Multi-Scale and Multi-Stream Fusion Network for Pansharpening

Lihua Jian 1, Shaowu Wu 2,*, Lihui Chen 3, Gemine Vivone 4,5, Rakiba Rayhana 6 and Di Zhang 1

1 School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China
2 School of Computer Science, Wuhan University, Wuhan 430072, China
3 School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China
4 National Research Council, Institute of Methodologies for Environmental Analysis (CNR-IMAA), 85050 Tito Scalo, Italy
5 NBFC (National Biodiversity Future Center), 90133 Palermo, Italy
6 School of Engineering, University of British Columbia, Kelowna, BC V1V 1V7, Canada

* Correspondence: wshaow@whu.edu.cn

Abstract: Pansharpening refers to the use of a panchromatic image to improve the spatial resolution of a multi-spectral image while preserving spectral signatures. However, existing pansharpening methods are still unsatisfactory at balancing the trade-off between spatial enhancement and spectral fidelity. In this paper, a multi-scale and multi-stream fusion network (named MMFN) that leverages the multi-scale information of the source images is proposed. The proposed architecture is simple, yet effective, and can fully extract various spatial/spectral features at different levels. A multi-stage reconstruction loss was adopted to recover the pansharpened images in each multi-stream fusion block, which facilitates and stabilizes the training process. The qualitative and quantitative assessment on three real remote sensing datasets (i.e., QuickBird, Pléiades, and WorldView-2) demonstrates that the proposed approach outperforms state-of-the-art methods.

Keywords: pansharpening; multi-scale; multi-stream fusion; multi-stage reconstruction loss; image enhancement; image fusion

1. Introduction

Remote sensing imaging is an important means by which to obtain Earth object information as high-resolution remote sensing images that are crucial in the interpretation of complex Earth objects. Due to physical limitations, a single type of satellite sensor cannot acquire a multi-spectral (MS) image with both high spatial and spectral resolution. An MS image is generally characterized by high spectral resolution but low spatial resolution, while a panchromatic (PAN) image has the opposite characteristic. However, in many practical applications, such as complex Earth feature interpretation [1,2], change detection [3,4], and land cover classification [5], high spatial and spectral resolutions are both crucial for a good analysis. Compared with hardware improvements, pansharpening provides a better solution to alleviate this problem and has gained much more attention in the remote sensing community. Pansharpening is referred to the generation of a desired high-resolution multi-spectral (HR MS) image from an MS image and a simultaneously acquired PAN image. Over the past decades, various pansharpening methods have been developed [6–8]. Classic and deep learning (DL)-based methods are two significant categories among the existing pansharpening methods.

Classic methods mostly use hand-crafted priors derived from basic assumptions or variational optimization procedures to fuse the low-resolution MS (LR MS) image and the PAN image. For instance, the component substitution (CS)-based methods assume that high-resolution PAN images can entirely or partially replace the principal component after the linear transformation of LR MS images. However, due to the differences in the spectral responses between the MS and PAN sensors, CS-based methods often suffer from severe
spectral distortions. Another historical category is based on the use of multi-resolution analysis (MRA) frameworks to extract PAN details. The hypothesis behind the approaches in this class is that the missing high spatial details of the LR MS image can be obtained by extracting them from the PAN image via multi-scale decomposition. However, MRA-based pansharpening results are affected by both detailed extraction and injection procedures. For instance, unreasonable extraction approaches and ineffective injection coefficients may produce insufficient details, which can cause blur effects or artifacts in the final results. To achieve a better tradeoff between spectral information and spatial details, variational model (VM)-based methods use data and regularization terms to guide the pansharpening process. However, several approaches rely upon complex variational models requiring time-consuming optimization procedures to be solved.

Recently, the advanced capabilities of getting non-linear representations using convolutional neural networks (CNNs) and deep learning have achieved breakthroughs in solving the pansharpening task. However, the existing deep learning (DL)-based pansharpening methods often encounter three challenges, which are stated as follows:

- Some DL-based methods combine all the input data together, but, in this way, they lose the possibility to independently analyze the inputs of the fusion process that are represented by the MS and PAN images;
- Most fusion models process the source images separately and neglect their spectral/spatial correlations;
- The use of the fixed-scale information of the input images limits the pansharpening performance.

Therefore, there is still room for improvement in pansharpening to enhance the spatial resolution while preserving the spectral information.

To address the above issues, in this paper, we propose a multi-scale and multi-stream fusion network (MMFN) for pansharpening. More specifically, we developed a multi-stream fusion block that gets the best from the input images’ spectral/spatial correlation and from the data, taking them as a separate source of information. The MS, PAN, and the concatenation of the two above-mentioned images were separately inputted to shallow convolution layers and deep convolution networks. Moreover, we downsampled the source images twice to extract multi-scale features, which can avoid the loss of spectral and spatial information. Additionally, we designed a multi-stage image reconstruction to recover the desired HR MS image. That is, for each multi-stream fusion block, a loss function was built, which can boost the training process and promote high-resolution MS images close to ground-truth images.

The main contributions of this paper can be summarized as follows:

- Multi-scale and multi-stream strategies. We combined the multi-scale and multi-stream strategies to build a hybrid network structure for pansharpening extracting both shallow and deep features at different scales;
- Multi-stream fusion network. On the basis of multi-scale and multi-stream strategies, we introduced a multi-stream fusion network, which separately leverages spectral and spatial information in the MS and PAN images. Simultaneously, we considered the pansharpening problem as a super-resolution task that concatenates the PAN and MS images to further extract spatial details;
- Multi-scale information injection. We make full use of the multi-scale information of the input (MS and PAN) images by exploiting downsampling and upsampling operations. At each scale, the information of the original MS image is injected through a multi-scale loss.

