Dynamic Cross Feature Fusion for Remote Sensing Pansharpening

Xiao Wu [1], Ting-Zhu Huang [1]*, Liang-Jian Deng [1]*, Tian-Jing Zhang [2]
[1]School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Yingcai Honors College, University of Electronic Science and Technology of China, Chengdu 611731, China
wxwsx1997@gmail.com; tingzhuhuang@126.com;
liangjian.deng@uestc.edu.cn; zhangtianjinguestc@163.com

Abstract

Deep Convolution Neural Networks have been adopted for pansharpening and achieved state-of-the-art performance. However, most of the existing works mainly focus on single-scale feature fusion, which leads to failure in fully considering relationships of information between high-level semantics and low-level features, despite the network is deep enough. In this paper, we propose a dynamic cross feature fusion network (DCFNet) for pansharpening. Specifically, DCFNet contains multiple parallel branches, including a high-resolution branch served as the backbone, and the low-resolution branches progressively supplemented into the backbone. Thus our DCFNet can represent the overall information well. In order to enhance the relationships of inter-branches, dynamic cross feature transfers are embedded into multiple branches to obtain high-resolution representations. Then contextualized features will be learned to improve the fusion of information. Experimental results indicate that DCFNet significantly outperforms the prior arts in both quantitative indicators and visual qualities.

1. Introduction

Pansharpening is a crucial technique in the field of remote sensing image processing, which aims at fusing a low-resolution multispectral (LRMS) image and a high-resolution (HR) panchromatic (PAN) image to generate a final HR image with the same spectral resolution as the MS image. The outcome of the pansharpening can provide a better visual interpretation, on the other hand, it is conducive to further processing, *e.g.*, land monitoring, mineral exploration, and change detection.

The major point for handling pansharpening task [33, 11, 19] is able to recover more spatial details while retaining more complete spectral information. The traditional methods can be roughly divided into three categories [18, 21, 14], *i.e.*, component substitution (CS) methods, multi-resolution analysis (MRA) methods, variational optimization (VO) approaches. Recently, with the impressive development driven by deep learning (DL), the existing convolutional neural network (CNN) based methods [4, 6, 27, 28, 29, 30, 32] for pansharpening have achieved encouraging performances. This is attributed to the strong nonlinear fitting ability of the CNN, which can well depict the relationship between LRMS image, PAN image, and the desired high-resolution multispectral (HRMS) image.

By observing the existing CNN-based methods, it is concluded that the PAN and LRMS images are used as the input

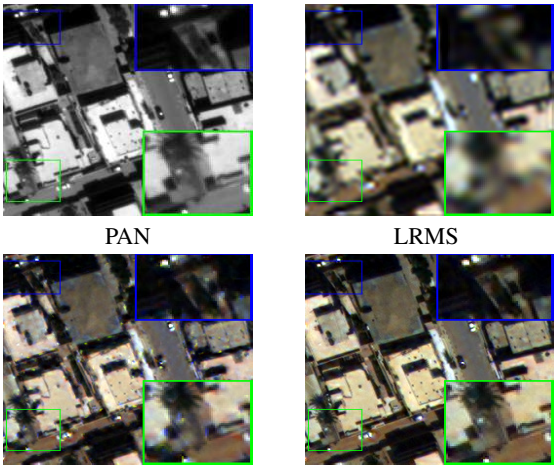


Figure 1: The visual comparison on an original-resolution WorldView-3 dataset. First row: the original PAN and upsampled low-resolution MS (LRMS) images. Second row: the pansharpened image by FusionNet [4] and DCFNet.

*Corresponding author.

of the network, and a number of different network architectures are designed to perform the fusion processing. Our intuitive reasoning is that whether the information in the data can be fully utilized and mined is closely related to the network structure. In recent years, many advanced networks have emerged for different computer vision tasks. A typical example is ResNet [9], which designs the module of residual learning and has become the basic feature extraction module in general computer vision problems. In [12], a feature pyramid network (FPN) is developed, which could efficiently extract various scale features. On its basis, its enhanced architecture provides us with more possibilities for feature fusion and characterization [12, 8, 17].

Although the effectiveness of deep convolutional networks has been proven in computer vision tasks, when it comes to the pansharpening, the defects of information distortion caused by deepening the network are what exactly required to be mitigated. And the existing networks have not adequately considered the cross-scale gap between low-resolution and high-resolution images well to coordinate the relationship between the main feature and supplementary information.

In this paper, we present a novel architecture for pansharpening, namely a dynamic cross feature fusion network (DCFNet). The proposed DCFNet contains three parallel branches, one branch maintains the same resolution as the PAN image and serves as the main branch, which is spatial reduction-free. One of the remaining two branches has the same spatial resolution as the MS image, and the other is twice that of the MS image. On the whole, the information between the three branches is dynamically fused. Features extracted from low spatial resolution are gradually injected into the main branch, maintaining high resolution while supplementing the information provided by low-resolution branch species. Extensive experiments demonstrate that DCFNet can generate reliable results.

To sum up, the contributions of this paper are summarized as follows:

1. We propose a novel architecture named DCFNet, which is the first network with cross-scale parallel branches designed for pansharpening. Benefit from the information fidelity capabilities of high-resolution branches, our model can perform the spatial reduction-free fusion.

2. We design a pyramid cross feature transition layer, which helps multi-resolution branches to capture inter-branches features. And dynamic branch fusion with few parameters is adopted to make the network more effective. As a result, DCFNet significantly outperforms the state-of-the-art methods on a wide range of datasets obtained by various satellite sensors.

3. The proposed DCFNet has a distinctive structure. It has two special variants, namely the famous U-Net and SegNet, which indicates our network can also be applied in more visual tasks.

2. Notations and Related Works

For better explanation, the notations used throughout this paper are first presented.

