 0.046	0.894 ± 0.117	0.907 ± 0.118
DiCNN [13]	3.981 ± 1.318	2.737 ± 1.016	0.952 ± 0.046	0.910 ± 0.112	0.911 ± 0.115
DMDNet [9]	3.971 ± 1.248	2.857 ± 0.966	0.953 ± 0.045	0.900 ± 0.114	0.913 ± 0.115
FusionNet [8]	3.744 ± 1.226	2.568 ± 0.994	0.958 ± 0.045	0.914 ± 0.112	0.914 ± 0.117
MUCNN	**3.495 ± 1.254**	**2.425 ± 0.956**	**0.963 ± 0.044**	**0.923 ± 0.109**	**0.921 ± 0.114**
Ideal value	**0**	**0**	**1**	**1**	**1**

Table 2: Average quantitative comparisons of the most representive approaches on 200 full-resolution WorldView-3 examples. Best results in boldface.

Method	*QNR*	D_λ	D_s
GS [18]	0.9026 ± 0.0453	0.0172 ± 0.0195	0.0821 ± 0.0322
SFIM [19]	0.9346 ± 0.0453	0.0216 ± 0.0210	0.0452 ± 0.0212
BDSD [11]	0.9352 ± 0.0389	0.0171 ± 0.0116	0.0488 ± 0.0309
BDSD-PC [28]	0.9166 ± 0.0495	0.0193 ± 0.0190	0.0660 ± 0.0357
PRACS [5]	0.9149 ± 0.0448	0.0174 ± 0.0165	0.0694 ± 0.0329
GLP-CBD [4]	0.9195 ± 0.0504	0.0278 ± 0.0242	0.0550 ± 0.0321
GLP-HPM [32]	0.8939 ± 0.0621	0.0470 ± 0.0322	0.0633 ± 0.0380
PNN [21]	0.9591 ± 0.0260	0.0163 ± 0.0149	0.0251 ± 0.0139
PanNet [36]	0.9581 ± 0.0199	0.0224 ± 0.0108	**0.0201 ± 0.0111**
DiCNN [13]	0.9460 ± 0.0325	0.0165 ± 0.0160	0.0385 ± 0.0201
DMDNet [9]	0.9460 ± 0.0196	0.0187 ± 0.0093	0.0214 ± 0.0122
fusionNet [8]	0.9559 ± 0.0276	0.0178 ± 0.0151	0.0269 ± 0.0161
MUCNN	**0.9629 ± 0.0215**	**0.0128 ± 0.0140**	0.0247 ± 0.0102
Ideal value	**1**	**0**	**0**

We give an example in Figure 7. As there is no GT image, we show the LR-MS image instead.

We evaluate the performances of different approaches with the assessment mentioned before, and the result is shown in Table 2. Our proposed model performs the best in the assessment of QNR and D_λ, except for D_s.

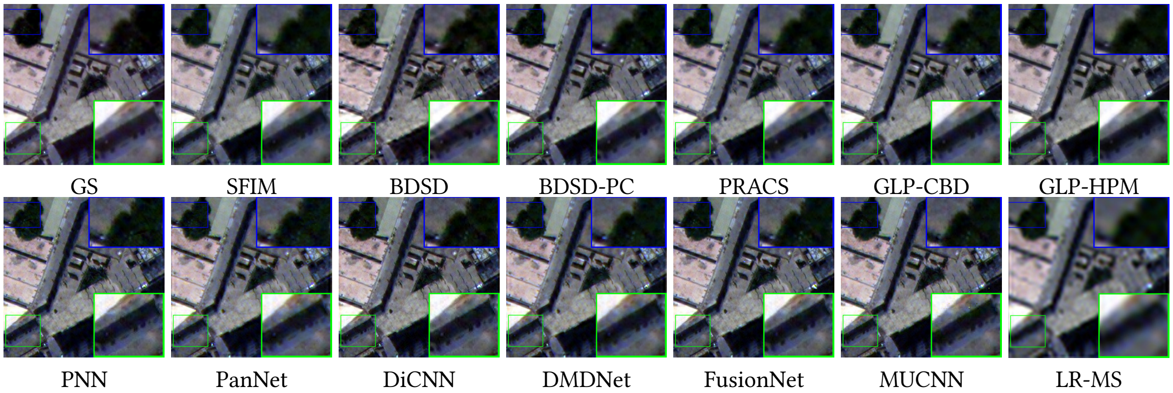


Figure 7: Visual comparisons in natural colors of the most representive approaches on full-resolution datasets (WorldView-3).