The remainder of this paper is organized as follows. Section 2 briefly reviews the existing pansharpening methods. Section 3 presents the proposed pansharpening approach. Section 4 shows the experimental results. Finally, in Section 5, the conclusions are drawn.
2. Related Works

2.1. Traditional Pansharpening

The first model-based methods devoted to solving the pansharpening problem belong to the component substitution (CS) and the multi-resolution analysis (MRA) classes [6]. Early stage CS-based techniques include the intensity–hue-saturation (IHS) [9], the principle component analysis (PCA) [10], and the Gram–Schmidt (GS) transform [11]. These methods assume that the HR PAN image can substitute the principal component of the LR MS image projected into a new domain using one of the above-mentioned transformations. Aiazzi et al. [12] proposed an adaptive CS-based pansharpening method using multivariate regression analysis of the two inputs. Garzelli et al. [13] designed a band-dependent-spatial-detail (BDSD) model that extracts the optimal detail image from the PAN image for each MS band. A robust estimator based on this model has recently been proposed in [14]. Choi et al. [15] exploited the idea of the partial replacement proposing the so-called PRACS approach. Instead, Kang et al. presented in [16] an image matting model-based (MMP) component substitution pansharpening method. To alleviate the well-known spectral distortion problem for CS outcomes, researchers developed MRA-based solutions. Representative MRA-based methods include the generalized Laplacian pyramid (GLP) [17], the contourlet transformation [18], the curvelet transformation [19], and the use of non-linear morphological filters [20]. Recently, a new view of seeing the MRA framework has been developed and the detail extraction is efficiently addressed by the simple difference between the PAN image and its low-pass version [21,22] leading to new solutions achieving a high performance with a limited computational burden [23–27]. In this case, to overcome the issue about the limited knowledge of the shape of the spatial filters to generate the low-pass versions of the PAN image, deconvolution approaches have recently been designed [28,29].

Some promising model-based methods are the VM ones [30], which use regularization terms based on images’ prior information to solve the pansharpening (ill-posed) problem. Palsson et al. [31] presented a total variation (TV)-based pansharpening method, which encourages noise removal and preserves the edge detail information of an image. To further reduce spectral distortion, Duran et al. [32] utilized the image self-similarity, which has been applied to the PAN image to establish the nonlocal relationships among the patches of the final result. Chen et al. [33] combined local spectral consistency and dynamic gradient sparsity to simultaneously implement image registration and fusion (SIRF). Liu et al. [34] exploited structural sparsity between the LR MS image and the desired HR MS image and spectral–spatial low-rank priors. In [35], Khademi et al. incorporated an adaptive Markov random field (MRF) prior into the Bayesian framework to restore pansharpened results. Finally, hyper-Laplacian prior-based [36], local gradient constraints-based [37], texture space-based [38], and gradient sparse representation-based [39] approaches have also achieved interesting performances.

2.2. Deep Learning-Based Pansharpening

Deep learning-based methods have recently shown great potential for pansharpening thanks to their powerful nonlinear mapping capability. A comprehensive review about the topic with a critical comparison of widely-used approaches together with a freely distributed toolbox can be found in [40]. The deep learning-based pansharpening methods can be divided roughly into two categories: supervised and unsupervised.

Supervised pansharpening methods require the generation of low-resolution MS images by exploiting the original MS images as ground-truth. Masi et al. [41] attempted first to apply a network for single-image super-resolution [42] to fuse PAN and MS images. This pansharpening method, the so-called PNN, uses a three-layer convolutional neural network (CNN) to construct the map between the inputs and the desired HRMS image. Similarly, Yang et al. [43] proposed a deep network architecture (PanNet) to improve the pansharpening accuracy in terms of spatial and spectral preservation. Scarpa et al. [44] proposed a fast and high-quality target-adaptive CNN (TACNN)-based pansharpening method.
To fully make use of the respective information of MS and PAN images, Liu et al. [45] adopted a two-stream fusion network (ResTFNet-l) for pansharpening. Xu et al. [46] also designed a shallow and deep feature-based spatial-spectral fusion network to enhance the pansharpened results. Liu et al. [47] attempted first to explore the combination of pansharpening and generative adversarial networks (GANs), i.e., the so-called PSGAN, to produce high-quality pansharpened results. Shao et al. [48] combined the idea of an encoder with GANs for pansharpening. A first attempt to integrate classical CS and MRA frameworks with deep convolutional neural networks has been provided in [49]. A more general framework that can fuse images with an arbitrary number of bands exploiting recurrent neural networks has recently been proposed for pansharpening [50]. Wang et al. [51] used an explicit spectral-to-spatial convolution that operates on the MS data to produce an HR MS image. However, the above methods only process single-scale images and do not perform multi-scale feature extraction. Exploiting multi-scale features is of crucial importance for pansharpening, even representing a hot topic for deep neural networks. Yuan et al. [52] introduced a multi-scale feature extraction and multidepth network for pansharpening. Wei et al. [53] proposed a two-stream coupled multi-scale network to fuse MS and PAN images. Huang et al. [54] utilized a non-subsampled contourlet transform to decompose MS and PAN images into low- and high-frequency components, then trained a network with high-frequency images and obtained the fused image by combining the output of the network and the low-frequency components of the MS image. In [55], multi-scale perception dense coding has been integrated into a neural network for pansharpening. Hu et al. [56] combined multi-scale features extraction and dynamic convolutional networks to fuse MS and PAN images. Wang et al. [57] employed a multi-scale deep residual network for pansharpening. In [58], a grouped multi-scale dilated convolution was designed to sharpen MS images. Peng et al. [59] adopted multi-scale dense blocks in a global dense connection network for pansharpening. Multi-scale feature extraction and multi-scale densely connection were employed in [60] to fuse MS and PAN images. Lai et al. [61] extracted multi-scale features in an encoder-decoder network for pansharpening. Zhou et al. [62] proposed a multi-scale invertible network and heterogeneous task distilling to fully utilize the information at full resolution. A multi-scale grouping dilated block was designed in [63] to obtain fine-grain representations of multi-scale features for pansharpening. Tu et al. [64] introduced a clique-structure-based multi-scale block and a multidistillation residual information block for MS and PAN image fusion. A parallel multi-scale attention network is presented in [65] to learn residuals to be injected into the LR MS image. Huang et al. [66] combined MRA and deep learning methods, achieving the injection coefficients with multi-scale residual blocks. Jin et al. [67] decomposed the input images using Laplacian pyramids, then exploited a multi-scale network to fuse each image scale. In [68], a pyramid attention was applied to capture multi-scale features and then the latter were fused by a feature aggregation module to obtain the fused product. A multi-scale transformer with an interaction attention module was introduced in [69]. Zhang et al. [70] proposed a 3D multi-scale attention network for MS and PAN image fusion, in which 2D and 3D convolutions were compared for this task. Although these methods all use multi-scale feature extraction, they either directly perform multi-scale processing on the concatenated MS and PAN images or separately extract features from the MS and PAN images and then perform multi-scale processing. None of the state-of-the-art approaches in the literature can effectively fuse the features extracted from the combined MS and PAN images and the features separately extracted from the MS and PAN images.

