remote sensing

MDPI

Article

Detail Information Prior Net for Remote Sensing Image Pansharpening

Yuchen Xie [1], Wei Wu [1,2,*], Haiping Yang [1], Ning Wu [2] and Ying Shen [1]

[1] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China; xieyuchen@zjut.edu.cn (Y.X.); yanghp@zjut.edu.cn (H.Y.); shenying@zjut.edu.cn (Y.S.)
[2] Research Center of Digital Space Technology, Southeast Digital Economic Development Institute, Quzhou 324014, China; qzxy_wn@foxmail.com
* Correspondence: wuwei@zjut.edu.cn

Abstract: Pansharpening, which fuses the panchromatic (PAN) band with multispectral (MS) bands to obtain an MS image with spatial resolution of the PAN images, has been a popular topic in remote sensing applications in recent years. Although the deep-learning-based pansharpening algorithm has achieved better performance than traditional methods, the fusion extracts insufficient spatial information from a PAN image, producing low-quality pansharpened images. To address this problem, this paper proposes a novel progressive PAN-injected fusion method based on superresolution (SR). The network extracts the detail features of a PAN image by using two-stream PAN input; uses a feature fusion unit (FFU) to gradually inject low-frequency PAN features, with high-frequency PAN features added after subpixel convolution; uses a plain autoencoder to inject the extracted PAN features; and applies a structural similarity index measure (SSIM) loss to focus on the structural quality. Experiments performed on different datasets indicate that the proposed method outperforms several state-of-the-art pansharpening methods in both visual appearance and objective indexes, and the SSIM loss can help improve the pansharpened quality on the original dataset.

Keywords: image fusion; pansharpening; feature fusion unit; superresolution

Citation: Xie, Y.; Wu, W.; Yang, H; Wu, N.; Shen, Y. Detail Information Prior Net for Remote Sensing Image Pansharpening. *Remote Sens.* **2021**, *13*, 2800. https://doi.org/10.3390/rs13142800

Academic Editor: Angel D. Sappa

Received: 2 June 2021
Accepted: 13 July 2021
Published: 16 July 2021

1. Introduction

The sensors onboard satellite platforms record the digital number of land surfaces in different spectral channels. The acquired images have formed the basis for mapping different land surfaces. Thus, the spectral parameters of an image, such as the number of spectral channels, channel width, and mid-bandwidth, are important for evaluating the quality of remote sensing imagery. The spatial resolution, which is the area of the land surface represented by a pixel in remote sensing imagery, is another important parameter. High-resolution remote sensing imagery can distinctly describe the distribution and structure in a land surface, which forms the basis for fine surface mapping. Therefore, obtaining imagery with high spatial and spectral resolutions will enrich the information content in imagery and enhance the capacity for identifying various land surfaces.

Due to the limitations imposed by the data volume collected by the sensor, the data transmission between a satellite and Earth, and the incoming radiation energy into sensors within surface units [1], it is exceedingly difficult to obtain imagery with high spatial and spectral resolution. To address these problems, one panchromatic (PAN) band and multiple multispectral (MS) bands can be used when installing several different spectrum monitors for a sensor. Pansharpening, which can overcome the shortcomings of sensors, increases the spectral resolution of a PAN band by integrating it with MS bands. This process can also be viewed as an enhancement of the spatial resolution of the MS bands, with the optimization objective of maintaining their spectral features while increasing their spatial resolution. To date, pansharpening has become an important technique for processing

remote sensing data. Based on whether deep learning (DL) is used, pansharpening methods can be categorized into conventional methods and DL-based methods.

Component substitution (CS)-based methods are one type of conventional pansharpening method. Considering that a PAN band receives a relatively broad range of wavelengths, generally covering the wavelength range of visible light, it is strongly correlated with the luminous intensity (I) component of the intensity-hue-saturation (IHS) color space. Based on a previous assumption, the IHS replacement method [2] first transforms an MS image into an IHS space and subsequently replaces the I component with a PAN band, thereby imparting the high-resolution information carried by the PAN band to the MS image and realizing pansharpening. As the formation mechanism of the reflectance varies between two land surfaces, the IHS replacement method is prone to causing color distortions in some land surface features. The principal component analysis (PCA) method [3] converts an MS image to multiple independent components that contain the main land surface information and subsequently replaces the first principal component with a PAN band to produce a sharpened MS image. However, high computational costs and poor real-time performance pose challenges for application of the PCA method to image pansharpening. The Gram–Schmidt adaptive (GSA) method [4] and the partial replacement adaptive component substitution (PRACS) method [5] are two improved CS-based pansharpening methods.

Another commonly employed pansharpening method is the multiresolution analysis (MRA)-based method. First, a PAN image and a low-resolution MS image are decomposed into one group of low-frequency images and one group of high-frequency images; second, the images at each corresponding scale are fused by using a combing algorithm; and last, the images are fused again at the original scale to produce a sharpened image. MRA-based combining methods require pyramid processing algorithms, including Laplacian pyramid transform, wavelet transforms, and so on. Two such algorithms include the generalized Laplacian pyramid with modulation transfer function-matched filter and regression-based injection model (MTF-GLP-CBD) [6] and the à Trous wavelet transform (ATWT) [7].

Because the differences among the regions, resolutions, spectral channels, and resolution conversion relationships for the channels are not adequately described by simple linear equations, spectral distortions appear in the pansharpened images. Characterized by nonlinear activation functions and multilayer convolution operations, a DL-based method, the convolutional neural network (CNN), has been extensively applied in areas that require the establishment of complex nonlinear relationships (e.g., pansharpening) in recent years. The CNN is capable of adaptively establishing complex relationships by supervised learning.

DL-based pansharpening methods have produced good results in applications. However, available methods upscale an MS image to the size of a PAN image simply by interpolation in a preprocessing procedure and fuse interpolated MS images with PAN images (e.g., CS-based and DL-based methods). The information content in features fused using multiscale fusion methods (e.g., MRA-based method) is limited, which generates some distorted results. Frontier research [8] is also exploring when to fuse or extract multiresolution features. In view of these two problems, this study presents a new progressive PAN-injected fusion method based on SR for remote sensing imagery, referred to as detail information prior net (DIPNet). The main contributions of this study are summarized as follows:

(1) We use two-stream PAN input to extract PAN features by using a convolution network.
(2) We use the feature fusion unit (FFU) to gradually inject low-frequency PAN features, and high-frequency PAN features are added after subpixel convolution to perfect an upsampled MS image.
(3) We use a plain autoencoder to inject the extracted PAN features.
(4) We use the structural similarity index measure (SSIM) [9] loss to guide the network during training, focusing on the structural quality.

The remainder of this paper is organized as follows: Section 2 presents related works in this study. Section 3 details the pansharpening method proposed in this study. Section 4 introduces the experimental data used in this study as well as the methods applied to

evaluate the results. In Section 5, experimental results and comparisons are presented. Section 6 focuses on the discussion and evaluation of the results. Section 7 concludes the paper.

2. Background and Related Work

2.1. Image Upscaling and Pyramid Processing Algorithm

Interpolation-based (e.g., nearest neighbor and bilinear) image upscaling algorithms are prone to blurring images after increasing their size. This phenomenon becomes more pronounced as the upscaling factor increases, mainly due to a lack of high-frequency, detailed spatial information after image upscaling. This blurring phenomenon is similarly associated with the CNN-based SR method. However, the CNN-based SR method is capable of adaptively adjusting the SR equation based on the image content and, consequently, inhibiting blurring to a certain extent. Nevertheless, the CNN-based SR method is unable to completely eliminate blurring.

The Gaussian–Laplacian pyramid-based processing algorithm integrates high- and low-frequency information at multiple scales and has shown relatively good performance in the merging and fusion of images. Similarly, for pansharpening, detailed features can be restored by adding the high-frequency information contained in a PAN image to an SR-upscaled MS image at multiple scales. This study presents a pansharpening method referred to as DIPNet that uses high-frequency, detailed information and the fusion of low-frequency PAN information.

2.2. Deep-Learning-Based Pansharpening

Based on the selected model, DL-based pansharpening methods can be categorized into four main types, namely image-feature-based methods, autoencoders, SR methods, and GANs.

Image-feature-based methods: Image-feature-based pansharpening methods have effective network architectures designed to correspond to the features of fused images. An MSDCNN [10], which involves the use of convolution kernels of varying sizes to extract multiscale features to improve the fusion performance, was designed to take full advantage of the multiscale information contained in image features. A network for pansharpening, referred to as PanNet [11], was proposed to improve the fusion of high-resolution satellite imagery. In the PanNet architecture, high-frequency image information is employed to train a residual network (ResNet) [12] to obtain the details missing in low-resolution images. Based on the general idea of PanNet, a deep multiscale detail network (DMDNet) [13] was designed by superseding the conventional residual module with a grouped, multiscale dilated convolutional residual module. The performance of DMDNet is superior to that of PanNet in migration, fusion, and reconstruction. Moreover, in the field of image restoration, You Only Look Yourself (YOLY) [14] uses image features to design an unsupervised image dehaze model. Therefore, the design of the network, which is based on image features, can achieve improved performance.

Autoencoders: An autoencoder converts an input image to deep features through a series of nonlinear mapping operations and subsequently restores the original image by decoding. Sparse [15] and convolutional [16] autoencoders encode the PAN and MS bands into sparse matrix features and subsequently enhance the spatial resolution of the MS image with the PAN image by decoding. TFNet [17] is a pansharpening network based on a convolutional autoencoder. Through a two-stream architecture, TFNet extracts features from an MS image and a PAN image and ultimately reconstructs a high-spatial-resolution, high-spectral-resolution image using a decoder.

SR methods: SR-based pansharpening methods view pansharpening as an SR problem in MS bands under the constraints of a PAN band. Using the SRCNN [18] architecture, the PNN method [19] integrates the spatial information in a PAN band during the SR process by introducing upsampled MS and PAN information and produces results superior to those produced by conventional methods. To further enhance the spatial resolution

of imagery, a deep residual PNN [20] method was introduced by improving the PNN method with the ResNet architecture. A bidirectional pyramid network [21] extracts features from a PAN image by convolution operations and produces good pansharpening results by subpixel convolutional SR fusion of MS and PAN image features at corresponding scales. The PCDRN [22] method progressively fuses images through ResNets and interpolation based on the scale relationship between MS images and PAN images. The PCDRN method has shown good fusion performance in high-resolution satellite imagery. The SR-guided progressive pansharpening based on a deep CNN (SRPPNN) [23] method upscales a low-resolution MS image by progressive SR and integrates it with a multiscale, high-frequency PAN image. This method has yielded good results in the pansharpening of remote sensing imagery.

