KNLConv: Kernel-space Non-local Convolution for Hyperspectral Image Super-resolution

Ran Ran, Liang-Jian Deng, *Senior Member, IEEE*, Tian-Jing Zhang,
Jianlong Chang, Xiao Wu, Qi Tian, *Fellow, IEEE*

Pixel-level adaptive convolution, which overcomes the deficiency of the spatial-invariance of standard convolution, is always limited to performing feature extraction from local patches and ignores the latent long-range dependencies imperceptible in the feature space, which are more significant in pixel-level tasks such as hyperspectral image super-resolution (HSISR). To handle such limitations, we propose kernel-space non-local convolution (KNLConv), which explores non-local dependencies in the generated kernel space, to leverage these global information to guide the network to extract image features more flexibly. Technically, the proposed KNLConv first decomposes the convolutional kernel space into spatial and channel dimensions, and designs a depth-wise non-local expansion convolution (NLEC) in the spatial dimension of the kernel-space to explore underlying global correlations. Then introduce an adaptive point-wise convolution (APC), generalizing the NLEC to the pixel-level while integrating features in the channel dimension. In addition, applying KNLConv, we design an effective network architecture for hyperspectral image super-resolution. Extensive experiments demonstrate that our approach performs favorably against current state-of-the-art HSISR methods, both on quantitative indicators and visual quality. The code is available at https://github.com/Evangelion09/KNLNet.

I. Introduction

Compared with common images with one or RGB channels, hyperspectral images (HSI) provide richer spectral information and can be used to explore more intrinsic properties of objects [14], [16], [30], [43], [69]. However, due to the hardware limitations of the imaging sensor, there are unavoidable tradeoffs between spatial resolution and spectral resolution. Therefore, hyperspectral image super-resolution (HSISR), which aims at fusing a high-resolution multispectral image (HR-MSI) and a low-resolution hyperspectral image (LR-HSI) that record the same scene to generate an ideal high-resolution hyperspectral image (HR-HSI), has received more and more interest in recent years [9], [20], [26], [58].

There has been notable progress in computer vision thanks to the current methods based on deep convolutional neural networks (CNNs), leveraging the powerful modeling capabilities for the relationship of images [2], [3], [7], [27], [49], [54], [55], [57]. Extensive studies have been conducted to improve the performance of HSISR [23], [32], [40]–[42], [56]. One prevailing direction is to change the structure of the network, such as increasing the depth [62], multi-scale structures [71] or attention mechanisms [23]. However, the enhancement of the feature representation capability of the CNNs by these approaches is limited.

Driven by the shortcoming of the standard convolution, *i.e.*, shared convolutional kernels are applied to each pixel of the input to extract features, some researchers have proposed adaptive convolutions that generate kernels based on the input information. These adaptive convolutions mainly generate unique convolution kernel correspond to each pixel, termed as pixel-level adaptive convolution. In [25] and [67], the convolution kernel of each pixel is generated directly. In [44], a shared convolutional kernel is modified by a learnable adapting kernel to obtain pixel-diverse kernels. Zhou *et al.* proposed the DDF, which produces spatial and channel dynamic filters decoupled by the adaptive kernel and combines them into the final kernel. In [29], Li *et al.* devised the Involution operation with only spatial specific, which generates the kernel that is identical in the channel direction from the input features.

Although these adaptive convolution methods achieve impressive performance, their adaptive kernels are generated from local information of the corresponding pixel neighborhoods that lack distant information as guidance to extract features comprehensively.

As shown in Fig. 1, non-local convolutional kernel similarities are prevalent, and capturing these non-local relations may enhance the perceptual field of adaptive convolution and contribute to the feature characterization of hyperspectral images. To obtain long-range dependencies hidden in kernel-space and pick up information ignored in the feature space, we develop non-local operations in adaptive convolution and propose a kernel-space non-local convolution (KNLConv). First, we design a pixel-level adaptive depth-wise convolution named non-local expansion convolution (NLEC), consisting of a channel weight generation (CWG) branch and a non-local spatial kernel generation (NSKG) branch, which can explore the latent relationships between the kernels at a lower-resolution. Afer that, the depth-wise convolution kernels are generated by expand the spatial kernels along the channel dimension with the learned weights. Then, a pixel-level adaptive point-wise convolution (APC) is designed to improve the representation of channel information and yield the final outputs. Finally, the above modules are integrated into kernel non-local residual block, thus building the complete KNLNet

The research is supported by NSFC (No. 12271083), and National Key Research and Development Program of China (No. 2020YFA0714001).

Ran Ran, Liang-Jian Deng, and Xiao Wu are with the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China (e-mails: ranran@std.uestc.edu.cn; liangjian.deng@uestc.edu.cn; wxwsx1997@gmail.com).

Tian-Jing Zhang is with the Department of Mathematics, National University of Singapore, (e-mail: zhangtianjinguestc@163.com).

Jianlong Chang and Qi Tian are with Cloud & AI, Huawei Technologies, (e-mail: {jianlong.chang, tian.qi1}@huawei.com).

Corresponding authors: Liang-Jian Deng

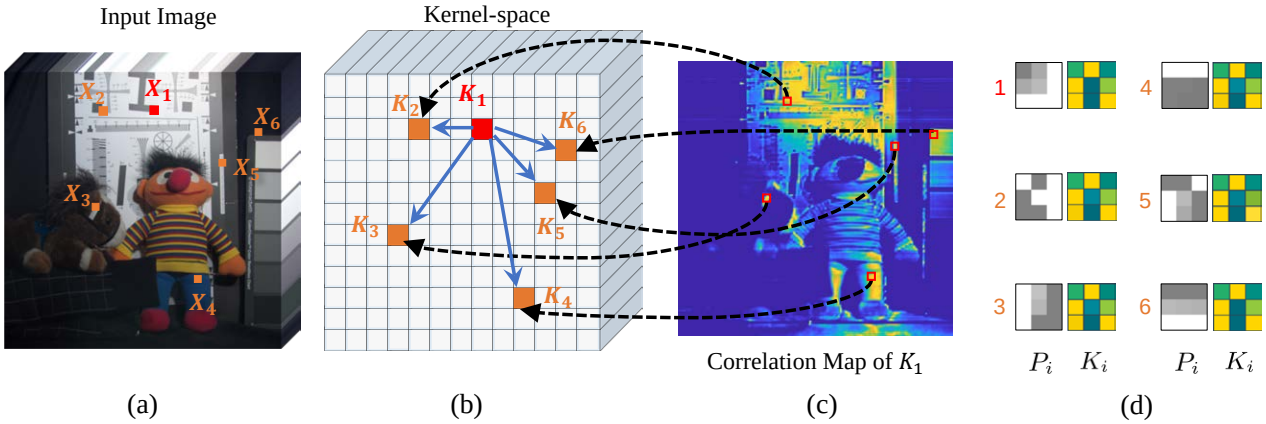


Fig. 1. The motivation of proposed KNLConv. (a) Input Image $\mathcal{I} \in \mathbb{R}^{H\times W\times C}$ (also can be a feature map) belonging to feature space; (b) Kernels $\mathcal{K} \in \mathbb{R}^{H\times W\times K^2C}$ yielded by our self-expansion module, belonging to kernel space; (c) The correlation map of K_1 obtained by calculating the similarity between the kernel K_1 and other kernel K_i $(i = 1, 2, \cdots, HW)$; By (c), it is clear that there exists global long-distance dependencies among kernels; (d) A toy example, in which the left part represents X_i-centered 3×3 patches P_i, and the right part is the corresponding kernel K_i; By (d), various patches in feature space may have similar kernels, which indicates the obvious correlation existed in the kernel space is imperceptible in the feature space.

for HSI super-resolution.

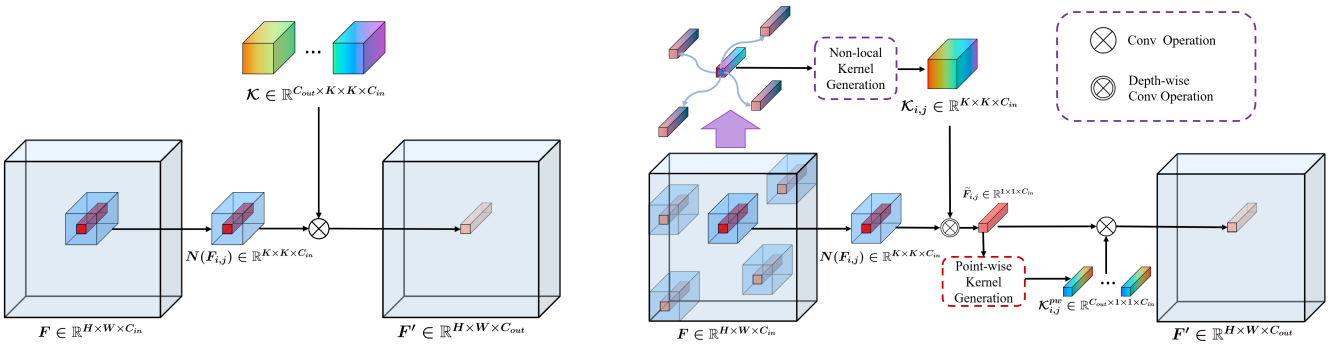
The key contributions can be summarized as follows:

- We propose a novel pixel-level adaptive KNLConv to build a deep neural network for the task of HSISR. The KNLConv overcomes the content-agnostic of standard convolutional with the capability of pixel-level representation. In addition, it is capable of exploring the latent dependencies in kernel space and more effectively extracting spatial and channel information from feature maps.
- A non-local expansion convolution module is proposed to transfer the input features into the kernel space, aiming to thoroughly capture the local information considering spatial and channel relationships. Moreover, we introduce the non-local technique into the spatial kernel, mitigating the discontinuous correlation caused by prior pixel-level adaptive convolutions.
- An adaptive point-wise convolution is designed by generalizing the dynamic convolution to the pixel level, generating a kernel for each pixel to explore the relationships in channel dimension, thus mitigating the lack of channel information caused by depth-wise convolution.
- To the best of our knowledge, we are the first to introduce non-local techniques into the kernel space and use pixel-level adaptive convolution for HSISR. Extensive experiments on several commonly used HSISR datasets demonstrate that the proposed method could achieve competitive performance compared with recent state-of-the-art (SOTA) HSISR techniques.