The other sub-class of deep learning-based pansharpening approaches relies upon unsupervised methods. Indeed, due to the lack of ground-truth MS images, Qu et al. [71] and Uezato et al. [72] presented unsupervised ways to train models for pan-sharpening.

Finally, there are the promoting methods, belonging to the class of hybrid pansharpening solutions, that combine variational optimization-based and deep learning models, thus benefiting from both the philosophies and increasing the generalization ability of deep learning approaches, see, e.g., [73–75].
3. Proposed Method

The proposed method utilizes multi-scale information and a multi-stream fusion strategy to fully use different levels of features in PAN and MS images. We will introduce the problem formulation of the pansharpening task and depict the overview framework. Afterwards, the network architecture and loss function are presented in detail.

3.1. Problem Formulation

The LR MS image is denoted as $M$ with a size of $w \times h \times c$ ($M \in \mathbb{R}^{w \times h \times c}$), while the high-resolution PAN image is denoted as $P$ with a size of $rw \times rh \times 1$ ($P \in \mathbb{R}^{rw \times rh \times 1}$), where $w$, $h$, and $c$ are the width, height, and number of channels of the MS image, $r$ is the spatial scale factor between the PAN and MS images. The goal of pansharpening is to obtain an HR MS image, $X$, as close as to the ground-truth image, $G$, with a size of $rw \times rh \times c$. We designed the overall framework, which learns the pansharpened process, namely, $X = f_{\text{net}}(M, P)$, where $f_{\text{net}}$ and $\Theta$ represent the MMFN architecture and its parameters, respectively.

3.2. Network Architecture

Figure 1 shows the overall framework of the proposed method, including the multi-scale feature extraction, the multi-stream feature fusion, and the multi-stage image reconstruction. To fully extract the features from the input PAN and MS images, we used a multi-scale strategy to downsample the input images twice, which allows the extracted features to represent original images at different levels. We used the same multi-stream block (MSB) for each scale to fuse the downsampled PAN and MS images. Finally, the pansharpened sub-images were upsampled to recover the HR MS image using a multi-stage strategy. A detailed description is given in the following sections.

3.2.1. Multi-Scale Feature Extraction

We used a fixed scale factor to downsample the PAN and MS images. This operation can extract different levels of spectral and spatial features in both PAN and MS images and it can maximize the information usage at different scales. We denote $D \downarrow (\cdot)$ as the downsampling operation. Thus, we have:

$$I_t = D \downarrow (I_{t-1}, s), t = 2, 3,$$

where $I_1$ is $\tilde{M}$ or $P$ when the downsampling of the upsampled version of the MS image or the PAN image is considered, respectively, $I_t$ (with $t > 1$) represents the downsampled version of the MS or PAN images, $t$ and $s$ are the scale index and the downsampling factor (equal to 2 in our case), respectively.
3.2.2. Multi-Stream Feature Fusion

Many researchers regard the pan-sharpening task as a super-resolution problem that concatenates the MS and PAN images to form a single stream network for spatial information improvement, as shown in Figure 2a. However, in this way, the related spectral characteristics of the input products are ignored. Some other researchers consider that the MS and PAN images convey various independent information. The typical operation is to use the high-frequency information of the PAN image to build the missing spatial details of the MS image. However, it is hard to state that the spatial information is a typical feature related to the sole PAN image and the spectral features are only related to the MS image. This challenge motivates us to focus on developing a dual-stream fusion network (see Figure 2b), which combines all the features of the MS and PAN images.

Figure 2. A comparison of three different stream-based fusion blocks. (a) The single-stream block is based on the concatenation of the MS and PAN images. (b) The dual-stream block separately inputs the two source images to the feature extraction network. (c) The multi-stream block (MSB) exploits three streams to separately extract the features from the PAN, the MS, and the concatenation of the two inputs (i.e., MS and PAN). © refers to the concatenation operation.

To avoid losing information as in the case of the single-stream or the dual-stream, we leveraged a multi-stream fusion strategy to comprehensively exploit information from the spectral and spatial domains. As shown in Figure 2c, we utilized three streams, namely, PAN, MS, and fusion streams, to extract the features from the PAN data, the MS image, and the concatenation of the two inputs (i.e., MS and PAN), respectively. First, the PAN and MS streams built by three sequential convolutional layers were used to extract spatial and spectral features as follows:

\[ F = \text{Conv}((\text{Conv}(\text{Conv}(P)))) , \]  

(2)

where \( \text{Conv}(\cdot) \) denotes the convolutional layers, \( P \) is the (input) MS/PAN image, and \( F \) is the corresponding output. The fusion stream is instead fed by the concatenation of the MS, \( M_i \), and the PAN, \( P_i \), images to extract spatial-spectral features by a ResNet between two convolutional layers. Hence, we have that the spatial–spectral features, \( F_{\text{PMS}} \), are obtained by:

\[ F_{\text{PMS}} = \text{Conv}(\text{ResNet}(\text{Conv}([P_i, M_i]))) , \]  

(3)
where \([\cdot]\) denotes the concatenation operation and \(ResNet(\cdot)\) is the function of ResNet as shown in Figure 2. Finally, we further fused the outputs from the three streams by the same structure as in the fusion stream, thus obtaining the output of the MSB by concatenating \(F_{\text{PMS}}\) with the further fused features, i.e.,

\[
H_t = [F_{\text{PMS}}, \text{Conv}(ResNet(Conv([P_t, M_t, F_{\text{PMS}}])))],
\]

where \(H_t\) is the output of the MSB.