GANs: The GAN architecture contains a generator coupled with a discriminator and achieves collaborative optimization through adversarial training. This architecture has achieved good results in areas such as image generation and style transfer. As pansharpening can be viewed as an image generation problem, deep networks based on the GAN architecture can be also employed in pansharpening. For example, the pansharpening GAN (PSGAN) [24] reconstructs high-spatial-resolution multiband images with TFNet as a generator and a conditional discriminator. Similarly, through the improvement of the PSGAN architecture, a residual encoder-decoder conditional GAN [25] was designed to further enhance the capacity to fuse remote sensing imagery. GAN-based pansharpening methods can help to describe the nonlinear mapping relationships among remote sensing images and produce relatively good results.

3. Method

3.1. Framework of the Method

The core ideas of the DIPNet are described as follows:

(1) A PAN band contains potential information in the MS bands. Low-frequency PAN information can reflect the main MS information. High-frequency PAN information can reflect the details in the PAN band.
(2) In this study, pansharpening is viewed as a PAN band-guided SR problem. High-frequency PAN information is added to ameliorate the SR-induced blurring problem.
(3) Multiple SR processes are required to obtain an MS image with the same spatial resolution in the PAN band. In conventional methods, single-scale fusion is inordinately simple, while features fused at multiple scales have limited information content. Multiscale high- and low-frequency deep PAN and MS features can be combined to better describe the mapping relationship between PAN bands and MS bands and to achieve higher-accuracy pansharpening.
(4) A multiscale auxiliary encoder with detailed PAN information and potential MS information in the PAN band is used to further reconstruct spatial information for the MS image.

For clarity, Figure 1 shows the workflow of our proposed work. Figure 2 shows the network architecture designed in this study based on the abovementioned ideas.

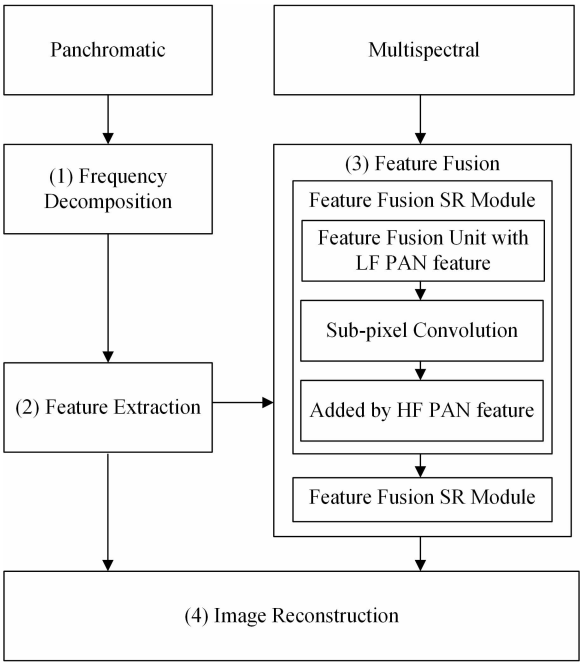

Figure 1. The workflow of the DIPNet.

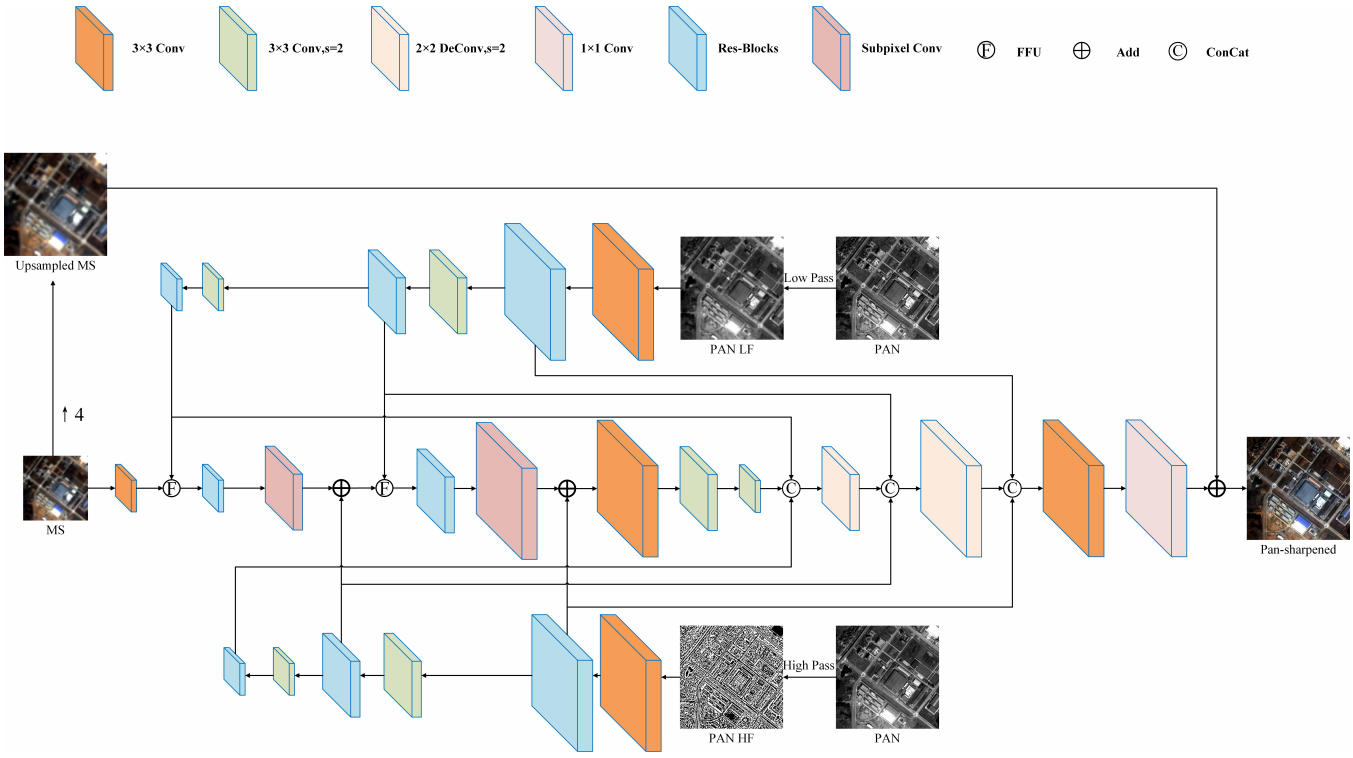

Figure 2. The detailed architecture of our proposed work in this study. The PAN image is decomposed into a high-pass component and low-pass component, which are two-stream PAN extracting inputs. They are then injected into the SR process and plain autoencoder of an MS image.

To facilitate the description of the problem, let h and l be the spatial resolutions of the PAN image P and the MS image M, respectively. To ensure a clear discussion, the PAN image and MS image are denoted P^h and M^l, respectively. Pansharpening fuses these two images into an MS image $\widetilde{M}^h$ with a spatial resolution of h. DIPNet involves four main steps:

(1) Frequency decomposition. In this step, P^h is decomposed into a high-pass component P_H^h and a low-pass component P_L^h. P_H^h reflects the high-frequency details (e.g., boundaries) of P^h. P_L^h reflects the complete spectral features (e.g., color features in a relatively large local area) of P^h. Frequency spectrum decomposition is achieved by Gaussian filtering. First, a Gaussian filter matrix with a window size of W_r is established. Second, P^h is filtered, and the result is treated as P_L^h. The difference between P^h and P_L^h is treated as P_H^h.

$$P_H^h = P^h - P_L^h \tag{1}$$

(2) Feature Extraction. Features are extracted from P_H^h and P_L^h using a 3 × 3 convolution operation followed by the ResNet module. Features $F(P_H^h)$ and $F(P_L^h)$, each with a spatial resolution of h and a total of K channels, are thus obtained. In addition, features are extracted from M^l using a 3 × 3 convolution operation. MS image features $F(M^l)$ with a total of K channels and a resolution of l are also obtained.

In many cases, the spatial-resolution multiples (e.g., two, four, or eight iterations) vary between a PAN image and MS image. In each convolutional downsampling operation, the output feature size is half the input feature size. The number of downsampling iterations required to downsample the PAN image to the spatial resolution in the MS bands also varies. For ease of the description of the problem, let m be the intermediate resolution. For example, when the resolution ratio of an MS image to a PAN image is 4, l is 4, m is 2, and h is 1; when the resolution ratio of an MS image to a PAN image is 6, l is 6, m is 2, and h is 1. This paper discusses a situation in which the resolution ratio of an MS image to a PAN image is 4, which is suitable for most high-resolution satellite images.

$F(P_H^h)$ and $F(P_L^h)$ are downsampled by a 3 × 3 convolution operation with a step size of 2. Features are further extracted using the ResNet module. A high-pass component and a low-pass component, each with a spatial resolution of m, are thus obtained; they are denoted as $F(P_H^m)$ and $F(P_L^m)$, respectively. Similarly, a high-pass component and a low-pass component, each with a spatial resolution of l, can be obtained; they are denoted as $F(P_H^l)$ and $F(P_L^l)$, respectively.

From this process, a low-frequency PAN information feature group $F(P_L^{h,m,l})$ and a high-frequency PAN information feature group $F(P_H^{h,m,l})$ are obtained:

$$F(P_L^{h,m,l}) = \{F(P_L^h), F(P_L^m), F(P_L^l)\} \tag{2}$$

$$F(P_H^{h,m,l}) = \{F(P_H^h), F(P_H^m), F(P_H^l)\} \tag{3}$$

(3) Feature Fusion (FF). $F(P_L^l)$ and $F(M^l)$ are fused using an FFU. Features $F(M_F^m)$ with a resolution of m are obtained by SR and subsequently added to $F(P_H^m)$ pixel by pixel. This process is repeated, and ultimately, MS features $F(M_F^h)$ with a resolution of h are obtained.

In this process, MS features are fused with low- and high-frequency PAN features at multiple scales. Thus, progressively fused MS features have more information content than features extracted from an interpolation-upscaled MS image.

(4) Image Reconstruction. An autoencoder is used to reconstruct the structure based on $F(P_L^{h,m,l})$, $F(P_H^{h,m,l})$, and $F(M_F^h)$ (fused features obtained by FF-based SR). A PAN image with an enhanced spatial resolution is thus obtained. In this process, multiscale PAN features are injected into the decoder to further increase the information content of the MS image.

Regarding the network activation function, a leaky rectified linear unit with a parameter of 0.2 is set as the activation function for all the convolutional layers, except for the ResNet module and the subpixel and output convolutional layers for SR.

