remote sensing

MDPI

Article

Differential Strategy-Based Multi-Level Dense Network for Pansharpening

Junru Yin *, Jiantao Qu, Qiqiang Chen, Ming Ju and Jun Yu

College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450001, China; qujt@email.zzuli.edu.cn (J.Q.); chenqq@zzuli.edu.cn (Q.C.); jmzw735@163.com (M.J.); yujun@zzuli.edu.cn (J.Y.)

* Correspondence: yinjr@zzuli.edu.cn; Tel.: +86-132-0382-0152

Abstract: Due to the discrepancy in spatial structure between multispectral (MS) and panchromatic (PAN) images, the general fusion scheme will lead to image error in the fused result. To solve this issue, a differential strategy-based multi-level dense network is proposed, and it regards the image pairs at different scales as the input of the network at different levels and is able to map the spatial information in PAN images to each band of MS images well by learning the differential information of different levels, which effectively solves the scale effect of remote sensing images. An improved dense network with the same hierarchical structure is used to obtain richer spatial features to enhance the spatial information of the fused result. Meanwhile, a hybrid loss strategy is used to constrain the network at different levels for obtaining better results. Qualitative and quantitative analyses show that the result has a uniform spectral distribution, a complete spatial structure, and optimal evaluation criteria, which fully demonstrate the superior performance of the proposed method.

Keywords: pansharpening; multi-level; differential; deep learning

Citation: Yin, J.; Qu, J.; Chen, Q.; Ju, M.; Yu, J. Differential Strategy-Based Multi-Level Dense Network for Pansharpening. *Remote Sens.* **2022**, *14*, 2347. https://doi.org/10.3390/rs14102347

Academic Editors: Thien Huynh-The, Sun Le and Huang Wei

Received: 1 April 2022
Accepted: 11 May 2022
Published: 12 May 2022

1. Introduction

Remote sensing images are widely used in agriculture, the military, and other fields. Due to the limitation of sensor hardware, the sensor cannot acquire high-resolution multispectral (HRMS) images directly. In general, sensors provide two types of images that include high-resolution panchromatic (PAN) images and low-resolution multispectral (MS) images. However, HRMS images are usually required in most applications, such as hyperspectral image classification [1], hyperspectral image mixed denoising [2], etc.; therefore, pansharpening is proposed for fusing the spatial information of PAN image with the spectral information of MS images to obtain HRMS images.

The existing pansharpening methods can be divided into four branches: component substitution (CS) [3–6] methods, multi-resolution analysis (MRA) [7–11] methods, variational optimization (VO) [12–16] methods, and deep-learning (DL) methods.

CS-based methods replace the intensity component of the MS images with the PAN image. The classical CS-based methods include the partial replacement adaptive component substitution (PRACS) [6], the intensity-hue-saturation (IHS) method [3], the principal component analysis method (PCA) [4], the Gram–Schmidt method (GS) [5], the Brovery method [17], and others. These CS-based methods are the most widely used, but they often suffer from spectral distortion.

MRA-based methods inject the high-frequency spatial details of the PAN image into MS images. The typical MRA methods include high-pass filtering (HPF) [18], the Laplace pyramid transform method [9], the wavelet transform method [6], the contour wavelet transform method [10], the curvilinear transform method [11], and more. In contrast to CS-based methods, MRA-based methods are usually sensitive to spatial distortion but have less spectral distortion.

VO-based methods model the pansharpening problem as an optimization problem using some prior spatial knowledge of the image, and the solution to the optimization problem comprises HRMS images. At the core of VO-based methods is selecting and designing a suitable prior-image model. Representative VO-based methods include the gradient descent algorithm [19], the split Bregman iteration algorithm [20], the alternating direction method of multipliers (ADMM) algorithm [15], and others. VO-based methods can reduce the spectral distortion effectively, but they usually cause fuzzy results due to the lack of an effective prior model for spatial information retention.

In recent years, with the continuous development of deep learning, DL-based methods have been used in pansharpening. In DL, convolution neural network (CNNs) are the most widely used models. Therefore, an increasing number of scholars have improved the CNN and designed many excellent pansharpening models. Inspired by image super-resolution using deep convolutional networks (SRCNN) [21], Masi et al. proposed a pansharpening method by CNN (PNN) [22], which is the first time that they have combined CNN with pansharpening. In PNN, the simple three-layer convolutional architecture obtains spatial detail information and spectral information from PAN image and MS images, and the information obtained is used to enhance the spatial resolution of the fused images. PNN is not sufficiently enhanced for spatial detail as its network architecture is too simple.

To design a deeper CNN model, Wei et al. [23] introduced residual networks that take advantage of the high nonlinearity of DL models and avoid gradient disappearance and gradient explosion occurring when the network is too deep. The deep CNN model extracts deep features from PAN and MS images to improve fusion, but it is similar to PNN in that it treats the pansharpening task as a black box and does not consider the retention of spectral and spatial information. Yang et al. proposed a deep network architecture for pansharpening (PanNet) [24]. The PanNet model is effective in reducing spectral distortion and is more obvious in spatial detail enhancement, but it ignores low-frequency information in PAN and MS images, which also plays a role in fusion. To make better use of the information in PAN and MS images and to further enhance the details of the fusion images, Deng et al. proposed a detailed injection-based deep CNN [25] that uses DL combined with traditional methods. Specifically, combining CS and MRA strategies, Deng et al. proposed a Fusion-Net. The differential information in Fusion-Net between PAN image and up-sampled MS images is trained in a residual network, and spectral preservation is performed at the output. Therefore, some scholars have worked on designing a multi-scale CNN to obtain richer feature information. To obtain different scale information from source images, Wang et al. proposed multi-scale deep residual network (MSDRN) [26]. In MSDRN, three-layer network architecture is used to extract and fuse features at different scales, and a three-layer loss function is designed to control the training process at each layer. Wang et al. [27] first introduced dense connected blocks and residual learning for pansharpening to better learn the nonlinear mapping relationship between the input image and the target image, reducing the number of parameters and preventing overfitting at the same time. Deng et al. [28] proposed an SSConv to implement spectral-to-spatial mapping by introducing sub-pixel convolution and to supervise network training by using a multi-layer loss strategy.