### 3.2.3. Multi-Stage Image Reconstruction

In our work, we used the original source images and their two downsampled versions to perform a three-stage pansharpening network. All the image pairs were fed into the same multi-stream fusion block using Equation (4) for the extraction of the multi-scale fused features. As shown in Figure 1, these multi-scale features were adopted to reconstruct the multi-scale image by using reconstruction blocks, which have the same structure as the one in the fusion stream (i.e., Net 1 in Figure 2c).

\[
F_t = \text{Conv}(ResNet(Conv([P_t, M_t, F_{\text{PMS}}]))),
\]

where \(F_t\) is the fused image at the \(t\)-th scale.

Afterwards, the fusion result of each RB was upsampled and concatenated with the corresponding downsampled MS image as follows:

\[
X_t = [U \uparrow (F_{t+1}, s), \widetilde{M}_t], t = 1, 2,
\]

where \(U \uparrow (\cdot)\) represents the upsampling operation. In our work, we have that \(X_3 = \widetilde{M}_3\).

After the multi-stage fusion, the final HRMS image can be reconstructed in a residual way, thus preserving the spectral information. Hence, we have:

\[
X = F_1 + \widetilde{M}_t
\]

where \(X\) is the fused (pansharpened) product.

### 3.3. Loss Function

During the training phase, we adopted the \(l_1\)-norm as loss function to measure the distance between the pansharpened product and the related ground-truth (reference) image. Comparing it with the \(l_2\)-norm, the \(l_1\)-norm can overcome local minima reaching stable training [76]. Our work had three MSB blocks that generated pansharpening images at three different resolutions. Thus, a loss function was adopted for each block to constrain its training. Hence, the overall loss function can be written as:

\[
L(\Theta) = \frac{1}{3} \sum_{t=1}^{3} |F_t - HRMS_t|_1,
\]

where \(\Theta\) is the set of parameters for the proposed framework, \(HRMS_t\) is the downsampled version of the ground-truth image at the \(t\)-th scale, and \(|\cdot|_1\) is the \(l_1\)-norm.

### 4. Experiment and Evaluations

In this section, we provide extensive experiments to validate the effectiveness of our proposed method. Moreover, three groups of ablation studies were investigated to further demonstrate the superiority of each claim.

#### 4.1. Datasets and Metrics

Three real remote sensing datasets, captured by QuickBird, Pléiades, and WorldView-2, were used. To boost the pansharpening capability of the proposed network, we adopted
data argumentation approaches, i.e., rotation, cropping, and flipping. The total number of training and testing datasets for each satellite were 4000 and 40, respectively.

Due to the lack of ground-truth datasets, we used Wald’s protocol [77] to obtain the downsampling version of both MS and PAN images (with a downsampling factor equal to 4). The degraded MS images were upscaled to the original size so that the MS image could serve as the HR MS (ground-truth) image for quality assessment.

In our study, to quantitatively evaluate the pansharpened results, six metrics with reference and a metric without reference were employed to evaluate the products at reduced resolution and at full resolution, respectively. For the metrics with reference, the correlation coefficient (CC) [12], the relative dimensionless error in synthesis (ERGAS) [78], the $Q_4$ [79], the spectral angle mapper (SAM) [80], the relative average spectral error (RASE) [81], and the root mean squared error (RMSE) were used. Additionally, the metric without reference consisted of the combination of a spectral distortion ($D_s$) index and a spatial distortion ($D_s$) index. It is commonly-used and known as a quality with no reference (QNR) index [82,83]. In general, the largest values of CC, $Q_4$, and QNR indicate the best performance. On the other hand, the smaller the ERGAS, SAM, RASE, and RMSE are, the closer the results are to the ground-truth (reference) image.

4.2. Implementation Details

We trained the proposed framework using PyTorch [84]. The configuration of the hardware device was composed of an NVIDIA GTX 1070Ti (GPU), 48-GB RAM (Memory), and an Intel Core i5-8500 (CPU). For the training phase, an Adam optimization algorithm is used. The initial learning rate, the moment decay, and the batch size were set to 0.0001, 0.9, and 4, respectively. The input size of the LR MS image and the PAN image were set to $64 \times 64$ and $256 \times 256$, respectively. The parameters of each convolutional layer in the MSB block and the RB block are described in Table 1. The training epochs were 400.

<table>
<thead>
<tr>
<th>Layer Name</th>
<th>Kernel Size</th>
<th>Stride</th>
<th>Padding</th>
<th>Input Channels</th>
<th>Output Channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv0</td>
<td>3 × 3</td>
<td>1</td>
<td>1</td>
<td>1 (PAN) or # MS bands (MS)</td>
<td>32</td>
</tr>
<tr>
<td>Conv1</td>
<td>3 × 3</td>
<td>1</td>
<td>1</td>
<td>32</td>
<td>64</td>
</tr>
<tr>
<td>Conv2</td>
<td>3 × 3</td>
<td>1</td>
<td>1</td>
<td>64</td>
<td>32</td>
</tr>
<tr>
<td>Conv1_1</td>
<td>3 × 3</td>
<td>1</td>
<td>1</td>
<td>1 + # MS bands</td>
<td>32</td>
</tr>
<tr>
<td>ResNet1_1</td>
<td>3 × 3</td>
<td>1</td>
<td>1</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Conv2_1</td>
<td>3 × 3</td>
<td>1</td>
<td>1</td>
<td>32</td>
<td># MS bands</td>
</tr>
<tr>
<td>ResNet2_1</td>
<td>3 × 3</td>
<td>1</td>
<td>1</td>
<td>64</td>
<td>32</td>
</tr>
<tr>
<td>Conv2_2</td>
<td>3 × 3</td>
<td>1</td>
<td>1</td>
<td>32</td>
<td># MS bands</td>
</tr>
<tr>
<td>Conv(RB)</td>
<td>3 × 3</td>
<td>1</td>
<td>1</td>
<td>2 × # MS bands</td>
<td>32</td>
</tr>
<tr>
<td>ResNet(RB)</td>
<td>3 × 3</td>
<td>1</td>
<td>1</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Conv(RB)</td>
<td>3 × 3</td>
<td>1</td>
<td>1</td>
<td>32</td>
<td># MS bands</td>
</tr>
</tbody>
</table>