The following section introduces an FF-based SR module and image reconstruction module into which high- and low-frequency PAN information is injected.

3.2. FF-Based SR Module (Step 3)

Prior to SR, the FFU is used to fuse $F(P_L^l)$ and $F(M^l)$ in the following manner:

$$F(M_F^l) = Conv_{1\times1}(F(M^l)©F(P_L^l)©(F(M^l) \oplus F(P_L^l))) \tag{4}$$

where © represents an operation that connects feature images in series, $\oplus$ denotes an operation that adds feature images pixel by pixel, $Conv_{1\times1}$ is a convolution function with a convolution kernel size of 1, and $F(M_F^l)$ is the fused MS features (the subscript F indicates fused features). The FFU produces combined features with a total of $3K$ channels through a serial connection operation and subsequently performs a 1×1 convolution operation on the combined features to produce fused features with a total of K channels. Thus, the extracted features are linearly fused by using rich per-added features.

Subsequently, the ResNet module is used to extract features from $F(M_F^l)$:

$$F_{res}(M_F^l) = RBs(F(M_F^l)) \oplus F(M_F^l) \tag{5}$$

where RBs represents the extraction operations by a total of L ResNet modules. The input features $F(M_F^l)$ are added for residual learning. Thus, $F(M_F^l)$-based deep features $F_{res}(M_F^l)$ are obtained. For the extraction of PAN features in Step 2, the residual module similarly consists of a total of L ResNet modules.

Subpixel convolution [26] is an upsampling method based on conventional convolution and pixel arrangement in feature images and can be used to achieve image SR. Let r be the upscaling factor and $c \times h \times w$ (c, h, and w are the number of channels, height, and width, respectively) be the size of the initial input feature image $F_{res}(M_F^l)$. First, through convolution operations on $F_{res}(M_F^l)$, a total of r^2c convolution kernels is extracted, and an output feature image with a size of $r^2c \times h \times w$ (i.e., a total of $h \times w$ vectors each with a length of r^2c) is obtained. Second, all the vectors, each with a length of r^2c, are arranged into a $c \times r \times r$ pixel matrix. Thus, a feature image with a size of $c \times hr \times wr$ is obtained. As the current resolution of this image is m, it is denoted by $F(M_\uparrow^m)$.

In conventional image SR, due to a lack of sufficient information for predicting the postupscaling pixel values, the post-SR image lacks detailed spatial features, i.e., the post-SR image is blurry. To address this problem, high-frequency features are fused to sharpen the blurry areas. The previously obtained $F(M_\uparrow^m)$ and the high-frequency information image $F(P_H^m)$ of the corresponding size extracted by convolution operations are added to restore a feature image that has become blurry after upscaling (i.e., $F(M^m)$), as shown in the following equation:

$$F(M^m) = F(M_\uparrow^m) \oplus F(P_H^m) \tag{6}$$

Based on these steps, the resolution of the MS image features is improved from l to m.

Similarly, the image features can be improved from m to h. Ultimately, fused MS and PAN image features $F(M^h)$ are obtained.

3.3. Image Reconstruction Module into Which High- and Low-Frequency PAN Information Is Injected (Step 4)

A convolutional autoencoder can effectively encode an image to produce high-dimensional coded information and decode deep information by reversing the encoding process to reconstruct the input image. Thus, an autoencoder is employed to reconstruct the image based on $F(M^h)$:

$$F(e_h), F(e_m), F(e_l) = E(F(M^h)) \tag{7}$$

where E represents a three-layer convolutional encoding operation (the first layer is a convolutional operation with a step size of 1 performed to produce coded features with a total of K channels and a resolution of h; the last two layers are convolutional operations with a step size of 2 performed to produce coded features with a total of $2K$ channels and a

resolution of m and coded features with a total of $4K$ channels and a resolution of l), and $F(e_h)$, $F(e_m)$, and $F(e_l)$ are coded features with scales of h, m, and l, respectively.

Conventional convolution operations produce feature images with specific sizes based on the convolution kernel size, weight, and step size of the sliding window. Generally, convolution operations reduce the feature size. To preserve the feature size, it is possible to fill numerical values at the boundaries of the feature image.

The encoder applied in this study encodes each feature image by taking advantage of the properties of convolution to recover multiscale feature information and thus facilitate the injection of multiscale PAN features.

To utilize important multiscale PAN information, the high-frequency PAN information feature group $F(P_H^{h,m,l})$ and low-frequency PAN information feature group $F(P_L^{h,m,l})$ are injected into the features that require decoding through the decoder architecture as follows:

$$F(d_m) = DeConv_1(F(e_l)©F(P_L^l)©F(P_H^l)) \tag{8}$$

$$F(d_h) = DeConv_2(F(d_m)©F(e_m)©F(P_L^m)©F(P_H^m)) \tag{9}$$

$$F(d) = Conv_{3\times3}(F(d_h)©F(e_h)©F(P_L^h)©F(P_H^h)) \tag{10}$$

where © represents an operation that connects feature images in a series based on the number of channels, and $DeConv$ represents a deconvolution operation, which is the reverse process of convolution and can upscale and output feature images with specific numbers of channels. Both $DeConv_1$ and $DeConv_2$ are 2×2 deconvolution operations with a step size of 2; $Conv_{3\times3}$ represents a conventional 3×3 convolution operation; and $F(d_m)$, $F(d_h)$, and $F(d)$ are fused high-frequency features, fused low-frequency features, and fused coded features, respectively, with resolutions of l, m, and h and $3K$, $3K$, and K channels, respectively. The decoder used in this study upscales and decodes coded features by taking advantage of the properties of deconvolution.

$F(d)$ is converted by a 1×1 convolution operation to the number of channels required for the MS image, to which the upsampled MS image $M^l_{\uparrow_{l/h}}$ is added. Thus, a high-resolution MS image $\widetilde{M}^h$ is obtained, as shown here:

$$\widetilde{M}^h = Conv_{1\times1}(F(d)) + M^l_{\uparrow_{l/h}} \tag{11}$$

3.4. Loss Function

Based on the abovementioned architecture, the whole pansharpening network architecture can be represented by the following equation:

$$\widetilde{M}^h = f_{AE}(f_{SR}(M^l, f_E(P^h)), f_E(P^h); \theta) \tag{12}$$

where θ is a network parameter, f_E represents the extraction of multiscale features from the high- and low-pass components of the PAN image, f_{SR} represents FF-based SR, and f_{AE} is a function of the autoencoder structure .

The SSIM function can quantitatively reflect the differences in brightness, contrast, and structure between two images. This function can make the network focus on the structural information of the image rather than the distance between the result and ground truth (e.g., MAE and MSE). The SSIM loss function is used to train the model in this study in the manner shown by the following equation:

$$\min_\theta \sum_i 1 - SSIM(f_{AE}(f_{SR}(M^l, f_E(P^h)), f_E(P^h); \theta), M_i^h) \tag{13}$$

where M_i^l, P_i^h, M_i^h represent the ith training sample.

As it is impossible to obtain true high-resolution MS images, the training data are preprocessed according to Wald's protocol [27]. Specifically, the downsampled MS and

PAN images are input into the network model; the original MS image is treated as the true-value image; and Equation (13) is applied to calculate and update the network.

4. Data and Evaluation Methods

4.1. Datasets

Three datasets produced by different satellites were selected for evaluating DIPNet and the comparison methods. The following subsection details information (i.e., sensors, wavelength, spatial resolution, and number of bands) about the datasets.

4.1.1. QuickBird Dataset

This dataset contains imagery for six regions in different geographic locations, which is from [23]. The surface cover types in these regions include forests, farmlands, buildings, and rivers. The MS images contain the visible-light band (RGB channels) and the near-infrared (NIR) band. The PAN images cover the RGB and NIR bands of the MS images. The spatial resolution (0.7 m) of the PAN images is four times that (2.8 m) of the MS images.

4.1.2. WorldView 2 Dataset

This dataset, provided by MAXAR (https://resources.maxar.com/optical-imagery (accessed on 1 June 2021)), contains data for Washington D.C. (an urban area), USA. The MS images have a spatial resolution of 1.6 m and contain eight (coastline, blue, green, yellow, red, red-edge, NIR-1, and NIR-2) bands. The PAN images have a spatial resolution of 0.4 m.

4.1.3. IKONOS Dataset

This dataset originated from Meng et al.'s [28] pansharpening evaluation dataset. The surface cover types include cities, vegetation, rivers, and lakes. The MS images have a spatial resolution of 4 m and contain four (blue, green, red, and NIR) bands. The PAN images have a spatial resolution of 1 m.

4.1.4. Dataset Preprocessing

The images have an 11-bit radiometric resolution, ranging from 0 to 2047. In this study, the images were not subjected to any relevant radiation corrections. The abovementioned images differ in size. To facilitate testing and training, the MS images and PAN images for the corresponding areas were cropped to 256 × 256 image blocks and 1024 × 1024 image blocks, respectively, which were then randomly divided into a training set and testing set. Table 1 summarizes the number of image blocks obtained.

Table 1. Number of image blocks for the experiment.

Dataset	Total Numbers	Training Numbers	Testing Numbers
QuickBird	714	514	200
WorldView 2	506	356	150
IKONOS	200	150	50

In this study, labels were prepared according to Wald's protocol for model training. The procedure is detailed as follows: first, the MS image and PAN image were downsampled fourfold based on the MTF low-pass filter of the corresponding sensor to a 64 × 64 image $M_{\downarrow 4}$ and a 256 × 256 image $P_{\downarrow 4}$, respectively. Eventually, a simulated image pair $(M_{\downarrow 4}, P_{\downarrow 4}, M)$ was obtained to allow for the use of the original MS image as a supervision objective for training. For the training set, each original MS–PAN image pair was similarly downsampled to obtain a simulated image pair $(M_{\downarrow 4}, P_{\downarrow 4}, M)$. The results were evaluated.

4.2. Experimental and Comparison Methods

During the training process, an Adam optimizer with an initial learning rate of 0.0001, a weight decay parameter of 10^{-8}, and other parameters set to their respective default values was employed to train 1000 epochs to compare DIPNet with other methods. The training parameters are detailed as follows: the training batch size was set to 16. Prior to training, all the initial weights of the neural network were initialized using a normal distribution with a mean of 0 and a variance of 0.02. All the other parameters were set to their respective default values. During the training process, several data augmentation techniques, including random horizontal flipping, random vertical flipping, random rotation by 90°, and random cropping, were used. In the random cropping process, each simulated image pair $(M_{\downarrow 4}, P_{\downarrow 4}, M)$ was cropped to a 32 × 32 $M_{\downarrow 4}$, a 128 × 128 $P_{\downarrow 4}$, and a 128 × 128 M.