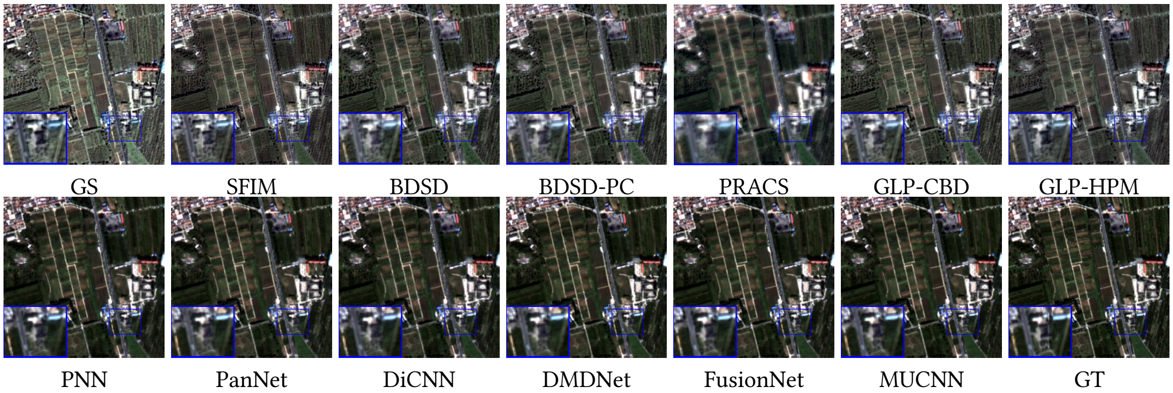


Figure 8: Visual comparisons in natural colors of the most representive approaches on reduced-resolution datasets (GaoFen-2).