4.3. Results and Discussion

To fairly compare the effectiveness of the proposed MMFN, we selected thirteen state-of-the-art pansharpening methods, including seven widely used traditional ones, i.e., Inter23 [6], which is the simple upsampling to the PAN scale using a 23-tap polynomial interpolator, AWLP [6,85], BDSD [13], GSA [12], MMP [16], MTF-GLP-HPM (MGH) [21], and SIRF [33], and six deep learning-based ones, i.e., PNN [41], PanNet [43], PSGAN [47], ResTFNet-l1 [45], TACNN [44], and MUCNN [51]. All the experiments were conducted using the set of parameters as originally indicated by the authors in the related papers. In the following, we will present the results for the two classical assessments, i.e., at reduced and full resolutions.
4.3.1. Assessment at Reduced Resolution

In this section, we compare the performance at reduced resolution of the proposed method with the proposed benchmark. Figure 3–5 present the pansharpened results for the three datasets acquired by QuickBird, Pléiades, and WorldView-2, respectively. For the visual comparison, all the images are shown in true (RGB) color. Additionally, mean absolute errors (MAEs, by using heatmaps), namely, residual maps between pansharpened results and the related reference images, are also given. The results show that the deep learning-based methods outperform the traditional ones in terms of MAEs.

Figure 3. Qualitative comparison at reduced resolution on the QuickBird dataset. (a) Inter23, (b) AWLP, (c) BDSD, (d) GSA, (e) MMP, (f) MGH, (g) SIRF, (h) PNN, (i) PanNet, (j) PSGAN, (k) ResTFNet-1, (l) TACNN, (m) MUCNN, (n) MMFN, (o) Ground-Truth (GT). The last two rows are MAEs between pansharpened results and the related ground-truth images.

Figure 4. Qualitative comparison at reduced resolution on the Pléides dataset. (a) Inter23, (b) AWLP, (c) BDSD, (d) GSA, (e) MMP, (f) MGH, (g) SIRF, (h) PNN, (i) PanNet, (j) PSGAN, (k) ResTFNet-1, (l) TACNN, (m) MUCNN, (n) MMFN, (o) Ground-Truth (GT). The last two rows are MAEs between pansharpened results and the related ground-truth images.
Figure 5. Qualitative comparison at reduced resolution on the WorldView-2 dataset. (a) Inter23, (b) AWLP, (c) BDSD, (d) GSA, (e) MMP, (f) MGH, (g) SIRF, (h) PNN, (i) PanNet, (j) PSGAN, (k) ResTFNet-1, (l) TACNN, (m) MUCNN, (n) MMFN, (o) Ground-Truth (GT). The last two rows are MAEs between pansharpened results and the related ground-truth images.

In Figure 3, compared with the ground-truth image (Figure 3o), the traditional approaches, i.e., Inter23, AWLP, BDSD, GSA, MMP, MGH, suffer from various degrees of spatial detail loss, especially some serious blurring effects generated by Inter23 and GSA methods. This is because the Inter23-based method directly utilizes a polynomial kernel (with 23 coefficients) to upsample the LRMS image without any information from the PAN image. The GSA-based method may fail to estimate the high-frequency details during the component replacement operation. Deep learning-based techniques also show poor performance in preserving spatial details, as seen from the results produced by PNN, PanNet, TACNN, and MUCNN. Although PSGAN and ResTFNet-1 can obtain promising pansharpened results that are close to the ground-truth image from a visual point of view, some differences still exist when observing the residual maps. By contrast, our MMFN method can achieve the best trade-off in preserving spatial details and spectral signatures, which are the closest with respect to the ground-truth ones. Table 2 reports the average quantitative performance on the QuickBird dataset including 40 groups of images by using different pansharpening methods. As shown in the table, the proposed MMFN achieves the best results for all the metrics, indicating that our approach has outstanding sharpening capabilities for reduced resolution images. The ResTFNet is also a promising method. From both a quantitative and qualitative point of view, we can state that our proposed method can achieve better pansharpening outcomes.

Similarly, in Figures 4 and 5, the pansharpening results for Pléiades and WorldView-2 are depicted. Again, it is easy to observe that the performance of traditional methods is inferior with respect to deep learning-based ones. In Figure 4, the conventional methods, except for Inter23, show various degrees of spectral distortion, mainly in the wooded areas. Moreover, the color of the road is also inconsistent with the reference image. It can be found that the BDSD and GSA methods generate block effects in the vegetation areas. Deep learning-based methods instead produce better results (closer to the ground-truth image). However, spatial distortion still exists, as seen from the MAE maps. Although the performance of PSGAN and the proposed approach are similar at first glance, the residual map of the proposed approach is closer to zero, see the red boxes in the figure. In Figure 5, the SIRF method can efficiently perform spatial details enhancement. However, it suffers from severe spectral distortion, see Figure 5g. Compared with the ground-truth image, the results of AWLP, BDSD, PNN, and TACNN generate visible color distortion. Additionally, the benchmarking methods still have a loss of spatial information on the WorldView-2
dataset in terms of MAEs, see, e.g., the MUCNN method. Thus, from this analysis, it is clear that the best performance is obtained by the proposed MMFN. This statement is further corroborated by the quantitative results in Tables 3 and 4. Indeed, the proposed MMFN method achieves the best results on both the Pliéades and WorldView-2 datasets and for all the metrics. In conclusion, the generalization ability of the proposed method for reduced-resolution images is better than that of the state-of-the-art methods.

Table 2. Average quantitative performance on the QuickBird dataset. The best and second best performances are shown in Red and blue, respectively.