With respect to the parameters of the experimental method, the size W_r and variance of the Gaussian filter kernel were set to 11 and 1, respectively, and the number of convolution kernels K and number of residual blocks L were set to 64 and 2, respectively. To prevent randomness from affecting the experimental results, the same seed was set for deterministic calculations to ensure that the experimental results were reproducible.

Four conventional methods (GSA, PRACS, ATWT, and MTF-GLP-CBD) and five DL methods (PNN, MSDCNN, PanNet, TFNet, and SRPPNN) were selected for comparison in this study. The MATLAB code for pansharpening provided by Vivone et al. [29] was used for the conventional pansharpening methods and comparison calculations. The experiment was conducted on a computer with an AMD Ryzen 5 3600 3.6 GHz processor, 32 GB of memory, and an NVIDIA RTX 2070 Super graphics card. The coding environment involved Windows 10 (64 bit), MATLAB (R2013a), Python 3.7.4, and PyTorch 1.6.0.

4.3. Quantitative Evaluation Indices

Several quantitative indices, including the relative dimensionless global error in synthesis (ERGAS) [30], spectral angle mapper (SAM) [31], universal image quality index (UIQI) [32] and its extended index $Q2^n$ [33], spatial correlation coefficient (SCC) [34], and quality without reference (QNR) [35], were employed in the experiment. According to the types of indicators, we divided them into three parts to provide a detailed description.

(1) Indices for spectrum: The ERGAS and SAM primarily reflect the spectral distortions in an enhanced image compared to a reference image. Lower values of ERGAS and SAM indicate that the spectral distribution of an enhanced image is similar to that of the reference image. The details are provided as follows:

$$RMSE(x,y) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2} \tag{14}$$

$$EDRAS(x,y) = 100\frac{h}{l}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{RMSE(x_i,y_i)}{MEAN(y_i)}\right)^2} \tag{15}$$

$$SAM(x,y) = \arccos\left(\frac{x \cdot y}{\|x\| \cdot \|y\|}\right) \tag{16}$$

where x and y are the pansharpened image and ground truth, respectively; m is the number of the pixels in the images; h and l are the spatial resolution of the PAN image and MS image, respectively; and $MEAN(y_i)$ is the mean of the ith band of the ground truth which has a total of N bands.

(2) Indices for structure: UIQI and $Q2^n$ represent the quality of each band and the quality of all the bands. High values of the UIQI and $Q2^n$ suggest that the quality of the resultant and reference images is similar. Their equations are expressed as follows:

$$UIQI(x,y) = \frac{4\sigma_{xy} \cdot \mu_x \cdot \mu_y}{(\sigma_x^2 + \sigma_y^2)(\mu_x^2 + \mu_y^2)} \tag{17}$$

$$Q2^n(x,y) = \frac{4\sigma_{XY} \cdot \mu_X \cdot \mu_Y}{(\sigma_X^2 + \sigma_Y^2)(\mu_X^2 + \mu_Y^2)} \tag{18}$$

For UIQI, μ_x and μ_y are the means of x and y, respectively; σ_x and σ_y are the variances of x and y, respectively; and σ_{xy} denotes the covariance between x and y. Generally, the index is calculated by a kernel.

For $Q2^n$, X and Y are the hypercomplex numbers of x and y, respectively; μ_X and μ_Y are the means of X and Y, respectively; σ_X and σ_Y are the variances of X and Y, respectively; and σ_{XY} denotes the covariance between X and Y.

The SCC is a spatial evaluation index that primarily reflects the difference in high-frequency details between two images, and a value of SCC near 1 indicates a good spatial resolution of the resultant image, as follows:

$$Filter = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \tag{19}$$

$$SCC(x,y) = \frac{\sum_{i=1}^{w}\sum_{j=1}^{h}(Filter(x)_{i,j} - \mu_{Filter(x)})(Filter(y)_{i,j} - \mu_{Filter(y)})}{\sqrt{\sum_{i=1}^{w}\sum_{j=1}^{h}(Filter(x)_{i,j} - \mu_{Filter(x)})^2 \sum_{i=1}^{w}\sum_{j=1}^{h}(Filter(y)_{i,j} - \mu_{Filter(y)})^2}} \tag{20}$$

where $Filter$ is a high frequency kernel, which is used to process images; $\mu_{Filter(x)}$ and $\mu_{Filter(y)}$ are the means of $Filter(x)$ and $Filter(y)$, respectively; and w and h are the weight and height, respectively, of an image.

(3) Indices for no reference: The QNR mainly reflects the fusion performance in the absence of true reference values, which consists of D_s and D_λ. An index of D_s near 0 represents good performance of a structure; an index of D_λ near 0 shows good fusion in a spectrum; and a value of QNR near 1 indicates a good original pansharpening performance.

$$D_\lambda(x,M) = \sqrt[p]{\frac{1}{C(C-1)}\sum_{c=1}^{C}\sum_{r=1(r\neq c)}^{C}|UIQI(x_c,x_r) - UIQI(M_c,M_r)|^p} \tag{21}$$

$$D_s(x,P) = \sqrt[q]{\frac{1}{C}\sum_{c=1}^{C}|UIQI(x_c,P) - UIQI(M_c,P\downarrow)|^q} \tag{22}$$

$$QNR(x,M,P) = (1 - D_\lambda(x,M))^i \cdot (1 - D_s(x,P))^j \tag{23}$$

where p and q denote positive integer exponents; M and P are the MS image and PAN image, respectively; i and j are the weighted parameters to quantify the spectral distortion and spatial distortions, respectively; and C is the number of the bands in an MS image. In our experiment, p, q, i, and j are set to 1.

5. Results and Evaluations

5.1. Experimental Results

This section presents a visual comparison of DIPNet and the comparison methods. To facilitate visualization, the RGB portion of each image was cropped and extended at 2% to an 8-bit color image. To clearly visually compare the reconstructed images, the absolute difference between the true-value and fused images was increased by factors of 10, 4, and 4 for the QuickBird dataset, WorldView 2 dataset, and IKONOS dataset, respectively.

Figure 3 shows the performance of each method on the QuickBird dataset. With respect to the original data, as shown in Figure 3I,IV, DIPNet performed the best in preserving both spectral information and structural information, whereas PanNet produced bright color spots at the edges of the buildings, and TFNet distorted the spectral information.

With respect to the simulated data, as shown in Figure 3II,V, the result produced by DIPNet was the closest to the true-value image, while the four DL-based methods, namely PNN, MSDCNN, TFNet, and SRPPNN, also performed considerably well. However, as shown by the residual images in Figure 3III,VI, the performance of PanNet was inferior to that of the other DL methods in data reconstruction on the QuickBird dataset.

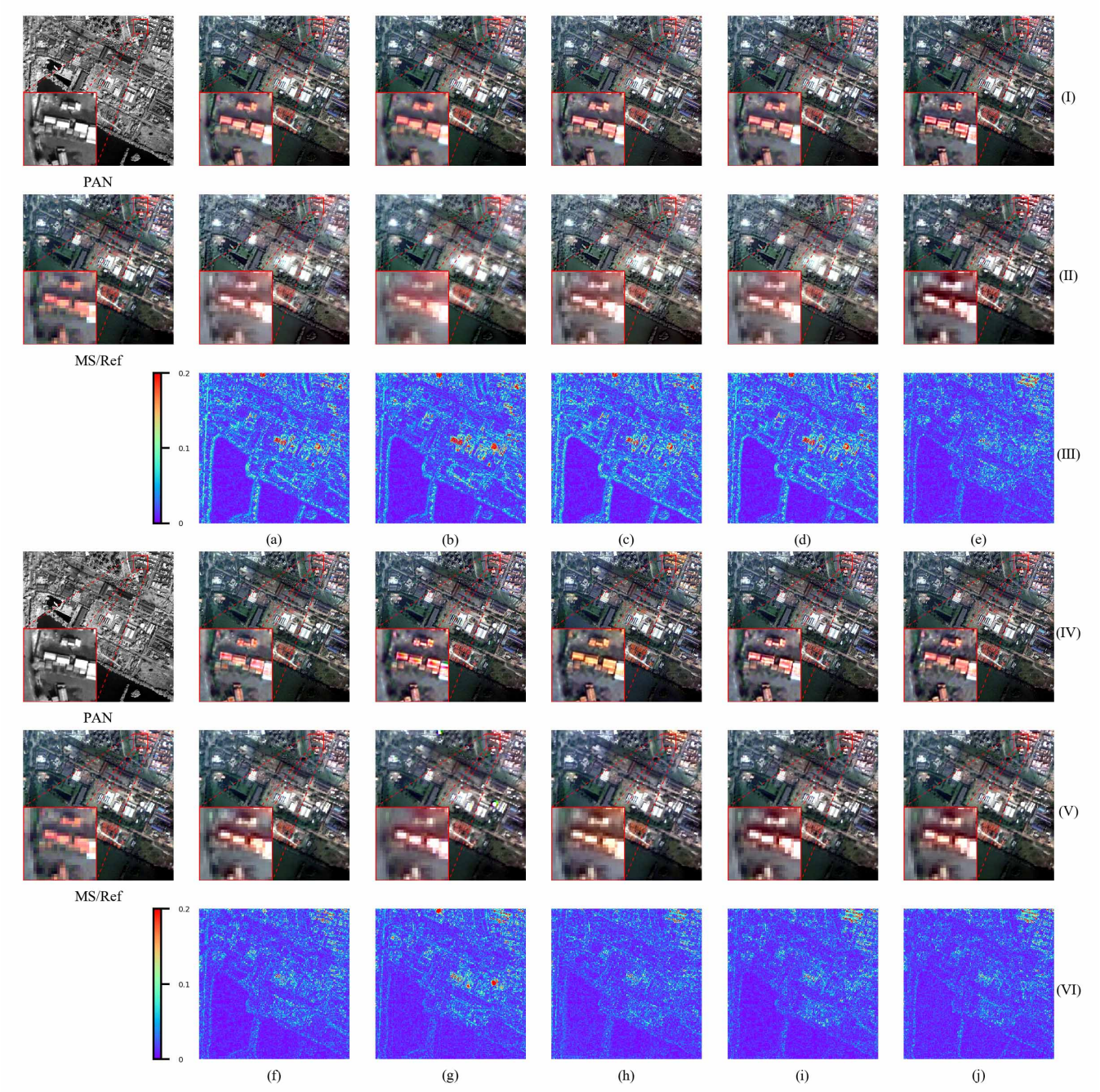

Figure 3. Results produced by DIPNet and the comparison methods on the QuickBird dataset and their residuals. (**a**) GSA; (**b**) PRACS; (**c**) ATWT; (**d**) MTF-GLP-CBD; (**e**) PNN; (**f**) MSDCNN; (**g**) PanNet; (**h**) TFNet; (**i**) SRPPNN; and (**j**) DIPNet. Rows (**I**,**IV**) show pansharpening on the original scale; rows (**II**,**V**) show pansharpening on a reduced scale; and rows (**III**,**VI**) show residuals on a reduced scale.

