Source-Adaptive Discriminative Kernels based Network for Remote Sensing Pansharpening

Si-Ran Peng, **Jin-Fan Hu**, **Yu-Wei Zhuo**, **Liang-Jian Deng***
University of Electronic Science and Technology of China, Chengdu, 611731
{Siran_Peng, hujf0206}@163.com, yuweii@yeah.net, liangjian.deng@uestc.edu.cn

Abstract

For the pansharpening problem, previous convolutional neural networks (CNNs) mainly concatenate high-resolution panchromatic (PAN) images and low-resolution multispectral (LR-MS) images in their architectures, which ignores the distinctive attributes of different sources. In this paper, we propose a convolution network with source-adaptive discriminative kernels, called ADKNet, for the pansharpening task. Those kernels consist of spatial kernels generated from PAN images containing rich spatial details and spectral kernels generated from LR-MS images containing abundant spectral information. The kernel generating process is specially designed to extract information discriminately and effectively. Furthermore, the kernels are learned in a pixel-by-pixel manner to characterize different information in distinct areas. Extensive experimental results indicate that ADKNet outperforms current state-of-the-art (SOTA) pansharpening methods in both quantitative and qualitative assessments, in the meanwhile only with about 60,000 network parameters. Also, the proposed network is extended to the hyperspectral image super-resolution (HSISR) problem, still yields SOTA performance, proving the universality of our model. The code is available at http://github.com/liangjiandeng/ADKNet.

1 Introduction

Because of hardware limitations, sensors of satellites cannot capture images of both spectral and high spatial resolution [Vivone *et al.*, 2021]. Only low-resolution multispectral (LR-MS) images and high-resolution panchromatic (PAN) images are captured respectively by satellites like IKONOS, WorldView-2, and WorldView-3. Thus, pansharpening, which aims to fuse a PAN image and an LR-MS image to obtain a high-resolution multispectral image (HR-MS), becomes a fundamental technique in the field of remote sensing image processing. In addition, pansharpening is proved to be popular by the contest in 2006 [Alparone *et al.*, 2007], and the increasing number of review papers published recently.

*Corresponding author

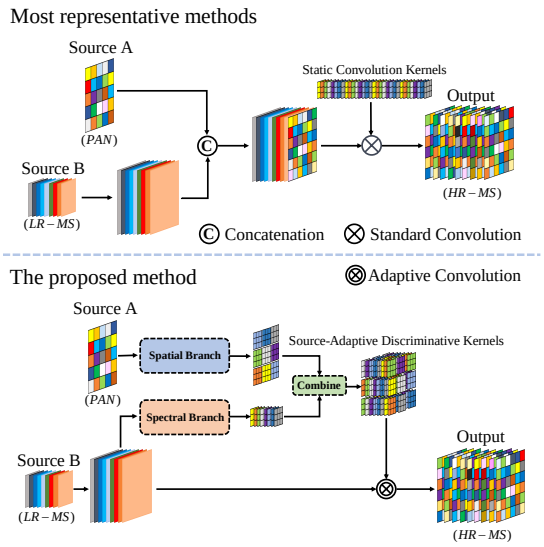

Figure 1: The comparison between most representative CNN-based methods and the proposed method. Note that the former concatenates different sources and applies standard convolution to them. The latter generates kernels discriminately from different sources and applies adaptive convolution to one source.

The traditional methods of pansharpening can be classified into three categories [Meng *et al.*, 2018], *i.e.*, component substitution (CS) methods, multi-resolution analysis (MRA) approaches, and variational optimization-based (VO) techniques. With the rapid growth of deep learning, methods based on convolutional neural networks (CNNs) [Zhang *et al.*, 2019; Fu *et al.*, 2020; Deng *et al.*, 2021] have been widely applied to the problem of pansharpening. Thanks to CNNs' excellent non-linear mapping and feature extraction capabilities, such methods yield satisfactory results.

In the field of pansharpening, since PAN images contain rich spatial details, while LR-MS images contain abundant spectral information, they are supposed to be processed discriminately. As shown in supplementary, most CNN-based methods simply concatenate PAN and LR-MS images, and throw them directly into meticulously designed networks, without considering the distinctive attributes of different sources. Thus, they may not extract features effectively,