II. Related Work

Hyperspectral Image Super-resolution. Recent years have witnessed tremendous strides in improving the quality of generated images by HSISR. Existing methods can be roughly categorized into variable optimization (VO) based approach and deep learning (DL) based approach. The VO-based approaches mainly emphasize proposing artificial priors to regularize the to-be-estimated HR-HSI [5], [12], [31], [33], [48], [59]. For instance, Grohnfeldt *et al.* [18] introduced a sparse prior between different channels in HSI. Dian *et al.* [10] exploited the non-local prior factors of HSI. Lanaras [28] *et al.* added a sparse prior to the abundance matrix from the HR-HSI decomposition. In [11], Dian *et al.* cluster patches in HSI and MSI with a low tensor-training rank (LTTR) constraint, achieving satisfactory results. The main disadvantage of these methods is that their effectiveness is dependent on manually pre-specified prior terms, and most manual prior assumptions may be incapable of comprehensively reflecting diverse and complicated images collected from real scenarios.

Recently, many DL-based approaches have shown promising performance in image fusion tasks [1], [19], [37], [38], [60], [63], [64], [74]–[77], mainly due to the powerful feature extraction and representation ability of CNNs. Palsson *et al.* introduced 3D convolution into the network [39], which facilitates the exploration of the spectral information of HSI. In [24], Hu *et al.* utilized the differential information to efficiently exploit the spectral features between consecutive bands. Xie *et al.* [56] proposed a deep network, CMHFnet, which can effectively capture spatial and spectral information using the prior knowledge learned by the network and producing satisfactory results. In [58], an innovative multi-level supervised learning paradigm was proposed for small object detection, which introduced more supervision to guide network training and significantly enhanced the network interpretability. This method effectively enhanced small targets detection performance in the low-resolution images and opens up some new ideas for remote sensing image small object detection. Dong *et al.* unfolded the iterative HSISR algorithm into a model-guided deep convolutional network (MoG-DCN) and achieved state-of-the-art results. Xu *et al.* proposed a coarse-to-fine unsupervised pansharpening method constrained by novel spectral and textural loss functions [57]. It designed a multi-level unsupervised framework to guide pansharpening



(a) Standard Convolution (b) Kernel-space Non-local Convolution (KNLConv)

Fig. 2. Illustration of the standard convolution and the proposed kernel-space non-local convolution (KNLConv). $N(\cdot)$ represents a pixel and its neighbors.

learning, so that can significantly reduce spectral and spatial distortions caused by the differences between PAN and MS images. These methods depend on the standard convolution, which is space-invariant and content-agnostic. As shown in Fig. 2 (a), the local patches of all pixels for the input features use the same kernel.

However, for HSISR that essentially requires pixel-level accuracy, the losses of different pixels will be superimposed since the same kernel is not optimal for all pixels, causing the efficiency of each convolution layer to be compromised [73]. Besides, a trained convolutional layer that uses the same kernel for any input is content-agnostic, resulting in limited image content adaptation and cannot be applied to every example in datasets. [44].

Pixel-level Adaptive Convolution. While standard convolution is simple and widely used, specialized techniques have been developed for adaptive convolution kernel to improve the flexibility of the standard convolution. One feasible approach is to use a semi-dynamic kernel [6], [29], [35], [45], [50], [61], [73], which obtains a combined convolutional kernel based on the input. Nevertheless, the semi-dynamic convolution is still space-invariant and has limited representation ability. Therefore, some approaches propose pixel-level convolution kernels [25], [44], [67]. For example, in [25], Jia *et al.* used a simple network to generate convolution kernels for all pixels. Wang *et al.* [50] proposed a CARAFE that introduces semantic information of the feature map in the convolution kernel branches to achieve adaptive upsampling convolution. In DDF [73], Zhou *et al.* generated a series of space-variant filters and a set of channel-variant filters in the convolutional kernel prediction module, which significantly reduces the computational burden.

III. The Proposed Approach

A. Motivation

Although the aforementioned pixel-level adaptive convolution methods can achieve content-adaptive kernel and respond to each pixel, the specific generation of their kernel actually only utilizes the local information of the corresponding pixel, thus ignoring the long-range dependencies which are crucial in CNNs [51]. Inspired by the non-local approaches [51], we try to explore the non-local dependencies in pixel-specific generated kernel space with a different perspective from the previous feature representation. This hypothesis is also verified in Fig. 1 and Fig. 2 (b), it is clear that there still exists latent correlation among pixels in kernel space, motivating us to model this hidden relationship by the non-local technique.

B. Kernel-space Non-local Convolution

To capture non-local dependencies in the kernel-space and extract features adaptively in the whole receptive field for different spatial locations, we propose the kernel-space non-local convolution (KNLConv) module (see Fig. 3). Specifically, our KNLConv comprises a NLEC module and an APC module.

Non-local Expansion Convolution.

Since the spatial and spectral information extraction and utilization are exactly essential for the HSISR task, The features' spatial and channel information needs to be given extra attention. For this purpose, we carefully design the convolutional kernel generation module of the NLEC module to generate kernels from both spatial and channel aspects. The kernel generation module of NLEC is divided into two branches, the NSKG branch and the CWG branch. The former aims to generate a spatial kernel that captures the non-local relationship for each pixel on the neighborhood of the convolution kernel size, and the latter is designed to predict the channel weights for the kernel.

Specifically, in the NSKG branch, we execute a non-local operation on kernel-space at the low resolution. The input feature F is first transformed into a kernel feature ($\mathcal{K}^{\downarrow_s} \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times C_{NL}}$) with the same space size as the input LR-HSI, and reduced feature channels by a convolution and a downsample module. This operation can significantly reduce the computational cost in the following steps and improve the efficiency of the network. The experimental results show that the downsample operation does not noticeably degrade the performance. Formally, the non-local attention of spatial kernel feature is defined as:

$$\widetilde{\mathcal{K}}_{i,j} = \sum_{m,n} \frac{f(\theta(\mathcal{K}^{\downarrow_s}_{i,j}), \phi(\mathcal{K}^{\downarrow_s}_{m,n}))}{\lambda_{i,j}(\mathcal{K}^{\downarrow_s})} g(\mathcal{K}^{\downarrow_s}_{m,n}), \tag{1}$$

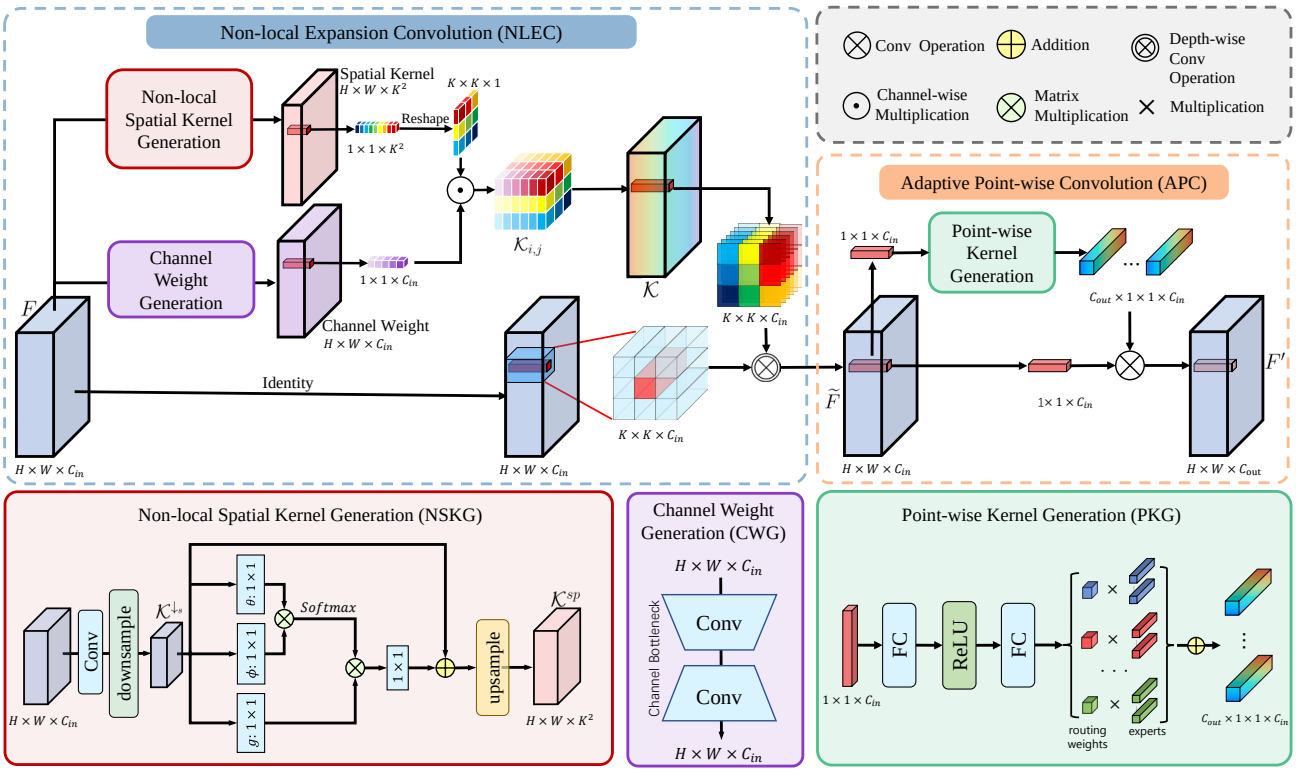


Fig. 3. The flowchart of the proposed Kernel-space Non-local Convolution (KNLConv).

where $\lambda_{i,j}(\mathcal{K}^{\downarrow_s}) = \sum_{m,n} f(\theta(\mathcal{K}^{\downarrow_s}_{i,j}), \phi(\mathcal{K}^{\downarrow_s}_{m,n}))$, (i,j) and (m,n) are the coordinates in $\mathcal{K}^{\downarrow_s}_{i,j}$, and $f(\cdot,\cdot)$ denotes the function to measure mutual similarity and is defined as $f(x,y) = \theta(x)^T\phi(y)$. $\theta(\cdot)$, $\phi(\cdot)$, and $g(\cdot)$ are three feature embeddings, $\widetilde{\mathcal{K}}$ is an output spatial kernel of equal size to $\mathcal{K}^{\downarrow_s}$. Finally, the used spatial kernel $\mathcal{K}^{sp} \in \mathbb{R}^{H\times W\times K^2}$ is obtained by an upsample module.