Figure 4 shows the performance of each method on the WorldView 2 dataset. With respect to the original data, as shown in Figure 4I,IV, DIPNet notably outperformed the other methods in fusion and reconstruction (evidenced, for example, by the structural edges

of the trees and the swimming pool to the right of the building within the red box). With respect to the simulated data, as shown in Figure 4II,V, the results produced by DIPNet were the closest to the true-value image. In addition, DIPNet far outperformed the other methods in representing the edge information for the swimming pool within the red box. These findings, with the residual images in Figure 4III,VI, show that DIPNet outperformed the other methods in the reconstruction of the structural details on the WorldView 2 dataset.

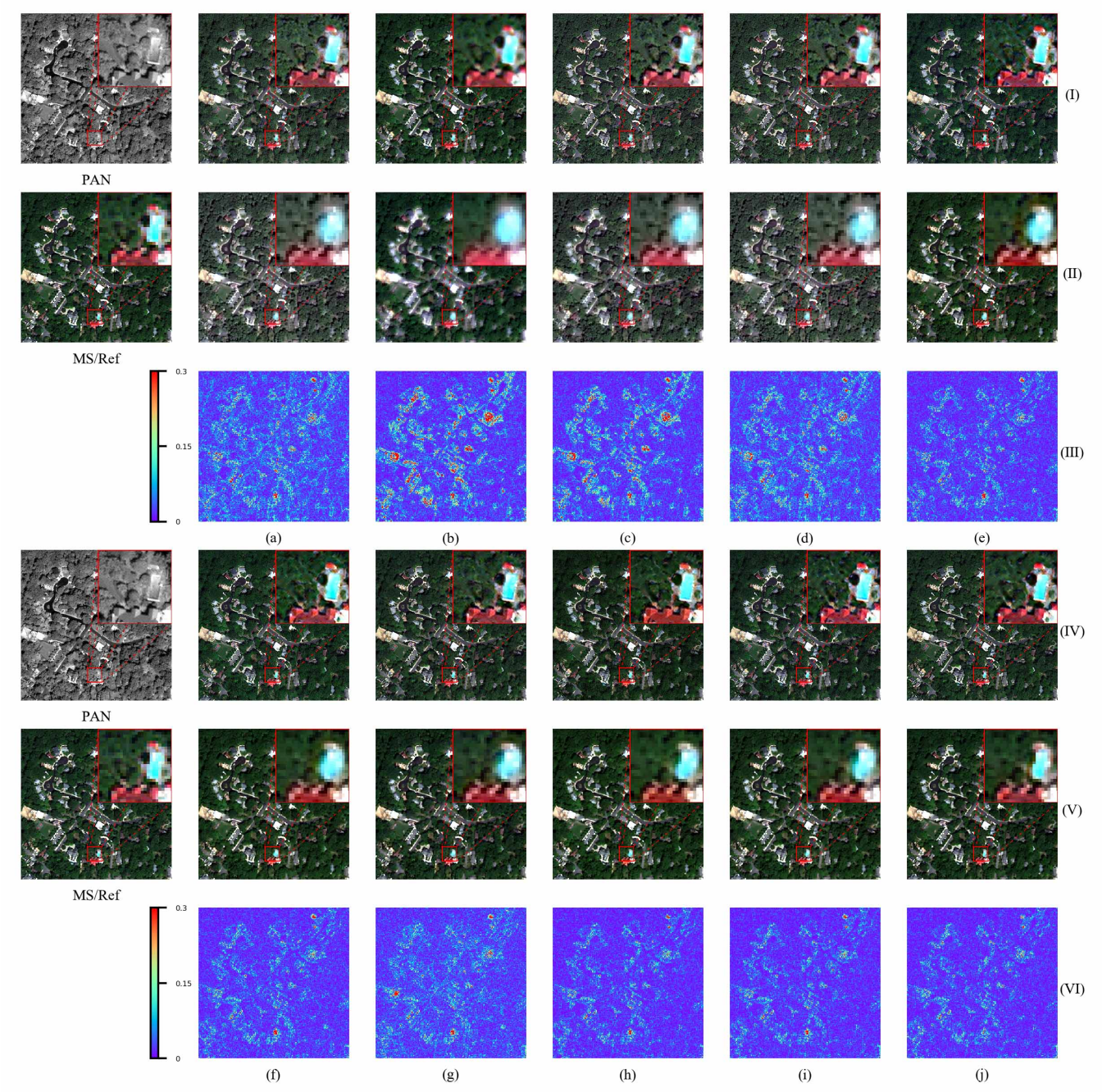

Figure 4. Results produced by DIPNet and the comparison methods on the WorldView 2 dataset and their residuals. (**a**) GSA; (**b**) PRACS; (**c**) ATWT; (**d**) MTF-GLP-CBD; (**e**) PNN; (**f**) MSDCNN; (**g**) PanNet; (**h**) TFNet; (**i**) SRPPNN; and (**j**) DIPNet. Rows (**I**,**IV**) show pansharpening on the original scale; rows (**II**,**V**) show pansharpening on a reduced scale; and rows (**III**,**VI**) show residuals on a reduced scale.

Figure 5 shows the performance of each method on the IKONOS dataset. With respect to the original data, as shown in Figure 5I,IV, the edges of the buildings within the red box

in the image produced by DIPNet were the smoothest and consistent with those in the original PAN image, whereas the edges of the buildings within the red box in the image produced by each of the other DL-based methods were distorted. While the buildings in the images produced by the conventional methods were structurally distinguishable, their colors differed from those in the true-value image. The simulated data in Figure 5II,V and the residual images in Figure 5III,VI show that DIPNet outperformed the other methods in the reconstruction of structural details on the IKONOS dataset.

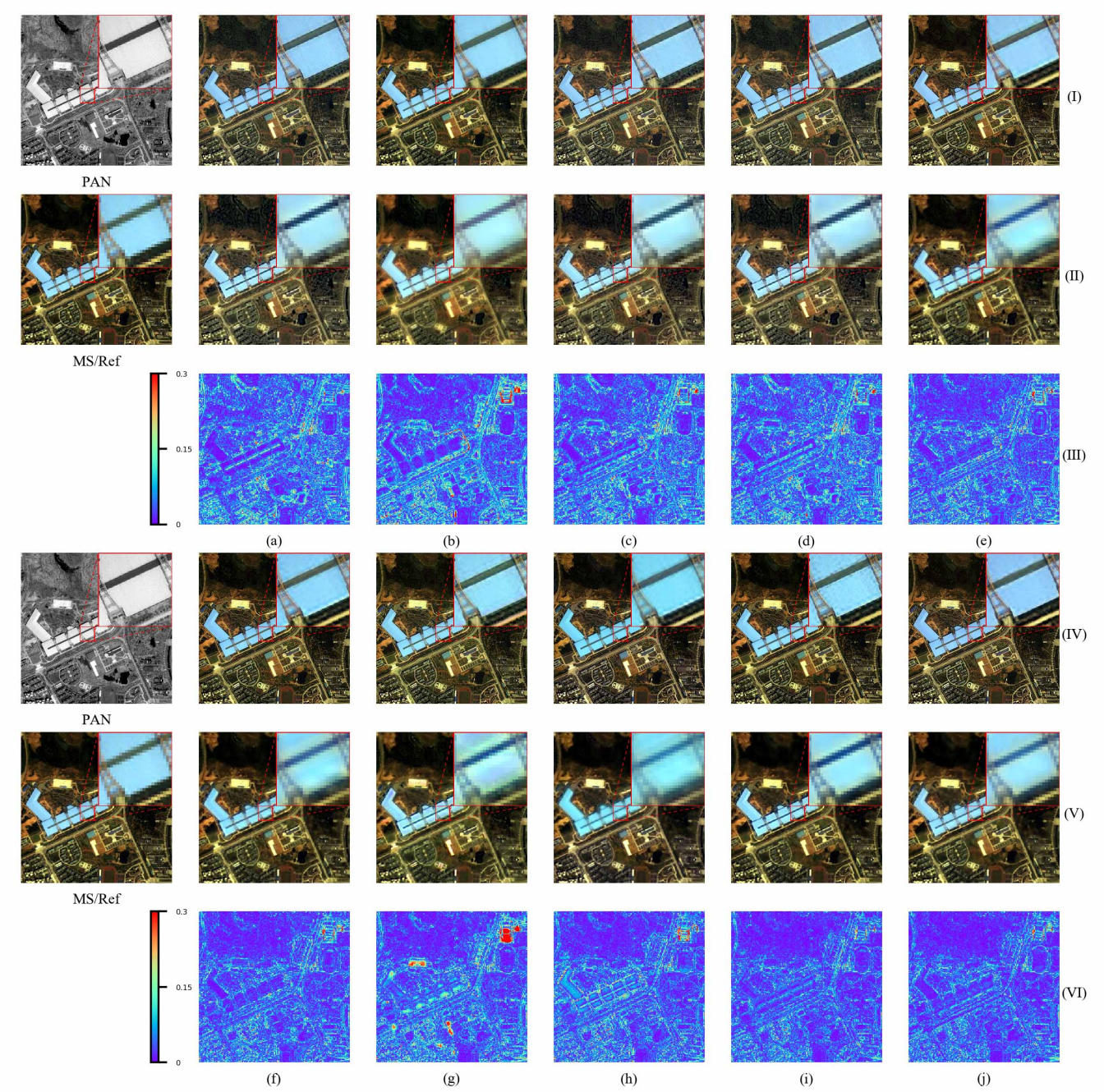

Figure 5. Results produced by DIPNet and the comparison methods on the IKONOS dataset and their residuals. (**a**) GSA; (**b**) PRACS; (**c**) ATWT; (**d**) MTF-GLP-CBD; (**e**) PNN; (**f**) MSDCNN; (**g**) PanNet; (**h**) TFNet; (**i**) SRPPNN; and (**j**) DIPNet. Rows (**I**,**IV**) show pansharpening on the original scale; rows (**II**,**V**) show pansharpening on a reduced scale; and rows (**III**,**VI**) show residuals on a reduced scale.