<table>
<thead>
<tr>
<th>Methods</th>
<th>CC (%)</th>
<th>ERGAS (%)</th>
<th>Q2E (%)</th>
<th>RASE (%)</th>
<th>RMSE (%)</th>
<th>SAM (%)</th>
<th>D1 (%)</th>
<th>QNR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter3</td>
<td>0.8746 ± 0.0070</td>
<td>2.2640 ± 0.0700</td>
<td>0.0599 ± 0.0726</td>
<td>9.2806 ± 5.3302</td>
<td>46.4564 ± 11.2309</td>
<td>0.0194 ± 0.0158</td>
<td>0.0234 ± 0.0153</td>
<td>0.0319 ± 0.0153</td>
</tr>
<tr>
<td>MMFN (ours)</td>
<td>0.8959 ± 0.0106</td>
<td>2.5690 ± 0.0700</td>
<td>0.0564 ± 0.0726</td>
<td>8.0769 ± 5.3623</td>
<td>22.5130 ± 5.3623</td>
<td>0.0274 ± 0.0146</td>
<td>0.0376 ± 0.0156</td>
<td>0.0504 ± 0.0156</td>
</tr>
<tr>
<td>MAVE</td>
<td>0.9127 ± 0.0135</td>
<td>1.9512 ± 0.0700</td>
<td>0.0516 ± 0.0726</td>
<td>6.5854 ± 5.2402</td>
<td>19.1297 ± 5.1857</td>
<td>0.0240 ± 0.0136</td>
<td>0.0315 ± 0.0136</td>
<td>0.0394 ± 0.0136</td>
</tr>
<tr>
<td>BOD</td>
<td>0.8279 ± 0.0075</td>
<td>1.8709 ± 0.0700</td>
<td>0.0511 ± 0.0407</td>
<td>7.5996 ± 5.2623</td>
<td>20.0394 ± 5.1296</td>
<td>0.0176 ± 0.0146</td>
<td>0.02176 ± 0.0150</td>
<td>0.0346 ± 0.0156</td>
</tr>
<tr>
<td>GSA</td>
<td>0.8221 ± 0.0082</td>
<td>1.8557 ± 0.0702</td>
<td>0.0324 ± 0.0434</td>
<td>7.3552 ± 5.2076</td>
<td>23.1829 ± 5.1290</td>
<td>0.0259 ± 0.0136</td>
<td>0.0316 ± 0.0136</td>
<td>0.0427 ± 0.0136</td>
</tr>
<tr>
<td>MMP</td>
<td>0.8805 ± 0.0126</td>
<td>1.4713 ± 0.0709</td>
<td>0.0344 ± 0.0607</td>
<td>6.1532 ± 2.4399</td>
<td>16.9579 ± 9.7879</td>
<td>0.0262 ± 0.0112</td>
<td>0.0302 ± 0.0112</td>
<td>0.0413 ± 0.0112</td>
</tr>
<tr>
<td>MGH</td>
<td>0.9038 ± 0.0112</td>
<td>1.6173 ± 0.0320</td>
<td>0.0301 ± 0.0156</td>
<td>6.2570 ± 2.1460</td>
<td>16.2500 ± 9.6539</td>
<td>0.0207 ± 0.0095</td>
<td>0.0267 ± 0.0095</td>
<td>0.0391 ± 0.0095</td>
</tr>
<tr>
<td>SRF</td>
<td>0.8226 ± 0.0104</td>
<td>2.4460 ± 0.0215</td>
<td>0.0037 ± 0.1247</td>
<td>9.1692 ± 5.6585</td>
<td>28.3468 ± 6.1354</td>
<td>0.0111 ± 0.0114</td>
<td>0.0666 ± 0.0141</td>
<td>0.0706 ± 0.0141</td>
</tr>
</tbody>
</table>

Table 3. Average quantitative performance on the Pliéades dataset. The best and second best performances are shown in Red and blue, respectively.

<table>
<thead>
<tr>
<th>Methods</th>
<th>CC (%)</th>
<th>ERGAS (%)</th>
<th>Q2E (%)</th>
<th>RASE (%)</th>
<th>RMSE (%)</th>
<th>SAM (%)</th>
<th>D1 (%)</th>
<th>QNR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter3</td>
<td>0.8576 ± 0.0075</td>
<td>1.9103 ± 0.0700</td>
<td>0.0546 ± 0.0202</td>
<td>8.5648 ± 5.2402</td>
<td>19.1297 ± 5.1857</td>
<td>0.0240 ± 0.0136</td>
<td>0.0315 ± 0.0136</td>
<td>0.0394 ± 0.0136</td>
</tr>
<tr>
<td>MMFN (ours)</td>
<td>0.8959 ± 0.0075</td>
<td>1.8709 ± 0.0700</td>
<td>0.0511 ± 0.0407</td>
<td>7.5996 ± 5.2623</td>
<td>20.0394 ± 5.1296</td>
<td>0.0176 ± 0.0146</td>
<td>0.02176 ± 0.0150</td>
<td>0.0346 ± 0.0156</td>
</tr>
<tr>
<td>MAVE</td>
<td>0.8953 ± 0.0156</td>
<td>1.6041 ± 0.0301</td>
<td>0.0060 ± 0.0156</td>
<td>6.0424 ± 5.2402</td>
<td>17.3820 ± 5.0567</td>
<td>0.0203 ± 0.0107</td>
<td>0.0248 ± 0.0108</td>
<td>0.0345 ± 0.0108</td>
</tr>
<tr>
<td>RSFNet1</td>
<td>0.8953 ± 0.0075</td>
<td>2.6738 ± 0.0700</td>
<td>0.0758 ± 0.0111</td>
<td>3.4364 ± 0.1020</td>
<td>33.0000 ± 4.4114</td>
<td>0.0203 ± 0.0104</td>
<td>0.0226 ± 0.0104</td>
<td>0.0335 ± 0.0104</td>
</tr>
<tr>
<td>TACNN</td>
<td>0.8902 ± 0.0025</td>
<td>1.1075 ± 0.0202</td>
<td>0.0359 ± 0.0171</td>
<td>4.2760 ± 0.1026</td>
<td>9.2120 ± 0.1026</td>
<td>0.0099 ± 0.0071</td>
<td>0.0125 ± 0.0071</td>
<td>0.0201 ± 0.0071</td>
</tr>
<tr>
<td>MUCNN</td>
<td>0.8910 ± 0.0135</td>
<td>1.0258 ± 0.0301</td>
<td>0.0060 ± 0.0103</td>
<td>4.0750 ± 0.1026</td>
<td>11.9302 ± 0.1026</td>
<td>0.0126 ± 0.0071</td>
<td>0.0126 ± 0.0071</td>
<td>0.0219 ± 0.0071</td>
</tr>
<tr>
<td>MMN (ours)</td>
<td>0.9033 ± 0.0049</td>
<td>0.9348 ± 0.0275</td>
<td>0.0779 ± 0.0102</td>
<td>3.6003 ± 0.0483</td>
<td>9.4264 ± 0.3155</td>
<td>0.0102 ± 0.0063</td>
<td>0.0205 ± 0.0072</td>
<td>0.0273 ± 0.0054</td>
</tr>
</tbody>
</table>