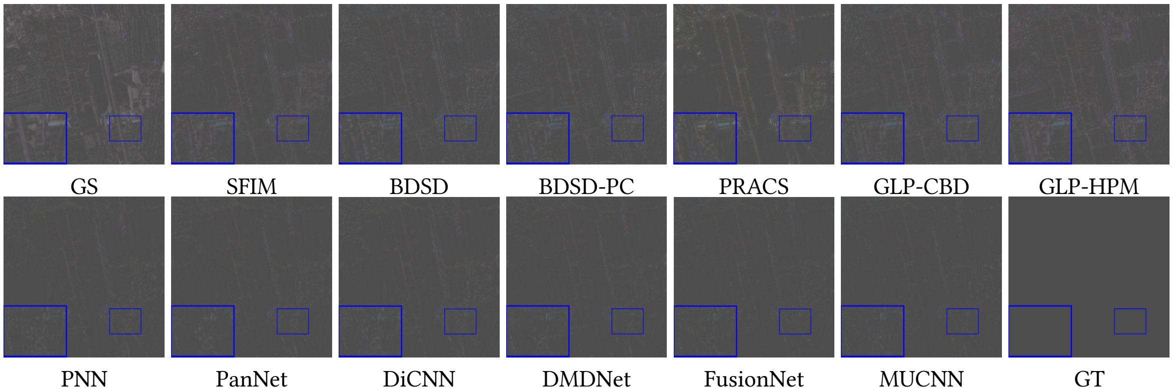


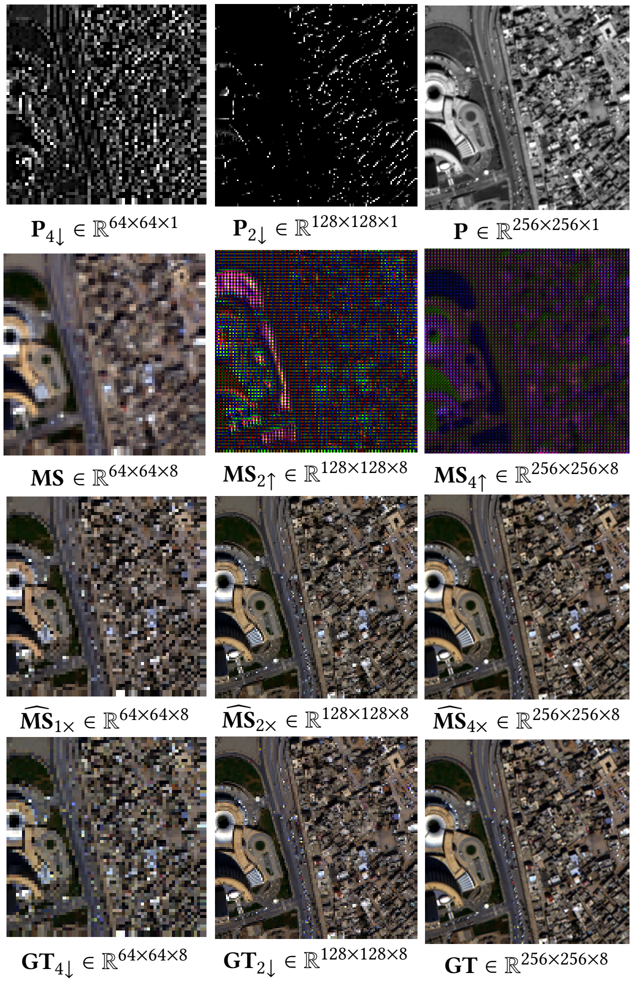
Figure 9: The AEMs of Figure 8.

4.6 Evaluation at 4-band datasets

From the indicators shown in Table 3, and the visual results shown in Figure 8 and Figure 9, the proposed MUCNN can recover more details in the spatial dimension without losing the spectral information, and its results far exceed the existing methods. This indicates that MUCNN can also be applied to 4-band data and its outcomes are satisfactory enough.

Table 3: Average quantitative comparisons of the most representive approaches on 81 reduced-resolution GaoFen-2 examples. Best results in boldface.

Method	*SAM*	*ERGAS*	*SCC*	*Q8*	*QAVE*
GS [18]	2.975 ± 1.111	2.966 ± 1.010	0.852 ± 0.062	0.787 ± 0.076	0.797 ± 0.076
SFIM [19]	2.297 ± 0.637	2.189 ± 0.695	0.861 ± 0.054	0.865 ± 0.040	0.876 ± 0.037
BDSD [11]	2.307 ± 0.669	2.070 ± 0.610	0.877 ± 0.052	0.876 ± 0.042	0.885 ± 0.020
BDSD-PC [28]	2.304 ± 0.643	2.075 ± 0.604	0.878 ± 0.051	0.878 ± 0.040	0.887 ± 0.039
PRACS [5]	2.311 ± 0.597	2.169 ± 0.599	0.867 ± 0.050	0.872 ± 0.035	0.876 ± 0.034
GLP-CBD [4]	2.274 ± 0.733	2.046 ± 0.620	0.873 ± 0.053	0.877 ± 0.041	0.880 ± 0.040
GLP-HPM [32]	0.552 ± 0.777	2.299 ± 0.713	0.867 ± 0.054	0.852 ± 0.045	0.852 ± 0.044
PNN [21]	1.460 ± 0.361	1.271 ± 0.324	0.948 ± 0.021	0.947 ± 0.020	0.949 ± 0.017
PanNet [36]	1.395 ± 0.326	1.224 ± 0.283	0.956 ± 0.012	0.947 ± 0.022	0.957 ± 0.015
DiCNN [13]	1.495 ± 0.381	1.320 ± 0.354	0.946 ± 0.022	0.945 ± 0.021	0.947 ± 0.018
DMDNet [9]	1.297 ± 0.316	1.128 ± 0.267	0.964 ± 0.010	0.953 ± 0.022	0.963 ± 0.014
FusionNet [8]	1.219 ± 0.292	1.037 ± 0.256	0.968 ± 0.010	0.962 ± 0.017	0.964 ± 0.015
MUCNN	**1.100 ± 0.274**	**0.937 ± 0.234**	**0.975 ± 0.008**	**0.970 ± 0.013**	**0.970 ± 0.013**
Ideal value	**0**	**0**	**1**	**1**	**1**