Compared with the other methods, DL-based pansharpening methods can easily achieve better overall fusion accuracy. However, most DL-based methods are insufficient in using source image information and have limitations in feature extraction. In this paper, a differential strategy-based multi-level dense network for pansharpening (DS-MDNP) is proposed that makes novel changes to the input and the structure of the network and effectively enhances the spatial detail of fusion results. The main contributions of the paper are as follows:

1. A DS-MDNP is proposed to solve the pansharpening problem, which combines the difference strategy with a multi-level structure. Using the difference strategy can map the spatial information of PAN image to each band of MS images at different levels, and then the features of different levels are fuses, reducing the global error caused by

the difference between the two images effectively and enhancing the spatial structure of the fusion results.

2. A hybrid loss strategy consisting of MSE and MAE was proposed to achieve a balance between convergence speed and robustness. This hybrid loss strategy supervises and optimizes different layers of DS-MDNP and is trained by back-propagation, making full use of the rich feature hierarchy.
3. To learn more discriminative deep spatial features, the improved DenseNet is used as a backbone feature extraction network, encouraging the reuse of spatial and spectral features and enhancing feature propagation. In the improved DenseNet, the structure of the transition layer is modified to better integrate the feature maps in the DenseBlocks, which reduces the amount of computation and renders the network more efficient.

The remainder of this paper is as follows. Section 2 introduces the different strategies of input, and Section 3 presents the proposed DS-MDNP and provides a detailed description of each part of the network. Section 4 shows the experimental results and compares them with other methods, Section 5 discusses the experimental results, and Section 6 concludes the paper.

2. Different Strategies of Input

In DL-based pansharpening methods, different strategies of input have a significant impact on fusion results. Generally, there are three main input strategies for Pansharpening's approach, which are the overlapping original information strategy, the overlapping high-pass information strategy, and the differential information mapping strategy.

Most CNN-based pansharpening methods first overlap up-sampled MS images and PAN image, then input the stacked images into the network for training. Alternatively, methods extract feature maps from PAN image and up-sampled MS images separately and then stack these feature maps for training. As shown in Figure 1a, this mapping strategy can achieve good spectral fidelity but have insufficient spatial detail enhancement due to the spatial detail required for HRMS images mainly being derived from the PAN image.

Another strategy is to obtain the high-pass part of MS and PAN images and then up-sample the high-passed MS images into the same size as the PAN image. Next, stack the up-sampled high-passed MS and PAN images into the network, and finally, add the results with the up-sampled MS images to obtain HRMS images, as shown in Figure 1b. This strategy does not use the low-pass portion of the PAN image, which may lead to the under-utilization of spatial information.

To balance spectral fidelity and spatial detail enhancement [29,30], we introduce a new strategy that uses differential information between PAN and MS, as shown in Figure 1c. Using the difference strategy can map the spatial information of PAN image to each band of the MS images. This strategy first copies the PAN image along the channel dimension into the same channel numbers as the MS images; it then differentiates with the corresponding band of the up-sampled MS images and the duplicated PAN image.

Each of these three strategies has its own advantages and disadvantages. Finally, the differential strategy is chosen as the input of DS-MDNP, and the differential information at different scales is feature extracted and fused in DS-MDNP, which can well preserve the spatial structure in the image; finally, the up-sampled MS is used for spectral preservation, which allows fused images to possess obvious spatial detail enhancement and spectral fidelity.

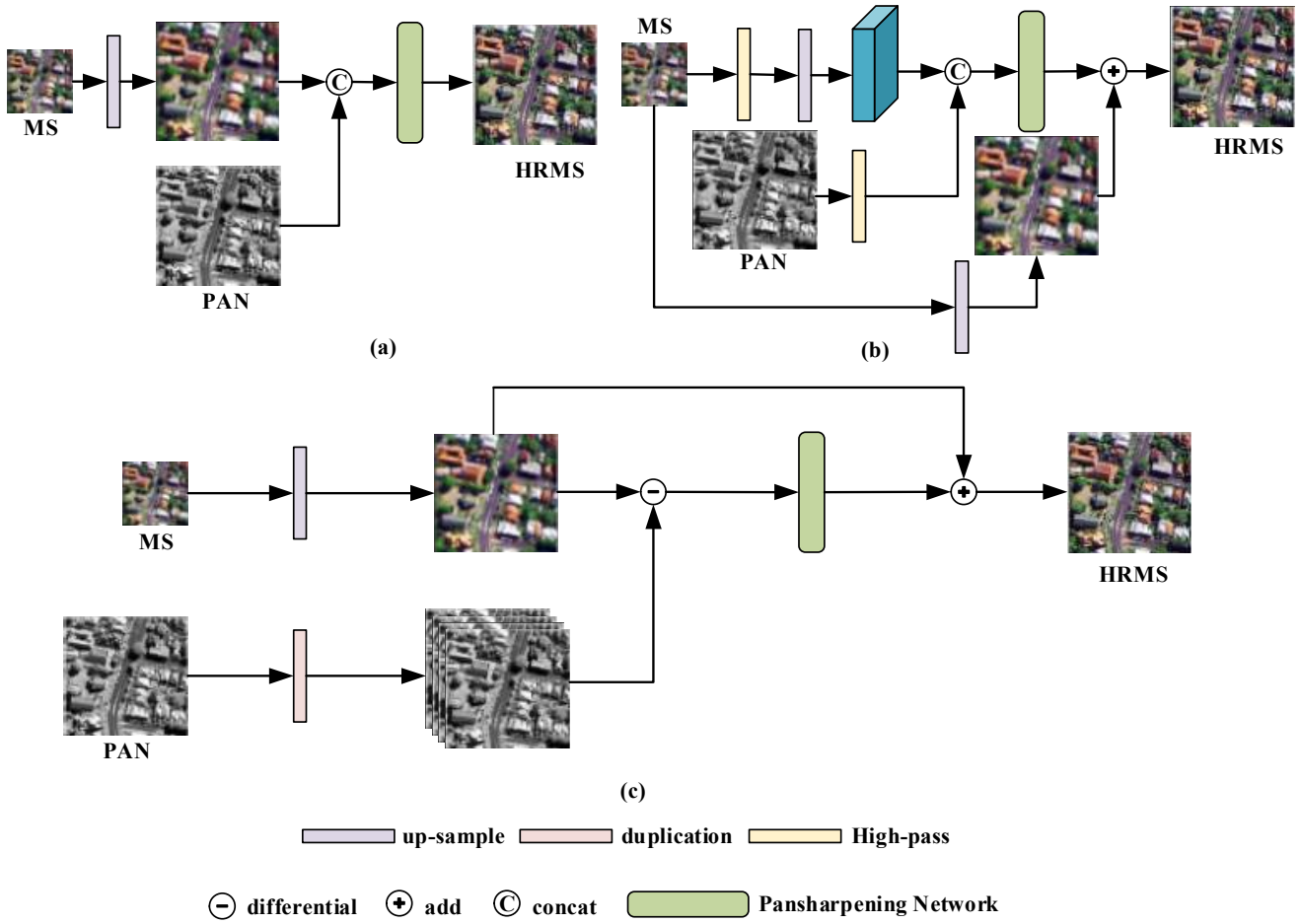

Figure 1. Different strategies of input: (**a**) the overlapping original information strategy, (**b**) the overlapping high-pass information strategy, and (**c**) the differential information mapping strategy.