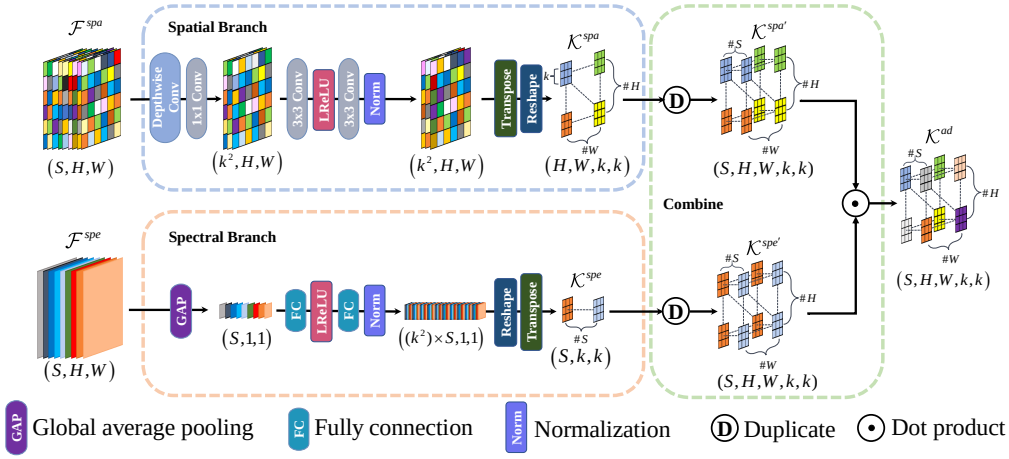

Figure 2: Schematic diagram of source-adaptive discriminative kernels generator (ADKG).

which may cause a lot of parameters to be wasted. Additionally, these methods apply standard convolution to achieve image fusion. In standard convolution, static kernels with the same weights are applied to different areas of given images, which is called content-agnostic [Su *et al.*, 2019]. Previous studies indicate that standard convolution may not be optimal to capture features in some computer vision tasks [Wu *et al.*, 2018]. Therefore, it is necessary to design adaptive kernels which dynamically change with the input.

To tackle the problems mentioned above, we propose a so-called ADKNet for pansharpening, which is composed of source-adaptive discriminative kernels generator (ADKG) modules in series. The spatial kernels and spectral kernels generated are combined through element-wise multiplication to form source-adaptive discriminative kernels, which are then applied to inject detailed information into the LR-MS image as shown in Fig. 1. Our contributions are as follows:

1. A novel ADKG for the task of pansharpening is designed to extract and process information from different sources discriminately and effectively, which guarantees the generalization ability and fewer parameters.
2. The source-adaptive discriminative kernels are generated in a pixel-by-pixel manner to characterize different information in distinct areas, which is proved to be optimal in capturing features for computer vision tasks.
3. Our network yields state-of-the-art (SOTA) outcomes on several datasets of pansharpening. Also, due to the effectiveness of ADKG, our ADKNet only costs about 60,000 network parameters for pansharpening.

2 Related Works

2.1 CNN-based Methods

With the rapid growth of deep learning, more and more CNN-based methods have emerged in the field of pansharpening, achieving competitive results. The initial work is the pansharpening neural network (PNN) by [Masi *et al.*, 2016], which fuses PAN and LR-MS images through three layers of standard convolution. After that, subsequent works, *e.g.*, PanNet by [Yang *et al.*, 2017], DMDNet by [Fu *et al.*, 2020], and FusionNet by [Deng *et al.*, 2021], further prove the potential of the CNNs by yielding remarkable results. However, most existing works do not fully consider the differences between spatial and spectral features as they simply concatenate PAN images and LR-MS images, and send them directly into the CNNs. Such a process cannot extract information discriminately and effectively, which may cause the waste of network parameters and the loss of generalization ability.

2.2 Adaptive Convolution

In standard convolution, static kernels with the same weights are shared across various areas of different images, leading to sub-optimal feature extraction and the loss of flexibility. Adaptive convolution replaces static kernels with adaptive ones generated from the input. Pioneering work is the dynamic filter networks (DFN) by [Jia *et al.*, 2016], where kernels are generated directly from input contents by a separate network branch. Thus, kernel weights vary as data input to the network. With the development of attention mechanisms in deep learning, this strategy is introduced to generate adaptive kernels [Wu *et al.*, 2018], which allows kernels to be learned from multiple neighboring areas of images. Later works, *e.g.*, pixel-adaptive CNNs (PAC) by [Su *et al.*, 2019], context-adaptive convolution for semantic segmentation (CAC) by [Liu *et al.*, 2020], and decoupled dynamic filter networks (DDF) by [Zhou *et al.*, 2021], further prove the superiority of adaptive convolution. It's notable that the above-mentioned adaptive convolutions are never used in the field of multi-source image fusion, *e.g.,* pansharpening, hyperspectral image super-resolution (HSISR), *etc*. Considering the advantages of adaptive convolution, we introduce it to the task of remote sensing pansharpening, by designing a generator that can extract features of distinct areas.

2.3 Motivation

For pansharpening, PAN images contain rich spatial details, while LR-MS images contain abundant spectral information.