In the CWG branch, a simple convolution with channel bottlenecks is employed to learn the weights of the spatial kernels on each pixel at each channel. For every feature pixel, a single spatial kernel is copied to all channels and multiplied by the corresponding channel's weight to obtain the kernel for the corresponding pixel. This process can be formulated as follows:

$$\begin{aligned} \mathcal{K}^{sp} &= \delta(F), W^{ch} = \varphi(F), \\ \mathcal{K}_{i,j} &= \psi(\mathcal{K}^{sp}_{i,j}) \odot W^{ch}_{i,j}, \end{aligned} \tag{2}$$

where $F \in \mathbb{R}^{H\times W\times C_{in}}$ denotes the input feature, $\delta(\cdot)$ denotes the non-local spatial kernel generation branch, and $\varphi(\cdot)$ denotes the CWG branch. $\mathcal{K}^{sp} \in \mathbb{R}^{H\times W\times K^2}$ is the spatial kernel with $\mathcal{K}^{sp}_{i,j} \in \mathbb{R}^{K^2}$ representing the spatial kernel at the corresponding coordinate (i,j) (K is the spatial size of the kernel), $W^{ch} \in \mathbb{R}^{H\times W\times C}$ is the channel weights with $W^{ch}_{i,j} \in \mathbb{R}^{C}$ representing the weight at the corresponding coordinate (i,j). Symbol ψ is the reshape operation, and $\odot$ denotes channel-wise multiplication. $\mathcal{K} \in \mathbb{R}^{H\times W\times K\times K\times C}$ is the final kernel.

The input features $F \in \mathbb{R}^{H\times W\times C_{in}}$ are convolved with the kernel $\mathcal{K}$ by a depth-wise convolution operation to get the output $\widetilde{F} \in \mathbb{R}^{H\times W\times C_{in}}$, as follows,

$$\widetilde{F}_{i,j,k} = \sum_{u=-\rho}^{\rho}\sum_{v=-\rho}^{\rho} \mathcal{K}_{i,j,u,v,k} F_{i+u,j+v,k}, \tag{3}$$

where $\rho = \lfloor K/2 \rfloor$, representing the neighborhood maximum offset of the convolution at the central pixel.

Adaptive Point-wise Convolution (APC) Since depth-wise convolution in NLEC lacks the exploration of channel information, a point-wise convolution is needed to enhance the representation of channels. We generalize the dynamic convolution to the pixel level to propose the APC, as shown in Fig. 3. First, each pixel in NLEC's output features is considered a vector and then fed to two FC layers for feature extraction to predict the routing weights. After that, the routing weights are used to linearly combine a set of experts to produce the kernels for point-wise convolution. Finally, each feature pixel is convolved with its corresponding predicted kernel, resulting in the final output features $F' \in \mathbb{R}^{H\times W\times C_{out}}$. The APC can be formulated as:

$$\begin{aligned} \mathcal{K}^{pw}_{i,j} &= \sum\nolimits_k \pi_k(\widetilde{F}_{i,j})\widetilde{\mathcal{K}}^{pw}_k, \\ F'_{i,j} &= \mathcal{K}^{pw}_{i,j} \otimes \widetilde{F}_{i,j}, \end{aligned} \tag{4}$$

where, $\pi(\cdot)$ denotes the network that generates the routing weights, $\{\widetilde{\mathcal{K}}^{pw}_k\}$ represent a set of parallel point-wise convolutional kernels. $\mathcal{K}^{pw}_{i,j} \in \mathbb{R}^{H\times W\times C_{out}}$ is the predicted kernel at the (i,j), and $\otimes$ indicates a convolution operation.

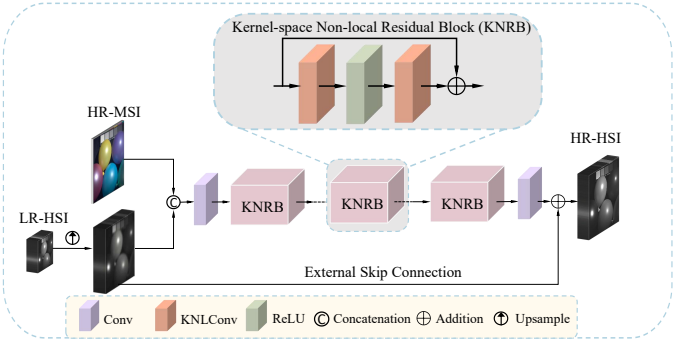


Fig. 4. Illustration of proposed Kernel-space Non-local Network (KNLNet). Please note that the size of all convolution kernels is uniformly set to 3×3. Besides, the number of channels in all convolutional layers is 64 except for the first layer where the input channel is $(S+s)$ and the last layer where the number of output channels is S.

C. Kernel-space Non-local Network (KNLNet)

KNLConv allows non-local correlation in the kernel-space to guide adaptive convolution to extract features, which can be effectively applied to fusion tasks. For the HSISR task, we present a kernel-space non-local network (KNLNet), and the architecture of the KNLNet is shown in Fig. 4. It mainly consists of several residual blocks, called kernel-space non-local residual blocks (KNRBs), which explore the latent kernel space prior with KNLConv to thoroughly extract features and exploit them.

The network takes LR-HSI $\mathcal{Y} \in \mathbb{R}^{h\times w\times S}$ and HR-MSI $\mathcal{Z} \in \mathbb{R}^{H\times W\times s}$ as inputs, where h, w are the spatial size of the LR-HSI and S is the number of spectral bands for the HSI, and H, W are the spatial size of the HR-MSI and s is the number of spectral bands for the MSI. We first upsample the LR-HSI to $\widehat{\mathcal{Y}} \in \mathbb{R}^{H\times W\times S}$ with the same spatial size as $\mathcal{Z}$ using bicubic interpolation. The HR-MSI is concatenated with the upsampled HSI to provide its potential spatial details. Then, a shallow feature of the concatenated image $F_0 \in \mathbb{R}^{H\times W\times C}$ is extracted via a convolution layer. This process can be expressed as:

$$F_0 = \mathcal{F}_{ex}(\{\widehat{\mathcal{Y}}, \mathcal{Z}\}), \tag{5}$$

where $\mathcal{F}_{ex}(\cdot)$ denotes a convolution layer. The shallow features F_0 are sent to the KNRB for deep feature extraction. The KNRB mainly consists of two proposed KNLConv layers, a $ReLU$ as the activation function, and a skip connection, *i.e.,* general ResBlock [21] with KNLConv replacing convolutional layers. The KNLConv serves to extract features in a comprehensive viewpoint using the kernel-space with a non-local prior. Several KNRBs compose the network backbone, and the features passing through the k-th KNRB are represented as:

$$F_k = \text{KNLConv}(\sigma(\text{KNLConv}(F_{k-1}))) + F_{k-1}, \tag{6}$$

where F_{k-1} represents the features of the last layer, and F_k is the output feature of k-th KNRB. KNLConv$(\cdot)$ denotes the operation of KNLConv, and symbols σ is the $ReLU$. In addition, after reconstructing the image using a convolution, we introduce an external skip connection that allows the rich low-frequency information to be directly bypassed. In this way, the network targets to recover more detailed information, and the network workload is significantly reduced. Thus, the HR-HSI reconstruction process can be expressed as:

$$\mathcal{X} = \mathcal{F}_{re}(F_n) + \widehat{\mathcal{Y}}, \tag{7}$$

where $\mathcal{X} \in \mathbb{R}^{H\times W\times S}$ denotes the output HR-HSI, and $\mathcal{F}_{re}(\cdot)$ is a convolution. F_n denotes the output feature of the last KNRB when the block's number is n.

IV. Experiments

In this section, we compare the performance of KNLNet with other state-of-the-art methods by performing comprehensive experiments on two widely used HSI datasets, *i.e.*, the CAVE [65] and Harvard [4] datasets, and the real-world pansharpening dataset of WorldView-3 (WV-3).

A. Network Training

Training Data. The CAVE dataset contains 32 HSIs of 512×512 spatial size with 31 spectral bands. We randomly selected 20 images for training the network and 11 images for testing. [1] According to Wald's protocol [68], we cropped 3920 patches of size $64 \times 64 \times 3$ from 20 images as HR-HSI ground truth and then used the modulation transfer function to downsample the HR-HSIs by $\times 4$. In addition, the spectral response functions of the commonly used Nikon D700 camera [10]–[12], [56] were used to generate RGB images (*i.e.*, HR-MSIs). Finally, 3920 training data pairs were obtained, including HR-MSIs of $64 \times 64 \times 3$, LR-HSIs of $16 \times 16 \times 31$, HR-HSIs of $64 \times 64 \times 31$, and $80\%/20\%$ of them were used as training/validation datasets, respectively.

Implementation Details. The overall network is trained by minimizing the following L_1 loss function:

$$\mathcal{L} = \frac{1}{N}\sum_{k=1}^{N} \left\| GT_k - \mathcal{F}_{KNLNet}(\mathcal{Y}_k, \mathcal{Z}_k) \right\|_1, \tag{8}$$

where N is the number of samples. We set the epoch to 1000, the learning rate to 1×10^{-4}, and the batch size to 32. The number of KNRBs defaults to 4. The network is implemented on PyTorch and runs on the GeForce GTX 3080.

B. Benchmark and Metrics

We compare the proposed KNLNet with several state-of-the-art methods for the HSISR task, which include traditional methods (*i.e.*, FUSE [53], CSTF [31], CNN-FUS [12]) and DL-based methods (*i.e.*, SSRNet [70], ResTFNet [34], CMHFnet [56], HSRnet [23], MoG-DCN [13]). Four quality indices to evaluate the performance are chosen, including PSNR, SAM [66], ERGAS [47], and SSIM [52]. Moreover, for the adaptive convolution methods, we selected DFN [25], DDF [73], and Involution [29] for the comparison.