3. Proposed Network

This section introduces the proposed network, which consists of three parts: acquisition of differential information, extraction and reconstruction of spatial features, and feature feedback. The structure of DS-MDNP is shown in Figure 2. The white blocks in Figure 2 represent differential images obtained from MS and PAN images at different scales, and these white blocks' size and number of channels are kept consistent with the corresponding input MS images.

For the convenience of modeling, the original MS images and PAN image are represented as MS and P, specifically, $MS \in \mathbb{R}^{H\times W\times s}$ and $P \in \mathbb{R}^{rH\times rW}$. H, W, and s represent the height, weight, and channels of MS images, and r represents the ratio of spatial resolution between the MS and PAN images. The network flowchart is for a 4-band MS image, and the size ratio of PAN and MS images are 4. The down-sampled P with 2×2 is denoted $P_{\downarrow 2}$, and the down-sampled P with 4×4 is denoted $P_{\downarrow 4}$. In contrast, the up-sampled MS with 2×2 is denoted $MS_{\uparrow 2}$, and the up-sampled MS with 4×4 is denoted $MS_{\uparrow 4}$. Overall, DS-MDNP can be summarized by Equation (1):

$$\left[\hat{MS}, \hat{MS}_{\downarrow 2}, \hat{MS}_{\downarrow 4}\right] = F_{\Theta_{DS\text{-}MDNP}}\left[\left(P, MS_{\uparrow 4}\right), \left(P_{\downarrow 2}, MS_{\uparrow 2}\right), \left(P_{\downarrow 4}, MS\right)\right] \quad (1)$$

where $F_{\Theta_{DS\text{-}MDNP}}$ represents the proposed network, and $\Theta_{DS-MDNP}$ denotes the parameters inside the network. $\hat{MS}$, $\hat{MS}_{\downarrow 2}$, and $\hat{MS}_{\downarrow 4}$ represent the outputs of DS-MDNP at different levels, and $\hat{MS}$ are the desired HRMS images.

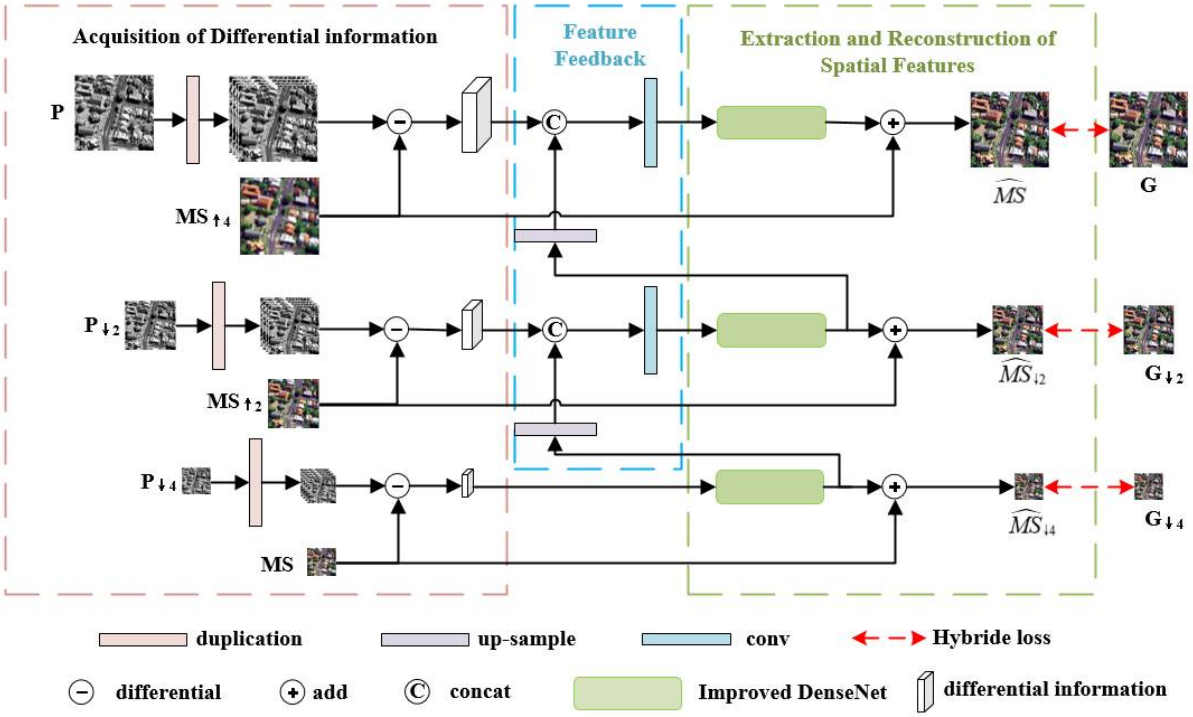

Figure 2. The framework of the proposed DS-MDNP.

In DS-MDNP, the PAN and MS images are pre-processed into three different levels that correspond to the three-layer architecture of DS-MDNP. The differential information is obtained from the three pairs of images at three different levels, and then the differential information at the bottom layer is input to an improved DenseNet for feature extraction; the output of the DenseNet is divided into two branches: One branch is injected into MS images to generate HRMS at the corresponding level, and the other branch is fused with differential information from the middle layer. The process in the middle layer is similar to that in the bottom layer—one branch of the middle layer output is fused with the differential information from the top layer. In the top layer of DS-MDNP, the output of DenseNet is injected into MS images to generate the desired HRMS images. At the same time, the training process is supervised by calculating the hybrid loss between the output of each layer of the proposed network and the corresponding scale of ground truth (GT) images.

