LAGConv: Local-Context Adaptive Convolution Kernels with Global Harmonic Bias for Pansharpening

Zi-Rong Jin[1*], **Tian-Jing Zhang** [1*], **Tai-Xiang Jiang** [2], **Gemine Vivone**[3], **Liang-Jian Deng** [1†]

[1] University of Electronic Science and Technology of China
[2] School of Economic Information Engineering, Southwestern University of Finance and Economics
[3] National Research Council - Institute of Methodologies for Environmental Analysis
2018051403016@std.uestc.edu.cn, zhangtianjinguestc@163.com,
taixiangjiang@gmail.com,gemine.vivone@gmail.com,liangjian.deng@uestc.edu.cn

Abstract

Pansharpening is a critical yet challenging low-level vision task that aims to obtain a higher-resolution image by fusing a multispectral (MS) image and a panchromatic (PAN) image. While most pansharpening methods are based on convolutional neural network (CNN) architectures with standard convolution operations, few attempts have been made with context-adaptive/dynamic convolution, which delivers impressive results on high-level vision tasks. In this paper, we propose a novel strategy to generate local-context adaptive (LCA) convolution kernels and introduce a new global harmonic (GH) bias mechanism, exploiting image local specificity as well as integrating global information, dubbed LAGConv. The proposed LAGConv can replace the standard convolution that is context-agnostic to fully perceive the particularity of each pixel for the task of remote sensing pansharpening. Furthermore, by applying the LAGConv, we provide an image fusion network architecture, which is more effective than conventional CNN-based pansharpening approaches. The superiority of the proposed method is demonstrated by extensive experiments implemented on a wide range of datasets compared with state-of-the-art pansharpening methods. Besides, more discussions testify that the proposed LAGConv outperforms recent adaptive convolution techniques for pansharpening.

Introduction

Pansharpening aims to fuse a low-resolution multispectral image (LR-MSI) and a high-resolution panchromatic image (HR-PANI) to make up for the deficiencies of certain kinds of remote sensing data, even promoting the applicability of remote sensing image for higher-level processing, such as classification (Cao et al. 2020), land monitoring (Du et al. 2013) and detection (Ying et al. 2017). Recently, there has been a considerable improvement for pansharpening thanks to new and complex CNN architectures, which are mainly based on the standard convolution operations (Vivone et al. 2021; Guo, Zhuang, and Guo 2020).

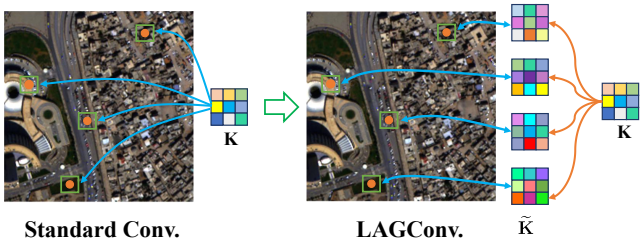

Figure 1: A toy example to motivate the use of the LAGConv. Left: The standard convolution operation, by which all the pixels in the image/feature map are convolved by the same kernel $\mathbf{K}$; Right: The LAGConv operation, by which all the pixels in the feature map are convolved by local-context adaptive (LCA) kernels $\widetilde{\mathbf{K}}$. The blue line indicates the classical convolution operation, and the orange line is the dot product between the kernel $\mathbf{K}$ and the weights adaptively learned from the local patches.

Standard convolution, however, is inherently limited by its spatial-invariance property when addressing pixel-wise tasks like image super-resolution and pansharpening. For a specific feature map, using a uniform convolution kernel on different locations that record different objects can lead to a limited ability in terms of image content adaptation (Su et al. 2019). To overcome this shortcoming, many adaptive convolution techniques have been designed to dynamically generate convolution kernels for different regions or pixels. They have yielded promising performance in several high-level vision tasks (Chen et al. 2020; Yang et al. 2019; Chen et al. 2021b). Nevertheless, existing adaptive convolution methods, either only focusing on the locality of small regions (Su et al. 2019) or full images (Yang et al. 2019), result in an undesired redundancy or neglect the details in the image. For this reason, they are hardly applicable to pansharpening. This paper proposes a novel adaptive convolution operation consisting of local-context adaptive (LCA) convolution kernels and global harmonic (GH) bias, specifically applied to remote sensing pansharpening. This method can fully extract and exploit the local and global information of the involved image/features to achieve superior performance. The

[*]Co-first authors contributed equally.
[†]Corresponding author.

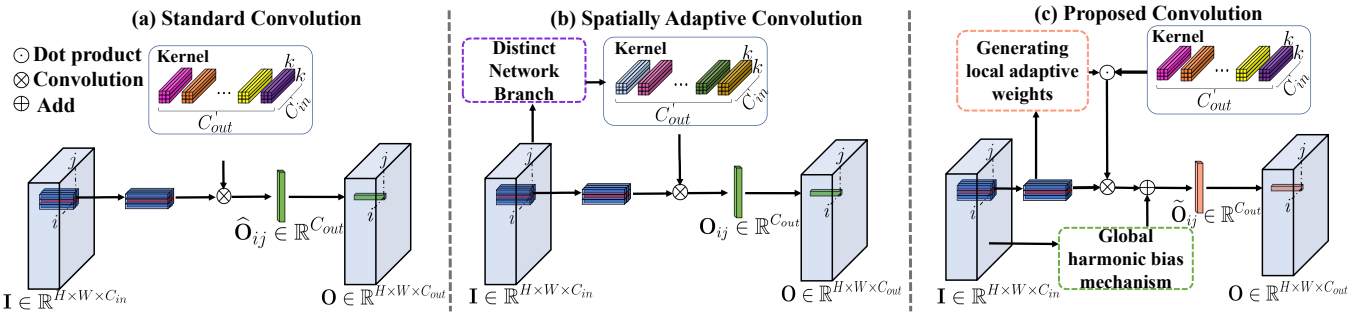

Figure 2: A comparison of the architectures of (a) the standard convolution, (b) the spatially adaptive convolution, and (c) the proposed adaptive convolution.