$\mathbf{P}_{4\downarrow} \in \mathbb{R}^{64\times64\times1}$ $\mathbf{P}_{2\downarrow} \in \mathbb{R}^{128\times128\times1}$ $\mathbf{P} \in \mathbb{R}^{256\times256\times1}$

$\mathbf{MS} \in \mathbb{R}^{64\times64\times8}$ $\mathbf{MS}_{2\uparrow} \in \mathbb{R}^{128\times128\times8}$ $\mathbf{MS}_{4\uparrow} \in \mathbb{R}^{256\times256\times8}$

$\widehat{\mathbf{MS}}_{1\times} \in \mathbb{R}^{64\times64\times8}$ $\widehat{\mathbf{MS}}_{2\times} \in \mathbb{R}^{128\times128\times8}$ $\widehat{\mathbf{MS}}_{4\times} \in \mathbb{R}^{256\times256\times8}$

$\mathbf{GT}_{4\downarrow} \in \mathbb{R}^{64\times64\times8}$ $\mathbf{GT}_{2\downarrow} \in \mathbb{R}^{128\times128\times8}$ $\mathbf{GT} \in \mathbb{R}^{256\times256\times8}$

Figure 10: The multiscale PAN/LR-MS/GT/output images. Please note that for the multispectral images, we show them in naturall colors. And the images in same column are of same resolution.

4.7 Comparison of the MUCNNs with different strategies

Since the multiscale injection branch, SSconv, the mixed multiscale loss is the core of our method. For demonstrating their indispensable and effectiveness, we provide the ablation study and compare our model with its seven variants on the datasets from WorldView-3 that is introduced in Section 4.1. As reported in Table 4, compared with our backbone model, adding the multiscale injection branch achieves an improvement on multiple quantitative metrics. Similarly, the mixed multiscale loss improves the results markedly. The best result can be obtained by combining the three proposed components, which exceed other network structures significantly.

4.8 Discussion on Multiscale Inputs and Outputs

The up/down-sampling method in prior approaches is fixed, which is inflexible for feature representation and extraction. In our work, the up/down-sampling modules we designed are equipped with learnable parameters, which are adaptively adjusted according to specific training examples. Therefore, we attempt to analyze these up/down-sampled images to explore how to help the network perform better.

As shown in Figure 10, we present the multiscale inputs and outputs. It is clear that $\mathbf{P}_{4\downarrow}$ maintains the outline. Although $\mathbf{P}_{2\downarrow}$ almost lost its original outline, which looks like a differential map that has be verified as useful feature in [15]. Besides, $\mathbf{MS}_{2\uparrow}$ and $\mathbf{MS}_{4\uparrow}$ are produced by the given SSconv operating on $\mathbf{MS}$, containing rich spectral information and conducive to feature extraction. Moreover, it can be seen that the multiscale outputs ($\widehat{\mathbf{MS}}_{1\times}$, $\widehat{\mathbf{MS}}_{2\times}$, $\widehat{\mathbf{MS}}_{4\times}$) are quite close to their corresponding GT images, which proves the effectiveness of the mixed multiscale loss.

5 CONCLUSIONS

This paper proposes MUCNN with the SSconv which is specially designed for pansharpening. The key difference from the prior techniques lies in that we map spectral feature to the spatial domain through SSconv so that the feature extraction of the MUCNN is more competent to the fusion of HR-PAN and LR-MS images. Besides, the multiscale injection branch is introduced to mitigate the distortion caused by the upsampling of LR-MS images. We choose U-Net as the backbone to construct MUCNN. Also, a mixed loss strategy is used to control the outputs hierarchically. A wide range of experiments demonstrates that our proposed method not only can capture the underlying details of HR-PAN and LR-MS images but also holds the powerful ability to balance spatial restoration and spectral preservation.

Certainly, there are still some drawbacks to our method, especially for the extremely bright spots on the images. For example, the reflection of the sun, which comes from the roof of a car towards the sensor, sometimes will be sharpened like a flock of scattered stars and lose the original outline of the car roof. Apart from that, sometimes pixel-wise noisy points will appear in the solid region. Through the experiments, we find that almost all DL-based methods had similar problems with uneven edges, while traditional methods

Table 4: Average quantitative comparisons of MUCNNs with different strategies on 1258 reduced-resolution WorldView-3 examples.