3.1. Acquisition of Differential Information

To adapt the input to the multi-level network structure, differential inputs at three different levels are required [31]. The differential objects must have the same size and dimension; thus, the PAN image must be processed to the same dimension as MS images in three layers [32]. The copied PAN image along the channel dimension is denoted P^D. P^D, ${P^D}_{\downarrow 2}$, and ${P^D}_{\downarrow 4}$ are directly copied by P, $P_{\downarrow 2}$, and $P_{\downarrow 4}$, respectively. Finally, we differentiate the images in the same resolution to obtain the following three differential inputs, as shown in Equations (2)–(4):

$$I = P^D - MS_{\uparrow 4} \tag{2}$$

$$I_{\downarrow 2} = {P^D}_{\downarrow 2} - MS_{\uparrow 2} \tag{3}$$

$$I_{\downarrow 4} = {P^D}_{\downarrow 4} - MS \tag{4}$$

where I, $I_{\downarrow 2}$, and $I_{\downarrow 4}$ are the differential inputs in three levels, which are used to extract features in the next improved DenseNet.

3.2. Extraction and Reconstruction of Spatial Features

We use the improved DenseNet to extract and reconstruct spatial features from differential inputs. In Figure 2, the three Improved DenseNets used in the three different branches of the overall proposed network comprise three distinct networks. However, these three improved DenseNets are designed to the same architecture to make better use of the different levels of features. Different levels of DenseNet yield output feature maps of corresponding sizes depending on the input, and these DenseNets are designed to extract deeper spatial features from source images. The idea of DenseNet is to enhance the spreading of features and to encourage feature reuse. The improved DenseNet consists of convolution (Conv), batch normalization (BN), DenseBlock, and the Transition layer, as shown in Figure 3.

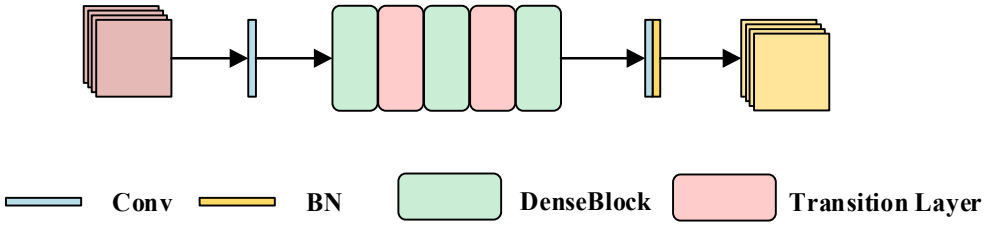

Figure 3. The framework of improved DenseNet.

The DenseBlock is used to stack features, which consists of three types of operation: BN, rectified linear unit (ReLU), and Conv, with a kernel size of 3×3, as shown in Figure 4.

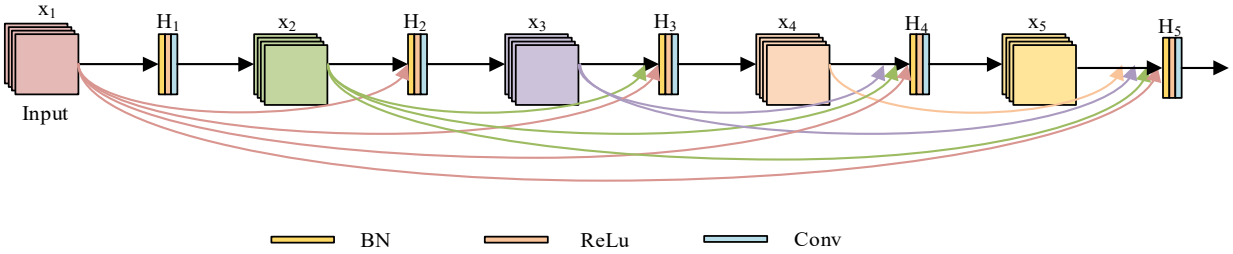

Figure 4. Architecture of DenseBlock.

In the DenseBlock, the feature maps are denoted x_l, $H_l(\cdot)$ represents the three successive operations: BN, ReLU, and Conv. The l-th layer feature map x_l is calculated as shown in Equation (5):

$$x_l = H_l([x_1, x_2, \ldots, x_{l-1}]) \tag{5}$$

where x_1, x_2, ..., x_{l-1} are the result of stitching the feature map from the first layer to the $(l-1)$-th layer. In Figure 4, the feature map after each $H_l(\cdot)$ is stacked with all previous feature maps; we denote $x_1 \in \mathbb{R}^{H \times W \times s}$, and then the final output is denoted as $x_l \in \mathbb{R}^{H \times W \times ls}$. For each layer of the DenseBlock, the feature maps of all previous layers are used as the input of the current layer, while their own feature maps are the input of the subsequent layers. The extracted feature maps from each layer are available for use in subsequent layers. This effectively prevents gradient disappearance, enhances feature propagation, and reduces the number of parameters. DenseBlock stacks all the layers into a feature map that has a large number of channels. This feature map contains much information but complicates computation, so the feature map needs to be compressed and is fused by a transition layer.

The transition layer is a module that connects different DenseBlocks, which integrate the features obtained from the previous DenseBlock. The structure of the transition layer is shown in Figure 5.

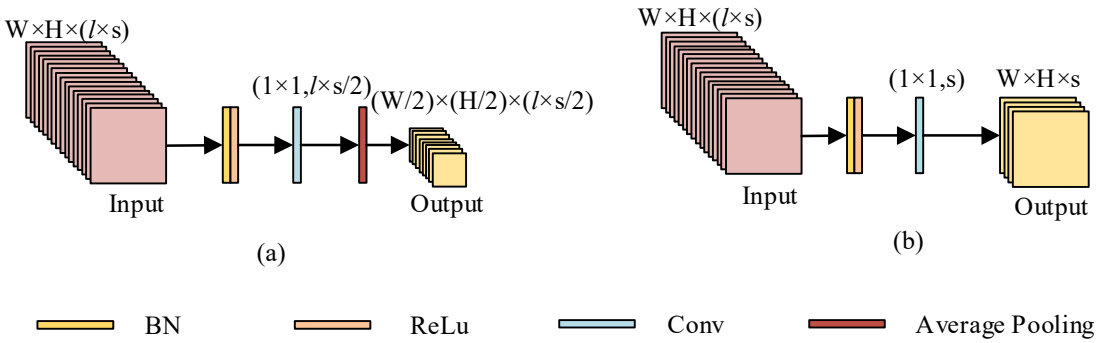

Figure 5. The transition layers: (**a**) the traditional transition layer and (**b**) the improved transition layer.