5.2. Comparison of the Quantitative Indices

This section presents a numeric assessment of DIPNet and the comparison methods. To facilitate a numeric comparison, the best performance, second-best performance, and third-best performance in Tables 2–4 are shown in red, green, and blue, respectively.

Table 2 summarizes a comparison of the values of the indices for the methods on 200 images from the QuickBird dataset. As demonstrated by the values of the first five indices, DIPNet outperformed the other methods in the preservation of the spectral and structural information. Regarding the last three indices, the D_λ value for PRACS was the lowest, indicating that the PRACS method outperformed the other methods in terms of preserving the spectral information on the original scale. The D_s value for DIPNet was the lowest, suggesting that the DIPNet far outperformed the other methods in preservation of the structural information and took full advantage of the multiscale high- and low-frequency feature information contained in the PAN image. When evaluated by the total index QNR, DIPNet was second only to SRPPNN. This finding is attributed to the notion that DIPNet does not adequately consider spectral information and that the D_λ value of DIPNet is higher than that of SRPPNN.

Table 3 summarizes a comparison of the values of the indices for the methods on 150 images from the WorldView 2 dataset. As demonstrated by the first five indices, DIPNet outperformed the other methods in preservation of the spectral and structural information. Regarding the last three indices, PRACS similarly performed the best in preservation of spectral information, followed by DIPNet, suggesting good performance of DIPNet in preservation of the original MS information on the WorldView 2 dataset. The D_s value for PanNet was the lowest, indicating good performance in the fusion of data acquired by the high-resolution satellites of the WorldView series. When evaluated by the total index QNR, DIPNet outperformed PanNet because DIPNet optimizes the nonlinear relationships between low-frequency PAN and MS information and high-frequency PAN and MS information at multiple scales.

Table 4 summarizes a comparison of the values of the indices for the methods on 50 images from the IKONOS dataset. As demonstrated by the first five indices, DIPNet outperformed the other methods in preservation of the spectral and structural information. With respect to the last three indices, the D_λ value for DIPNet was the lowest, suggesting that DIPNet performed the best regarding preservation of the MS information in a low-resolution satellite. The D_s value for DIPNet was also the lowest. Thus, DIPNet performed the best in preservation of the structural information and total index QNR.

Table 2. Quantitative results of various methods on QuickBird.

Method	ERGAS↓	SAM↓	SCC↑	$Q2^n$↑	UIQI↑	D_λ↓	D_s↓	QNR↑
GSA	1.8814	2.4336	0.8984	0.8126	0.7906	0.0506	0.0916	0.8640
PRACS	1.8067	2.3154	0.9020	0.7974	0.7810	0.0231	0.0774	0.9015
ATWT	1.8478	2.4240	0.9022	0.8068	0.8029	0.0929	0.1182	0.8030
MTF-GLP-CBD	1.8584	2.4074	0.8992	0.8146	0.7916	0.0426	0.0693	0.8921
PNN	1.3068	1.7034	0.9438	0.8783	0.8792	0.0470	0.0493	0.9072
MSDCNN	1.2697	1.6403	0.9463	0.8836	0.8847	0.0402	0.0537	0.9096
PanNet	1.4373	1.8319	0.9393	0.8687	0.8738	0.0477	0.0482	0.9072
TFNet	1.0756	1.3809	0.9731	0.8929	0.9039	0.0618	0.0419	0.9000
SRPPNN	1.0090	1.3318	0.9736	0.9052	0.9058	0.0403	0.0434	0.9190
DIPNet	0.9203	1.2373	0.9797	0.9128	0.9180	0.0672	0.0264	0.9096

Table 3. Quantitative results of various methods on WorldView 2.

Method	ERGAS↓	SAM↓	SCC↑	$Q2^n$↑	UIQI↑	D_λ↓	D_s↓	QNR↑
GSA	4.4955	7.7368	0.8198	0.8525	0.8275	0.0758	0.1488	0.7877
PRACS	5.4284	7.5695	0.7489	0.7705	0.7569	**0.0149**	0.0887	0.8980
ATWT	4.7309	7.2916	0.8405	0.8344	0.8217	0.0902	0.1420	0.7819
MTF-GLP-CBD	4.6106	7.6830	0.8178	0.8446	0.8212	0.0788	0.1347	0.7983
PNN	3.2127	5.3787	0.9229	0.9073	0.9049	0.0434	0.0614	0.8986
MSDCNN	3.0133	4.9910	0.9347	0.9164	0.9138	0.0539	0.0629	0.8886
PanNet	3.4526	5.6125	0.9250	0.8985	0.9044	0.0359	0.0519	0.9145
TFNet	2.7642	4.5205	0.9471	0.9265	0.9237	0.0538	0.0751	0.8777
SRPPNN	2.7717	4.5596	0.9461	0.9267	0.9245	0.0478	0.0631	0.8937
DIPNet	2.7368	4.4226	0.9518	0.9313	0.9288	0.0287	0.0533	0.9210

Table 4. Quantitative results of various methods on IKONOS.

Method	ERGAS↓	SAM↓	SCC↑	$Q2^n$↑	UIQI↑	D_λ↓	D_s↓	QNR↑
GSA	1.5683	2.2253	0.9172	0.8252	0.8199	0.1382	0.2044	0.6944
PRACS	1.7359	2.2594	0.9040	0.7993	0.7959	0.0748	0.1432	0.7979
ATWT	1.6331	2.2373	0.9173	0.8188	0.8208	0.1606	0.2104	0.6714
MTF-GLP-CBD	1.5929	2.2356	0.9166	0.8245	0.8179	0.1411	0.1923	0.7025
PNN	1.4814	2.1211	0.9310	0.8348	0.8429	0.0826	0.1032	0.8304
MSDCNN	1.3676	1.9286	0.9411	0.8462	0.8594	0.0969	0.1134	0.8123
PanNet	1.5728	2.3126	0.9262	0.8402	0.8428	0.0769	0.1124	0.8232
TFNet	1.4147	1.9169	0.9406	0.8459	0.8561	0.0971	0.0785	0.8401
SRPPNN	1.2263	1.6641	0.9500	0.8693	0.8745	0.0831	0.1043	0.8302
DIPNet	1.2300	1.6446	0.9521	0.8725	0.8764	0.0739	0.0664	0.8679

6. Discussion

6.1. Ablation Experiment

In DIPNet, the high- and low-frequency PAN feature groups and MS features are progressively fused through means such as an SR module, an FFU module, and feature addition. An ablation experiment was conducted to examine the effectiveness of DIPNet. The SR of the MS features, fusion with low-frequency PAN information, fusion with high-frequency PAN information, low-frequency PAN information autoencoder, and high-frequency PAN information autoencoder are denoted as MSR, PL, PH, AEL, and AEH, respectively. The network parameters for the ablation experiment were set to L = 2 and K = 64. To facilitate numeric comparison, the best performance, second-best performance, and third-best performance in Tables 5–7 are shown in red, green, and blue, respectively.

Table 5. Quantitative results produced by various modules on the QuickBird dataset.

Row	Method	ERGAS↓	SAM↓	SCC↑	$Q2^n$↑	UIQI↑	D_λ↓	D_s↓	QNR↑
1	MSR+MSE	2.3216	2.5047	0.7618	0.7639	0.7667	0.0399	0.1050	0.8592
2	MSR+PL+MSE	1.0015	1.3234	0.9752	0.9049	0.9082	0.0529	0.0277	0.9219
3	MSR+PH+MSE	1.3479	1.7686	0.9482	0.8620	0.8758	0.0424	0.0384	0.9212
4	MSR+PL+PH+MSE	1.0656	1.4245	0.9726	0.8973	0.8993	0.0474	0.0281	0.9265
5	MSR+PL+PH+AEL+MSE	0.9127	1.2211	0.9788	0.9788	0.9148	0.0596	0.0280	0.9146
6	MSR+PL+PH+AEH+MSE	0.9574	1.2716	0.9771	0.9122	0.9154	0.0510	0.0260	0.9252
7	DIPNet-FFU+MSE	0.9303	1.2446	0.9772	0.9122	0.9136	0.0421	0.0327	0.9271
8	DIPNet+MSE	0.9378	1.2392	0.9781	0.9043	0.9129	0.0447	0.0356	0.9217
9	DIPNet+MAE	0.8980	1.1908	0.9799	0.9160	0.9193	0.0460	0.0288	0.9275
10	DIPNet	0.9203	1.2373	0.9797	0.9128	0.9180	0.0672	0.0264	0.9096

Table 6. Quantitative results produced by various modules on the WorldView 2 dataset.

Row	Method	ERGAS↓	SAM↓	SCC↑	$Q2^n$↑	UIQI↑	D_λ↓	D_s↓	QNR↑
1	MSR+MSE	5.7362	5.9000	0.7018	0.7762	0.7757	0.0440	0.0865	0.8730
2	MSR+PL+MSE	2.8065	4.4839	0.9470	0.9265	0.9241	0.0317	0.0606	0.9113
3	MSR+PH+MSE	2.9345	4.7894	0.9430	0.9211	0.9185	0.0279	0.0668	0.9082
4	MSR+PL+PH+MSE	2.8237	4.5604	0.9465	0.9239	0.9223	0.0275	0.0593	0.9161
5	MSR+PL+PH+AEL+MSE	2.7183	4.4147	0.9502	0.9279	0.9257	0.0282	0.0610	0.9141
6	MSR+PL+PH+AEH+MSE	2.6557	4.3223	0.9530	0.9294	0.9273	0.0227	0.0596	0.9205
7	DIPNet-FFU+MSE	2.6529	4.3328	0.9529	0.9304	0.9278	0.0279	0.0611	0.9148
8	DIPNet+MSE	2.6550	4.3369	0.9528	0.9293	0.9270	0.0274	0.0624	0.9136
9	DIPNet+MAE	2.6374	4.3099	0.9534	0.9316	0.9289	0.0252	0.0625	0.9157
10	DIPNet	2.7368	4.4226	0.9518	0.9313	0.9288	0.0287	0.0533	0.9210

Table 7. Quantitative results produced by various modules on the IKONOS dataset.