Method	*multiscale injection branch*	*SSconv*	*mixed multiscale loss*	*SAM*	*ERGAS*	*SCC*	*Q8*	*QAVE*
UMCNN	×	×	×	3.755 ± 1.264	2.565 ± 0.954	0.959 ± 0.045	0.915 ± 0.111	0.914 ± 0.115
	×	×	✓	3.660 ± 1.262	2.511 ± 0.946	0.961 ± 0.044	0.918 ± 0.109	0.916 ± 0.115
	×	✓	×	4.265 ± 1.381	2.973 ± 1.107	0.945 ± 0.049	0.900 ± 0.118	0.900 ± 0.122
	✓	×	×	3.593 ± 1.234	2.441 ± 0.938	0.962 ± 0.044	0.920 ± 0.109	0.918 ± 0.114
	✓	✓	×	3.985± 1.282	2.699 ± 0.987	0.955 ± 0.045	0.909 ± 0.116	0.908 ± 0.115
	✓	×	✓	3.676± 1.229	2.507 ± 0.932	0.960 ± 0.044	0.917 ± 0.110	0.915 ± 0.116
	×	✓	✓	3.745± 1.270	2.567 ± 0.954	0.959 ± 0.044	0.915 ± 0.445	0.914 ± 0.116
	✓	✓	✓	**3.495 ± 1.254**	**2.425 ± 0.956**	**0.963 ± 0.044**	**0.923 ± 0.109**	**0.921 ± 0.114**
Ideal value				**0**	**0**	**1**	**1**	**1**

did not. This drawback reminds us to look for the characteristics of conventional methods and combine them with DL-based methods.

REFERENCES

[1] Bruno Aiazzi, Luciano Alparone, Stefano Baronti, and Andrea Garzelli. 2002. Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. *IEEE Transactions on geoscience and remote sensing* 40, 10 (2002), 2300–2312.

[2] B Aiazzi, L Alparone, S Baronti, A Garzelli, and M Selva. 2006. MTF-tailored multiscale fusion of high-resolution MS and Pan imagery. *Photogrammetric Engineering & Remote Sensing* 72, 5 (2006), 591–596.

[3] Luciano Alparone, Lucien Wald, Jocelyn Chanussot, Claire Thomas, Paolo Gamba, and Lori Mann Bruce. 2007. Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S data-fusion contest. *IEEE Transactions on Geoscience and Remote Sensing* 45, 10 (2007), 3012–3021.

[4] Luciano Alparone, Lucien Wald, Jocelyn Chanussot, Claire Thomas, Paolo Gamba, and Lori Mann Bruce. 2007. Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S data-fusion contest. *IEEE Transactions on Geoscience and Remote Sensing* 45, 10 (2007), 3012–3021.

[5] Jaewan Choi, Kiyun Yu, and Yongil Kim. 2010. A new adaptive component-substitution-based satellite image fusion by using partial replacement. *IEEE Transactions on Geoscience and Remote Sensing* 49, 1 (2010), 295–309.

[6] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*. Springer, 424–432.

[7] Mauro Dalla Mura, Saurabh Prasad, Fabio Pacifici, Paulo Gamba, Jocelyn Chanussot, and Jón Atli Benediktsson. 2015. Challenges and opportunities of multimodality and data fusion in remote sensing. *Proc. IEEE* 103, 9 (2015), 1585–1601.

[8] Liang-Jian Deng, Gemine Vivone, Cheng Jin, and Jocelyn Chanussot. 2020. Detail Injection-Based Deep Convolutional Neural Networks for Pansharpening. *IEEE Transactions on Geoscience and Remote Sensing* (2020).

[9] Xueyang Fu, Wu Wang, Yue Huang, Xinghao Ding, and John Paisley. 2020. Deep multiscale detail networks for multiband spectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[10] Andrea Garzelli and Filippo Nencini. 2009. Hypercomplex quality assessment of multi/hyperspectral images. *IEEE Geoscience and Remote Sensing Letters* 6, 4 (2009), 662–665.