The traditional transition layer comprises BN, ReLU, Conv, and Average Pooling, as shown in Figure 5a. Average Pooling has a kernel size of 2 × 2 and a step size of 2. The channel numbers of the output are half of the input of the traditional transition layer. The improved transition layer comprises BN, ReLU, and Conv, as shown in Figure 5b. The 1 × 1 Conv used in the improved transition layer is different from the traditional transition layer. Specifically, the number of 1 × 1 Conv kernels in the traditional transition layer is half of the channel number of the input feature map. The number of 1 × 1 Conv kernels in the improved transition layer is the same as the channel number of MS image. This method aimed to reduce the number of channels in the output feature map, which reduces the computational effort and renders the network more efficient. In addition, in the improved transition layer, we remove Average Pooling to keep the size of the output feature map consistent with the input, and we change the number of 1 × 1 Conv kernels to reduce the number of channels in the output feature map. The purpose of the transition layer is to integrate the large number of feature maps obtained from DenseBlock. The purpose of the improved transition layer is to integrate features while keeping the feature size constant, reducing the loss of spatial information and reducing the computational effort, thus improving network efficiency.

In Section 3.1, the three differential inputs are obtained, which are then inputted into the improved DenseNet. The bottom output of DS-MDNP can be calculated as shown in Equation (6):

$$\hat{MS}_1 = f\left(I_{\downarrow 4}\right) \oplus MS \tag{6}$$

where f denotes the improved DenseNet, and '$\oplus$' represents pixel-by-pixel addition. The output of improved DenseNet $f\left(I_{\downarrow 4}\right)$ has two branches: one is constructed in HRMS, as shown in Equation (6), and the other is fused with the differential information of the middle layer by feature feedback connection.

3.3. Feature Feedback

To make full use of the features obtained through DenseNet at each layer, feedback connections were made among the three layers of DS-MDNP. The feedback connection consists of three parts: up-sampling, stacking, and Conv [33]. The middle output of DS-MDNP can be calculated as shown in Equation (7):

$$\hat{MS}_2 = f\left(\left[f\left(I_{\downarrow 4}\right)_{\uparrow 2}, I_{\downarrow 2}\right]_C\right) \oplus MS_{\uparrow 2} \tag{7}$$

where C represent the Conv with the kernel size of 3 × 3 and the kernel number of 4. The output of bottom DenseNet is up-sampled to $f(I_4)_{\uparrow 2}$ and then stacked with $I_{\downarrow 2}$ along the channel dimension to obtain the fused feature map, which is convoluted to obtain the output of the middle layer. The output DenseNet has two branches in the middle layer of DS-MDNP, which is the same as the bottom layer. One branch is constructed into HRMS, as shown in Equation (7), and the other branch is fused with the differential information

of the top layer by feature feedback connection. In the top layer of DS-MDNP, the desired HRMS can be calculated as shown in Equation (8).

$$\hat{MS} = f\left(\left[f\left(\left[f\left(I_{\downarrow 4}\right)_{\uparrow 2}, I_{\downarrow 2}\right]_C\right)_{\uparrow 2}, I\right]_C\right) \oplus MS_{\uparrow 4} \tag{8}$$

Specifically, the output of DenseNet in the middle layer is up-sampled and then stacked with I to obtain the fused feature; the fused feature is convoluted and then inputted into DenseNet. Finally, the output of DenseNet is added with $MS_{\uparrow 4}$ to generate the desired HRMS.

Overall, the main purpose of feature feedback is to fuse the multi-level features. $\hat{MS}_{\downarrow 2}$ and $\hat{MS}_{\downarrow 4}$ are used for the proposed hybrid loss strategy to obtain better fusion results, and the details are described in Section 3.4.

3.4. Hybrid Loss Strategy

A hybrid loss strategy is proposed to supervise the fusion of features at different layers. We compare the three outputs of DS-MDNP with the GT images of corresponding size, respectively. The GT image is denoted G, the medium-size GT is denoted $G_{\downarrow 2}$, and the low-size GT is denoted $G_{\downarrow 4}$. Finally, the hybrid loss of DS-MDNP is defined as Equation (9):

$$Loss(\Theta_{DS-MDNP}) = \lambda_1 \Phi\left(G, \hat{MS}\right) + \lambda_2 \Phi\left(G_{\downarrow 2}, \hat{MS}_{\downarrow 2}\right) + \lambda_3 \Phi\left(G_{\downarrow 4}, \hat{MS}_{\downarrow 4}\right) \tag{9}$$

where $\Phi(\cdot)$ is composed of Mean Squared Error (MSE) and Mean Absolute Error (MAE), as shown in Equation (10). λ_1, λ_2, and λ_3 are three proportionality coefficients, which are set as [0.5,0.3,0.2] and are inspired by [28] and experimentally validated. Details of the experiments are shown in Section 4.3.1:

$$\Phi\left(G, \hat{MS}\right) = \frac{1}{N}\sum_{n=1}^{N}\left(G^{\{n\}} - \hat{MS}^{\{n\}}\right)^2 + \frac{1}{N}\sum_{n=1}^{N}\left|G^{\{n\}} - \hat{MS}^{\{n\}}\right| \tag{10}$$

where $\hat{MS}^{\{n\}}$ are the desired HRMS images generated by the PAN and MS images in a set of training samples, and $G^{\{n\}}$ is the GT image in this set of training samples. Some studies have shown that MSE is sensitive to outliers, producing results that cannot represent the overall error of the fused image. MAE is more robust to outliers but converges more slowly. Therefore, the proposed hybrid loss strategy uses a combination of these two losses.

4. Experiments

4.1. Datasets

The network architecture in this study uses the TensorFlow deep learning framework and is trained on an NVIDIA GeForce RTX 2070 GPU. We set the epochs to 200,000 and batch size to 64, and we use the Adam optimization algorithm to optimize the model by setting the learning rate to 0.001. Three datasets from different satellites were used to evaluate the performance of DS-MDNP. The three datasets are QuickBird, GaoFen-1, and WorldView-2, which are described in detail next. Our experiments are performed on three datasets, which is trained first, respectively, and then tested on each dataset.

4.1.1. QuickBird Dataset

The QuickBird satellite is one of the world's first commercial satellites to offer a sub-meter resolution, and it has high geolocation accuracy. The QuickBird dataset provides a PAN image resolution of 0.61 m and MS images resolution of 2.44 m. The QuickBird dataset can be used for drawing maps, changing detection, and image analysis. Generally, we use the QuickBird dataset with four bands: blue, green, red, and near-infrared (NIR).