2.1. Notations

LRMS and PAN images captured by the remote sensing satellite are denoted as $\mathbf{MS} \in \mathbb{R}^{h\times w\times c}$ and $\mathbf{P} \in \mathbb{R}^{H\times W}$, respectively. The desired high-resolution multi-spectral (HRMS) image is defined as $\widehat{\mathbf{MS}} \in \mathbb{R}^{H\times W\times c}$, and the ground truth is represented as $\mathbf{GT} \in \mathbb{R}^{H\times W\times c}$, where $H = 4h, W = 4w$. Moreover, we adopt the interpolation method by a polynomal kernel with 23 coefficients to upsample the $\mathbf{MS} \in \mathbb{R}^{h\times w\times c}$ to obtain the $2\times$ and $4\times$ MS images, defined as $\mathbf{MS}_{2\times} \in \mathbb{R}^{2h\times 2h\times c}$, and $\mathbf{MS}_{4\times} \in \mathbb{R}^{H\times W\times c}$.

2.2. Related Works

CNN-based methods. Pioneering work for pansharpening based on CNN is the pansharpening neural network (PNN) [13], learning the mapping relationship between images only through a simple three-layer CNN. After PNN, a noteworthy work called PanNet [28] proposes a simple structure with a certain degree of physical interpretability. To be more specific, PAN and MS images are passed through a low-pass filter firstly, and their high frequency components are obtained as the input of the network. Spatial information is learned through convolution layers, and the shortcut connection operation in ResBlock is used for spectral preservation. Subsequent works, *e.g.*, DMDNet [6], and FusionNet [4] further prove that the residual learning module is an effective choice for pansharpening. However, existing works do not fully consider the difference in spatial resolution between the MS and PAN images. The most common strategy is to directly resize the MS image to match the spatial resolution of the PAN image and perform convolution operations only at the single scale of high spatial resolution. Such strategy will cause spectral distortion during upsampling, and cannot make full use of the known LRMS image and PAN image.

Motivation. For pansharpening, the supplement of contextual information is conducive to recovering more desired information. However, a family of pansharpening networks mentioned before only adopts single-scale feature fusion to generate final HRMS, lacking contextual guidances to feature representations. And existing feature pyramid network (FPN) provides us a framework for extracting contextual information. Regrettably, since FPN always reduces the spatial resolution of features in the process of feature extrac-

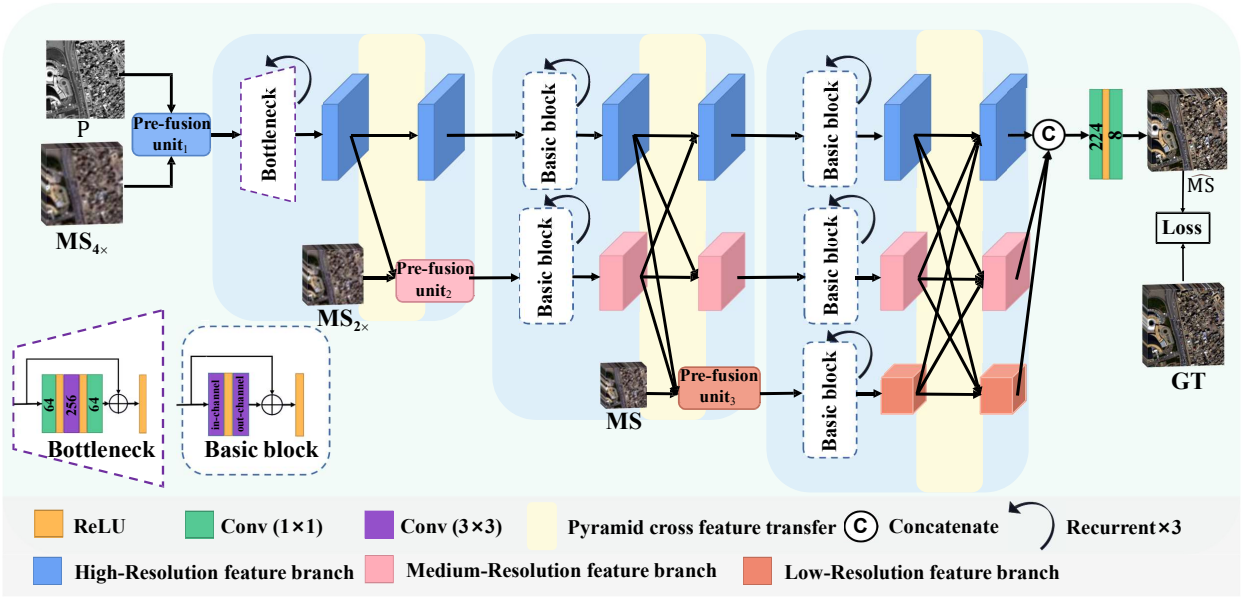


Figure 2: Flowchart of the proposed DCFNet.

tion, it is not wise to adopt FPN for pansharpening. To alleviate the above problems, we propose DCFNet inspired by the HRNet [16], which aims to obtain inter-branch feature fusions from the pyramidal module while always maintaining high resolution in the main branch. Moreover, we adopt a dynamic fusion strategy to coordinate the information fusion between multi-scale branches, which improves the redundancy and conflicts in information supplementation, so that our network can achieve satisfactory results.