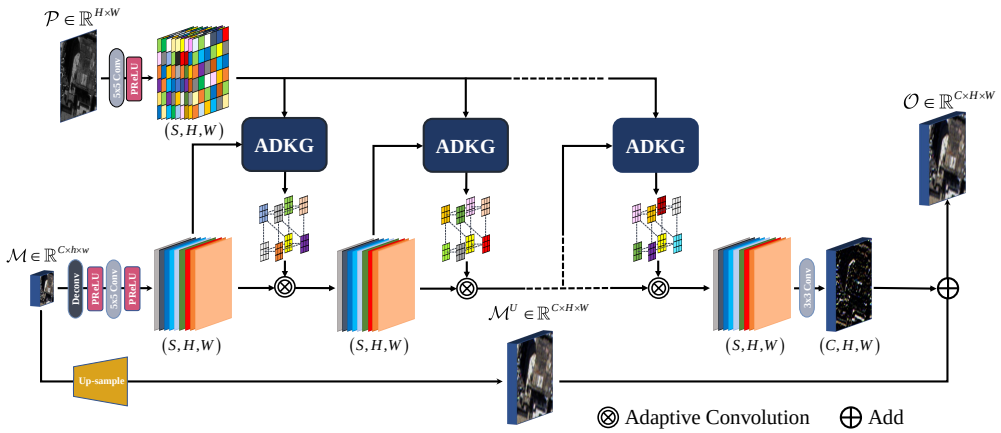

Figure 3: The architecture of source-adaptive discriminative kernels based network (ADKNet).

However, most existing CNN-based methods overlook the differences between spatial and spectral features that they treat PAN and LR-MS images equally, which may cause weak capability of feature representation and the loss of generalization ability. Thus a great number of network parameters are needed to reach considerable performance. Though adaptive convolution has been proved optimal in other computer vision tasks, regrettably, it has not been fully applied to the field of pansharpening to improve network performance. To alleviate the problems above, we propose ADKNet, which extracts features through specially designed ADKG and injects details into LR-MS images in a pixel-by-pixel manner, aiming to obtain better generalization ability while significantly reducing network parameters.

3 The Proposed Method

3.1 Notations

The LR-MS image obtained directly from a remote sensing satellite is denoted as $\mathcal{M} \in \mathbb{R}^{C \times h \times w}$, where C, h, w represents spectral band, height and weight. $\mathcal{P} \in \mathbb{R}^{H \times W}$ denotes the PAN image, in which $H = 4h, W = 4w$ because the scaling factor is generally 4 in pansharpening. The up-sampled LR-MS, the desired HR-MS and the ground-truth (GT) image are defined as $\mathcal{M}^U, \mathcal{O}, \mathcal{X} \in \mathbb{R}^{C \times H \times W}$, respectively.

3.2 ADKG

To extract and process information discriminately while being lightweight, we carefully design ADKG, shown in Fig. 2, for the task of pansharpening. ADKG is generally composed of two branches, *i.e.,* spatial branch, and spectral branch. The former learns spatial details of various areas from PAN images, forming spatial kernels. And the latter extracts spectral information among different channels of LR-MS images to form spectral kernels.

For the spatial branch, we first apply a 1×1 standard convolution layer to alter the number of input channels. Then, we learn in-depth spatial features via 3×3 standard convolution layers. After that, we transpose the feature maps to the size of $H \times W \times k^2$ (k is the kernel size) and reshape them into spatial kernels in a pixel-by-pixel organized manner. Given the input spatial feature representation $\mathcal{F}^{spa} \in \mathbb{R}^{S \times H \times W}$ (S is the channel number of input feature maps), the formation of spatial kernels can be simply represented as:

$$\mathcal{K}^{spa} = \mathrm{B}_{\mathrm{spa}}(\mathcal{F}^{spa}), \tag{1}$$

where $\mathrm{B}_{\mathrm{spa}}(\cdot)$ denotes the operation of spatial branch in Fig. 2, and $\mathcal{K}^{spa} \in \mathbb{R}^{H \times W \times k \times k}$ denotes the spatial kernels. The formed spatial kernels can be seen as a group of convolution kernels with the size of $k \times k$, and each kernel corresponds to a pixel of a distinct position.

As for the spectral branch, global average pooling is first used to aggregate spectral information, while dislodging useless spatial details. Then, fully connected layers are applied to further extract spectral features of higher levels. After that, the feature maps are reshaped to the size of $S \times k \times k$ and transposed to the spectral kernels. Given the input spectral feature representation $\mathcal{F}^{spe} \in \mathbb{R}^{S \times H \times W}$, the formation of spectral kernels can be simply presented as:

$$\mathcal{K}^{spe} = \mathrm{B}_{\mathrm{spe}}(\mathcal{F}^{spe}), \tag{2}$$

where $\mathrm{B}_{\mathrm{spe}}(\cdot)$ denotes the operation of spectral branch in Fig. 2, and $\mathcal{K}^{spe} \in \mathbb{R}^{S \times k \times k}$ denotes spectral kernels.

To obtain source-adaptive discriminative kernels, we first duplicate the produced spatial kernels and spectral kernels $\mathcal{K}^{spa}$ and $\mathcal{K}^{spe}$, forming $\mathcal{K}^{spa\prime}$, $\mathcal{K}^{spe\prime} \in \mathbb{R}^{S \times H \times W \times k \times k}$, respectively. Then we operate element-wise product between $\mathcal{K}^{spa\prime}$ and $\mathcal{K}^{spe\prime}$ to combine them, producing kernels of rich spatial details and abundant spectral information, which can be represented as:

$$\mathcal{K}^{ad} = \mathcal{K}^{spa\prime} \odot \mathcal{K}^{spe\prime}, \tag{3}$$

where $\mathcal{K}^{ad} \in \mathbb{R}^{S \times H \times W \times k \times k}$ denotes the desired source-adaptive discriminative kernels, and $\odot$ denotes element-wise product.