main contributions of this paper can be summarized as follows:

1. We propose a novel strategy to generate LCA convolution kernels based on each pixel and its neighbors, which not only inherits the advantages of standard convolution, but also enhances the ability to focus on local features and overcomes the limitation of context-agnostic.
2. An GH bias mechanism is introduced to supplement the global information into the local features, thus mitigating the subtle distortion caused by spatial discontinuities, further making the network more flexible and achieving a balance between global and local relationships.
3. The standard convolution layer can be replaced by the combination of the LCA convolution kernels and the GH bias mechanism. We adopt the structure of the residual block, then designing a simple network. To the best of our knowledge, this is the first attempt of using the adaptive convolution to address the pansharpening task.
4. Our network is advantageous thanks to a simple implementation, the end-to-end learning, and the computational efficiency. Experiments show that our model achieves outstanding performance with respect to the state-of-the-art methods in spite of the absence of deep layers and a huge number of parameters.

Related Works and Motivations

In this section, we review first several state-of-the-art works on pansharpening and adaptive convolution methods. Then, our motivations are presented.

Pansharpening: The State of Art

Existing pansharpening approaches can be divided in model-driven and data-driven methods. Model-driven methods take into account the imaging mechanism, which is predictable and theoretically reasonable. Some representative instances of model-driven methods are the smoothing filter-based intensity modulation (SFIM) (J. Liu 2000), the generalized Laplacian pyramid (GLP) (Aiazzi et al. 2002) with modulation transfer function (MTF)-matched filters (Aiazzi et al. 2006), the GLP with a regression-based injection model (GLP-CBD) (Alparone et al. 2007), and the band-dependent spatial-detail with local parameter estimation (BDSD) (Garzelli, Nencini, and Capobianco 2007). Nonetheless, they are incapable of modeling complex nonlinear situations in an efficient way.

Several CNN-based data-driven techniques have recently emerged, pushing the task of pansharpening to a new era and alleviating the issues arisen by model-driven methods. Some representative instances of works in this class are the PNN (Masi et al. 2016), the PanNet (Yang et al. 2017), the DiCNN1 (He et al. 2019), the DMDNet (Fu et al. 2020), and the FusionNet (Deng et al. 2021). They have in common the use of the uniform convolution kernel and conventional bias for feature extraction, resulting in limited learning capabilities of the network.

Adaptive Convolution Techniques

Recently, adaptive convolution techniques, in which sampling locations and/or kernel values are adapted or inferred depending on the inputs, have gained much attention in the field of computer vision (Zhou et al. 2021; Chen et al. 2021a). Existing techniques can be classified into the following three categories:

Adaptive Receptive Fields: To tackle the demand for hand-crafted modifications for receptive field sizes, a scale-adaptive convolution method is proposed for acquiring receptive fields of variable size (Zhang et al. 2017). Moreover, Tabernik et al. present the displaced aggregation units to learn spatial displacements, also adapting the receptive field sizes (Tabernik, Kristan, and Leonardis 2020). Besides, Dai et al. provide the idea of dilating the spatial sampling locations with additionally learned offsets, thus enhancing the geometric transformation modeling ability of the CNN (Dai et al. 2017).

Learning Specialized Convolutional Kernels for Each Example: In (Yang et al. 2019), researchers propose conditionally parametrized convolutions (CondConv) breaking the traditional standard convolution characteristics by calculating the convolution kernel parameters through the input samples. Another notable work is the dynamic convolution (DYConv) proposed in (Chen et al. 2020), which aggregates multiple convolution kernels according to their customized attention degree to each sample. Similar works include the

WeightNet (Ma et al. 2020) and the DYNet (Zhang et al. 2020), in which the convolution kernel is spatially shared.

Spatially Adaptive Convolution Kernel: To overcome the context-agnostic nature of the standard convolution, a deeply explored direction in adaptive convolution is to learn an independent kernel at each pixel by using distinct network branches as illustrated in Fig. 2 (Jia et al. 2016; Zamora Esquivel et al. 2019; Tian, Shen, and Chen 2020), which leads to a huge amount of parameters. Due to computational limitations, these adaptive convolutions are only used to replace a few convolutional layers or in small frameworks. Furthermore, Sun et al. propose a pixel-adaptive convolutional neural network (PAC) that adjusts the filters in a pixel-specific manner (Su et al. 2019). The PAC has a pre-defined form. Limited by the fixed form, it is prone to overfit when applied to pansharpening. By employing decoupled spatial and channel adaptive kernels, the decoupled dynamic filter network (Zhou et al. 2021) is lightweight even compared with the standard convolution. These spatially adaptive methods abandon the kernel sharing mechanism of the standard convolution. Although these spatially adaptive methods are useful for many applications, they are often viewed as a way to increase the kernel redundancy.

Motivations

Based on the related works, we know that standard convolution operations have the defect of context-agnostic. Different positions in the same feature map use a uniform convolution kernel for feature extraction, even if these positions contain different semantic information. However, for pansharpening, a pixel-wise convolution kernel needs to achieve a more effective feature representation. Most of the existing pixel-by-pixel adaptive convolution kernels completely abandon the global-sharing properties of standard convolution and directly introduce convolution kernels by designing network branches, which can generate excessive calculations or redundancy problems. Therefore, we retain the standard spatial-shared convolution kernel, and, according to the local content, we estimate their adaptive weights.