3. Network Architecture

The overall pipeline of DCFNet is presented in Fig. 2, it consists of three parallel branches: the main high-resolution (HR) feature branch, the medium-resolution (MR) feature branch, and the low-resolution (LR) feature branch. The three branches are arranged in parallel and are combined progressively to form three convolution stages. Specifically, the main high-resolution feature branch starts from the feature maps obtained by concatenating $\mathbf{MS}_{4\times}$ and $\mathbf{P}$; the medium-resolution feature branch starts from $\mathbf{MS}_{2\times}$ and the feature maps passed by the high-resolution branch. Similarly, the low-resolution feature branch takes the $\mathbf{MS}$ and the feature maps passed by the above two branches as the input of head structure. The pyramid cross feature transfer (PCFT) layer is designed to realize the transfer of feature information between different scales. And the three branches are cross-fused by the proposed dynamic branch fusion (DBF) between each stage.

3.1. Pre-fusion units and building blocks

For the input of the MS image in each branch, we design pre-fusion units as the head structure of each branch to aggregate feature maps transferred from other branches with newly MS images as shown in Fig. 3. In particular, pre-fusion is conducive to network learning of multi-modal information and preliminary feature fusion.

As shown in Fig. 2, we choose the residual block and bottleneck as the building block, which has been proved effective in pansharpening. The convolution kernel of the residual block of each branch is the same. Finally, the stacking of residual blocks is arranged behind pre-fusion units. Therefore, a complete stage is constructed to makes the network deeper so that it can extract better features.

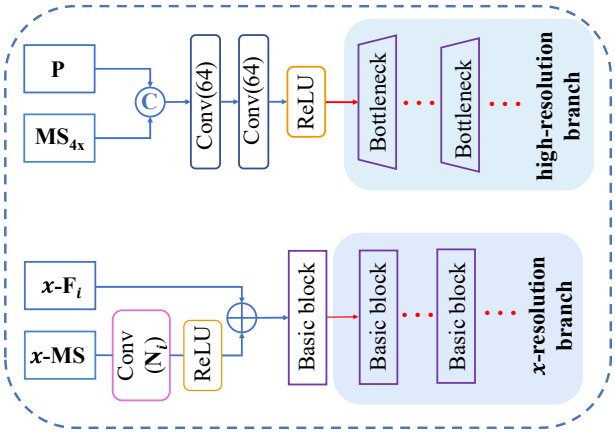


Figure 3: Flowchart of the pre-fusion units. Please note that $\boldsymbol{x}$ refers to medium or low. $\boldsymbol{x} - \mathbf{F}_i$ represents one or more feature maps transferred from other branches. N_i equals to 64/128 for medium/low-resolution feature branch, respectively.

3.2. Pyramid Cross Feature Transfer

Compared with the previous feature pyramid network, the proposed feature transition layer simplifies upsampling process and transfers feature maps to different scale branches as shown in Fig. 4. The PCFT includes two steps: 1) Downsample and transfer the feature maps of higher resolution to lower resolution. 2) Upsample and transfer the feature maps of lower resolution to higher resolution. The form of fusion is the weighted addition of corresponding elements via 3x3 Conv. Notably, the operations of upsampling and downsampling are not symmetrical owing to boost slightly. Specifically, for the path of the high-resolution branch feature to the low-resolution branch feature, the high-resolution feature is first transferred to the medium-resolution feature, and then the medium-resolution feature is transferred to the low-resolution feature, which is a progressive process. However, the path from the LR branch to the HR branch is directly realized, and there is no intermediate process.

DCFNet always maintain the high-resolution branch, which is spatial reduction-free. The PCFT aggregates feature maps from high-to-low and low-to-high branches and transfers the cross-scale feature maps back to high-resolution branches through the above operations, and high-level semantic information is fed into high-resolution branches. The PCFT makes it easier for parallel branches to capture contextual information.

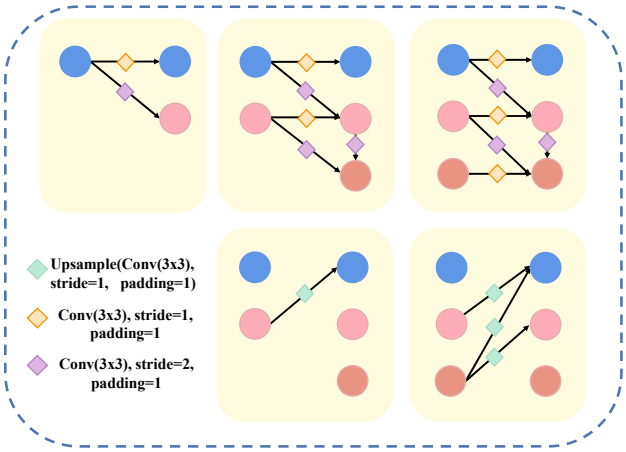


Figure 4: Diagram of the pyramid cross feature transfer layer, corresponding to the yellow part of Fig. 2. The circles in the figure represent the feature maps in each branch, and are color-coded to distinguish the resolution of the feature maps.

3.3. Dynamic branch fusion

Regarding the fusion of features of different resolutions, the method adopted by HRNet [16] is to first adjust their sizes to the same resolution, and then add them accordingly. However, considering the unequal effects of different resolutions on the final result, feature maps of different resolutions should be weighted before being added. Inspired by the weighted feature fusion (WFF) proposed in [17], we adopt the following weighting method:

$$O = \sum_i \frac{w_i}{\sum_j w_j + \epsilon} \cdot I_i, \tag{1}$$

where $w_i > 0$ are the weight learned dynamically, its non-negativity is guaranteed by a layer of ReLU. The value of ϵ is set to 0.0001 to ensure numerical stability.

3.4. Diverse structural deformation

In this section, we devote ourselves to exploring the particularities and possibilities of the DCFNet structure. DCFNet has diverse transformations and connection paths in the process of transferring feature maps. The highlight is that it can be degenerated into two well-known networks, *i.e.*, (a) U-Net [15]; (b) SegNet [3]. For the sake of intuition, we present its degenerate form in Fig.5. In the framework of the convolutional network, the features extracted from deep layers provide contextual semantic information in the entire image, and the features extracted from the shallow network provide more refined information. Whether it is U-Net or SegNet, they can combine information from deep and shallow layers. Both structures are variants of DCFNet. This also indicates DCFNet has a superior foundation for feature extraction and fusion.