[1] One image in CAVE dataset, i.e., the "Watercolor", is discarded because it is unavailable. [23]

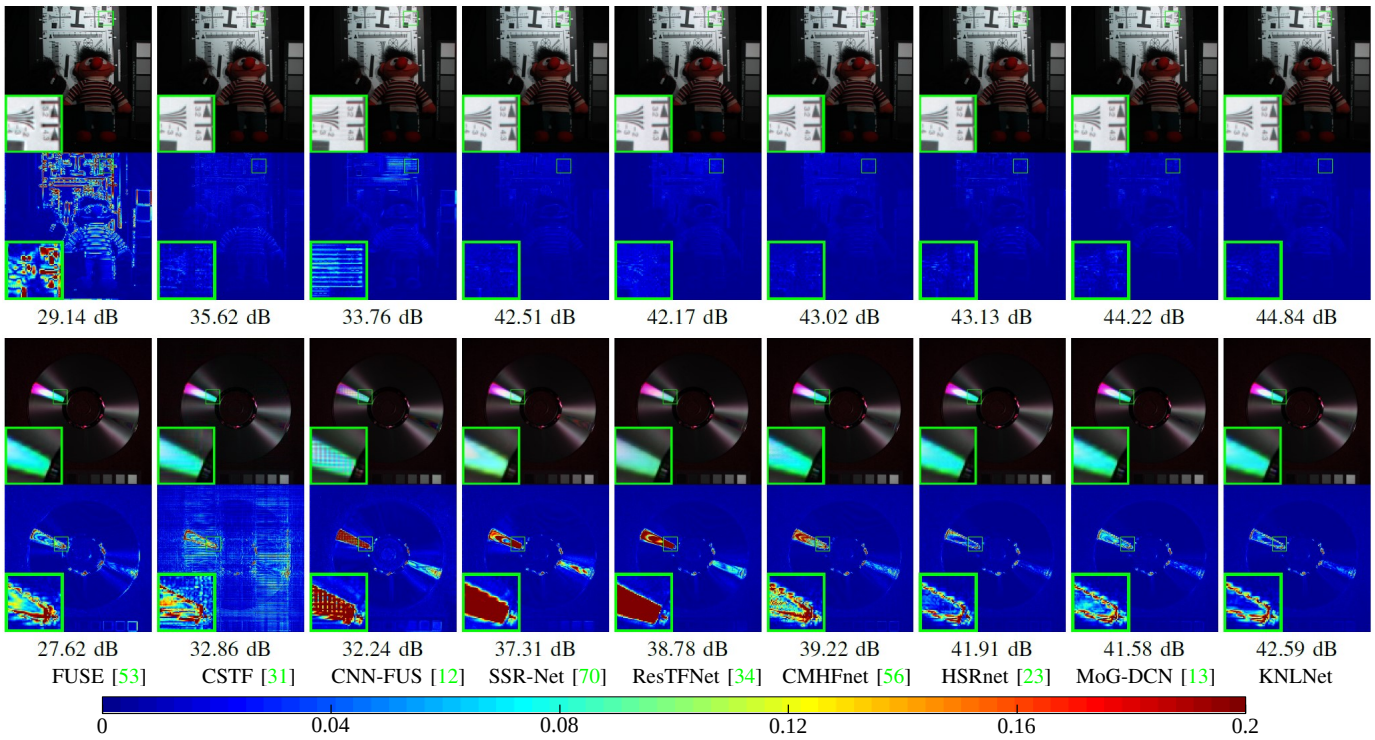


Fig. 5. The 1st and 3rd rows are the visual results for all the compared approaches at the *chart and stuffed toy* (R-24, G-14, B-13) and the *cd* (R-35, G-21, B-28) test cases from the CAVE dataset, respectively. The 2rd and 4th rows: the error maps for all the compared approaches. Furthermore, we show the corresponding PSNR metrics.

TABLE I
A QUANTITATIVE COMPARISON OF 11 CAVE AND 10 HARVARD SAMPLES. THE MEAN AND STANDARD DEVIATION OF THE RESULTS ARE SHOWN BEFORE AND AFTER THE ± SIGN, RESPECTIVELY. IN ADDITION, THE NUMBER OF PARAMETERS BASED ON DEEP LEARNING METHODS IS GIVEN. (**BOLD**: BEST; UNDERLINE: SECOND BEST)

Method	*CAVE dataset*				*Harvard dataset*				*Params*
	PSNR	*SAM*	*ERGAS*	*SSIM*	*PSNR*	*SAM*	*ERGAS*	*SSIM*	
FUSE [53]	39.72 ± 3.5	5.83 ± 2.0	4.18 ± 3.1	0.975 ± 0.02	42.06 ± 2.9	3.23 ± 0.9	3.14 ± 1.5	0.977 ± 0.01	-
CSTF [31]	42.14 ± 3.0	9.92 ± 4.1	3.08 ± 1.6	0.964 ± 0.03	42.97 ± 3.3	3.30 ± 1.3	2.43 ± 1.1	0.972 ± 0.02	-
CNN-FUS [12]	42.66 ± 3.5	6.44 ± 2.3	2.95 ± 2.2	0.982 ± 0.01	43.61 ± 4.7	3.32 ± 1.2	2.78 ± 1.6	0.978 ± 0.02	-
SSRNet [70]	45.28 ± 3.1	4.72 ± 1.8	2.06 ± 1.3	0.990 ± 0.00	44.40 ± 3.5	2.61 ± 0.7	2.39 ± 1.0	0.985 ± 0.01	**0.03M**
ResTFNet [34]	45.35 ± 3.7	3.76 ± 1.3	1.98 ± 1.6	0.993 ± 0.00	44.47 ± 4.0	2.56 ± 0.7	2.21 ± 0.9	0.985 ± 0.01	2.26M
CMHFnet [56]	46.32 ± 2.8	4.33 ± 1.5	1.74 ± 1.4	0.992 ± 0.01	43.10 ± 3.9	2.76 ± 0.8	3.28 ± 1.5	0.977 ± 0.01	3.63M
HSRnet [23]	47.82 ± 2.7	2.66 ± 0.9	1.34 ± 0.8	0.995 ± 0.00	45.01 ± 3.0	2.56 ± 0.7	2.11 ± 0.8	0.985 ± 0.01	1.90M
MoG-DCN [13]	48.30 ± 2.6	**2.62** ± 0.9	1.36 ± 0.8	0.995 ± 0.00	45.82 ± 3.5	**2.22** ± 0.6	1.99 ± 0.9	0.987 ± 0.01	47.97M
KNLNet	**48.45** ± 2.4	2.80 ± 0.9	**1.32** ± 0.8	**0.995** ± 0.00	**46.15** ± 3.5	2.23 ± 0.6	**1.98** ± 0.9	**0.986** ± 0.01	1.52M
Ideal value	$+\infty$	0	0	1	$+\infty$	0	0	1	0

C. Quantitative Results

The average PSNR, SAM, ERGAS, and SSIM results on CAVE [65] and Harvard [4] datasets are shown in Tab. I. It should be noted that the numbers after the ± signs in the tables represent the standard deviations of the results. We first assessed the performance of our proposed KNLNet on the CAVE dataset. Specifically, KNLNet achieved outstanding PSNR values, indicating a high degree of similarity between the reconstructed images and their originals. Additionally, the ERGAS metric showed that KNLNet minimizes the differences between the reconstructed images and the ground truth, underscoring its effectiveness in preserving spectral information. Moreover, KNLNet exhibited remarkable SSIM scores, signifying a strong structural similarity between its generated images and the reference images in the CAVE dataset. However, the proposed method showed a weakness in the SAM metric.

Turning our attention to the Harvard dataset, KNLNet continued to demonstrate its capabilities in remote sensing image processing. Specifically, our method showcased favorable performance in terms of PSNR, ERGAS, and SSIM, highlighting its ability to maintain both structural and spectral fidelity in reconstructed images. It is worth noting that when comparing KNLNet and MoG-DCN, KNLNet exhibited a slightly lower

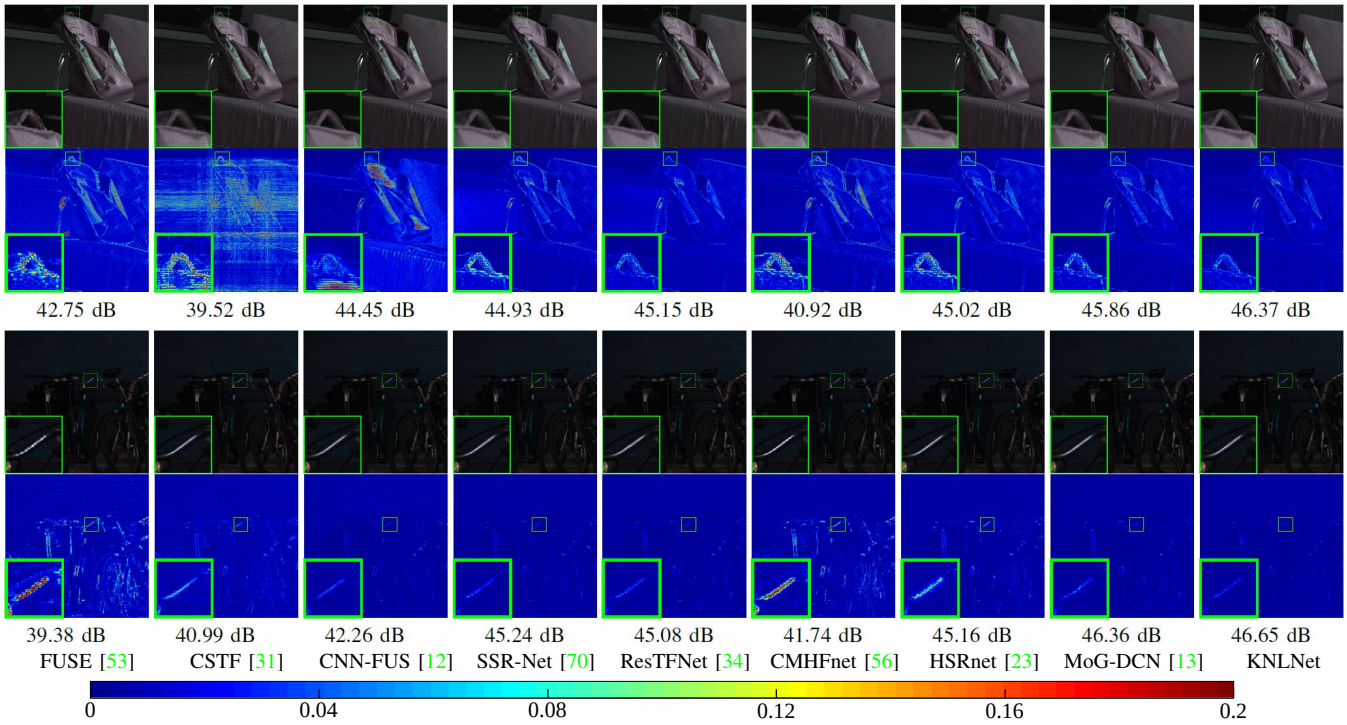
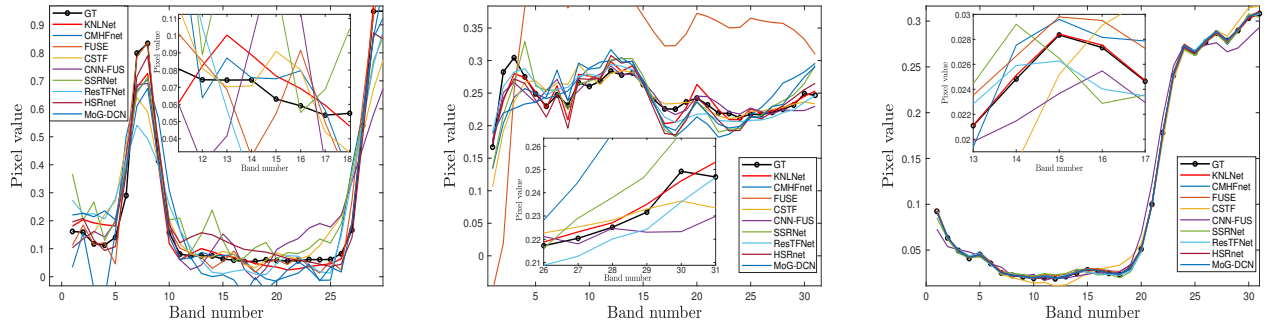