However, while focusing on local uniqueness, global information cannot be ignored. To reconcile the local and global balance, we design a global harmonic bias mechanism, thus integrating the representation of global and local features into a convolution module to replace the standard convolution.

Proposed Method

In this section, we introduce first the designed LAGConv. Then, this LAGConv is further embedded into a residual network architecture, which is able to transfer image details from shallow layers to deep layers to sharpen the low-resolution multispectral image, see, e.g., (Yang et al. 2017).

LAGConv

In pansharpening, the value of each pixel should be accurately determined and the pixel reconstruction is closely related to its neighbors. Therefore, we made a change in the design of the convolution kernel. While retaining the standard convolution kernel, we dynamically learn the weight for each pixel and, finally, realize the adaptive convolution by the dot product of the standard convolution kernel and the weight. The specific operation is detailed below.

Standard Convolution First, let us review the standard convolution. As shown in Fig. 2, a standard convolution without bias operates on a pixel $\mathbf{I}_{ij} \in \mathbb{R}^{1\times1\times C_{in}}$ located at spatial coordinates (i, j). Its local patch is defined as $\mathbf{A}_{ij} \in \mathbb{R}^{k\times k\times C_{in}}$, where C_{in} and k indicate the channels of the input feature map and the patch size, respectively. During the standard convolution operation, all the local patches of the input feature map use the same kernel $\mathbf{K}$. Thus, the operation can be expressed as follows:

$$\widehat{\mathbf{O}}_{ij} = \mathbf{A}_{ij} \otimes \mathbf{K}, \tag{1}$$

where $\mathbf{K} \in \mathbb{R}^{C_{in}\times k\times k\times C_{out}}$ can be viewed as C_{out} convolution kernels with size $k \times k \times C_{in}$ on one layer, $\otimes$ represents the convolution operation, $\widehat{\mathbf{O}}_{ij} \in \mathbb{R}^{1\times1\times C_{out}}$ is the result after the convolution, with C_{out} denoting the channels of the output feature map.

Local-context Adaptive Kernels Different from the standard convolution, the kernel in our LAGConv is automatically adjusted depending on the local patch. Let $\widetilde{\mathbf{K}}_{ij} \in \mathbb{R}^{C_{in}\times k\times k\times C_{out}}$ represents the kernel that is used to perform the convolution on $\mathbf{A}_{ij}$. The proposed LAGConv can be expressed as follows:

$$\widehat{\mathbf{O}}_{ij} = \mathbf{A}_{ij} \otimes \widetilde{\mathbf{K}}_{ij}. \tag{2}$$

In particular, the generation of $\widetilde{\mathbf{K}}_{ij}$ consists of the following three steps, as shown in the top part of Fig. 3. First, $\mathbf{A}_{ij}$ is sent to the convolutional layer with the ReLU activation function to yield its shallow feature. Second, the shallow feature is sent to the fully connected (FC) layers with ReLU and sigmoid activations. A weight $\widetilde{\mathbf{W}}_{ij} \in \mathbb{R}^{1\times k^2}$ is learned, which can perceive the potential relationship between the central pixel $\mathbf{I}_{ij}$ and its neighbors. Finally, the $\widetilde{\mathbf{W}}_{ij} \in \mathbb{R}^{1\times k^2}$ is reshaped to $\mathbf{W}_{ij} \in \mathbb{R}^{k\times k}$ used as the scaling factor for every kernel in $\mathbf{K}$. The scaled kernel is denoted as $\widetilde{\mathbf{K}}_{ij}$ and it can be calculated as follows:

$$\widetilde{\mathbf{K}}_{ij} = \mathbf{W^D}_{ij} \odot \mathbf{K}, \tag{3}$$

where $\odot$ represents the dot product and $\mathbf{W^D}_{ij}$ is the duplicated version of $\mathbf{W}_{ij}$ along the C_{in} channels. The obtained local-context adaptive kernel allows to the network to produce distinctive predictions that consider the local content inconsistencies of the feature map.

Global Harmonic Bias Mechanism We design a global harmonic bias mechanism for our LAGConv. The motivation of this mechanism is to impose an overall continuity of the output feature map. The whole operation process of the LAGConv can be expressed as follows:

$$\mathbf{O}_{ij} = \widetilde{\mathbf{O}}_{ij} + \mathbf{D}, \tag{4}$$

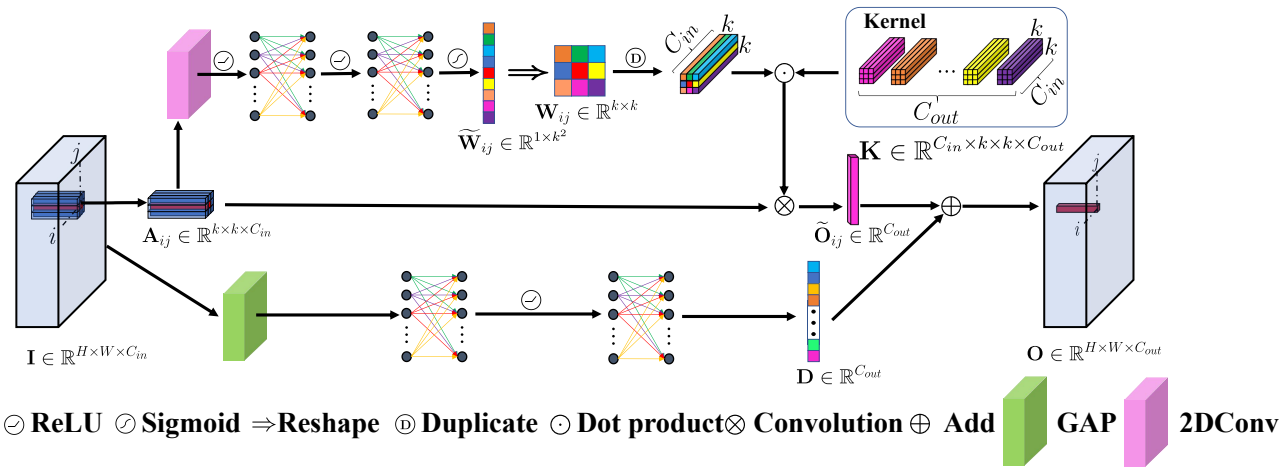

Figure 3: The overview of the LAGConv architecture. The upper part is the local-context adaptive kernel (LAC) and the bottom part is the global harmonic (GH) bias mechanism. The pink block is a 2D convolutional (2D Conv) layer, where the kernel size and padding are set to k and p, respectively, and the input channel and the output channel are C_{in} and C_{out}, respectively. For better understanding, we take the kernel size $k = 3$ and the padding $p = 1$.

where $\mathbf{D} \in \mathbb{R}^{1\times C_{out}}$ is defined as the global harmonic bias generated by the following two steps. First, the input feature, $\mathbf{I}$, is passed through the global average pooling layer (GAP) to obtain $\widetilde{\mathbf{I}} \in \mathbb{R}^{1\times C_{in}}$. Second, $\widetilde{\mathbf{I}}$ is sent to the FC layers with the ReLU activation function to get the output $\mathbf{D}$. This mechanism allows the LAGConv to yield a coherent output that considers all the pixels.

In contrast to previous works, we propose to dynamically adapt the feature map within the network. On one hand, the specificity of each pixel is not ignored. On other hand, since we do not directly discard the kernel shared in the standard convolution operation, no computational resource is wasted in the processing of redundant information.

Local-context Adaptive Residual Network

Based on the proposed LAGConv, we construct a local-context adaptive residual block (LCA-ResBlock) to form the overall network as shown in Fig. 4. We denote the LR-MSI as "**LR**" and the HR-PANI as "**HR**". We want to develop a simple but effective image fusion network that takes an up-sampled "**LR**" (denoted as $\widetilde{\mathbf{LR}}$) image and an "**HR**" data as input. "**SR**" is instead the fused image in output.

LCA-ResBlock is exactly the same as the original ResBlock (He et al. 2016), except that the standard convolution in ResBlock is substituted by the proposed LAGConv. In what follows, we will introduce the proposed overall architecture. As shown in Fig. 4, the proposed network has three steps. The first one contains a LAGConv layer and a ReLU activation layer, then followed by several stacked LCA-ResBlocks. The last step is also an LAGConv layer. Specifically, the **HR** and the $\widetilde{\mathbf{LR}}$ are concatenated together to obtain a feature map $\mathbf{M}$ containing the two input images. After that, $\mathbf{M}$ is passed through the network. Finally, the output of the network is added to the $\widetilde{\mathbf{LR}}$ to get the final **SR** image. The whole procedure can be expressed by the following equation:

$$\mathbf{SR} = \widetilde{\mathbf{LR}} + \mathcal{F}_\Theta(\widetilde{\mathbf{LR}}; \mathbf{HR}), \tag{5}$$

where $\mathcal{F}_\Theta(\cdot)$ represents the mapping function with its parameters Θ that is updated to minimize the distance between the **SR** and the ground-truth (**GT**) image. We chose the simple mean square error (MSE) loss function, since it is enough to yield good outcomes:

$$\mathcal{L}(\Theta) = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathcal{F}_\Theta(\widetilde{\mathbf{LR}}^{(i)}; \mathbf{HR}^{(i)}) + \widetilde{\mathbf{LR}}^{(i)} - \mathbf{GT}^{(i)}\right\|_F^2, \tag{6}$$

where N is the number of training examples and $\|\cdot\|_F$ represents the Frobenius norm.

Experiments

Datasets and Metrics

To benchmark the effectiveness of our network for pan-sharpening, we adopt a wide range of datasets including 8-band data captured by the WorldView-3 (WV3) sensor and 4-band datasets captured by the GaoFen-2 (GF2) and the QuickBird (QB) sensors. Since ground-truth (GT) images are not available, Wald's protocol (B. Aiazzi and Garzelli 2002) is applied. All the source data can be downloaded from the public websites[1] [2]. As in the case of (Deng et al. 2021), for WV3 data, we obtain 12580 PAN/MS/GT image pairs (70%/20%/10% as training/validation/testing datasets) with size 64×64×1, 16×16×8, and 64×64×8, respectively; For GF2 data, we use 10000 PAN/MS/GT image pairs (70%/20%/10% as training/validation/testing datasets) with size 64×64×1, 16×16×4, and 64×64×4, respectively; For QB data, 20000 PAN/MS/GT image pairs (70%/20%/10%