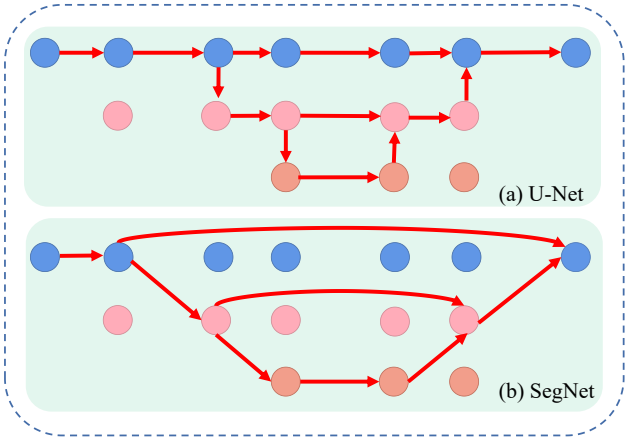


Figure 5: Schematic diagram of DCFNet deformation.

3.5. Loss fuction

We expect to get an ideal HRMS image close to the GT image for achieving good performance. The following experiments (see Sect. 4) prove the significant advantages of the DCFNet structure, though here only choose the simple

mean squared error (MSE) as the loss function,

$$\mathcal{L}\text{oss} = \frac{1}{n}\sum_{k=1}^{n}\left\|\mathcal{F}_{\Theta_{DCFNet}}\left(\mathbf{I}^{\{k\}}\right) - \mathbf{GT}^{\{k\}}\right\|_F^2, \quad (2)$$

where $\mathbf{I}^{\{k\}} = \{\mathbf{P}^{\{k\}}, \mathbf{MS}_{4\times}^{\{k\}}, \mathbf{MS}_{2\times}^{\{k\}}, \mathbf{MS}^{\{k\}}\}$, which represents the input of the DCFNet. n is the number of training examples, and $\|\cdot\|_F$, is Frobenius norm.

4. Experiments

In this section, we gauge the performance of DCFNet[1] by comparing it with other state-of-the-art pansharpening methods through a series of experiments on various datasets obtained by WorldView-2(8-bands), WorldView-3(8-bands), QuickBird(QB, 4-bands), and GaoFen-2(GF-2, 4-bands).

4.1. Network training

In this work, we mainly conduct experiments on data obtained by WorldView-3. We render 8806 sets of data (size: $256 \times 256 \times 8$) from the public website and use $70\%/20\%/10\%$ of them as the training/validation/test datasets. However, due to the lack of the ground truth image, we are required to follow Wald's protocol [26] to get them. The specific data generation steps are: 1) Use modulation transfer function (MTF) for 4x downsampling of original PAN and MS images; 2) Take the downsampled PAN image and the downsampled MS image as the simulated PAN image and the MS image, respectively; 3) Take the original MS image as the simulated GT image.

4.2. Benchmark and Metrics

We compare the proposed DCFNet with several state-of-the-art methods containing the traditional methods (*i.e.*, MS image interpolation using a polynomial kernel with 23 coefficients (EXP) [1], BDSD-PC [20], GLP-HPM [2, 24], GLP-Reg [2, 23][2], CVPR19 [5]), and five competitive CNN-based methods (*i.e.*, PNN [13], PanNet [28], DiCNN1 [10], DMDNet [6], and FusionNet [4]). The evaluation calculates four metrics for simulation (reduced-resolution) experiment, and three metrics for real (full-resolution) experiment. The former includes the SAM [31], ERGAS [25], SCC [34], Q4 (for 4-band data) or Q8 (for 8-band data) [7]. Accordingly, the latter includes the QNR [22], the D_λ, and the D_s [21].

4.3. Evaluation on reduced-resolution datasets

Comparation of CNN-based methods. The results obtained by the CNN-based methods are based on large data set training. The traditional method does not have this prior work. Therefore, the comparison on the test datasets (mentioned in Sect. 4.1) only includes other advanced CNN-based methods. We calculate the average and standard deviation of each indicator on the test dataset and show them in Tab. 1. Obviously, our method far exceeds the performance of other methods on all indicators, which fully proves that DCFNet has a strong learning ability.

Table 1: Quantitative metrics the compared CNN-based methods on 1258 reduced-resolution test datasets (WorldView-3). Best results are in boldface.

Method	*SAM* ($\pm$ std)	*ERGAS* ($\pm$ std)	*Q8* ($\pm$ std)	*SCC* ($\pm$ std)
PNN [13]	4.401 ± 1.329	3.228 ± 1.004	0.888 ± 0.112	0.921 ± 0.046
DiCNN1 [10]	3.980 ± 1.318	2.736 ± 1.015	0.909 ± 0.111	0.951 ± 0.047
PanNet [28]	4.092 ± 1.273	2.952 ± 0.977	0.894 ± 0.117	0.949 ± 0.046
DMDNet [6]	3.971 ± 1.248	2.857 ± 0.966	0.900 ± 0.114	0.952 ± 0.044
FusionNet [4]	3.743 ± 1.225	2.567 ± 0.944	0.913 ± 0.112	0.958 ± 0.045
DCFNet	$\mathbf{3.377 \pm 1.200}$	$\mathbf{2.257 \pm 0.910}$	$\mathbf{0.926 \pm 0.107}$	$\mathbf{0.967 \pm 0.043}$
Ideal value	**0**	**0**	**1**	**1**