Table 4. Average quantitative performance on the WorldView-2 dataset. The best and second best performances are shown in Red and blue, respectively.

<table>
<thead>
<tr>
<th>Methods</th>
<th>CC (%)</th>
<th>ERGAS (%)</th>
<th>Q2E (%)</th>
<th>RASE (%)</th>
<th>RMSE (%)</th>
<th>SAM (%)</th>
<th>D1 (%)</th>
<th>QNR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter3</td>
<td>0.8959 ± 0.0075</td>
<td>2.1371 ± 0.3277</td>
<td>0.9351 ± 0.0317</td>
<td>5.1701 ± 1.1225</td>
<td>20.3643 ± 6.6001</td>
<td>0.2281 ± 0.0401</td>
<td>0.1723 ± 0.0401</td>
<td>0.2235 ± 0.0401</td>
</tr>
</tbody>
</table>

4.3.2. Full Resolution Evaluation

We further assess the proposed method at full resolution using datasets at the original scale with no spatial degradation to generate the ground-truth (reference) image. Figures 6–8 show the qualitative comparison. The D3, D4, and QNR metrics were used to evaluate the pansharpened performance. Inter3 obtains the best D3 value, which indicates the absence of spectral distortion. Therefore, we used the results of the Inter3-based method as the basis images from which to observe the detailed information of PAN images injected into the output images. For visual comparison, we still adopted a residual representation calculating the difference between the pansharpened result and Inter3 to display the injected details. It is observed that the DL-based methods inject more structures and textures than the compared traditional methods, see, e.g., the residual images for the
QuickBird dataset in Figure 6. BDSD transfers the wrong details into the pansharpened product, leading to noise effects on green areas, see Figure 6c. From a visual point of view, lawn areas should be smooth without containing any detail. Significant spectral distortion still exists in the result of the SIRF-based method, see Figure 6g. Among the traditional methods, the AWLP, GSA, MMP, and MGH-based approaches can hardly transfer precise edges into the pansharpened results, as shown in the related MAEs in Figure 6. It is clear from Figure 6j that PSGAN produces spectral distortion in the buildings with an orange-like color (the right side of the image). A similar phenomenon is that partial edges around green areas can be retained in the MAE results of PNN, PanNet, and TACNN-based methods. Although the ResTFNet, MUCNN, and the proposed method have highly similar results, the blurring effects and the noise cannot be removed in green vegetation areas. Objective evaluations are reported in the last three columns of Table 2. We find that the proposed method obtains a high performance in terms of $D_S$, and almost the best QNR metrics. Overall, the full-resolution experiments on the QuickBird dataset are satisfying.

A similar visual analysis can be performed in Figures 7 and 8 for the other two datasets (i.e., Pléiades and WorldView-2). It is easy to see that the proposed MMFN achieves the best trade-off between spatial and spectral consistency. Furthermore, objective results mark the high performance of the proposed approach; see the last three columns of Tables 3 and 4, again. Indeed, the proposed MMFN has a promising fusion performance, balancing the trade-off between spatial enhancement and spectral fidelity, thanks to the use of the multi-scale information and the multi-stream fusion strategy.

4.3.3. Parameter Numbers vs. Model Performance

We added a comparative experiment evaluating the running times of deep learning methods. We conducted experiments on a hardware device with an Intel(R) Xeon(R) Gold 5117 CPU @ 2.00 GHz and 128 GB memory. We used Python 3.9 and PyTorch 1.11 as the programming language and deep learning framework, respectively. We tested running times for both full and reduced resolution MS images with a batch size of 4. We measured the average execution times for processing a single input on the CPU. We used the Python built-in timer to measure the execution times, repeating each measurement 10 times to
ensure statistical significance. The results are shown in Table 5. Although our method is not the most time-efficient one (because it processes more features generated at different scales), it exploits fewer parameters than PSGAN and ResTFNet-1, thus having a shorter running time. Furthermore, while our method is not as time-efficient as PNN, PanNet, MUCNN, and TACNN, it achieves a better pansharpening performance.

Figure 7. Qualitative comparison at full resolution on the Pléiades dataset. (a) Inter23, (b) AWLP, (c) BDSD, (d) GSA, (e) MMP, (f) MGH, (g) SIRF, (h) PNN, (i) PanNet, (j) PSGAN, (k) ResTFNet-1, (l) TACNN, (m) MUCNN, (n) MMFN (Our). The last two rows are MAEs between the pansharpened results and the Inter23 outcome.