Fig. 6. The 1st and 3rd rows are the visual results for all the compared approaches at the *backpack* (R-31, G-29, B-30) and the *bikes* (R-15, G-22, B-30) test cases from the Harvard dataset, respectively. The 2rd and 4th rows: the error maps for all the compared approaches. Furthermore, we show the corresponding PSNR metrics.



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

Fig. 7. The comparison of spectral vectors by different methods. (a) Spectral vectors in *cd* located at $(289, 458)$; (b) Spectral vectors in *chart and stuffed toy* located at $(266, 507)$; (c) Spectral vectors in *fake and real tomatoes* located at $(456, 438)$.

SAM score. However, KNLNet's average performance metric surpassed that of all other methods except for MoG-DCN, emphasizing its competitiveness in the context of this dataset.

A remarkable aspect of KNLNet is its efficient model design, with the number of parameters being only 3.2% of MoG-DCN. This efficiency is particularly noteworthy when considering its competitive performance on the ×4 HSISR task. These findings underscore KNLNet's potential for applications in HSISR, where balancing model complexity and performance is crucial.

D. Visual Results

We have demonstrated visual comparisons using the CAVE dataset [65] and the Harvard dataset [4] in Fig.5 and Fig.6, respectively. From the visualizations, it is evident that the results obtained by our method are significantly superior to those of other methods. While other methods exhibit discontinuities at the boundaries between objects, our method effectively captures both high-frequency and smooth parts of the image. This comparison aligns with our initial expectations that KNLConv is content-adaptive. Specifically, upon closer examination of the results in Fig. 5, it becomes clear that on the CAVE dataset, other methods introduce noticeable artifacts and discontinuities at object boundaries, whereas our method excels in preserving image continuity, resulting in smoother transitions between objects and the background. Likewise, the comparison using the Harvard dataset, as shown in Fig. 6, further validates the superiority of our method. Other methods may suffer from loss of image details or the

presence of artifacts, while our method effectively captures high-frequency information in the image while maintaining overall smoothness, resulting in more natural and realistic outcomes.

In addition, it's important not to overlook the significance of spectral fidelity in the context of HSISR tasks. Demonstrating the capability of various methods to preserve spectral information is crucial. To illustrate this, we have included spectral vector comparisons for three specific examples from the CAVE dataset in Fig. 7, offering close-up shots for a more detailed view. The spectral vectors, which represent the distribution of pixel values across different wavelengths, provide valuable insights into how well each method maintains spectral information. Upon careful examination, it becomes evident that the spectral vectors of our proposed KNLNet closely align with the ground truth. This alignment serves as a strong indicator of our technique's superior spectral preservation ability. In other words, our method excels in faithfully reproducing the original spectral characteristics of the images, ensuring that the fine details and nuances in the spectral domain are accurately retained. This is a crucial aspect, especially in applications where the preservation of spectral fidelity is of paramount importance, such as remote sensing, medical imaging, and various scientific domains.

These visual comparisons not only provide qualitative insights but also underscore the practical effectiveness of our method in handling a variety of image content. This reinforces our belief that KNLConv can adapt to content and deliver outstanding image enhancement across diverse scenarios.

E. Experiment with the Scale Factor of 8

In the manuscript, our HSISR experiments are conducted on the scale factor of 4. To further verify the advantages of our approach at different scales, we perform a comparison of HSISR with a larger scale factor of 8. Especially, the data simulation is similar to that conducted on the experiment of $4\times$. By the simulation, we finally obtain 2172 and 242 pairs of LR-HSI $(10 \times 10 \times 31)$, HR-MSI $(80 \times 80 \times 3)$, and HR-HSI $(80 \times 80 \times 31)$ as training and validation datasets from CAVE dataset, respectively, in the meanwhile keep 11 images as testing examples. For a fair comparison, all networks are trained and tested using $8\times$ data.

Table II presents a comprehensive overview of the comparative performance of eight competitive approaches on the CAVE dataset, specifically with a scale factor of 8. Upon reviewing the table, it becomes evident that our KNLNet continues to exhibit superior performance when compared to the other techniques. This observation underscores the robustness and effectiveness of KNLNet in the context of high-resolution image enhancement, particularly at an 8x scale factor, with all four metrics - PSNR, ERGAS, SAM, and SSIM - leading the field.

F. Ablation Study and Discussions

Effectiveness of KNLConv. We first explore the validity of the proposed KNLConv. Taking the standard convolution kernel as the baseline, we also compared KNLConv with three other similar pixel-level adaptive convolution kernel generation methods, *i.e.*, DFN, DDF, and Involution. We replace the KNLConv in the KNLNet with these convolution operations, and the results on the CAVE dataset are presented in Tab. III. Note that we also compare the number of parameters, floating point operations (FLOPs) and run time. It can be seen that the proposed KNLNet surpasses the current state-of-the-art convolutional structures in HSISR performance, and achieves satisfactory results in hardware consumption. The consumption of parameters and computations is only second to Involution, reaching the suboptimal level. This further proves the effectiveness of KNLNet.

TABLE II
A QUANTITATIVE COMPARISONS ON 11 CAVE EXAMPLES WITH A SCALE FACTOR OF 8. THE MEAN AND STANDARD DEVIATION OF THE RESULTS ARE SHOWN BEFORE AND AFTER THE $\pm$ SIGN, RESPECTIVELY. (**BOLD**: BEST; UNDERLINE: SECOND BEST)

Method	PSNR	SAM	ERGAS	SSIM
FUSE [53]	36.18 $\pm$ 3.4	9.87 $\pm$ 4.8	1.57 $\pm$ 1.0	0.927 $\pm$ 0.07
CSTF [31]	39.13 $\pm$ 3.3	15.61 $\pm$ 6.2	1.19 $\pm$ 0.6	0.946 $\pm$ 0.03
CNN-FUS [12]	38.20 $\pm$ 3.5	9.57 $\pm$ 2.8	2.32 $\pm$ 1.5	0.955 $\pm$ 0.02
SSRNet [70]	43.79 $\pm$ 3.5	5.09 $\pm$ 1.7	1.23 $\pm$ 0.8	0.988 $\pm$ 0.01
ResTFNet [34]	43.21 $\pm$ 4.0	4.69 $\pm$ 1.5	1.32 $\pm$ 1.0	0.990 $\pm$ 0.00
CMHFnet [56]	45.00 $\pm$ 3.1	4.88 $\pm$ 1.9	0.99 $\pm$ 0.7	0.990 $\pm$ 0.00
HSRnet [23]	44.97 $\pm$ 3.4	3.33 $\pm$ 1.0	0.94 $\pm$ 1.1	0.992 $\pm$ 0.00
MoG-DCN [13]	<u>46.08</u> $\pm$ 3.6	<u>3.33</u> $\pm$ 0.9	<u>0.90</u> $\pm$ 0.6	<u>0.993</u> $\pm$ 0.00
KNLNet	**46.48** $\pm$ 3.4	**3.29** $\pm$ 1.0	**0.86** $\pm$ 0.6	**0.994** $\pm$ 0.00
Ideal value	$+\infty$	0	0	1

Discussion of Structural Components. To better understand the role of the proposed modules in KNLConv, we compare the method without these modules on the CAVE dataset. The model without (w/o) non-local just builds by removing the non-local attention operation. When without NSKG, the channel weights of CWG perform depth-wise convolution as a 1×1 kernel. When without CWG, the spatial kernels in NSKG are replicated to each channel. Tab. IV shows the results, includes HSISR performance, number of parameters, FLOPs, and runtime. Compared to without the non-local, the proposed method improves significantly by 0.7 dB, which shows the validity of the proposed exploring Non-local Dependencies in the kernel-space. In addition, the modules used in the method significantly enhance the network performance compared without these modules, which verifies the validity of these designed modules.

Considering reducing the consumption of computing resources, we adopt the operation of downsampling for the input feature map in NSKG. To see how this affects the results, we removed the downsampling and tested it on the CAVE dataset in Tab. V. Clearly, the method of non-local operation without downsampling gets a slight improvement but with a substantial computational consumption.

Number of KNRB. We also investigated how the number of KNRB affects the experimental results. As seen from Tab. VI, with the increase of KNRB, the metrics on the test data set gradually improved. This ablation study demonstrates the effectiveness of KNRB, including KNLConv.