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

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

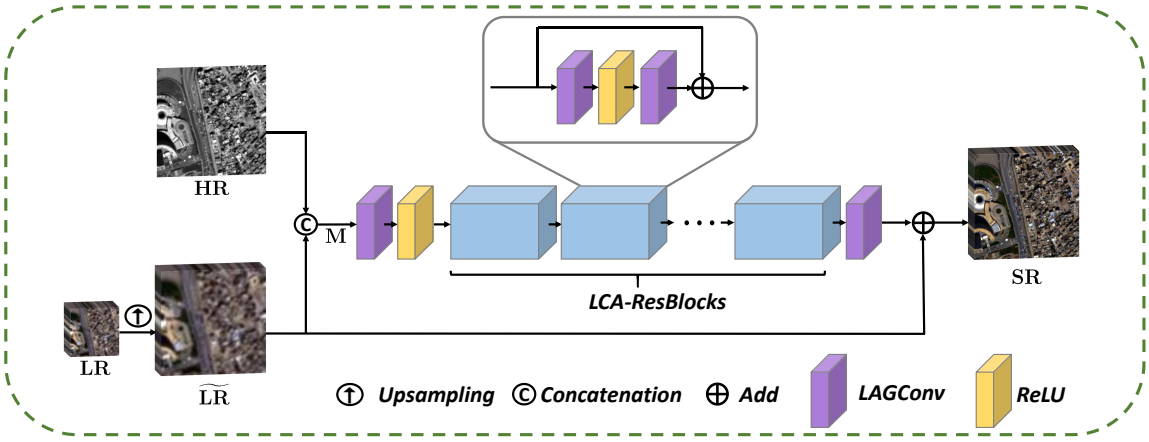

Figure 4: The overall architecture of the proposed network for pansharpening. The network consists of several LCA-ResBlocks, in which the proposed LAGConv is adopted to exploit both local and global information.

.

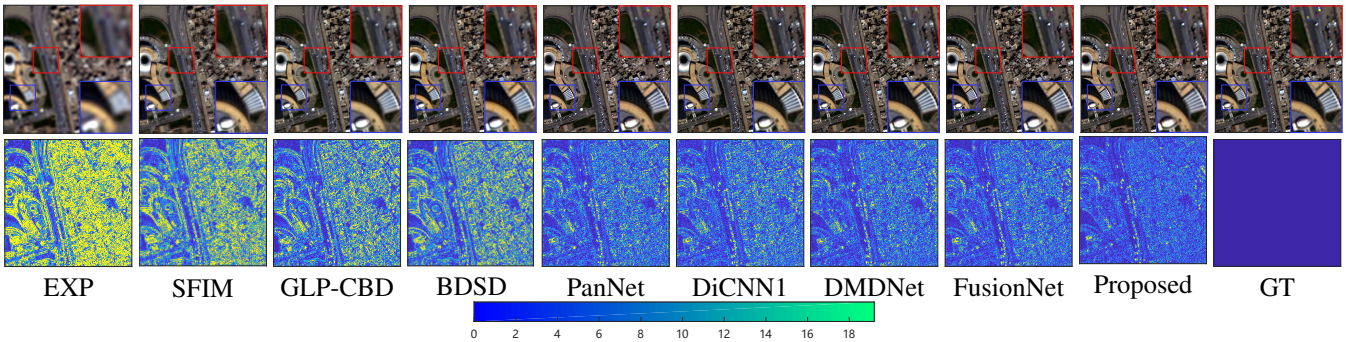

Figure 5: Qualitative comparison on the reduced resolution Rio dataset (source: WV3). The first row presents the RGB visualization, while the second row displays the corresponding absolute error maps (AEMs).

as training/validation/testing datasets) with size 64×64×1, 16×16×4, and 64×64×4 are adopted.

The quality evaluation is conducted both at reduced and full resolutions. For reduced resolution tests, the widely used SAM (Yuhas, Goetz, and Boardman 1992), ERGAS (Wald 2002), SCC (Zhou, Civco, and Silander 1998), and Q-index for 4-band (Q4) and 8-band data (Q8) (Garzelli and Nencini 2009) are adopted to assess the quality of the results. To evaluate the performance at full resolution, the QNR, the D_λ, and the D_s (Vivone et al. 2015) indexes are considered.

Training Details and Parameters

The models are implemented with PyTorch on NVIDIA GeForce GTX 2080Ti. For the parameters of the proposed model, the number of the LCA-ResBlocks is set to 5, while the channels of the LAGConv and the kernel size are 32 and $k \times k$ (with $k = 3$), respectively. Besides, we set 1000 epochs for the network training, while the learning rate is 1×10^{-3} in the first 500 epochs and 1×10^{-4} in the last 500 epochs. The FC layers used in the LAGConv consist of two dense layers with k^2 neurons, and the FC layers in the GH bias consist of two dense layers with C_{out} neurons. Adam optimizer is used for training with a batch size equal to 32, while β_1 and β_2 are set to 0.9 and 0.999, respectively. Besides, the code is available at https://github.com/liangjiandeng/LAGConv.

Comparison with State of Art

We evaluate the proposed method comparing it with several state-of-the-art approaches, including model-driven and data-driven methods.

Evaluation on 8-band Reduced Resolution Dataset. Table 1 reports the average results of all the metrics for the compared methods on the WV3 dataset. By using the same training dataset, the proposed method overcomes the FusionNet clearly getting better performance. Remarkably, the proposed method achieves an elevate spatial fidelity measured by the SCC. The visual quality comparison of the pansharpening methods for the Rio dataset captured by the WV3 sensor is shown in Fig. 5. We can easily see that all the model-driven methods produce some artifacts. Data-driven methods instead get images with finer details. To aid the visual inspection, we also show the absolute error maps (AEMs). It can be observed that our result is the closest to the GT image with cleaner edges than the other compared techniques.