Figure 8. Qualitative comparison at full resolution on the WorldView-2 dataset. (a) Inter23, (b) AWLP, (c) BDSD, (d) GSA, (e) MMP, (f) MGH, (g) SIRF, (h) PNN, (i) PanNet, (j) PSGAN, (k) ResTFNet-1, (l) TACNN, (m) MUCNN, (n) MMFN (Our). The last two rows are MAEs between the pansharpened results and the Inter23 outcome.
Table 5. Parameter numbers and running times for the compared deep learning methods on datasets at different scales.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Scale (Time)</th>
<th># Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RR (s)</td>
<td>FR (s)</td>
</tr>
<tr>
<td>PNN</td>
<td>0.14</td>
<td>1.67</td>
</tr>
<tr>
<td>PanNet</td>
<td>0.26</td>
<td>3.84</td>
</tr>
<tr>
<td>PsGan</td>
<td>1.11</td>
<td>13.77</td>
</tr>
<tr>
<td>ResTFNet-1</td>
<td>0.85</td>
<td>12.00</td>
</tr>
<tr>
<td>TACNN</td>
<td>0.15</td>
<td>1.72</td>
</tr>
<tr>
<td>MUCNN</td>
<td>0.51</td>
<td>7.21</td>
</tr>
<tr>
<td>MMFN (ours)</td>
<td>0.62</td>
<td>8.97</td>
</tr>
</tbody>
</table>

4.4. Ablation Studies

In this section, we provide three groups of ablation studies to demonstrate each claim of the proposed framework.

4.4.1. Ablation Study about Different Scales

To investigate the influence of the number of scales on the pansharpening performance, we compared the results of six different scales by using a fixed multi-stream fusion network. From Figure 9, it can be observed that increasing the number of scales cannot continuously improve the pansharpening performance in terms of objective evaluation. Most evaluation metric values indicate that the performance of the multi-scale (scale number = 3) results is better than that of the other scales. Additionally, as the number of scales increases, the processing time increases. Therefore, the scale number equal to three represents a good choice for our approach.

As shown in Figure 10, by observing the residual maps between the results of the different scales and the reference image, the spectral distortion and spatial distortion can be significantly reduced, increasing the number of scales. It is also straightforward to see that three scales represent the best result among the three presented in the figure.

4.4.2. Ablation Study about Different Streams

To verify the effectiveness of the choice of a number of streams equal to three, the single-stream (stream number = 1) fusion and the dual-stream (stream number = 2) fusion were investigated, fixing the multi-scale network.

The quantitative comparison is presented in Figure 11. The results of the multi-stream block using three streams are better than those of single-stream and dual-stream fusion blocks. This is because the sole use of an architecture with single-stream or dual-stream fusion cannot guarantee the full merge of the spatial details of the PAN image together with the spectral information of the MS image. In addition, Figure 12 shows the qualitative results on the three real remote sensing datasets, varying the number of streams. As the number of streams increases, the spectral and spatial distortions decrease. Thus, the above-mentioned ablation studies confirm the validity of the choices made (the number of scales and streams both equal to three).
Figure 9. Quantitative fusion results by varying the number of scales on the three real remote sensing datasets.

Figure 10. Pansharpening results by using a different number of scales on the three real remote sensing datasets. From top to bottom: fusion results using one, two, and three scales, respectively, and the ground-truth images. From left to right: (a) QuickBird dataset, (b) MAE map of (a), (c) Pléiades dataset, (d) MAE map of (c), (e) WorldView-2 dataset, and (f) MAE map of (e).
Figure 11. Quantitative fusion results by varying the number of streams on the three real remote sensing datasets.

Figure 12. Pansharpening results by using a different number of streams on the three real remote sensing datasets. From top to bottom: fusion results using one, two, and three streams, respectively, and the ground-truth images. From left to right: (a) QuickBird dataset, (b) MAE map of (a), (c) Pléiades dataset, (d) MAE map of (c), (e) WorldView-2 dataset, (f) MAE map of (e).
4.4.3. Ablation Study about Different numbers of Residual Blocks

To assess the effect of the number of residual blocks in the MSB module on the results, we experimentally tested the impact of using one, two, and three residual blocks. The experimental results shown in Figure 13 indicate that increasing the number of residual blocks leads to a slight improvement in performance on WorldView-2 and QuickBird datasets. However, on the Pléiades dataset, there was a decrease in performance as the number of residual blocks increases. This is because the simplest model structure is enough for the Pléiades dataset. Considering that increasing the number of residual blocks also increases the complexity and computational cost of the network, leading to overfitting and longer training times, the simplest structure was chosen.

![Figure 13](image-url)

**Figure 13.** Quantitative fusion results by varying the number of residual blocks on the three real remote sensing datasets.

5. Conclusions

This paper proposed a multi-scale and multi-stream fusion network (MMFN), which is simple but effective for pansharpening. Different levels of spectral/spatial information contained in the MS and PAN images was obtained first by using a multi-scale strategy. Afterwards, a multi-stream fusion block was adopted to fully fuse the MS and PAN images, preserving their spatial and spectral characteristics. Additionally, to constrain the training process, we developed a multi-stage reconstruction approach, setting the same loss function for each fusion block. The proposed method was assessed on three real remote sensing datasets.
datasets acquired by three different sensors. The reduced and full resolution assessments demonstrated the validity of the proposed approach.

**Author Contributions:** Conceptualization, L.J. and S.W.; methodology, L.J.; software, S.W.; validation, R.R., L.C. and D.Z.; formal analysis, S.W.; investigation, G.V.; resources, D.Z.; data curation, S.W.; writing—original draft preparation, L.J.; writing—review and editing, L.J., R.R. and G.V.; visualization, L.J. and R.R.; supervision, S.W. and G.V.; project administration, L.J.; funding acquisition, L.J. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by National Natural Science Foundation of China (NSFC) under Grant 62101502, in part by China Postdoctoral Science Foundation under Grant 2022T150596, and in part by Postdoctoral Research Grant in Henan Province under Grant 202101012.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**


36. Deng, L.J.; Feng, M.; Tai, X.C. The fusion of panchromatic and multispectral remote sensing images via tensor-based sparse modeling and hyper-Laplacian prior. *Inf. Fusion* 2019, 52, 76–89. [CrossRef]


45. Liu, X.; Liu, Q.; Wang, Y. Remote sensing image fusion based on two-stream fusion network. *Inf. Fusion* 2020, 55, 1–15. [CrossRef]


Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.