Since the generated kernels may contain extremely large or small values, the normalization method in [Zhou *et al.*, 2021] is applied to enhance the stability of training.

3.3 Adaptive Convolution for Pansharpening

Standard convolution. In standard convolution, given $F = \mathcal{F}_{(1)} \in \mathbb{R}^{S \times N}$ ($N = H \times W$) that denotes the mode-1 un-

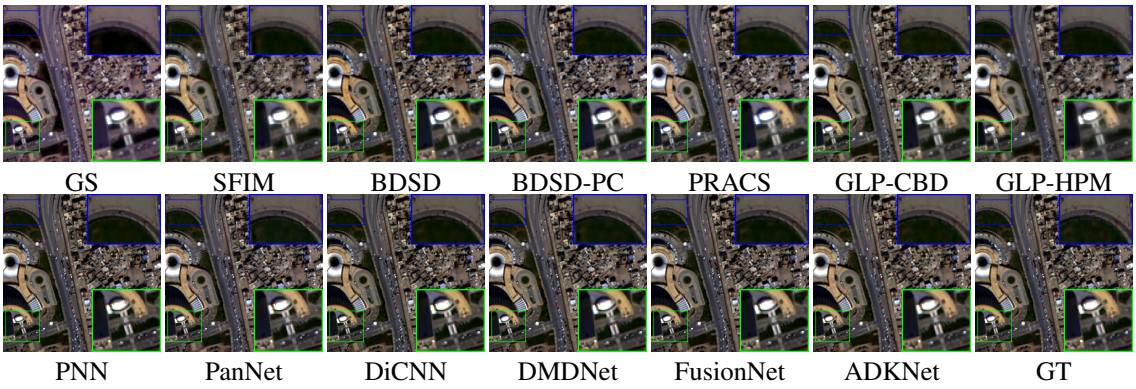

Figure 4: Visual comparisons in natural colors of the most representative methods on Rio dataset of WV3.

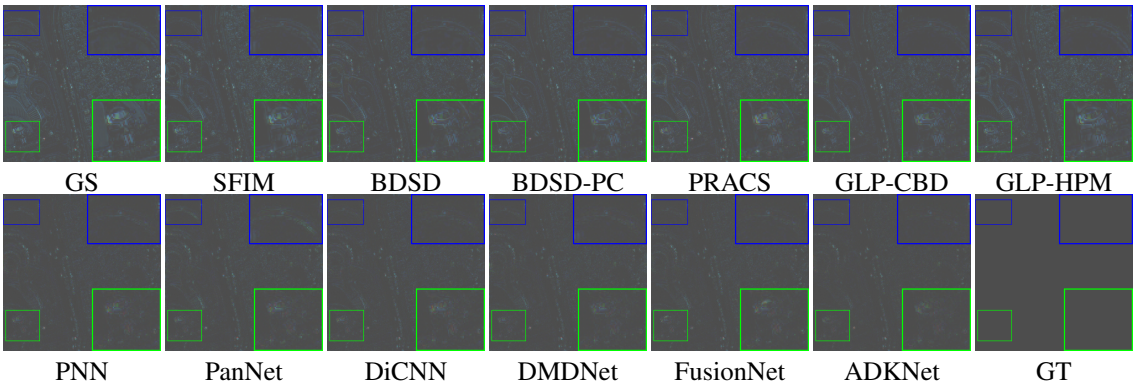

Figure 5: Absolute error maps of Fig. 4.

folding[1] of input feature maps $\mathcal{F} \in \mathbb{R}^{S\times H\times W}$, the i^{th} pixel of the unfolded output feature maps can be written as a combination of input features:

$$F'_{(\cdot,i)} = \sum_{j\in\Omega(i)} \mathcal{W}[p_i - p_j]F_{(\cdot,j)} + b, \tag{4}$$

in which $F_{(\cdot,j)} \in \mathbb{R}^S$ denotes the input feature vector at the j^{th} pixel, and $F'_{(\cdot,i)} \in \mathbb{R}^{S'}$ (S' is the channel number of output feature maps) represents the i^{th} pixel of the output feature vector. $b \in \mathbb{R}^{S'}$ defines the bias vector. $\Omega(i)$ is a $k \times k$ convolution window around the i^{th} pixel, and $\mathcal{W} \in \mathbb{R}^{S'\times S\times k\times k}$ is a bank of static kernels with the size of $k \times k$. Since p_i denotes 2D pixel coordinates, $[p_i - p_j]$ represents indexing of the spatial dimensions of an array with 2D offsets, which makes $\mathcal{W}[p_i - p_j] \in \mathbb{R}^{S'\times S}$ the kernels at position offset between the i^{th} and j^{th} pixels. Thus, the same bank of kernels is shared across various areas of different images in standard convolution, leading to sub-optimal feature extraction.