[11] Andrea Garzelli, Filippo Nencini, and Luca Capobianco. 2007. Optimal MMSE pan sharpening of very high resolution multispectral images. *IEEE Transactions on Geoscience and Remote Sensing* 46, 1 (2007), 228–236.

[12] Kaiming He and Jian Sun. 2015. Convolutional neural networks at constrained time cost. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5353–5360.

[13] Lin He, Yizhou Rao, Jun Li, Jocelyn Chanussot, Antonio Plaza, Jiawei Zhu, and Bo Li. 2019. Pansharpening via detail injection based convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12, 4 (2019), 1188–1204.

[14] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. 2019. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2024–2032.

[15] Menghui Jiang, Huanfeng Shen, Jie Li, Qiangqiang Yuan, and Liangpei Zhang. 2020. A differential information residual convolutional neural network for pansharpening. *ISPRS Journal of Photogrammetry and Remote Sensing* 163 (2020), 257–271.

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.

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

[19] JG Liu. 2000. Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing* 21, 18 (2000), 3461–3472.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[21] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. 2016. Pansharpening by convolutional neural networks. *Remote Sensing* 8, 7 (2016), 594.

[22] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957* (2018).

[23] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. 2018. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999* (2018).

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[25] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1874–1883.

[26] Karasawa Takumi, Kohei Watanabe, Qishen Ha, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Multispectral object detection for autonomous vehicles. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. 35–43.

[27] Claire Thomas, Thierry Ranchin, Lucien Wald, and Jocelyn Chanussot. 2008. Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics. *IEEE Transactions on Geoscience and Remote Sensing* 46, 5 (2008), 1301–1312.

[28] Gemine Vivone. 2019. Robust band-dependent spatial-detail approaches for panchromatic sharpening. *IEEE transactions on Geoscience and Remote Sensing* 57, 9 (2019), 6421–6433.

[29] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A Licciardi, Rocco Restaino, and Lucien Wald. 2014. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing* 53, 5 (2014), 2565–2586.

[30] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A Licciardi, Rocco Restaino, and Lucien Wald. 2014. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing* 53, 5 (2014), 2565–2586.

[31] Gemine Vivone, Mauro Dalla Mura, Andrea Garzelli, Rocco Restaino, Giuseppe Scarpa, Magnus Orn Ulfarsson, Luciano Alparone, and Jocelyn Chanussot. 2020. A New Benchmark Based on Recent Advances in Multispectral Pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods. *IEEE Geoscience and Remote Sensing Magazine* (2020).

[32] Gemine Vivone, Rocco Restaino, Mauro Dalla Mura, Giorgio Licciardi, and Jocelyn Chanussot. 2013. Contrast and error-based fusion schemes for multispectral image pansharpening. *IEEE Geoscience and Remote Sensing Letters* 11, 5 (2013), 930–934.

[33] Lucien Wald. 2002. *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES.

[34] Zhou Wang and Alan C Bovik. 2002. A universal image quality index. *IEEE signal processing letters* 9, 3 (2002), 81–84.

[35] Shuang Wu, Guanrui Wang, Pei Tang, Feng Chen, and Luping Shi. 2019. Convolution with even-sized kernels and symmetric padding. *arXiv preprint arXiv:1903.08385* (2019).

[36] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. 2017. PanNet: A deep network architecture for pan-sharpening. In *Proceedings of the IEEE international conference on computer vision*. 5449–5457.

[37] Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. 1992. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, Vol. 1. 147–149.

[38] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

[39] Huanrong Zhang, Zhi Jin, Xiaojun Tan, and Xiying Li. 2020. Towards Lighter and Faster: Learning Wavelets Progressively for Image Super-Resolution. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2113–2121.

[40] J Zhou, DL Civco, and JA Silander. 1998. A wavelet transform method to merge Landsat TM and SPOT panchromatic data. *International journal of remote sensing* 19, 4 (1998), 743–757.

[41] Qiang Zhou, Shifeng Chen, Jianzhuang Liu, and Xiaoou Tang. 2011. Edge-preserving single image super-resolution. In *Proceedings of the 19th ACM International Conference on Multimedia*. 1037–1040.

[42] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 3–11.