4.1.2. GaoFen-1 Dataset

GaoFen-1 satellites are the first satellite in the National High-Resolution Earth Observation System (NHROS) Major Project Space-Based System. NHROS provides PAN image resolutions of 2 m and MS images resolutions of 8 m. The GaoFen1 dataset we used has four bands: blue, green, red, and NIR.

4.1.3. WorldView-2 Dataset

WorldView-2 satellites are the first commercial satellites in the world to use Controlled Moment Gyros (CMGs). This high-performance technology provides up to 10 times more acceleration for attitude control operations, allowing for more accurate targeting and scanning. The WorldView-2 dataset provides a PAN image resolution of 0.5 m and MS images resolution of 1.8 m. Normally, the WorldView-2 dataset we use has four bands, which are blue, green, red, and NIR.

4.1.4. Dataset Preprocess

We divided the dataset into a training set and a validation set in a 4:1 ratio by cutting, disrupting, and randomly selecting the original satellite images, as shown in Table 1.

Table 1. Size of training and validation sets for GaoFen-1, QuickBird, and WorldView-2.

Dataset	Training Set	Validation Set	Size of Original PAN Image	Size of Original MS Images
QuickBird	2187	546	16,251 × 16,004	4063 × 4001 × 4
GaoFen-1	2779	694	18,192 × 18,000	4548 × 4500 × 4
WorldView-2	1603	400	16,384 × 16,384	4096 × 4096 × 4

To facilitate training and testing, we cropped the images of all three datasets to the same size, as shown in Table 2. To reduce training time, we use images with a size of 64 × 64 for training. To show more details of the fused image, we use images with a size of 256 × 256 for testing and evaluation indices with reference assessment in comparative experiments. To verify the robustness of the model on real experiments and to demonstrate the variability of different methods for spectral and spatial enhancement, we use images with sizes of 1024 × 1024 for real experiments and non-referenced evaluation index assessment.

Table 2. Size of PAN image, MS images, and GT images for GaoFen-1, QuickBird, and WorldView-2.

Dataset	PAN	MS	GT
QuickBird	64 × 64	16 × 16 × 4	64 × 64 × 4
GaoFen-1	64 × 64	16 × 16 × 4	64 × 64 × 4
WorldView-2	64 × 64	16 × 16 × 4	64 × 64 × 4

4.2. Quantitative Evaluation Indices

The evaluation method can employ subjective visual evaluation as well as objective evaluation indices. Specifically, subjective visual evaluation relies on the human eye to make subjective judgments on the effect of fusion images. However, many objective evaluation metrics can be used to accurately evaluate fusion results; thus, we selected five commonly used evaluation indices with references, which are the spectral angle mapper (SAM) [34], Erreur Relative Global Adimensionnelle de Synthèse (ERGAS) [35], correlation coefficient (CC) [36], universal image quality index (Q) [37], and an extended version of Q ($Q2^n$) [38]. We have also chosen a non-referenced evaluation index, namely quality with no reference (QNR) [39].

Specifically, SAM is an evaluation indicator that measures the spectral distortion of the fused image compared to the reference image, which is expressed as the absolute value of the spectral angle between the two images. The smaller SAM is, the lower the spectral

distortion; that is, if SAM is zero, then there is no spectral distortion. ERGAS represents the synthetic error for all bands. The smaller ERGAS is, the better the spectral quality of the fused image over the spectral range. CC is the most widely used similarity metric in the field of pansharpening, quantifying the proximity between the fused image and the reference image based on a correlation function. The higher the CC, the more spatial information present in the fused image. Q is a universal objective quality index that is simple to calculate and is suitable for the quality assessment of various image applications: the value of Q is -1 to 1, and the closer the value of Q is to 1, the higher quality of the fused image. $Q2^n$ is a common quality evaluation index and global metric in the field of pansharpening. QNR is a method for the quantitative evaluation of pansharpening images without reference. QNR is calculated from two factors, the spectral distortion index and the spatial distortion index, and these two indexes are based on the Q. A maximum value of 1 is obtained for QNR when both the spatial and spectral distortions of the fused image are zero; the formula is shown in (11):

$$QNR = (1 - D_\lambda)^\alpha \cdot (1 - D_S)^\beta \tag{11}$$

where spectral distortion and spatial distortion are quantified by D_λ and D_S.

4.3. Experiments and Analysis

In this section, we present ablation experiments and comparative experiments. The ablation experiments demonstrate the process of optimizing the proposed framework and ultimately identifying the option with the best fusion performance. The comparative experiments are designed to demonstrate the superiority of our proposed framework compared to traditional methods and classical approaches of DL-based methods.

4.3.1. Ablation Experiments

To design an optimal scheme for fusion, we designed five experimental approaches based on different input information, feature extraction networks, and loss functions. In detail, the input can use differential information or stacked information, the loss function can use MSE or a hybrid loss of MSE and MAE, and the feature extraction network can use ResNet [40] or DenseNet. The experimental results are shown in Table 3, and experimental results are plotted in Figure 6.

Table 3. Quantitative evaluation results of ablation experiments on the WorldView-2 dataset. The values in bold represent the best results.

Method	Strategies			Quantitative Evaluation Indices				
	Differential	**DenseNet**	**Hybrid Loss**	**SAM**	**ERGAS**	**Q**	**$Q2^n$**	**CC**
1				3.5142	2.1057	0.9342	0.9309	0.96439
2	√			2.6416	1.8835	0.9432	0.9473	0.9754
3		√		2.6377	1.8916	0.9595	0.9479	0.9785
4	√	√		2.6085	1.8952	0.9456	0.9509	0.9757
5	√	√	√	**2.5606**	**1.8755**	**0.9751**	**0.9624**	**0.9865**

By comparing ablation experiments 1 and 2, it can be found that the differential input is more effective in improving model performance because more spatial structure information can be obtained from differential information, which is more obvious for the detail enhancement of fusion results. By comparing ablation experiments 1 and 3, using the dense network as the backbone network to extract features has a better advantage on feature enhancement than the residual network, because the network in the dense network is deep enough to obtain deep features and, at the same time, can reuse shallow features to enhance feature propagation, which allows the model to perform better. By comparing ablation experiments 4 and 5, using the hybrid loss strategy in DS-MDNP has better results

than single loss. In Figure 6, by comparing the fused images of the ablation experiments, DS-MDNP has the best performance in spatial detail enhancement.