Row	Method	ERGAS↓	SAM↓	SCC↑	$Q2^n$↑	UIQI↑	D_λ↓	D_s↓	QNR↑
1	MSR+MSE	2.4161	2.3731	0.7590	0.7025	0.7077	0.0757	0.1232	0.8113
2	MSR+PL+MSE	1.3314	1.8612	0.9442	0.8505	0.8561	0.0696	0.0799	0.8615
3	MSR+PH+MSE	1.5921	2.2924	0.9215	0.8191	0.8320	0.0688	0.0973	0.8439
4	MSR+PL+PH+MSE	1.5200	2.1522	0.9325	0.8252	0.8431	0.0708	0.1087	0.8330
5	MSR+PL+PH+AEL+MSE	1.2163	1.6557	0.9528	0.8703	0.8770	0.0836	0.0717	0.8566
6	MSR+PL+PH+AEH+MSE	1.1551	1.5570	0.9564	0.8787	0.8821	0.0866	0.0759	0.8501
7	DIPNet-FFU+MSE	1.2595	1.6962	0.9510	0.8650	0.8719	0.0945	0.0902	0.8320
8	DIPNet+MSE	1.2227	1.6230	0.9533	0.8583	0.8764	0.0754	0.0776	0.8578
9	DIPNet+MAE	1.2186	1.6399	0.9525	0.8711	0.8769	0.0782	0.0766	0.8567
10	DIPNet	1.2300	1.6446	0.9521	0.8725	0.8764	0.0739	0.0664	0.8679

6.1.1. Network Architecture

To verify the performance improvement resulting from the integration of the high- and low-frequency PAN feature groups with MSR, the architecture of DIPNet was split into the following: MSR; FFU-based SR of the MS and low-frequency PAN information (MSR+PL); MSR combined with high-frequency PAN information (MSR+PH); FFU-based SR of the MS and low-frequency PAN information combined with high-frequency PAN information (MSR+PL+PH); FFU-based SR of the MS and low-frequency PAN information combined with high-frequency PAN information and low-frequency PAN information autoencoder (MSR+PL+PH+AEL); and FFU-based SR of the MS and low-frequency PAN information combined with high-frequency PAN information and high-frequency PAN information autoencoder (MSR+PL+PH+AEH). Rows 1, 2, 3, 4, 5, 6, and 8 in each of Tables 5–7 show the results produced by these six components and the complete DIPNet under the same training conditions. As demonstrated in these three tables, MSR alone could not produce relatively good quantitative results due to a lack of PAN information. Adding PH to MSR slightly improved the indices due to the addition of some high-frequency information after upsampling of the network. Adding PL to MSR improved the indices to a far greater extent than adding PH, due to the inclusion of MS-band information in the PAN image. The improvement in the indices from integrating a combination of PL and PH with MSR differed insignificantly from that from integrating PL alone with MSR. Introducing the features into the autoencoder for reconstruction further improved the indices compared to those with the integration of MSR and PL, suggesting that an autoencoder with multiscale high- and low-frequency PAN features can improve the robustness of the network. On the IKONOS dataset, however, using the low-frequency PAN information autoencoder can increase the number of reduced indices.

6.1.2. Function of the FFU

Equation (2) details the fusion method for the FFU. We believe that a simple feature addition damages the detailed outline features at the edges that can be potentially extracted.

A comparison of rows 7 and 8 in Tables 5–7 under the same conditions shows that the proposed FFU significantly improved the fusion performance on the IKONOS dataset. While the FFU did not improve the fusion performance on the images acquired by the other satellites, it did not have a significant impact. Figure 6 visualizes the effects of the FFU on the network. A comparison of Figure 6a,c with Figure 6d,f reveals that the fused feature image produced by the network with the FFU showed no distortion compared to the extracted MS features, that the network with the FFU exhibited good stability, and that the fused feature image produced by simply adding the feature images pixel by pixel was overexposed, affecting network learning.

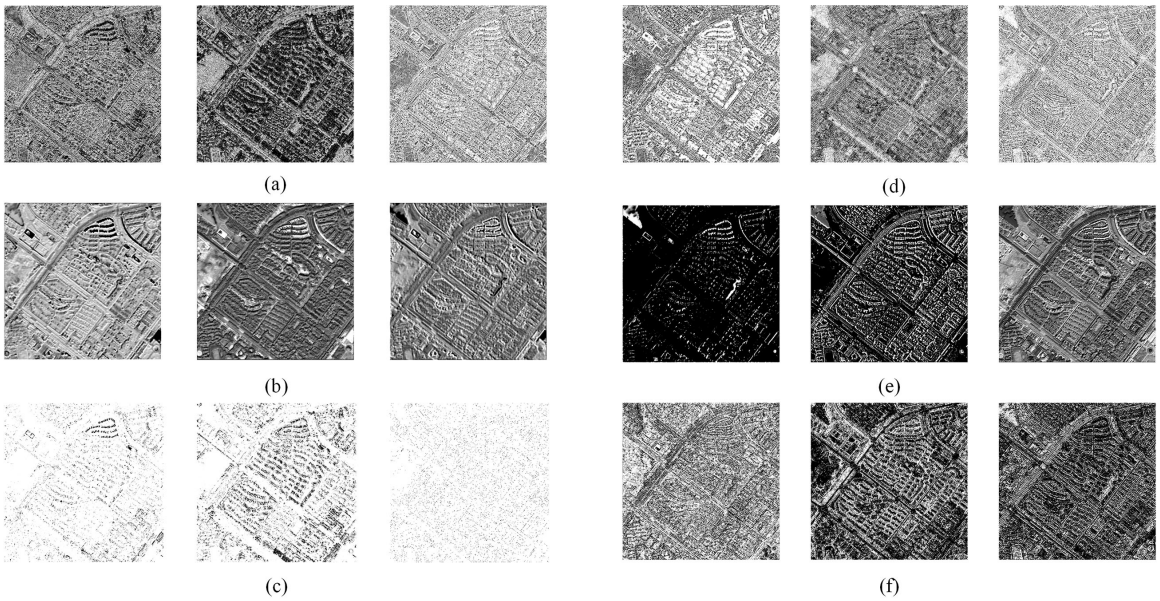

Figure 6. Feature images produced with and without FFU-based fusion. (**a**) MS features extracted without FFU-based fusion. (**b**) PAN features extracted without FFU-based fusion. (**c**) Fused features extracted with the ResNet but without FFU-based fusion. (**d**) MS features extracted with FFU-based fusion. (**e**) PAN features extracted with FFU-based fusion. (**f**) Fused features extracted with FFU-based fusion and ResNet.

6.1.3. Loss Function

Equations (12) and (13) were used to optimize and fit DIPNet. A comparison of rows 8, 9, and 10 in each of Tables 5–7 shows that the D_s value for DIPNet was far smaller than those for the methods that use the MSE and MAE losses in image reconstruction on the QuickBird, WorldView 2, and IKONOS datasets and that the SSIM loss enhanced the image fusion performance at the original size. A comparison of the first five indices in rows 9 and 10 in Table 6 reveals that the use of the MAE loss was superior to that of the SSIM loss on the simulated WorldView 2 data. However, a comparison of the last three indices indicates that the use of the SSIM loss led to better fusion performance on the original data. Thus, based on the evaluation of the simulated and original data and by taking into account the ultimate fusion application needs and performance, the SSIM loss-based DIPNet was selected as the ultimate method proposed in this study.

6.2. Experiments on the Network Performance

6.2.1. The Setting of the Parameters

The effects of K on the fusion performance were investigated. While the architecture of the network was maintained and the SSIM loss function was used in each case under the same experimental conditions, K was set to 16, 32, 48, and 64 and L was set to 2 on all three datasets. In addition, the effects of L on the fusion performance were examined. While the architecture of the network was maintained and the SSIM loss function was used in each case under the same experimental conditions, L was set to 0, 1, 2, and 3

and K was set to 64 on all three datasets. To facilitate a numeric comparison, the best, second-best, and third-best performance in Table 8 are shown in red, green, and blue, respectively. Table 8 summarizes the quantitative results. Clearly, increasing K could effectively improve the fusion performance. However, if L was too high or too low, the performance deteriorated. Thus, by comprehensively considering the experimental results, computational expenditure, and performance on different datasets, a K of 64 and an L of 2 were selected as the parameter settings for the proposed method.

Table 8. Quantitative results produced using various K and L values on each dataset.

Satellite	Setting	ERGAS↓	SAM↓	SCC↑	$Q2^n$↑	UIQI↑	D_λ↓	D_s↓	QNR↑
QuickBird	K = 16, L = 2	1.2379	1.7441	0.9678	0.8824	0.8936	0.0753	0.0397	0.8885
	K = 32, L = 2	1.0561	1.4008	0.9739	0.9051	0.909	0.0693	0.0298	0.9039
	K = 48, L = 2	0.9563	1.2997	0.9783	0.9102	0.9162	0.0734	0.0298	0.8993
	K = 64, L = 2	0.9203	1.2373	0.9797	0.9128	0.9180	0.0672	0.0264	0.9096
	K = 64, L = 0	0.9907	1.3051	0.9750	0.9107	0.9113	0.0373	0.0329	0.9318
	K = 64, L = 1	0.9371	1.2326	0.9788	0.9114	0.9167	0.0516	0.0277	0.9229
	K = 64, L = 3	0.9417	1.2574	0.9788	0.9065	0.9184	0.0785	0.0301	0.8944
WorldView 2	K = 16, L = 2	3.1457	5.0293	0.9316	0.9175	0.915	0.0394	0.0582	0.9071
	K = 32, L = 2	2.7602	4.5033	0.9491	0.9280	0.9262	0.0281	0.0586	0.9172
	K = 48, L = 2	2.7013	4.4241	0.9520	0.9318	0.9290	0.0294	0.0567	0.9174
	K = 64, L = 2	2.7368	4.4226	0.9518	0.9313	0.9288	0.0287	0.0533	0.9210
	K = 64, L = 0	2.9046	4.7474	0.9418	0.9251	0.9226	0.0409	0.0657	0.8982
	K = 64, L = 1	2.6723	4.3898	0.9531	0.9304	0.9295	0.0243	0.0582	0.9204
	K = 64, L = 3	2.6614	4.3438	0.9527	0.9317	0.9295	0.0263	0.0590	0.9177
IKONOS	K = 16, L = 2	1.7051	2.2201	0.9293	0.7845	0.8249	0.1314	0.1491	0.7497
	K = 32, L = 2	1.2711	1.7153	0.9496	0.8657	0.8713	0.0871	0.0599	0.8605
	K = 48, L = 2	1.7708	2.1385	0.9478	0.7866	0.8595	0.1318	0.0703	0.8121
	K = 64, L = 2	1.2300	1.6446	0.9521	0.8725	0.8764	0.0739	0.0664	0.8679
	K = 64, L = 0	1.3735	1.7784	0.9484	0.8574	0.8758	0.0793	0.1006	0.8374
	K = 64, L = 1	1.3626	1.8421	0.9483	0.8459	0.8724	0.0825	0.0985	0.8350
	K = 64, L = 3	1.2676	1.6887	0.9539	0.8572	0.8792	0.0700	0.0725	0.8667

6.2.2. Number of Parameters

The number of parameters for our proposed work is compared with other prior networks according to Table 9. As shown in Tables 2–4 and 8, the setting of K of 32 and L of 2 has fewer parameters than prior networks but achieves the same performance.