Feature-space v.s. Kernel-space. To confirm our idea of

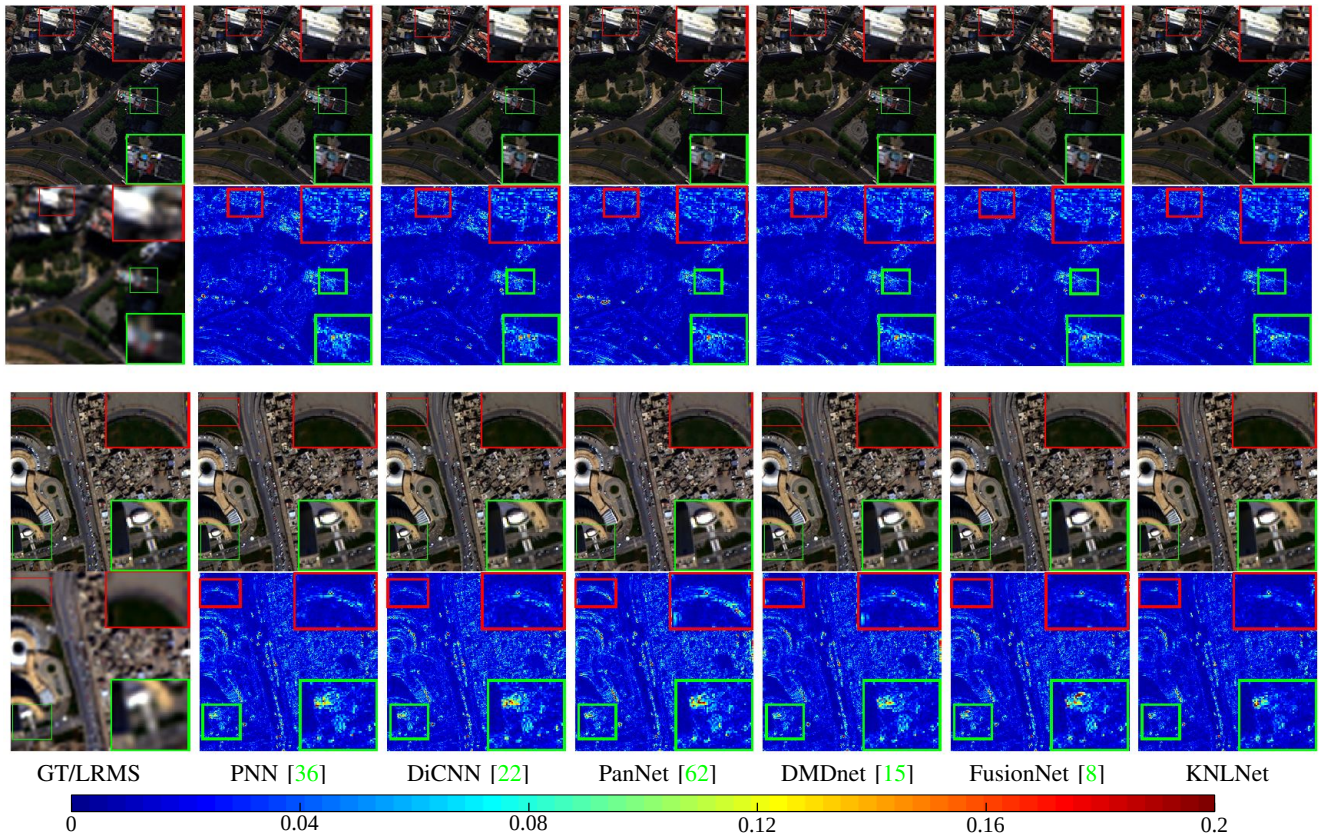


Fig. 8. Visual comparison of two simulated cases for pansharpening task on the WV-3 dataset, with two regions enlarged for detailed observation.

TABLE III
COMPARISON OF KNLCONV WITH OTHER ADAPTIVE CONVOLUTIONS ON THE CAVE DATASET, INCLUDES HSISR PERFORMANCE, NUMBER OF PARAMETERS, FLOPS, AND RUNTIME.

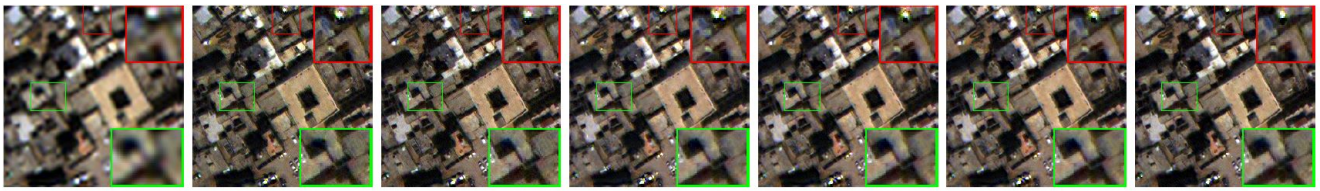
Method	PSNR	SAM	ERGAS	SSIM	Parameters	FLOPs (G)	Time (ms)
Standard Conv	46.87	3.45	1.53	0.991	2.97M	28.1	**72**
DFN [25]	46.90	3.86	1.49	0.992	5.35M	39.3	324
DDF [73]	47.62	3.14	1.45	1.02M	2.76M	27.7	136
Involution [29]	47.56	3.34	1.45	0.994	**1.25M**	**20.3**	106
KNLNet	**48.45**	**2.80**	**1.32**	**0.995**	1.52M	23.6	171
Ideal value	$+\infty$	0	0	1	0	0	0

exploring the kernel space, we also compared it with the traditional non-local representation of feature space. When implementing this comparison, KNLConv is replaced by the common non-local operation. The qualitative results are shown in Tab. VII. Clearly, the results give us a positive response that exploration in kernel space is superior to that in feature space. This also proves that there exists information that is difficult to be perceived in feature space, and kernel space is worth to be concerned.

G. Application Extension.

As a substitute for standard convolution, the application scope of the proposed KNLConv is beyond HSISR. It can be applied to other low-level visual tasks, including image de-raining, image de-hazing, *etc.*, in which accuracy is down to the pixel level while required to maintain the continuity of the overall image. In particular, we adopted the same network architecture for another task similar to HSISR, *i.e.*, the widely studied remote sensing pansharpening [8], [62], [78]. Previous pansharpening experiments are mainly performed on WorldView-3 (WV-3) dataset, which can be easily downloaded from a public website[2]. The WV-3 dataset could provide a high-resolution PAN image (0.3m) and a low resolution multispectral (MS) image (1.2m) with eight bands and 11 bits of radiometric resolution. After downloading the dataset, we simulate 12580 PAN/MS/GT image pairs of sizes 64×64, $16 \times 16 \times 8$, and $64 \times 64 \times 8$, respectively, then divide them into $70\%/20\%/10\%$ for training/validation/testing. Note that, similar with the simulation of HSISR experiments, we follow Wald's protocol [68] to obtain these image pairs. The specific

[2]https://www.maxar.com/



LRMS PNN [36] DiCNN1 [22] PanNet [62] DMDnet [15] FusionNet [8] KNLNet

Fig. 9. Visual comparisons in the RGB visualization of six approaches on a WV-3 full-resolution example.

TABLE IV
THE EFFECT OF THE MODULES IN KNLCONV, INCLUDES HSISR PERFORMANCE, NUMBER OF PARAMETERS, FLOPS, AND RUNTIME.

Method	PSNR	SAM	ERGAS	SSIM	Parameters	FLOPs (G)	Time (ms)
w/o Non-local	47.73	3.18	1.44	0.994	1.30M	16.3	119
w/o NSKG	47.22	3.61	1.49	0.992	1.24M	15.7	113
w/o CWG	47.98	3.07	1.38	0.994	1.43M	22.4	165
w/o NLEC	46.87	3.82	1.55	0.992	**0.46M**	**13.9**	**101**
w/o APC	48.26	2.99	1.33	**0.995**	1.08M	20.1	162
KNLNet	**48.45**	**2.80**	**1.32**	**0.995**	1.52M	23.6	171
Ideal value	$+\infty$	0	0	1	0	0	0

TABLE V
ABLATION STUDY ABOUT THE DOWNSAMPLING ON THE CAVE DATASET.

Scale	PSNR	SAM	ERGAS	SSIM	Time (ms)	FLOPs
$\downarrow 4\times$	48.45	2.80	1.32	**0.995**	**171**	**11.2G**
w/o	**48.52**	**2.75**	**1.31**	**0.995**	736	17.5G
Ideal value	$+\infty$	0	0	1	0	0

TABLE VI
ABLATION STUDY ABOUT USING DIFFERENT NUMBERS OF KNRBS.

Block number	PSNR	SAM	ERGAS	SSIM	Time (ms)
1	46.52	3.68	1.57	0.992	**92**
2	47.72	3.32	1.39	0.994	127
3	48.35	2.86	1.34	0.994	139
4 (default)	**48.45**	**2.80**	**1.32**	**0.995**	171
Ideal value	$+\infty$	0	0	1	0

TABLE VII
COMPARISON WITH FEATURES SPACE NON-LOCAL OPERATIONS ON THE CAVE DATASET.

Method	PSNR	SAM	ERGAS	SSIM	Time (ms)	FLOPs
Feature-space	47.72	3.23	1.42	0.994	**165**	**10.6G**
Kernel-space	**48.45**	**2.80**	**1.32**	**0.995**	171	11.2G
Ideal value	$+\infty$	0	0	1	0	0

steps are: 1) downsampling original PAN and MS images with a scale factor of 4 after conducting a modulation transfer function (MTF)-based filter to the original PAN and MS images; 2) Taking the downsampled PAN and MS images as the inputs of network training, and the original MS image as the ground-truth.

We compare the proposed KNLNet with several state-of-the-art (SOTA) CNN-based methods (*i.e.*, PNN [36], DiCNN1 [22], PanNet [62], DMDNet [15], and FusionNet [8]). For simulation (*i.e.*, reduced-resolution) experiment, the performance assessment is conducted using the spectral angle mapper (SAM) [66], the relative dimensionless global error in synthesis (ERGAS) [47], the universal image quality index for the eight-band image (Q8) [17], and the spatial correlation coefficient (SCC) [72]. The ideal values of Q8 and SCC are 1, and the ideal values of SAM and ERGAS are 0. Furthermore, for real (*i.e.*, full-resolution) experiment, we employ the quality without reference (QNR), D_λ, and D_s indexes [46]. The ideal value of QNR is 1, and the ideal values of D_λ, D_s are 0.