Method	*(a) Reduced resolution WV3 dataset*				*(b) Full resolution WV3 dataset*		
	SAM	*ERGAS*	*SCC*	*Q8*	*QNR*	D_λ	D_s
SFIM	5.452 ± 1.90	4.690 ± 6.574	0.866 ± 0.067	0.798 ± 0.122	0.9282 ± 0.051	0.025 ± 0.028	0.0485 ± 0.028
GLP-CBD	5.286 ± 1.95	4.163 ± 1.775	0.890 ± 0.070	0.854 ± 0.114	0.9113 ± 0.067	0.033 ± 0.033	0.0590 ± 0.043
BDSD	7.000 ± 2.85	5.167 ± 2.248	0.871 ± 0.080	0.813 ± 0.123	0.9300 ± 0.049	0.017± 0.013	0.0537 ± 0.040
PanNet	4.092 ± 1.27	2.952 ± 0.978	0.949 ± 0.046	0.894 ± 0.117	0.9521 ± 0.021	0.0260 ± 0.011	0.0226 ± 0.012
DiCNN1	3.981 ± 1.31	2.737 ± 1.016	0.952 ± 0.047	0.910 ± 0.112	0.9436 ± 0.045	0.018 ± 0.021	0.0392 ± 0.029
DMDNet	3.971 ± 1.24	2.857 ± 0.966	0.953 ± 0.045	0.913 ± 0.115	0.9554 ± 0.020	0.021 ± 0.009	0.0237 ± 0.011
FusionNet	3.744 ± 1.22	2.568 ± 0.944	0.958 ± 0.045	0.914 ± 0.112	0.9556 ± 0.031	0.0198 ± 0.016	0.0254 ± 0.018
Proposed	**3.473 ± 1.19**	**2.338 ± 0.911**	**0.965 ± 0.043**	**0.923 ± 0.114**	**0.9637 ± 0.011**	**0.0147 ± 0.007**	**0.0220 ± 0.006**
Ideal value	**0**	**0**	**1**	**1**	**1**	**0**	**0**

Table 1: Average results on 1258 reduced resolution WV3 data and 50 full resolution WV3 images, respectively. (Bold: best; Underline: second best)

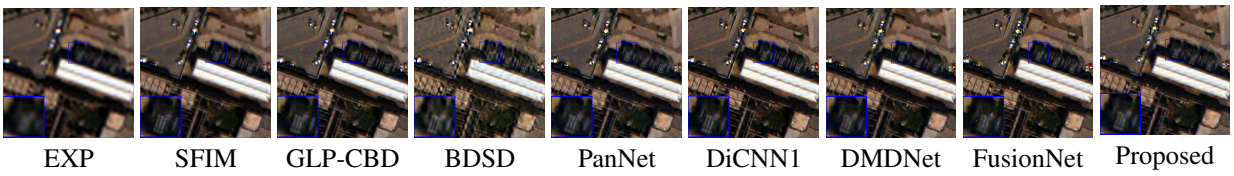

Figure 6: Qualitative comparison on a full resolution WV3 dataset.

Evaluation on 8-band Full Resolution Dataset. The goal of pansharpening is related to real-world applications. Therefore, we further perform a full resolution experiment on 50 WV3 examples. The quantitative results are reported in Table 1 and the visual results are shown in Fig. 6. Again, our method overcomes the other compared approaches both quantitatively and qualitatively.

Evaluation on 4-band Reduced Resolution Dataset. To prove the wide applicability of the proposed method, we also conduct experiments on the 4-band GF2 and QB datasets. Table 2 reports the outcomes for the whole benchmark. It is clear to see the proposed approach gets the best results.

Ablation Study

To verify the effectiveness of the LCA kernel (LCAK) and the GH bias, we perform a wide ablation study on the Tripoli dataset captured by the WV3 sensor. The specific settings for the five variants of the LAGConv are as follows: 1) only conventional kernels (CK); 2) CK and bias; 3) CK with GH bias; 4) only LCAK; 5) LCAK and bias. The experimental results are shown in Fig. 7 and Table 3. It is can be observed that the proposed LAGConv works better than the network with standard convolutions. Besides, the comparison between the conventional bias and the absence of bias demonstrates that the conventional bias is not suitable for this image fusion task. On other hand, the network with GH bias clearly shows better performance supporting the fact that the GH bias makes coherent outputs.

Comparison with Spatially Adaptive Kernels

We also compare the deployment of the LAGConv with respect to some existing spatially adaptive filters for pansharpening on 1258 WV3 data. In particular, we compare the proposed LAGConv with the pixel-adaptive convolutional (PAC) (Su et al. 2019) and the decoupled dynamic filter (DDF) (Zhou et al. 2021). Since they are not originally designed for pansharpening, we replace the LAGConv in the residual network with PAC and DDF, as well as retraining them with the same training set. Table 4 shows the numerical comparison. It is clear that the proposed LAGConv achieves the best results. It also proves that it is worth to design specific methods for specific tasks.

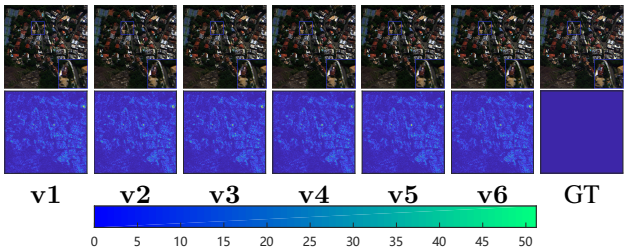

Figure 7: Qualitative comparison of the ablation study on the Tripoli dataset (source: WV3). The first row is about the RGB products and the second one represents the corresponding AEMs.

Extension to Another Application

To demonstrate the robustness and the adaptability of our model, we tested it on another application, i.e., the hyperspectral image super-resolution (HISR), which fuses an LR hyperspectral image (LR-HSI) and an HR multispec-