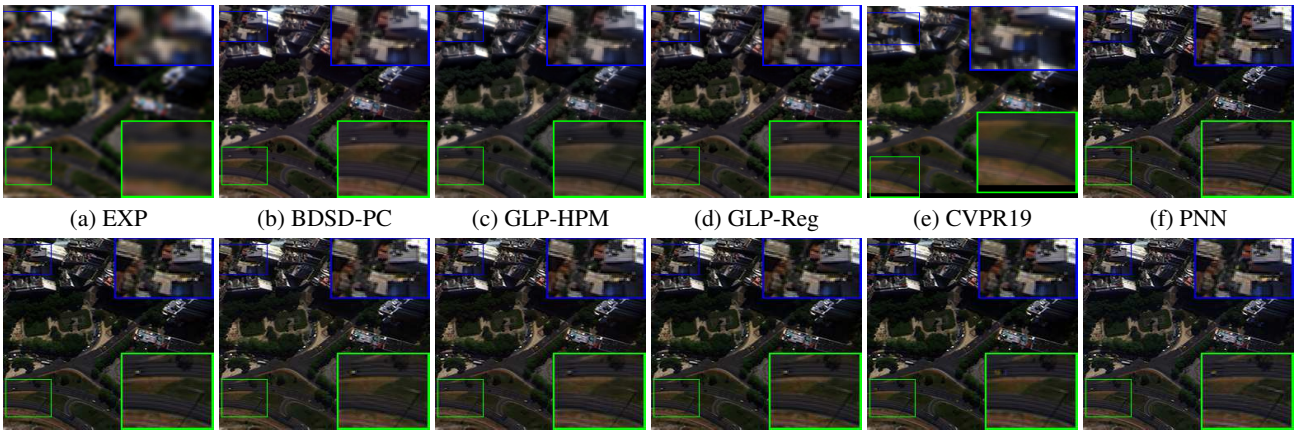
Evaluation on Tripoli dataset. We further carry out the test on new data captured by WorldView-3, which records the local data of Tripoli. In this comparison, all the methods in the benchmark are included. The quantitative evaluation results are shown in Tab. 2, which again indicates the superiority of the DCFNet. In addition, considering real-world applications and observations, it is necessary to compare visual perception. Therefore, we present natural color maps and the absolute error maps with GT as the reference image in Fig. 6 and Fig. 7, respectively. Since the darker the the absolute error map is, the closer the result is to the GT image, it is obvious that DCFNet surpasses other representative methods.

Table 2: Quantitative results for Tripoli dataset (WorldView-3). Best results are in boldface.

Method	*SAM*	*ERGAS*	*Q8*	*SCC*
EXP [1]	6.7883	8.5719	0.7235	0.5129
BDSD-PC [20]	6.4985	6.7186	0.8475	0.7313
GLP-HPM [2, 24]	6.8196	6.8881	0.8393	0.7350
GLP-Reg [2, 23]	6.4100	6.5463	0.8548	0.7394
CVPR19 [5]	6.2395	7.0669	0.8152	0.7321
PNN [13]	5.0778	3.9614	0.9214	0.9242
DiCNN1 [10]	4.7552	3.4978	0.9444	0.9482
PanNet [28]	4.6079	3.4227	0.9395	0.9516
DMDNet [6]	4.4282	3.1972	0.9458	0.9613
FusionNet [4]	4.2764	3.0568	0.9522	0.9646
DCFNet	**3.8666**	**2.8208**	**0.9594**	**0.9718**
Ideal value	**0**	**0**	**1**	**1**

[1]Our model is implemented in the Pytorch framework.

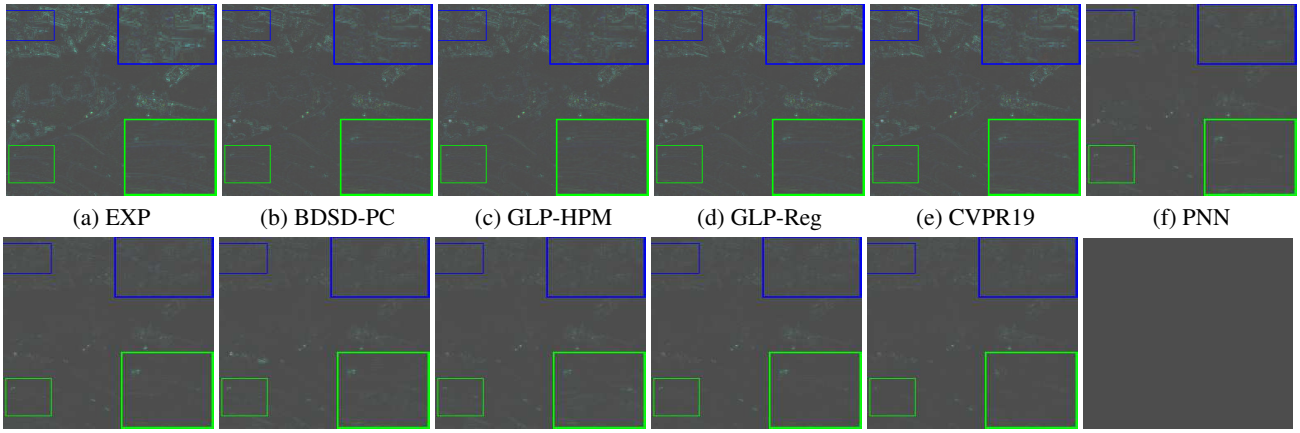
[2]http://openremotesensing.net/kb/codes/pansharpening/



(a) EXP (b) BDSD-PC (c) GLP-HPM (d) GLP-Reg (e) CVPR19 (f) PNN

(g) DiCNN1 (h) PanNet (i) DMDNet (j) FusionNet (k) DCFNet (l) GT

Figure 6: Visual comparisons in natural colors of all the methods on Tripoli dataset (WorldView-3).