Table 9. The parameters of different networks.

Method	PNN	MSDCNN	PanNet	TFNet	SRPPNN
Parameter (M)	0.08	0.19	0.08	2.36	1.83
Method	**DIPNet (K = 16, L = 2)**	**DIPNet (K = 32, L = 2)**	**DIPNet (K = 48, L = 2)**	**DIPNet (K = 64, L = 2)**	
Parameter (M)	0.18	0.73	1.65	2.92	

6.2.3. Efficiency of the Network

Due to the limitation of our computing resources, in the prediction and pansharpening stage, we divide the high-resolution image into small pieces of a certain size for pansharpening and then combine them into the whole image.

In the original resolution evaluation experiment for QuickBird, we record the average running time of the different DL methods. The corresponding results are summarized in Table 10. As shown in Tables 2–4 and 8, the setting of K of 32 and L of 2 has been as fast as prior networks. Although the running time of our proposed method, which has deeper features, is slower than that of other DL methods because a larger number of parameters reduces the efficiency of the network, our method outperforms other methods.

Table 10. Efficiency of different networks on 200 full-resolution satellite images from QuickBird.

Method	PNN	MSDCNN	PanNet	TFNet	SRPPNN
Time (s)	0.38	0.83	0.39	0.64	0.65
Method	**DIPNet (K = 16, L = 2)**	**DIPNet (K = 32, L = 2)**	**DIPNet (K = 48, L = 2)**	**DIPNet (K = 64, L = 2)**	
Time (s)	0.49	0.72	1.15	1.45	

7. Conclusions

This study presents a new DL-based pansharpening method referred to as DIPNet. DIPNet addresses two difficult problems, namely the need for upsampling and serial fusion at a single scale and limited information content in multiscale fused features. In regard to preprocessing, different from other methods that fuse upsampled MS images, DIPNet separates the frequency information contained in a PAN image and then obtains the corresponding features by convolution operations as prior information. To achieve improved fusion performance, DIPNet uses an SR module to fuse the prior PAN information and the MS features and learns nonlinear mapping relationships through the conventional encoder–decoder architecture to produce an enhanced remote sensing image. To enable the network to focus on the structural quality, the SSIM loss function is applied instead of the conventional MSE loss function to train the network to facilitate the high-quality fusion of remote sensing images. The experimental results demonstrate the superiority of DIPNet to the other methods.

Although we have achieved gratifying results, the method of frequency decomposition, in which we simply use a Gaussian filter, is common. We did not discuss the impact of other backbones (in this paper, we use ResNet) for extracting PAN features on network performance and efficiency or even design a better module for pansharpening. In the near future, we will focus on a novel way to pre-extract PAN image features and the design of a novel panchromatic image feature extraction network. For reconstruction, we will develop a new method to reconstruct an image to further improve the quality. As an application, we will also apply this method to other low-resolution satellites, such as Landsat 8 and Sentinel 2.

Author Contributions: Conceptualization, W.W.; methodology, W.W. and Y.X.; software, Y.X.; validation, Y.X.; formal analysis, W.W.; investigation, Y.X.; resources, Y.X.; data curation, Y.X.; writing—original draft preparation, W.W. and Y.X; writing—review and editing, W.W., H.Y. , N.W., and Y.S.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (no. 2017YFB0503603) and the National Natural Science Foundation of China under (no. 41631179).

Data Availability Statement: The dataset and source code are available from https://github.com/xyc19970716/Deep-Learning-PanSharpening (accessed on 1 June 2021).

Acknowledgments: The authors would like to thank all the reviewers for their valuable contributions to our work.

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

References

1. Meng, X.; Shen, H.; Li, H.; Zhang, L.; Fu, R. Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges. *Inf. Fusion* **2019**, *46*, 102–113. [CrossRef]
2. Carper, W.J.; Lillesand, T.M.; Kiefer, P.W. The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data. *Photogramm. Eng. Remote Sens.* **1990**, *56*, 459–467 .
3. Jr, P.S.C.; Kwarteng, A.Y. Extracting spectral contrast in Landsat Thematic Mapper image data using selective principal component analysis. *Photogramm. Eng. Remote Sens.* **1989**, *55*, 339–348.
4. Aiazzi, B.; Baronti, S.; Selva, M. Improving Component Substitution Pansharpening Through Multivariate Regression of MS $+$Pan Data. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 3230–3239. [CrossRef]

5. Choi, J.; Yu, K.; Kim, Y. A New Adaptive Component-Substitution-Based Satellite Image Fusion by Using Partial Replacement. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 295–309. [CrossRef]
6. Aiazzi, B.; Alparone, L.; Baronti, S.; Garzelli, A. Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 2300–2312. [CrossRef]
7. Nunez, J.; Otazu, X.; Fors, O.; Prades, A.; Pala, V.; Arbiol, R. Multiresolution-based image fusion with additive wavelet decomposition. *IEEE Trans. Geosci. Remote Sens.* **1999**, *37*, 1204–1211. [CrossRef]
8. Gou, Y.; Li, B.; Liu, Z.; Yang, S.; Peng, X. CLEARER: Multi-Scale Neural Architecture Search for Image Restoration. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 17129–17140.
9. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc.* **2004**, *13*, 600–612. [CrossRef]
10. Yuan, Q.; Wei, Y.; Meng, X.; Shen, H.; Zhang, L. A Multiscale and Multidepth Convolutional Neural Network for Remote Sensing Imagery Pan-Sharpening. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 978–989. [CrossRef]
11. Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Paisley, J. PanNet: A Deep Network Architecture for Pan-Sharpening. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
12. Gao, H.; Liu, Z.; Weinberger, K.; van der Maaten, L. Deep residual learning for image recognition. In Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017.
13. Fu, X.; Wang, W.; Huang, Y.; Ding, X.; Paisley, J. Deep Multiscale Detail Networks for Multiband Spectral Image Sharpening. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**. [CrossRef] [PubMed]
14. Li, B.; Gou, Y.; Gu, S.; Liu, J.Z.; Peng, X. You Only Look Yourself: Unsupervised and Untrained Single Image Dehazing Neural Network. *Int. J. Comput. Vis.* **2021**, *129*, 1754–1767. [CrossRef]
15. Huang, W.; Xiao, L.; Wei, Z.; Liu, H.; Tang, S. A New Pan-Sharpening Method With Deep Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1037–1041. [CrossRef]
16. Azarang, A.; Manoochehri, H.E.; Kehtarnavaz, N. Convolutional Autoencoder-Based Multispectral Image Fusion. *IEEE Access* **2019**, *7*, 35673–35683. [CrossRef]
17. Liu, X.; Liu, Q.; Wang, Y. Remote sensing image fusion based on two-stream fusion network. *Inf. Fusion* **2020**, *55*, 1–15. [CrossRef]
18. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [CrossRef] [PubMed]
19. Masi, G.; Cozzolino, D.; Verdoliva, L.; Scarpa, G. Pansharpening by Convolutional Neural Networks. *Remote Sens.* **2016**, *8*, 594. [CrossRef]
20. Wei, Y.; Yuan, Q. Deep residual learning for remote sensed imagery pansharpening. In Proceedings of the International Workshop on Remote Sensing with Intelligent Processing, Shanghai, China, 19–21 May 2017; pp. 1–4.
21. Zhang, Y.; Liu, C.; Sun, M.; Ou, Y. Pan-Sharpening Using an Efficient Bidirectional Pyramid Network. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5549–5563. [CrossRef]
22. Yang, Y.; Tu, W.; Huang, S.; Lu, H. PCDRN: Progressive Cascade Deep Residual Network for Pansharpening. *Remote Sens.* **2020**, *12*, 676. [CrossRef]
23. Cai, J.; Huang, B. Super-Resolution-Guided Progressive Pansharpening Based on a Deep Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, 1–15. [CrossRef]
24. Liu, X.; Wang, Y.; Liu, Q. PSGAN: A Generative Adversarial Network for Remote Sensing Image Pan-Sharpening. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018.
25. Shao, Z.; Lu, Z.; Ran, M.; Fang, L.; Zhou, J.; Zhang, Y. Residual Encoder-Decoder Conditional Generative Adversarial Network for Pansharpening. *IEEE Geosci. Remote Sens. Lett.* **2019**, 1–5. [CrossRef]
26. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Wang, Z. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In Proceedings of the CVPR 2016, Las Vegas, NV, USA, 26 June–1 July 2016.
27. Wald, L.; Ranchin, T.; Mangolini, M. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogramm. Eng. Remote Sens.* **1997**, *63*, 691–699.
28. Meng, X.; Xiong, Y.; Shao, F.; Shen, H.; Sun, W.; Yang, G.; Yuan, Q.; Fu, r.; Zhang, H. A Large-Scale Benchmark Data Set for Evaluating Pansharpening Performance: Overview and implementation. *IEEE Geosci. Remote Sens. Mag.* **2020**. [CrossRef]
29. Vivone, G.; Alparone, L.; Chanussot, J.; Mura, M.D.; Garzelli, A.; Licciardi, G.A.; Restaino, R.; Wald, L. A Critical Comparison Among Pansharpening Algorithms. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2565–2586. [CrossRef]
30. Wald, L. *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions*; Presses des MINES: Paris, France, 2002 .
31. Yuhas, R.H.; Goetz, A.F.H.; Boardman, J.W. Discrimination among Semi-Arid Landscape Endmembers Using the Spectral Angle Mapper (SAM) Algorithm. 1992. Available online: https://ntrs.nasa.gov/search.jsp?R=19940012238 2020-06-16T20:21:56+00:00Z (accessed on 1 July 2021)
32. Wang, Z.; Bovik, A.C. A universal image quality index. *IEEE Signal Process. Lett.* **2002**, *9*, 81–84. [CrossRef]
33. Garzelli, A.; Nencini, F. Hypercomplex Quality Assessment of Multi/Hyperspectral Images. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 662–665. [CrossRef]

34. Zhou, J.; Civco, D.; Silander, J. A wavelet transform method to merge Landsat TM and SPOT panchromatic data. *Int. J. Remote Sens.* **1998**, *19*, 743–757. [CrossRef]
35. Alparone, L.; Aiazzi, B.; Baronti, S.; Garzelli, A.; Nencini, F.; Selva, M. Multispectral and Panchromatic Data Fusion Assessment without Reference. *Photogramm. Eng. Remote Sens.* **2008**, *74*, 193–200. [CrossRef]