Adaptive convolution. To alleviate the limitation of standard convolution, we apply adaptive convolution on spectral feature maps with kernels generated by ADKG for pansharpening, which can be written as:

$$F^{spe\,\prime}_{(r,i)} = \sum_{j\in\Omega(i)} \mathcal{K}^{ad}_{(r,i)}[p_i - p_j]F^{spe}_{(r,j)}, \tag{5}$$

where $F^{spe}_{(r,j)} \in \mathbb{R}$ denotes the value at the j^{th} pixel of the r^{th} channel of the unfolded input spectral feature maps, and $F^{spe\,\prime}_{(r,i)} \in \mathbb{R}$ denotes the output one. $\mathcal{K}^{ad}_{(r,i)} \in \mathbb{R}^{k\times k}$ defines a single kernel at the i^{th} pixel-wise position of the r^{th} channel of the unfolded $\mathcal{K}^{ad}$. Thus, the kernels applied in adaptive convolution are element-wise generated from the input, considering the spatial details and spectral information.

3.4 ADKNet

To prove the effectiveness of ADKG and adaptive convolution for pansharpening, the ADKNet, as shown in Fig. 3, is designed to be a simple series network architecture with a skip connection.

Firstly, we apply the basic forward propagation section to process $\mathcal{P}$ and $\mathcal{M}$, obtaining $\mathcal{F}^{spa}$ and $\mathcal{F}^{spe}$. Then, the acquired feature maps are sent into ADKG to form source-adaptive discriminative kernels in a pixel-by-pixel manner. The generated kernels are utilized to apply adaptive convolution on $\mathcal{F}^{spe}$, delivering spatial and spectral information.

[1]The mode-1 unfolding of a tensor $\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times n_3}$ can be defined as a matrix $\mathbf{X} \in \mathbb{R}^{n_1\times n_2 n_3}$, where the tensor's (i,j,k)-th element maps to the matrix's (i,l)-th element satisfying $l = (j-1)n_2 + k$.

Table 1: Quality assessment of the most representative methods on 1258 reduced-resolution samples of WV3. Best results are in bold.

Method	SAM	ERGAS	SCC	Q8	QAVE	Parameters
GS	5.70±2.01	5.28±2.19	0.873±0.071	0.766±0.139	0.768±0.146	/
SFIM	5.45±1.90	5.20±6.57	0.866±0.067	0.798±0.122	0.811±0.130	/
BDSD	7.00±2.85	5.17±2.25	0.871±0.080	0.813±0.123	0.817±0.126	/
BDSD-PC	5.43±1.97	4.25±1.86	0.891±0.069	0.853±0.116	0.852±0.124	/
PRACS	5.59±1.98	4.69±1.85	0.866±0.081	0.813±0.129	0.811±0.137	/
GLP-CBD	5.29±1.96	4.16±1.78	0.890±0.070	0.854±0.114	0.849±0.123	/
GLP-HPM	5.60±1.97	4.76±1.94	0.873±0.065	0.817±0.128	0.810±0.139	/
PNN	4.00±1.33	2.73±1.00	0.952±0.046	0.908±0.112	0.911±0.114	1.0×10^5
PanNet	4.09±1.27	2.95±0.98	0.949±0.046	0.894±0.117	0.907±0.118	2.5×10^5
DiCNN	3.98±1.32	2.74±1.02	0.952±0.046	0.910±0.112	0.911±0.115	1.5×10^5
DMDNet	3.97±1.25	2.86±0.97	0.953±0.045	0.900±0.114	0.913±0.115	3.1×10^5
FusionNet	3.74±1.23	2.57±0.99	0.958±0.045	0.914±0.112	0.914±0.117	1.5×10^5
ADKNet	**3.56±1.22**	**2.43±0.93**	**0.962±0.043**	**0.921±0.108**	**0.920±0.112**	$\mathbf{0.6\times10^5}$
Ideal value	**0**	**0**	**1**	**1**	**1**	**0**

We called the procedure above an adaptive convolution layer. In our network, the layer number is set as 7, which preferably simulates a physical process of keeping injecting detailed information of different levels into maps with insufficient features. After these layers, we restore the shape of output feature maps and add them with $\mathcal{M}^U$ derived through operating deconvolution on $\mathcal{M}$, obtaining $\mathcal{O}$ which contains rich spatial details and abundant spectral information.

3.5 Loss Function

To further verify the effectiveness of our network, we choose the simple mean square error (MSE) as the loss function:

$$\mathcal{L}oss = \frac{1}{M}\sum_{m=1}^{M} \| f_\Theta(\mathcal{M}^{\{m\}}, \mathcal{P}^{\{m\}}) - \mathcal{X}^{\{m\}} \|_2^2, \qquad (6)$$

where $\mathcal{M}^{\{m\}}$, $\mathcal{P}^{\{m\}}$ and $\mathcal{X}^{\{m\}}$ denote the m^{th} LR-MS and PAN training pair, and GT image, respectively. $f_\Theta(\cdot)$ represents our network and Θ defines the involved model parameters. M is the number of training examples, and $\|\cdot\|_2$ indicates the ℓ_2 norm.