(a) EXP (b) BDSD-PC (c) GLP-HPM (d) GLP-Reg (e) CVPR19 (f) PNN

(g) DiCNN1 (h) PanNet (i) DMDNet (j) FusionNet (k) DCFNet (l) GT

Figure 7: Absolute error maps of Fig. 6.

(a) EXP (b) PNN (c) DiCNN1 (d) PanNet (e) DMDNet (f) Fusion-Net (g) DCFNet

Figure 8: Visual comparisons in natural colors of the most representative 6 approaches on Tripoli-OS dataset (WorldView-3) at the original scale.

4.4. Evaluation on full-resolution datasets

In order to demonstrate the application value of DCFNet, we further perform experiments on 50 sets of full-resolution data obtained by WorldView. The quantitative results of compared CNN-based methods[3] are shown in Tab. 3. Moreover, we exhibit the six most competitive methods' results on one example from 50 sets of data (called Tripoli-OS) in Fig. 8. It can be easily seen that whether it is quantitative indicators or visual comparisons, DCFNet is the best.

[3]Please note that traditional methods are relatively poor according to the CNN-based methods. Hence, for the sake of space-saving, we exclude traditional methods from the analysis. Furthermore, for the same reason, we only show the results of the six CNN-based methods.

Table 3: Average values of QNR, D_λ and D_s with the related standard deviations (std) for the 50 full-resolution samples (WorldView-3). Best results are in boldface.

Method	*QNR* ($\pm$ std)	D_λ ($\pm$ std)	D_s ($\pm$ std)
PNN [13]	0.946 ± 0.022	0.023 ± 0.014	0.032 ± 0.012
DiCNN1 [10]	0.939 ± 0.024	0.026 ± 0.016	0.035 ± 0.011
PanNet [28]	0.948 ± 0.017	0.029 ± 0.011	$\mathbf{0.022 \pm 0.009}$
DMDNet [6]	0.945 ± 0.020	0.024 ± 0.012	0.030 ± 0.013
FusionNet [4]	0.941 ± 0.022	0.024 ± 0.013	0.031 ± 0.013
DCFNet	$\mathbf{0.956 \pm 0.013}$	$\mathbf{0.022 \pm 0.009}$	$\mathbf{0.022 \pm 0.006}$
Ideal value	**1**	**0**	**0**

4.5. More experiments on extensive satellite data

In order to further prove the effectiveness of DCFNet, we expand the type of experimental data, including data acquired by GF-2 and QB sensors (see Sect 4). For the GF-2 case, we adopt a huge image (size: $6907 \times 7300 \times 4$) captured over the city of Beijing from the open website [4] to generate 21607 training data (size: $64 \times 64 \times 4$), and another large image acquired over the Guangzhou city to simulate 81 testing data (size: $256 \times 256 \times 4$). For the QB case, we adopt a large image (size: $4906 \times 4906 \times 4$) captured over the city of Indianapolis to generate 20685 training data (size: $64 \times 64 \times 4$) and 48 testing data (size: $256 \times 256 \times 4$). From the indicators shown in Tab. 4, and the visual results shown in Fig. 9 and Fig. 10, the proposed DCFNet can recover more spatial details without losing the spectral information, and its results far exceed the existing methods. This indicates that DCFNet can also be applied to 4-bands data and its outcomes are satisfactory enough.

4.6. Network generalization

To prove the generalization of the network, we apply a ready-made model trained on WorldView-3 data to another dataset obtained by WorldView-2. For a reasonable experiment, we implement the same data generation steps as WorldView3 (see Sect 4.1). The quantitative results are displayed in Tab. 5. Since it is difficult to keep the consistency of spectrum information between branches, the SAM obtained by the compared approaches is slightly better. Overall, our network has produced satisfactory results, which are the best for other indicators except SAM. Experimental results demonstrate that DCFNet has a reliable generalization ability.

4.7. Ablation study

We ablate our various methods for DCFNet by taking a sample from Tripoli dataset. The PCFT (mentioned in Sect 3.2) plays an important role in improving inter-branch fusions. Specifically, we arrange the module of PCFT to conduct cross-scale fusions. Without PCFT, inter-branch fusions degenerate into the sum of low-to-high and high-to-low cross-scale features, then the current branches generate a new branch via Conv2D with a stride of 2. Moreover, we employ learnable parameters to fuse features, which adjusts the effects of branches. With the dynamic branch fusion, the results can be slightly improved, but DBF (mentioned in Sect 3.3) can keep conformity of fusions that is progressively supplemented between branches. From Tab. 6, DCFNet has better results on SAM and slightly better on ERGAS and SCC.

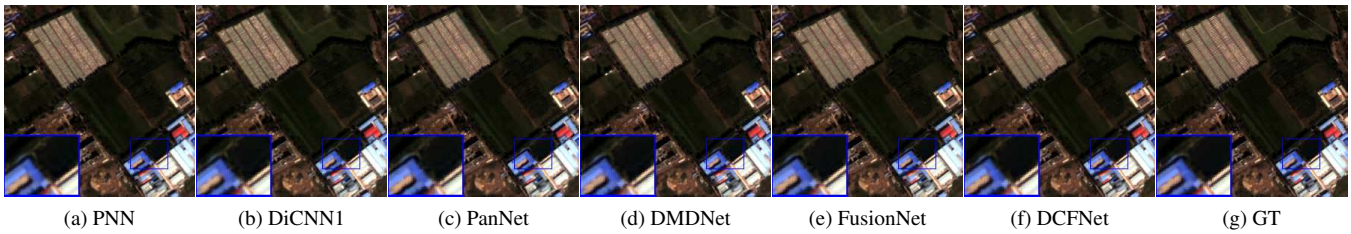
[4]data link: `http://www.rscloudmart.com/dataProduct/sample`

Table 4: Quantitative metrics of the compared CNN-based methods for the GF-2 testing dataset (81 samples) and the QB testing dataset (48 samples). Best results are in boldface.