Method	*(a) GF dataset*				*(b) QB dataset*			
	SAM	*ERGAS*	*SCC*	*Q4*	*SAM*	*ERGAS*	*SCC*	*Q4*
SFIM	2.29 ± 0.63	2.18 ± 0.69	0.861 ± 0.054	0.865 ± 0.04	7.71 ± 1.87	8.77 ± 2.38	0.832 ± 0.105	0.767 ± 0.11
GLP-CBD	2.27 ± 0.73	2.04 ± 0.62	0.873 ± 0.053	0.877 ± 0.04	7.39± 1.78	7.29 ± 0.93	0.854 ± 0.064	0.819 ± 0.12
BDSD	2.30 ± 0.67	2.07 ± 0.61	0.877 ± 0.052	0.876 ± 0.04	7.67 ± 1.91	7.46 ± 0.99	0.851 ± 0.062	0.813 ± 0.13
PanNet	1.40 ± 0.32	1.22 ± 0.28	0.956 ± 0.012	0.947 ± 0.02	5.31 ± 1.01	5.16 ± 0.68	0.930 ± 0.059	0.883 ± 0.14
DiCNN1	1.49 ± 0.38	1.32 ± 0.35	0.946 ± 0.022	0.945 ± 0.02	5.30 ± 0.99	5.23 ± 0.54	0.922 ± 0.051	0.882 ± 0.14
DMDNet	1.29 ± 0.31	1.12 ± 0.26	0.964 ± 0.010	0.953 ± 0.02	5.12 ± 0.94	4.73 ± 0.64	0.935 ± 0.065	0.891 ± 0.14
FusionNet	1.18 ± 0.27	1.00 ± 0.20	0.971 ± 0.007	0.963 ± 0.01	4.54 ± 0.77	4.05 ± 0.26	0.955 ± 0.046	0.910 ± 0.13
Proposed	**1.08 ± 0.23**	**0.91 ± 0.20**	**0.977 ± 0.006**	**0.970 ± 0.01**	**4.37 ± 0.72**	**3.74 ± 0.29**	**0.959 ± 0.047**	**0.916 ± 0.13**
Ideal value	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**1**

Table 2: Average results on 81 GF and 48 QB examples, respectively. (Bold: best; Underline: second best)

Method	*SAM*	*ERGAS*	*SCC*	*Q8*
CK (v1)	4.2564	3.1026	0.9628	0.9511
CK + bias (v2)	4.3483	3.1302	0.9614	0.9511
CK + GH bias (v3)	4.2267	3.0575	0.9637	0.9524
LCAK (v4)	4.0354	2.9495	0.9676	0.9571
LCAK + bias (v5)	4.0264	2.9129	0.9684	0.9568
LCAK + GH bias (v6)	**3.9740**	**2.9010**	**0.9692**	**0.9584**
Ideal value	**0**	**0**	**1**	**1**

Table 3: Quantitative comparison of the ablation study on the Tripoli dataset (source: WV3).

tral image (HR-MSI) to obtain an HR-HSI. We adopt the same evaluation framework as in (Xie et al. 2020). Furthermore, we compare our network with three state-of-the-art data-driven methods, including the SSRNet (Zhang et al. 2020), the ResTFNet (Liu, Liu, and Wang 2020), and the MHFNet (Xie et al. 2020). Table 5 shows the quantitative performance on a widely used dataset, i.e., the CAVE dataset (Yasuma et al. 2010). Our model gets the best overall outcomes. For the sake of brevity, we only show the visual comparison with the MHFNet in Fig. 8. More results can be found in the supplementary material. The AEMs of our approach are displayed in dark blue indicating a better spatial and spectral preservation than the MHFNet.

Conclusions

We have presented a novel adaptive convolution operation, called LAGConv. The adaptive local and translation-invariance properties of the LAGConv guarantee its huge potential for pixel-level vision tasks. Besides, the global information is added to the output as a bias, making the results more reasonable. We further adopt a simple residual structure network equipped with the LAGConv for the task of remote sensing pansharpening. The experiments prove that the proposed method could achieve the best results compared with state-of-the-art approaches, and it can be easily extended to another similar tasks, e.g., the challenging hyperspectral image super-resolution problem.

Method	*SAM*	*ERGAS*	*SCC*	*Q8*
PAC	4.10± 1.7	3.01 ± 0.9	0.94± 0.05	0.90 ± 0.12
DDF	3.87 ± 1.2	2.87 ± 0.9	0.95 ± 0.04	0.91 ± 0.11
LAGConv	**3.47 ± 1.1**	**2.33 ± 0.9**	**0.96 ± 0.04**	**0.92 ± 0.11**

Table 4: Comparisons with two state-of-the-art spatially adaptive kernel-based methods on 1258 reduced resolution WV3 test cases.

Method	*PSNR*	*SAM*	*ERGAS*	*SSIM*
SSRNet	45 ± 3	4.7 ± 1.7	2.0 ± 1.3	0.990 ± 0.004
ResTFNet	45 ± 3	3.7 ± 1.3	1.9 ± 1.6	0.993 ± 0.003
MHFNet	46 ± 2	4.3 ± 1.4	1.7 ± 1.4	0.992 ± 0.006
Proposed	**47 ± 3**	**3.0 ± 0.9**	**1.4 ± 0.9**	**0.995 ± 0.002**
Ideal value	∞	**0**	**0**	**1**

Table 5: Average results on 11 CAVE examples.

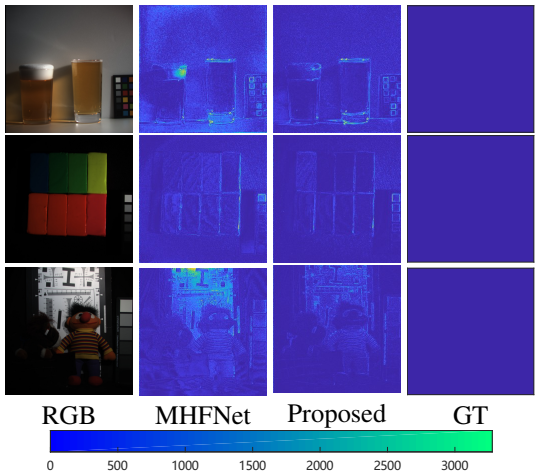

Figure 8: AEMs for the HISR task on three CAVE examples.

Acknowledgements. This research was supported by by NSFC (12171072), Key Projects of Applied Basic Research in Sichuan Province (2020YJ0216), and National Key Research and Development Program of China (2020YFA0714001).