4 Experiments

In this section, we measure the performance of the proposed method by comparing it with some recent state-of-the-art (SOTA) pansharpening approaches belonging to the CS-based, MRA-based, and CNN-based methods through a series of experiments on various datasets acquired by WorldView-3 (WV3) and WorldView-2 (WV2) satellites.

4.1 Experiment Settings

Datasets. In this work, we mainly conduct our experiments on WV3 with a spatial resolution of about 0.3 m for the PAN and 1.2 m for the LR-MS images. The spatial resolution ratio is equal to 4 and the radiometric resolution is 11 bits. The MS bands are composed of four standard colors (RGB and near-infrared) and four new bands (coastal, yellow, red edge, and near-infrared). The dataset is downloaded from the public website[2], which contains 12580 samples. We process the dataset to PAN/LR-MS/GT image pairs (70%/20%/10% as training/validation/testing dataset) with the size of 64×64, $64 \times 64 \times 8$ and $16 \times 16 \times 8$ following Wald's protocol by [Wald *et al.*, 1997], same as FusionNet by [Deng *et al.*, 2021].

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

Table 2: Quality assessment on Rio dataset of WV3.

Method	SAM	ERGAS	SCC	Q8	QAVE
GS	4.061	3.896	0.897	0.866	0.867
SFIM	3.913	3.563	0.888	0.885	0.890
BDSD	3.957	2.849	0.907	0.936	0.936
BDSD-PC	3.807	2.849	0.906	0.936	0.935
PRACS	4.026	3.250	0.897	0.906	0.899
GLP-CBD	3.707	2.773	0.909	0.935	0.934
GLP-HPM	4.135	3.492	0.882	0.894	0.891
PNN	3.073	1.908	0.961	0.969	0.970
PanNet	3.005	1.951	0.964	0.965	0.969
DiCNN	3.025	1.912	0.963	0.969	0.970
DMDNet	2.936	1.812	0.970	0.969	0.973
FusionNet	2.834	1.751	0.971	0.973	0.974
ADKNet	**2.713**	**1.533**	**0.980**	**0.977**	**0.978**
Ideal value	**0**	**0**	**1**	**1**	**1**

Table 3: Quality assessment on 30 full-resolution samples of WV3.

Method	QNR	D_λ	D_s
GS	0.896±0.067	0.021±0.032	0.085±0.046
SFIM	0.932±0.058	0.024±0.033	0.045±0.033
BDSD	0.941±0.055	0.016±0.014	0.044±0.045
BDSD-PC	0.915±0.063	0.020±0.026	0.066±0.046
PRACS	0.912±0.070	0.019±0.030	0.071±0.050
GLP-CBD	0.916±0.074	0.031±0.039	0.055±0.049
PNN	0.957±0.035	0.016±0.020	0.026±0.018
PanNet	0.961±0.026	0.019±0.013	0.019±0.015
DiCNN	0.942±0.056	0.017±0.026	0.041±0.036
DMDNet	0.963±0.019	0.016±0.010	0.020±0.010
FusionNet	0.951±0.038	0.018±0.019	0.031±0.022
ADKNet	**0.972±0.012**	**0.010±0.006**	**0.018±0.007**
Ideal value	**1**	**0**	**0**

Benchmarks. We compare our method with several state-of-the-art (SOTA) approaches consist of four CS-based methods: GS by [Laben and Brower, 2000], BDSD by [Garzelli *et al.*, 2008], BDSD-PC by [Vivone, 2019] and PRACS by [Choi *et al.*, 2010]; three MRA-based methods: SFIM by [Liu and J., 2000], GLP-HPM by [Vivone *et al.*, 2014], and GLP-CBD by [Alparone *et al.*, 2007]; and five CNN-based methods: PNN by [Masi *et al.*, 2016], PanNet by [Yang *et al.*, 2017], DiCNN by [He *et al.*, 2019], DMDNet by [Fu *et al.*, 2020] and FusionNet by [Deng *et al.*, 2021]. For a fair comparison, all CNN-based approaches are trained on the same Nvidia GPU-2080Ti and Pytorch environments.

Evaluation Metrics. According to the pansharpening research standard, we choose five quality indexes for the reduced resolution, including SAM, ERGAS [Wald, 2002], SCC [Zhou *et al.*, 1998], QAVE [Zhou and Bovik, 2002] and Q8 [Garzelli and Nencini, 2009]. And we apply QNR, D_λ and D_s indexes [Vivone *et al.*, 2015] for the full resolution.