Method	*SAM* ($\pm$ std)	*ERGAS* ($\pm$ std)	*Q8* ($\pm$ std)	*SCC* ($\pm$ std)
Guangzhou (GF-2)				
PNN [13]	1.659 ± 0.360	1.570 ± 0.324	0.927 ± 0.020	0.928 ± 0.020
DiCNN1 [10]	1.494 ± 0.381	1.320 ± 0.354	0.944 ± 0.021	0.945 ± 0.022
PanNet [28]	1.395 ± 0.326	1.223 ± 0.282	0.946 ± 0.022	0.955 ± 0.012
DMDNet [6]	1.297 ± 0.315	1.128 ± 0.266	0.952 ± 0.021	0.964 ± 0.010
FusionNet [4]	1.179 ± 0.271	1.002 ± 0.227	0.962 ± 0.016	0.971 ± 0.007
DCFNet	$\mathbf{0.994 \pm 0.185}$	$\mathbf{0.811 \pm 0.144}$	$\mathbf{0.971 \pm 0.016}$	$\mathbf{0.982 \pm 0.004}$
Indianapolis dataset (QB)				
PNN [13]	5.799 ± 0.947	5.571 ± 0.458	0.857 ± 0.148	0.902 ± 0.048
DiCNN1 [10]	5.307 ± 0.995	5.231 ± 0.541	0.882 ± 0.143	0.922 ± 0.050
PanNet [28]	5.314 ± 1.017	5.162 ± 0.681	0.883 ± 0.139	0.929 ± 0.058
DMDNet [6]	5.119 ± 0.939	4.737 ± 0.648	0.890 ± 0.146	0.134 ± 0.065
FusionNet [4]	4.540 ± 0.778	4.050 ± 0.266	0.910 ± 0.136	0.954 ± 0.045
DCFNet	$\mathbf{4.342 \pm 0.719}$	$\mathbf{3.749 \pm 0.266}$	$\mathbf{0.920 \pm 0.129}$	$\mathbf{0.961 \pm 0.046}$
Ideal value	**0**	**0**	**1**	**1**

Table 5: Quantitative results on Stockholm dataset (WorldView2). Best results are in boldface.

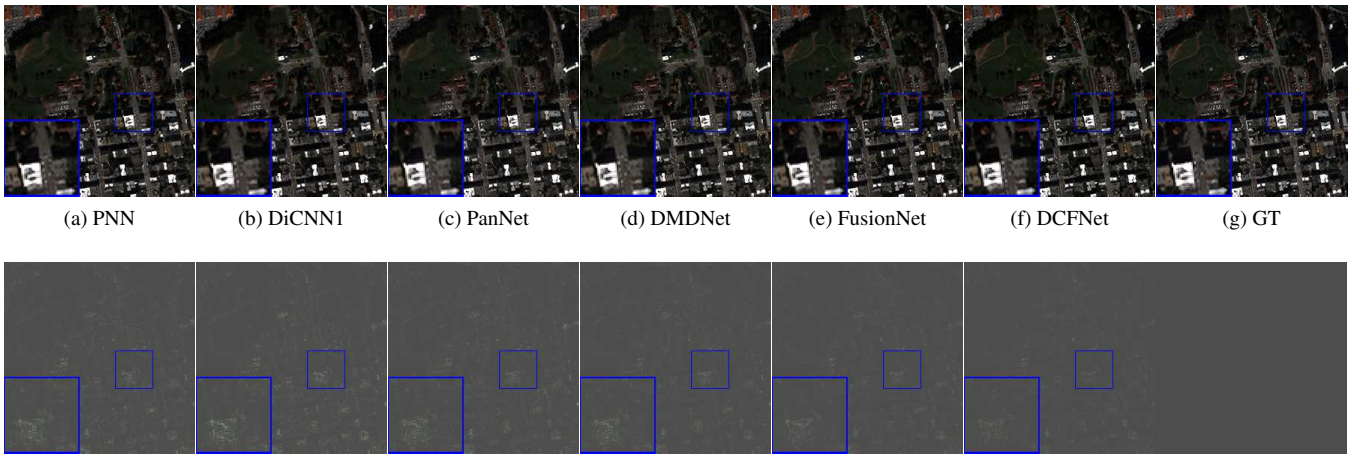
Method	*SAM*	*ERGAS*	*Q8*	*SCC*
EXP [1]	7.8500	9.6793	0.6540	0.4505
BDSD-PC [20]	7.0953	6.3233	0.8819	0.8578
GLP-HPM [2, 24]	7.2988	6.9965	0.8527	0.8355
CVPR19 [5]	7.1098	6.5434	0.8752	0.8457
GLP-Reg [2, 23]	7.1195	6.4998	0.8776	0.8453
PNN [13]	6.8624	5.6259	0.8642	0.8539
DiCNN1 [10]	6.8159	5.9773	0.8802	0.8797
PanNet [28]	6.3916	5.6302	0.8897	0.8895
DMDNet [6]	**6.1986**	5.5692	0.8903	0.8965
FusionNet [4]	6.2784	5.5499	0.8969	0.8897
DCFNet	6.6871	**5.1682**	**0.9175**	**0.9125**
Ideal value	**0**	**0**	**1**	**1**



(a) PNN (b) DiCNN1 (c) PanNet (d) DMDNet (e) FusionNet (f) DCFNet (g) GT

(h) PNN (i) DiCNN1 (j) PanNet (k) DMDNet (l) FusionNet (m) DCFNet (n) GT

Figure 9: Visual comparisons in natural colors of the most representative 6 approaches on the Guangzhou dataset (sensor: GF-2). First row: visual results; Second row: absolute error maps.