References

Aiazzi, B.; Alparone, L.; Baronti, S.; and Garzelli, A. 2002. Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 40(10): 2300–2312.

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

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

B. Aiazzi, S. B., L. Alparone; and Garzelli, A. 2002. Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 40(10): 2300–2312.

Cao, X.; Yao, J.; Xu, Z.; and Meng, D. 2020. Hyperspectral Image Classification With Convolutional Neural Network and Active Learning. *IEEE Trans. Geosci. Remote Sens.*, 58(7): 4604–4616.

Chen, F.; Wu, F.; Xu, J.; Gao, G.; Ge, Q.; and Jing, X.-Y. 2021a. Adaptive deformable convolutional network. *Neurocomputing*, 453: 853–864.

Chen, J.; Wang, X.; Guo, Z.; Zhang, X.; and Sun, J. 2021b. Dynamic region-aware convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8064–8073.

Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; and Liu, Z. 2020. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11030–11039.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 764–773.

Deng, L.-J.; Vivone, G.; Jin, C.; and Chanussot, J. 2021. Detail Injection-Based Deep Convolutional Neural Networks for Pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8): 6995–7010.

Du, P.; Liu, S.; Xia, J.; and Zhao, Y. 2013. Information fusion techniques for change detection from multi-temporal remote sensing images. *Information Fusion*, 14(1): 19–27.

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

Garzelli, A.; and Nencini, F. 2009. Hypercomplex Quality Assessment of Multi-/Hyper-Spectral Images. *IEEE Geoscience and Remote Sensing Letters*, 6(4): 662–665.

Garzelli, A.; Nencini, F.; and Capobianco, L. 2007. Optimal MMSE pan sharpening of very high resolution multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(1): 228–236.

Guo, P.; Zhuang, P.; and Guo, Y. 2020. Bayesian pansharpening with multiorder gradient-based deep network constraints. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13: 950–962.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

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

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

Jia, X.; De Brabandere, B.; Tuytelaars, T.; and Gool, L. V. 2016. Dynamic filter networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 29: 667–675.

Liu, X.; Liu, Q.; and Wang, Y. 2020. Remote sensing image fusion based on two-stream fusion network. *Information Fusion*, 55: 1–15.

Ma, N.; Zhang, X.; Huang, J.; and Sun, J. 2020. WeightNet: Revisiting the Design Space of Weight Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Masi, G.; Cozzolino, D.; Verdoliva, L.; and Scarpa, G. 2016. Pansharpening by Convolutional Neural Networks. *Remote Sensing*, 8: 594.

Su, H.; Jampani, V.; Sun, D.; Gallo, O.; Learned-Miller, E. G.; and Kautz, J. 2019. Pixel-Adaptive Convolutional Neural Networks.

Tabernik, D.; Kristan, M.; and Leonardis, A. 2020. Spatially-adaptive filter units for compact and efficient deep neural networks. *International Journal of Computer Vision*, 128(8): 2049–2067.

Tian, Z.; Shen, C.; and Chen, H. 2020. Conditional convolutions for instance segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 282–298.

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

Vivone, G.; Dalla Mura, M.; Garzelli, A.; Restaino, R.; Scarpa, G.; Ulfarsson, M. O.; Alparone, L.; and Chanussot, J. 2021. 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.

Wald, L. 2002. Data fusion: definitions and architectures: Fusion of images of different spatial resolutions. *Presses des MINES*.

Xie, Q.; Zhou, M.; Zhao, Q.; Xu, Z.; and Meng, D. 2020. MHF-net: An interpretable deep network for multispectral and hyperspectral image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi:10.1109/TPAMI.2020.3015691.

Yang, B.; Bender, G.; Le, Q. V.; and Ngiam, J. 2019. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in Neural Information Processing Systems (NeurIPS)*, 32: 1307–1318.

Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; and Paisley, J. 2017. PanNet: A deep network architecture for pansharpening. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 5449–5457.

Yasuma, F.; Mitsunaga, T.; Iso, D.; and Nayar, S. K. 2010. Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. *IEEE Transactions on Image Processing*, 19(9): 2241–2253.

Ying, Q.; Qi, H.; Ayhan, B.; Kwan, C.; and Kidd, R. 2017. DOES multispectral / hyperspectral pansharpening improve the performance of anomaly detection? In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*.

Yuhas, R. H.; Goetz, A. F. H.; and Boardman, J. W. 1992. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. *JPL Airborne Geoscience Workshop; AVIRIS Workshop: Pasadena, CA, USA*, 147–149.

Zamora Esquivel, J.; Cruz Vargas, A.; Lopez Meyer, P.; and Tickoo, O. 2019. Adaptive convolutional kernels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (CVPRW)*.

Zhang, R.; Tang, S.; Zhang, Y.; Li, J.; and Yan, S. 2017. Scale-Adaptive Convolutions for Scene Parsing. In *IEEE International Conference on Computer Vision (ICCV)*, 2050–2058.

Zhang, X.; Huang, W.; Wang, Q.; and Li, X. 2020. SSR-NET: Spatial-Spectral Reconstruction Network for Hyperspectral and Multispectral Image Fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7): 1–13.

Zhang, Y.; Zhang, J.; Wang, Q.; and Zhong, Z. 2020. DyNet: Dynamic Convolution for Accelerating Convolutional Neural Networks. arXiv:2004.10694.

Zhou, J.; Civco, D. L.; and Silander, J. A. 1998. A wavelet transform method to merge Landsat TM and SPOT panchromatic data. *International Journal of Remote Sensing*, 19: 743–757.

Zhou, J.; Jampani, V.; Pi, Z.; Liu, Q.; and Yang, M. H. 2021. Decoupled Dynamic Filter Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.