Tab. Tab. VIII displays the performance comparison results between our Pansharpening method and other methods. We employed a range of standard metrics, including SAM, ERGAS, Q8, and SCC, to assess the quantitative differences in image quality. It is evident from the table that our method exhibits a significant advantage across all metrics. When compared to other methods, our approach consistently outperforms them with lower SAM and ERGAS, higher Q8 and SCC values, indicating superior image quality. In addition to the tabulated results, Fig. 8 illustrates the visual comparison of our Pansharpening method against other methods. The chart visually showcases the difference in image quality, highlighting that our method consistently produces images with fewer artifacts and improved clarity. The visual comparisons further reinforce the effectiveness of our approach in improving pansharpening performance and overall visual quality.

To demonstrate the promising pansharpening ability of KNLNet on practical examples, we conduct experiments on real (*i.e.*, full-resolution) dataset. To fairly evaluate the performance of all compared approaches, we employ 50 full-resolution examples (including MS and PAN images) obtained

TABLE VIII
AVERAGE QUANTITATIVE METRICS WITH RELATED STANDARD DEVIATIONS ON 1258 TESTING IMAGES ON THE WV-3 DATASET. (**BOLD**: BEST; UNDERLINE: SECOND BEST)

Method	SAM	ERGAS	Q8	SCC
PNN [36]	4.00 ± 1.3	2.73 ± 1.0	0.908 ± 0.11	0.922 ± 0.05
DiCNN [22]	3.98 ± 1.3	2.74 ± 1.0	0.910 ± 0.11	0.952 ± 0.05
PanNet [62]	4.09 ± 1.3	2.95 ± 1.0	0.894 ± 0.12	0.949 ± 0.05
DMDnet [15]	3.97 ± 1.2	2.95 ± 1.0	0.894 ± 0.12	0.953 ± 0.04
FusionNet [8]	3.74 ± 1.2	2.57 ± 0.9	0.914 ± 0.11	0.958 ± 0.05
KNLNet	**3.36** ± 1.2	**2.33** ± 0.9	**0.926** ± 0.11	**0.965** ± 0.04
Ideal value	0	0	1	1

TABLE IX
AVERAGE QUANTITATIVE METRICS AND THE RELATED STANDARD DEVIATIONS ON 50 WV-3 FULL-RESOLUTION EXAMPLES. (**BOLD**: BEST; UNDERLINE: SECOND BEST)

Method	QNR	D_λ	D_s
PNN [36]	0.942±0.03	0.029±0.02	0.030±0.02
DiCNN1 [22]	0.926±0.04	0.032±0.02	0.048±0.03
PanNet [62]	0.942±0.03	0.029±0.01	0.025±0.02
DMDnet [15]	0.935±0.03	0.027±0.01	0.037±0.02
FusionNet [8]	0.944±0.04	0.024±0.02	0.029±0.03
KNLNet	**0.952**±0.03	**0.021**±0.02	**0.024**±0.01
Ideal value	1	0	0

from WV-3 sensor to implement the experiments.

In Table IX, we provide a comprehensive overview of the quantitative results obtained across three common metrics, namely QNR, D_λ, and D_s. A closer examination of the table reveals that our method consistently ranks first across all of these metrics, solidly establishing its efficacy when applied to real images. Furthermore, the visual results, as depicted in Figure 9, showcase the performance of all approaches on a full-resolution WV-3 example. It is worth noting that our method exhibits superior performance in the visual results as well, reaffirming its capability to produce outstanding results in real-world scenarios.

V. CONCLUSION

This paper introduces a novel convolution operation based on pixel-level adaptive convolution kernel and non-local technique, named KNLConv, including two parts,*i.e.*, NLEC and APC. NLEC is able to explore the long-range dependencies in the generated kernel space to enhance the perception of the convolution. Besides, APC that is capable of exploiting dynamic convolution at the pixel level is introduced to boost the representation of channel information. Finally, we propose a simple residual structure network utilizing KNLConv, called KNLNet, to solve the HSISR task. Benefiting from this novel feature representation approach, our network on several benchmark datasets of HSISR and pansharpening outperforms recent competitive methods. In addition, more ablation studies are implemented to prove that exploring kernel space is worthwhile and effective.

REFERENCES

[1] Cao, X., Fu, X., Hong, D., Xu, Z., Meng, D.: Pancsc-net: A model-driven deep unfolding method for pansharpening. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2021)

[2] Cao, X., Fu, X., Xu, C., Meng, D.: Deep spatial-spectral global reasoning network for hyperspectral image denoising. IEEE Trans. Geosci. Remote Sens. **60**, 1–14 (2022)

[3] Cao, X., Yao, J., Xu, Z., Meng, D.: Hyperspectral image classification with convolutional neural network and active learning. IEEE Trans. Geosci. Remote Sens. **58**(7), 4604–4616 (2020)

[4] Chakrabarti, A., Zickler, T.: Statistics of real-world hyperspectral images. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 193–200 (2011)

[5] Chen, L., Liu, J., Chen, W., Du, B.: A glrt-based multi-pixel target detector in hyperspectral imagery. IEEE Transactions on Multimedia **25**, 2710–2722 (2023). https://doi.org/10.1109/TMM.2022.3150185

[6] Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic convolution: Attention over convolution kernels. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 11030–11039 (2020)

[7] Chen-Yu Zhao, TianJing Zhang, R.R.Z.X.C.L.J.D.: Lgpconv: Learnable gaussian perturbation convolution for lightweight pansharpening. International Joint Conference on Artificial Intelligence (IJCAI) (2023)

[8] Deng, L.J., Vivone, G., Jin, C., Chanussot, J.: Detail injection-based deep convolutional neural networks for pansharpening. IEEE Trans. Geosci. Remote Sens. **59**(8), 6995–7010 (2021)

[9] Dian, R., Guo, A., Li, S.: Zero-shot hyperspectral sharpening. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12650–12666 (2023). https://doi.org/10.1109/TPAMI.2023.3279050

[10] Dian, R., Li, S.: Hyperspectral image super-resolution via subspace-bascomputed low tensor multi-rank regularization. IEEE Trans. Image Process. **28**(10), 5135–5146 (2019)

[11] Dian, R., Li, S., Fang, L.: Learning a low tensor-train rank representation for hyperspectral image super-resolution. IEEE Trans. Neural Net. Learn. Syst. **30**(9), 2672–2683 (2019)

[12] Dian, R., Li, S., Kang, X.: Regularizing hyperspectral and multispectral image fusion by cnn denoiser. IEEE Trans. Neural Net. Learn. Syst. **32**(3), 1124–1135 (2021)

[13] Dong, W., Zhou, C., Wu, F., Wu, J., Shi, G., Li, X.: Model-guided deep hyperspectral image super-resolution. IEEE Transactions on Image Processing **30**, 5754–5768 (2021)

[14] Du, B., Zhang, M., Zhang, L., Hu, R., Tao, D.: Pltd: Patch-based low-rank tensor decomposition for hyperspectral images. IEEE Transactions on Multimedia **19**(1), 67–79 (2017). https://doi.org/10.1109/TMM.2016.2608780

[15] Fu, X., Wang, W., Huang, Y., Ding, X., Paisley, J.: Deep multiscale detail networks for multiband spectral image sharpening. IEEE Trans. Neural Net. Learn. Syst. **32**(5), 2090–2104 (2020)

[16] Fu, X., Xiao, J., Zhu, Y., Liu, A., Wu, F., Zha, Z.J.: Continual image deraining with hypergraph convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(8), 9534–9551 (2023). https://doi.org/10.1109/TPAMI.2023.3241756

[17] Garzelli, A., Nencini, F.: Hypercomplex quality assessment of multi/hyperspectral images. IEEE Geosci. Remote Sens. Letters **6**(4), 662–665 (2009)

[18] Grohnfeldt, C., Zhu, X.X., Bamler, R.: Jointly sparse fusion of hyperspectral and multispectral imagery. In: IEEE Int. Geosci. and Remote Sens. Symp. (IGARSS). pp. 4090–4093 (2013)

[19] Guo, A., Dian, R., Li, S.: A deep framework for hyperspectral image fusion between different satellites. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 7939–7954 (2023). https://doi.org/10.1109/TPAMI.2022.3229433

[20] Guo, P., Zhuang, P., Guo, Y.: Bayesian pan-sharpening with multiorder gradient-based deep network constraints. IEEE Jour. Selec. Topics Applied Earth Obser. & Remote Sens. **13**, 950–962 (2020)

[21] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 770–778 (2016)

[22] He, L., Rao, Y., Li, J., Chanussot, J., Plaza, A., Zhu, J., Li, B.: Pansharpening via detail injection based convolutional neural networks. IEEE Jour. Selec. Topics Applied Earth Obser. & Remote Sens. **12**(4), 1188–1204 (2019)

[23] Hu, J.F., Huang, T.Z., Deng, L.J., Jiang, T.X., Vivone, G., Chanussot, J.: Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks. IEEE Trans. Neural Net. Learn. Syst. pp. 1–15 (2021). https://doi.org/10.1109/TNNLS.2021.3084682

[24] Hu, J., Jia, X., Li, Y., He, G., Zhao, M.: Hyperspectral image super-resolution via intrafusion network. IEEE Trans. Geosci. Remote Sens. **58**(10), 7459–7471 (2020)

[25] Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. Adv. Neural Inform. Process. Syst. (NIPS) **29**, 667–675 (2016)

[26] Jin, C., Deng, L.J., Huang, T.Z., Vivone, G.: Laplacian pyramid networks: A new approach for multispectral pansharpening. Info. Fusion **78**, 158–170 (2022)

[27] Junming Hou, Qi Cao, R.R.C.L.J.L.L.J.D.: Bidomain modeling paradigm for pansharpening. ACM International Conference on Multimedia (ACMMM) (2023)

[28] Lanaras, C., Baltsavias, E., Schindler, K.: Hyperspectral super-resolution by coupled spectral unmixing. In: Int. Conf. Comput. Vis. (ICCV). pp. 3586–3594 (2015)

[29] Li, D., Hu, J., Wang, C., Li, X., She, Q., Zhu, L., Zhang, T., Chen, Q.: Involution: Inverting the inherence of convolution for visual recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 12321–12330 (2021)

[30] Li, J., Zheng, K., Yao, J., Gao, L., Hong, D.: Deep unsupervised blind hyperspectral and multispectral data fusion. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022). https://doi.org/10.1109/LGRS.2022.3151779