Parameters Tuning. In our ADKNet, we set the initial learning rate, epoch, and batch size as 0.003, 1000, and 32, respectively. Thus, the number of iterations is 2.5×10^5. In addition, the learning rate is reduced by half every 100 epochs and Adam is used as the optimizer. In particular, for the setting of other compared CNN-based methods, we apply the default asset in related papers and codes.

4.2 Reduced Resolution Assessment

We train our network and other CNN-based methods on the training dataset acquired by the WV3 satellite. Then, we

Table 4: Quality assessment on Stockholm dataset of WV2.

Method	SAM	ERGAS	SCC	Q8	QAVE
GS	7.730	7.364	0.844	0.808	0.818
SFIM	7.115	6.957	0.856	0.843	0.848
BDSD	7.182	6.377	0.860	0.879	0.881
BDSD-PC	7.095	6.323	0.857	0.881	0.883
PRACS	7.589	7.408	0.812	0.831	0.826
GLP-CBD	7.110	6.543	0.845	0.875	0.873
GLP-HPM	7.299	6.997	0.835	0.852	0.850
PNN	6.862	5.626	0.884	0.894	0.903
PanNet	6.348	5.683	0.879	0.893	0.899
DiCNN	6.816	5.977	0.880	0.880	0.890
DMDNet	6.199	5.369	0.890	0.906	0.910
FusionNet	7.536	6.392	0.840	0.875	0.880
ADKNet	**6.000**	**4.935**	**0.909**	**0.924**	**0.927**
Ideal value	**0**	**0**	**1**	**1**	**1**

Table 5: Ablation study for ADKNet on Stockholm dataset of WV2.

Method	SAM	ERGAS	SCC	Q8	QAVE	Parameters
ConvNet	7.169	5.760	0.869	0.895	0.905	2.5×10^5
AKNet	7.827	6.853	0.804	0.855	0.858	$\mathbf{0.6\times10^5}$
ADKNet	**6.000**	**4.935**	**0.909**	**0.924**	**0.927**	$\mathbf{0.6\times10^5}$
Ideal value	**0**	**0**	**1**	**1**	**1**	**0**

evaluate the most representative methods on 1258 reduced-resolution images. The quantitative assessment results are presented in Tab. 1 which shows our ADKNet obtains the best average quantitative performance for all the quality indexes. Moreover, the ADKNet possesses the fewest network parameters compared with other CNN-based approaches.

We further implement the test on a new dataset acquired by WV3 named Rio with the size of 256×256 for a PAN image. The quantitative assessment results are presented in Tab. 2, which proves the priority of our network. Besides, we present the qualitative assessment results in Fig. 4 and Fig. 5. The darker the absolute error map, the better, indicating our ADKNet outperforms other methods.

4.3 Full Resolution Assessment

To demonstrate the application value of our proposed method, we perform experiments on 30 full-resolution samples on WV3 with the size of 256×256 for PAN images. The quantitative assessment results are shown in Tab. 3. Obviously, the ADKNet performs the best on all three indexes, which firmly indicates the superiority of our method.

4.4 Generalization

One significant problem of CNN-based methods for pansharpening is the generalization ability. Once the testing dataset varies a lot, some CNN-based approaches may not perform well. Owing to the distinctive features extraction process, our ADKNet possesses a stronger generalization ability over other CNN-based methods. We verify this by applying models trained on the dataset of WV3 to another dataset named Stockholm acquired by WV2, with the size of 256×256 for PAN image. The quantitative assessment results are shown in Tab. 4, from which we can observe that the ADKNet far exceeds the performance of other methods on all five indexes, which strongly proves the superior generalization ability of our network. More relevant visual results can be found in the supplementary material.

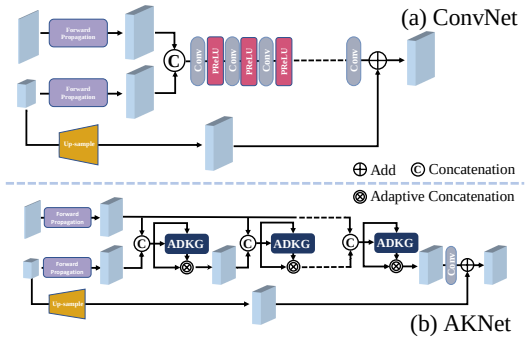

Figure 6: Networks designed for ablation study.

4.5 Ablation Study

To prove the effectiveness of our ADKNet, we design two networks in Fig. 6 to carry out the ablation study.

For the first network (called ConvNet), we concatenate PAN and LR-MS images and throw them into seven standard convolution layers with the kernel size of 3×3. As for the second network (called AKNet), the concatenation of the PAN and LR-MS images is thrown into both spatial and spectral branches of ADKG, ignoring the distinction between spatial and spectral features.

ConvNet is designed to verify the superior of adaptive convolution for pansharpening, while AKNet is designed to prove the correctness of the strategy of processing different sources discriminately. We train ConvNet and AKNet on WV3 and test on WV2. The results are shown in Tab. 5. It is obvious that the ADKNet achieves far better performance than ConvNet and AKNet, which proves that extracting spatial details and spectral information discriminately, injecting the details learned into spectral feature maps through source-adaptive kernels can greatly improve the performance for pansharpening.