(a) PNN (b) DiCNN1 (c) PanNet (d) DMDNet (e) FusionNet (f) DCFNet (g) GT

(h) PNN (i) DiCNN1 (j) PanNet (k) DMDNet (l) FusionNet (m) DCFNet (n) GT

Figure 10: Visual comparisons in natural colors of the most representative 6 approaches on the Indianapolis dataset (sensor: QB). First row: visual results; Second row: absolute error maps.

Table 6: Abalation study of DCFNet with/without some fusion operations on Tripoli dataset.

Method	*SAM*	*ERGAS*	*Q8*	*SCC*
w/o DFB	3.893	2.836	0.971	0.959
w/o PCFT	4.001	2.852	**0.972**	0.959
DCFNet	**3.852**	**2.825**	**0.972**	**0.960**
Ideal value	**0**	**0**	**1**	**1**

5. Conclusion

In this paper, we propose a novel network called DCFNet for pansharpening. DCFNet consists of three parallel branches, where the main branch maintains an end-to-end high-resolution representation, and the remaining two branches continuously inject feature maps into the main branch and adopt the designed pre-fusion units and pyramid cross transfer to balance spatial-reduction and spectral recovering. Extensive experiments on various datasets verify DCFNet achieves significant superiority results and a reliable generalization capability over other advanced methods.

6. ACKNOWLEDGMENT

This work is supported by NSFC (61772003, 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

[1] Bruno Aiazzi, Luciano Alparone, Stefano Baronti, and Andrea Garzelli. 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, 2002.

[2] Bruno Aiazzi, L. Alparone, Stefano Baronti, Andrea Garzelli, and Massimo Selva. Mtf-tailored multiscale fusion of high-resolution ms and pan imagery. *Photogrammetric Engineering and Remote Sensing*, 72(5):591–596, 2015.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[4] Liang-Jian Deng, Gemine Vivone, Cheng Jin, and Jocelyn Chanussot. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 2020, DOI: 10.1109/TGRS.2020.3031366.

[5] Xueyang Fu, Zihuang Lin, Yue Huang, and Yue Huang. A variational pan-sharpening with local gradient constraints. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10257–10266, 2019.

[6] Xueyang Fu, Wu Wang, Yue Huang, Xinghao Ding, and John Paisley. Deep multiscale detail networks for multi-band spectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, PP(99):1–15, 2020.

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

[8] Golnaz Ghiasi, Tsung-Yi Lin, Ruoming Pang, and Quoc V. Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

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

[11] Weihong Guo Mauro Dalla Mura Jocelyn Chanussot Liang-Jian Deng, Gemine Vivone. A variational pansharpening approach based on reproducible kernel hilbert space and heaviside function. *IEEE Transactions on Image Processing*, 27(9):4330–4344, 2018.

[12] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Ross Girshick. Feature pyramid networks for object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 936–944, 2017.

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

[14] Xiangchao Meng, Huanfeng Shen, Huifang Li, Liangpei Zhang, and Randi 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, 2019.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[16] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[17] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10778–10787, 2019.

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

[19] Liang-Jian Deng, Xi-Le Zhao, Jie Huang Ting-Xu, Ting-Zhu Huang. Hyperspectral image superresoluition using unidirectional total variation with tucker decomposition. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:4381–4398, 2020.

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

[21] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A. Licciardi, Rocco Restaino, and Lucien Wald. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2565–2586, 2015.

[22] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A. Licciardi, Rocco Restaino, and Lucien Wald. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2565–2586, 2015.

[23] Gemine Vivone, Rocco Restaino, and Jocelyn Chanussot. Full scale regression-based injection coefficients for panchromatic sharpening. *IEEE Transactions on Image Processing*, 27:3418–3431, 2018.

[24] Gemine Vivone, Rocco Restaino, Mauro Dalla Mura, Giorgio Licciardi, and Jocelyn Chanussot. Contrast and error-based fusion schemes for multispectral image pansharpening. *IEEE Geoscience and Remote Sensing Letters*, 11(5):930–934, 2014.

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

[26] Lucien Wald, Thierry Ranchin, and Marc Mangolini. Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images. *Photogrammetric Engineering and Remote Sensing*, 63:691–699, 1997.

[27] Yancong Wei, Qiangqiang Yuan, Huanfeng Shen, and Liangpei Zhang. Boosting the accuracy of multispectral image pansharpening by learning a deep residual network. *IEEE*

Geoscience and Remote Sensing Letters, 14(10):1795–1799, 2017.

[28] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *IEEE International Conference on Computer Vision*, pages 1753–1761, 2017.

[29] Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, and Liangpei Zhang. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3):978–989, 2018.

[30] Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, and Liangpei Zhang. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3):978–989, 2018.

[31] Roberta H. Yuhas, Alexander F. H. Goetz, and Joe W. Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. *JPL Airborne Geoscience Workshop; AVIRIS Workshop: Pasadena, CA, USA*, pages 147–149, 1992.

[32] Yongjun Zhang, Chi Liu, Mingwei Sun, and Yangjun Ou. Pan-sharpening using an efficient bidirectional pyramid network. *IEEE Transactions on Geoence and Remote Sensing*, 57(8):5549–5563, 2019.

[33] Liang-Jian Deng, Gemine Vivone, Jia-Qing Miao, Jin-Fan Hu, Xi-Le Zhao, Zhong-Cheng Wu, Ting-Zhu Huang. A new variational approach based on proximal deep injection and gradient intensity similarity for spatio-spectral image fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:6277–6290, 2020.

[34] Jie Zhou, Daniel L. Civco, and John A. Silande. A wavelet transform method to merge landsat tm and spot panchromatic data. *International Journal of Remote Sensing*, 19:743–757, 1998.