[31] Li, S., Dian, R., Fang, L., Bioucas-Dias, J.M.: Fusing hyperspectral and multispectral images via coupled sparse tensor factorization. IEEE Trans. Image Process. **27**(8), 4118–4130 (2018)

[32] Li, Y., Xu, Q., He, Z., Li, W.: Progressive task-based universal network for raw infrared remote sensing imagery ship detection. IEEE Transactions on Geoscience and Remote Sensing (2023)

[33] Lin, X., Zhou, Y., Zhang, X., Liu, Y., Zhu, C.: Illumination-insensitive binary descriptor for visual measurement based on local inter-patch invariance. IEEE Transactions on Instrumentation and Measurement (2023)

[34] Liu, X., Liu, Q., Wang, Y.: Remote sensing image fusion based on two-stream fusion network. Info. Fusion **55**, 1–15 (2020)

[35] Ma, N., Zhang, X., Huang, J., Sun, J.: Weightnet: Revisiting the design space of weight networks. In: Eur. Conf. Comput. Vis. (ECCV). pp. 776–792 (2020)

[36] Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G.: Pansharpening by convolutional neural networks. Remote Sens. **8**(7), 594 (2016)

[37] Meng, X., Shen, H., Yuan, Q., Li, H., Zhang, L., Sun, W.: Pansharpening for cloud-contaminated very high-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **57**(5), 2840–2854 (2018)

[38] Meng, X., Wang, N., Shao, F., Li, S.: Vision transformer for pansharpening. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–11 (2022)

[39] Palsson, F., Sveinsson, J.R., Ulfarsson, M.O.: Multispectral and hyperspectral image fusion using a 3-D-convolutional neural network. IEEE Geosci. and Remote Sens. Let. **14**(5), 639–643 (2017)

[40] Ran, R., Deng, L.J., Jiang, T.X., Hu, J.F., Chanussot, J., Vivone, G.: Guidednet: A general cnn fusion framework via high-resolution guidance for hyperspectral image super-resolution. IEEE Transactions on Cybernetics pp. 1–14 (2023). https://doi.org/10.1109/TCYB.2023.3238200

[41] Shang-Qi Deng, Liang-Jian Deng, X.W.R.R.D.H.G.V.: Psrt: Pyramid shuffle-and-reshuffle transformer for multispectral and hyperspectral image fusion. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–15 (2023). https://doi.org/10.1109/TGRS.2023.3244750

[42] Shang-Qi Deng, Liang-Jian Deng, X.W.R.R.R.W.: Bidirectional dilation transformer for multispectral and hyperspectral image fusion. International Joint Conference on Artificial Intelligence (IJCAI) (2023)

[43] Shi, C., Pun, C.M.: Multiscale superpixel-based hyperspectral image classification using recurrent neural networks with stacked autoencoders. IEEE Transactions on Multimedia **22**(2), 487–501 (2020). https://doi.org/10.1109/TMM.2019.2928491

[44] Su, H., Jampani, V., Sun, D., Gallo, O., Learned-Miller, E., Kautz, J.: Pixel-adaptive convolutional neural networks. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 11166–11175 (2019)

[45] Su, Z., Fang, L., Kang, W., Hu, D., Pietikäinen, M., Liu, L.: Dynamic group convolution for accelerating convolutional neural networks. In: Eur. Conf. Comput. Vis. (ECCV). pp. 138–155 (2020)

[46] Vivone, G., Alparone, L., Chanussot, J., Dalla Mura, M., Garzelli, A., Licciardi, G.A., Restaino, R., Wald, L.: A critical comparison among pansharpening algorithms. IEEE Trans. Geosci. Remote Sens. **53**(5), 2565–2586 (2015)

[47] Wald, L.: Data Fusion. Definitions and Architectures - Fusion of Images of Different Spatial Resolutions. Presses des MINES (2002)

[48] Wan, W., Guo, W., Huang, H., Liu, J.: Nonnegative and nonlocal sparse tensor factorization-based hyperspectral image super-resolution. IEEE Trans. Geosci. Remote Sens. **58**(12), 8384–8394 (2020)

[49] Wang, B., Niu, H., Zeng, J., Bai, G., Lin, S., Wang, Y.: Latent representation learning model for multi-band images fusion via low-rank and sparse embedding. IEEE Transactions on Multimedia **23**, 3137–3152 (2021). https://doi.org/10.1109/TMM.2020.3020695

[50] Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C.C., Lin, D.: Carafe: Content-aware reassembly of features. In: Int. Conf. Comput. Vis. (ICCV). pp. 3007–3016 (2019)

[51] Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 7794–7803 (2018)

[52] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)

[53] Wei, Q., Bioucas-Dias, J., Dobigeon, N., Tourneret, J.Y., Godsill, S.: Blind model-based fusion of multi-band and panchromatic images. In: IEEE Int. Conf. on Multisensor Fusion Integ. Intel. Syst. (MFI). pp. 21–25 (2016)

[54] Wen, R., Deng, L.J., Wu, Z.C., Wu, X., Vivone, G.: A novel spatial fidelity with learnable nonlinear mapping for panchromatic sharpening. IEEE Transactions on Geoscience and Remote Sensing . https://doi.org/10.1109/TGRS.2023.3265404

[55] Wu, X., Huang, T.Z., Deng, L.J., Zhang, T.J.: Dynamic cross feature fusion for remote sensing pansharpening. In: Int. Conf. Comput. Vis. (ICCV). pp. 14687–14696 (October 2021)

[56] Xie, Q., Zhou, M., Zhao, Q., Xu, Z., Meng, D.: Mhf-net: An interpretable deep network for multispectral and hyperspectral image fusion. IEEE Trans. Pattern Anal. Mach. Intell. **44**(3), 1457–1473 (2022)

[57] Xu, Q., Li, Y., Nie, J., Liu, Q., Guo, M.: Upangan: Unsupervised pansharpening based on the spectral and spatial loss constrained generative adversarial network. Information Fusion **91**, 31–46 (2023)

[58] Xu, Q., Li, Y., Zhang, M., Li, W.: Coco-net: A dual-supervised network with unified roi-loss for low-resolution ship detection from optical satellite image sequences. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2022)

[59] Xu, T., Huang, T.Z., Deng, L.J., Zhao, X.L., Huang, J.: Hyperspectral image superresolution using unidirectional total variation with tucker decomposition. IEEE Jour. Selec. Topics Applied Earth Obser. & Remote Sens. **13**, 4381–4398 (2020)

[60] Yan, K., Zhou, M., Liu, L., Xie, C., Hong, D.: When pansharpening meets graph convolution network and knowledge distillation. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2022). https://doi.org/10.1109/TGRS.2022.3168192

[61] Yang, B., Bender, G., Le, Q.V., Ngiam, J.: Condconv: Conditionally parameterized convolutions for efficient inference. In: Adv. Neural Inform. Process. Syst. (NIPS). vol. 32 (2019)

[62] Yang, J., Fu, X., Hu, Y., Huang, Y., Ding, X., Paisley, J.: Pannet: A deep network architecture for pan-sharpening. In: Int. Conf. Comput. Vis. (ICCV). pp. 5449–5457 (2017)

[63] Yang, Y., Lu, H., Huang, S., Tu, W.: Pansharpening based on joint-guided detail extraction. Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing **14**, 389–401 (2020)

[64] Yang, Y., Wu, L., Huang, S., Wan, W., Tu, W., Lu, H.: Multiband remote sensing image pansharpening based on dual-injection model. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **13**, 1888–1904 (2020)

[65] Yasuma, F., Mitsunaga, T., Iso, D., Nayar, S.K.: Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum. IEEE Trans. Image Process. **19**(9), 2241–2253 (2010)

[66] Yuhas, R.H., Goetz, A.F., Boardman, J.W.: Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In: JPL Airborne Geosci. Workshop. vol. 1, pp. 147–149 (1992)

[67] Zamora Esquivel, J., Cruz Vargas, A., Lopez Meyer, P., Tickoo, O.: Adaptive convolutional kernels. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 1998–2005 (2019)

[68] Zeng, Y., Huang, W., Liu, M., Zhang, H., Zou, B.: Fusion of satellite images in urban area: Assessing the quality of resulting images. In: Int. Conf. Geoinfo. (ICG). pp. 1–4 (2010)

[69] Zhang, J., Jiao, L., Ma, W., Liu, F., Liu, X., Li, L., Chen, P., Yang, S.: Transformer based conditional gan for multimodal image fusion. IEEE Transactions on Multimedia pp. 1–14 (2023). https://doi.org/10.1109/TMM.2023.3243659

[70] Zhang, X., Huang, W., Wang, Q., Li, X.: SSR-NET: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion. IEEE Trans. Geosci. Remote Sens. **59**(7), 5953–5965 (2021)
[71] Zhang, Y., Liu, C., Sun, M., Ou, Y.: Pan-sharpening using an efficient bidirectional pyramid network. IEEE Trans. Geosci. Remote Sens. **57**(8), 5549–5563 (2019)
[72] Zhou, J., Civco, D.L., Silander, J.: A wavelet transform method to merge landsat tm and spot panchromatic data. International journal of remote sensing **19**(4), 743–757 (1998)
[73] Zhou, J., Jampani, V., Pi, Z., Liu, Q., Yang, M.H.: Decoupled dynamic filter networks. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 6647–6656 (2021)
[74] Zhou, M., Yan, K., Fu, X., Liu, A., Xie, C.: Pan-guided band-aware multi-spectral feature enhancement for pan-sharpening. IEEE Transactions on Computational Imaging **9**, 238–249 (2023). https://doi.org/10.1109/TCI.2023.3248956
[75] Zhou, M., Yan, K., Pan, J., Ren, W., Xie, Q., Cao, X.: Memory-augmented deep unfolding network for guided image super-resolution. International Journal of Computer Vision **131**(1), 215–242 (2023)
[76] Zhu, Z., Cao, X., Zhou, M., Huang, J., Meng, D.: Probability-based global cross-modal upsampling for pansharpening. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14039–14048 (2023)
[77] Zhuang, P., Liu, Q., Ding, X.: Pan-ggf: A probabilistic method for pan-sharpening with gradient domain guided image filtering. Signal Processing **156**, 177–190 (2019)
[78] Zhuang, P., Liu, Q., Ding, X.: Pan-ggf: A probabilistic method for pan-sharpening with gradient domain guided image filtering. Sign. Process. **156**, 177–190 (2019)