5 Conclusion

In this work, we proposed a novel scheme named ADKNet, which consists of ADKG modules in series. Through ADKG block, spatial details from PAN images and spectral information from LR-MS images can be extracted effectively, thereby forming source-adaptive discriminative kernels which can inject detailed information into LR-MS images in a pixel-by-pixel manner. ADKNet yields state-of-the-art (SOTA) outcomes on various datasets of remote sensing pansharpening with the fewest network parameters, proving the strong feature learning ability of the proposed method. In addition, the excellent generalization capability of ADKNet indicates that it is more reliable and robust than other advanced methods.

6 Acknowledgment

This research is supported by NSFC (12171072, 61876203, 61702083), Key Projects of Applied Basic Research in Sichuan Province (Grant No. 2020YJ0216), and National Key Research and Development Program of China (Grant No. 2020YFA0714001)

References

[Alparone *et al.*, 2007] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce. Comparison of pansharpening algorithms: Outcome of the 2006 grss data-fusion contest. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3012–3021, 2007.

[Choi *et al.*, 2010] J. Choi, K. Yu, and Y. Kim. A new adaptive component-substitution-based satellite image fusion by using partial replacement. *IEEE Transactions on Geoscience and Remote Sensing*, 2010.

[Deng *et al.*, 2021] L. J. Deng, G. Vivone, C. Jin, and J. Chanussot. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8):6995–7010, 2021.

[Fu *et al.*, 2020] X. Fu, W. Wang, Y. Huang, X. Ding, and J. Paisley. Deep multiscale detail networks for multiband spectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2090–2104, 2020.

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

[Garzelli *et al.*, 2008] A. Garzelli, F. Nencini, and L. Capobianco. Optimal mmse pan sharpening of very high resolution multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(1):228–236, 2008.

[He *et al.*, 2019] L. He, Y. Rao, J. Li, J. Chanussot, A. Plaza, J. Zhu, and B. Li. Pansharpening via detail injection based convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(4):1188–1204, 2019.

[Jia *et al.*, 2016] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29:667–675, 2016.

[Laben and Brower, 2000] C. A. Laben and B. V. Brower. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening. 2000.

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

[Liu *et al.*, 2020] J. Liu, J. He, J. S. Ren, Y. Qiao, and H. Li. Learning to predict context-adaptive convolution for semantic segmentation. *European Conference on Computer Vision (ECCV)*, 2020.

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

[Meng *et al.*, 2018] X. Meng, H. Shen, H. Li, L. Zhang, and R. Fu. Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges. *Information Fusion*, 46:102–113, 2018.

[Su *et al.*, 2019] H. Su, V. Jampani, D. Sun, O Gallo, E. Learned-Miller, and J. Kautz. Pixel-adaptive convolutional neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11158–11167, 2019.

[Vivone *et al.*, 2014] G. Vivone, R. Restaino, M. Dalla Mura, G. Licciardi, and J. Chanussot. Contrast and error-based fusion schemes for multispectral image pansharpening. *IEEE Geoscience and Remote Sensing Letters*, 11(5):930–934, 2014.

[Vivone *et al.*, 2015] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. A. Licciardi, R. Restaino, and L. Wald. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2565–2586, 2015.

[Vivone *et al.*, 2021] G. Vivone, M. Dalla Mura, A. Garzelli, R. Restaino, and J. Chanussot. A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods. *IEEE Geoscience and Remote Sensing Magazine*, 9(1):53–81, 2021.

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

[Wald *et al.*, 1997] L. Wald, T. Ranchin, and M. Mangolini. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric Engineering and Remote Sensing*, 63:691–699, 1997.

[Wald, 2002] L. Wald. Data fusion. definitions and architectures - fusion of images of different spatial resolutions. *Presses des MINES*, 2002.

[Wu *et al.*, 2018] J. Wu, Dai, Li, Y. Yang, and X. Ji. Dynamic filtering with large sampling field for convnets. *European Conference on Computer Vision (ECCV)*, 2018.

[Yang *et al.*, 2017] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley. Pannet: A deep network architecture for pan-sharpening. *IEEE International Conference on Computer Vision (ICCV)*, pages 1753–1761, 2017.

[Zhang *et al.*, 2019] Y. Zhang, C. Liu, M. Sun, and Y. Ou. Pan-sharpening using an efficient bidirectional pyramid network. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5549–5563, 2019.

[Zhou and Bovik, 2002] W. Zhou and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, 2002.

[Zhou *et al.*, 1998] J. Zhou, D. L. Civco, and J. A. Silander. A wavelet transform method to merge landsat tm and spot panchromatic data. *International Journal of Remote Sensing*, 19(4):743–757, 1998.

[Zhou *et al.*, 2021] J. Zhou, V. Jampani, Z. Pi, Q. Liu, and M. H. Yang. Decoupled dynamic filter networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.