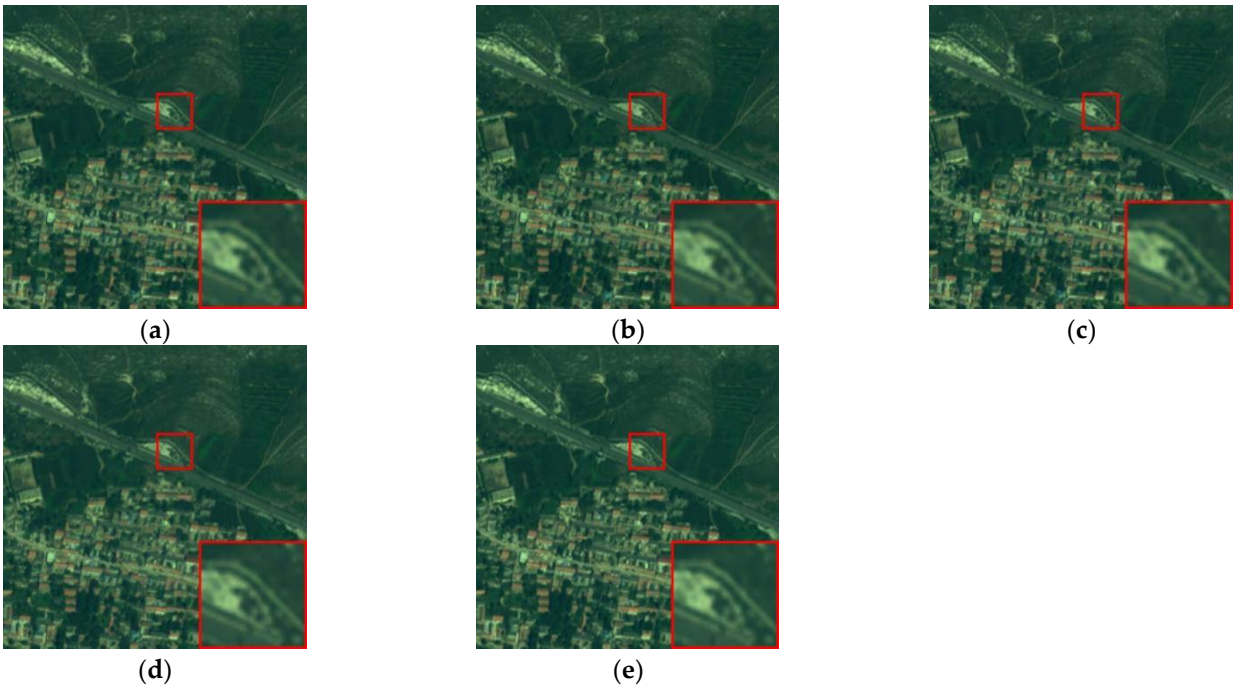

Figure 6. Fused image of WorldView-2. (**a**) Base line, (**b**) differential, (**c**) DenseNet, (**d**) differential + DenseNet, (**e**) differential + DenseNet + Hybrid Loss.

To verify the parameter settings in the mixture loss, we performed some experiments. The experimental setup was to first select one of the parameters to be kept constant and then change the other two parameters using Q2n, CC, and QNR as evaluation metrics. Since the size of the first layer of the network is the same as the size of the fusion result, we assumed that the feature contribution of the first layer is the largest, and we set the weight of the first layer to 0.5, and the remaining 0.5 is divided into two parts and allocated to the other two parameters. In Figure 7, the experimental results show that the model performs best when λ1, λ2, and λ3 are set to 0.5, 0.3, and 0.2, respectively.

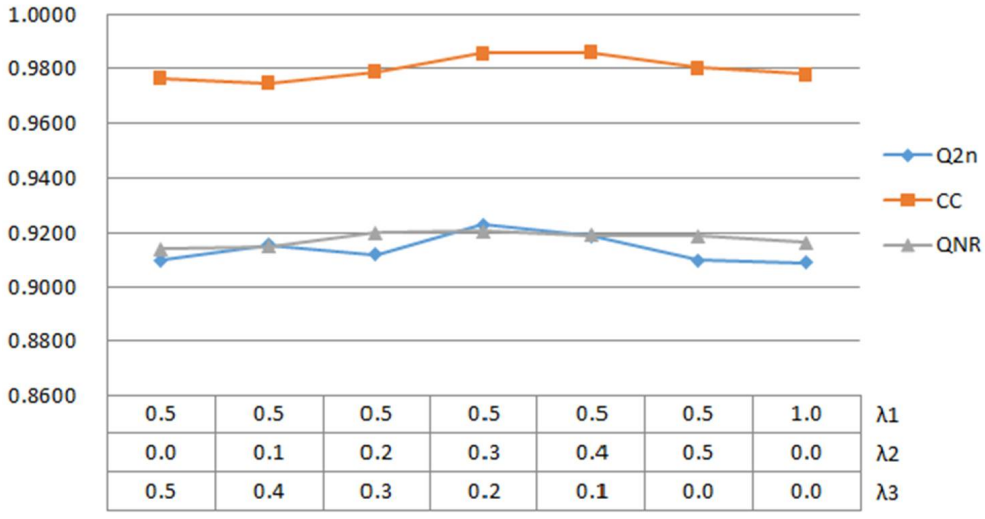

Figure 7. Line graph for mixed loss parameter validation.

4.3.2. Comparative Experiments

In this subsection, to demonstrate the effectiveness of DS-MDNP, we compare several traditional MRA-based and CS-based methods and some classical CNN-based pansharpening methods. Specifically, these methods are IHS [3], Wavelet [8], HPF [18], PRACS [6], PNN [22], Fusion-Net [25], SSConv [28], and DS-MDNP. To verify the robustness of DS-MDNP, we use three datasets for our experiments, namely, QuickBird, GaoFen-1, and WorldView-2. For QuickBird, the experimental results with reference evaluation indicators are shown in Table 4, and the experimental results are shown in Figure 8.

Table 4. Quantitative evaluation comparison of fusion results on the QuickBird dataset. The values in bold represent the best result.

Method	SAM	ERGAS	Q	$Q2^n$	CC
IHS	4.6574	2.6945	0.9137	0.6521	0.9141
Wavelet	4.2887	3.4220	0.9068	0.6884	0.8771
HPF	4.5232	2.7628	0.9134	0.6617	0.9022
PRACS	2.7181	1.8431	0.9640	0.7682	0.9668
PNN	2.6906	1.7205	0.9635	0.8718	0.9687
Fusion-Net	**1.7285**	1.2036	0.9497	0.9172	0.9840
SSConv	1.7875	1.2364	0.9818	0.9086	0.9825
DS-MDNP	1.7560	**1.1964**	**0.9834**	**0.9231**	**0.9857**

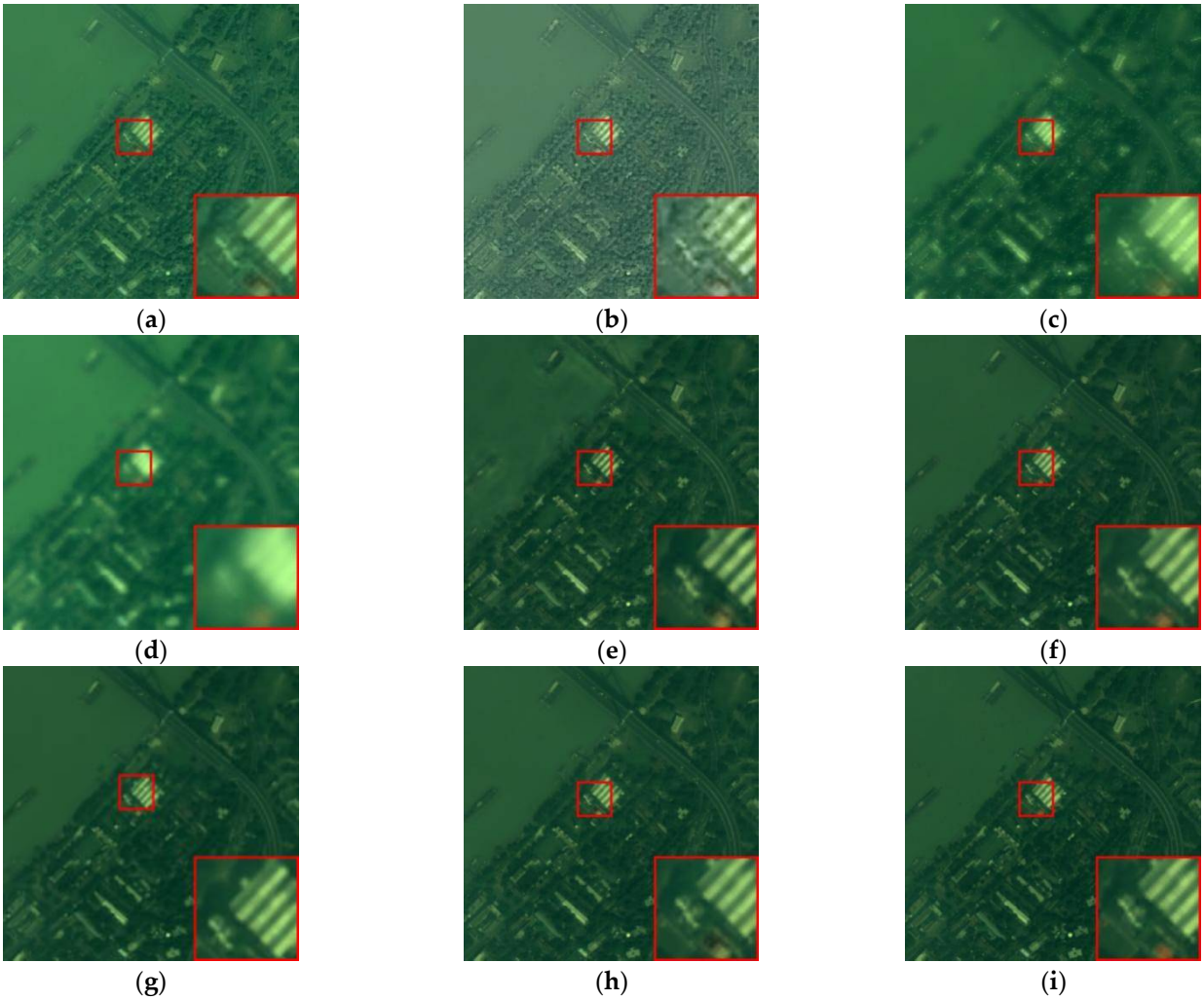

Figure 8. Fused images of QuickBird where (**a**) IHS, (**b**) Wavelet, (**c**) HPF, (**d**) PRACS, (**e**) PNN, (**f**) Fusion-Net, (**g**) SSConv, (**h**) DS-MDNP, and (**i**) GT.

In Table 4, it is obvious to see that SAM of the traditional methods, such as IHS, Wavelet, and HPF, are higher, which indicates that the spectral distortion of these traditional methods is more pronounced in contrast to the DL-based method, which has less spectral distortion, while DS-MDNP and Fusion-Net have the lowest spectral distortion. For multiple band spectral quality, ERGASs of the DL-based methods are lower than these traditional methods, which indicates that the spectral quality of the fusion results by DL-based methods is better, and DS-MDNP achieves the best spectral quality with a value of 1.1964, which validates the advantages of DS-MDNP in terms of spectral retention. Finally, Q and $Q2^n$ are global quality evaluation criteria that can objectively reflect the overall effectiveness of the fusion results. In Table 4, we can see that the values of $Q2^n$ obtained by the DL-based methods are all above 0.80, which indicates that the DL-based methods are more effective in the overall evaluation than these traditional methods, and DS-MDNP is able to obtain the best value of 0.9231. CC reflects the correlation between the fusion results and the reference image. According to Table 4, we can see that the correlation between the fusion results and the reference image is not very different for all methods, except for Wavelet. DS-MDNP and Fusion-Net were able to achieve the best correlation, reaching above 0.98. In summary, DS-MDNP is able to obtain good fusion performance for the reference evaluation. In addition, in Figure 8, we clearly see that DL-based methods (e), (f), (g), and (h) are the closest to reference image (i). The fused images obtained by the traditional methods are slightly less spatially detailed.

To verify the robustness of DS-MDNP, we also experimented in GaoFen-1. In detail, the experimental results with reference evaluation indicators are shown in Table 5, and the experimental results are shown in Figure 9.

Table 5. Quantitative evaluation comparison of fusion results on the GaoFen-1 dataset. The values in bold represent the best result.

Method	SAM	ERGAS	Q	$Q2^n$	CC
IHS	1.2739	1.1711	0.9438	0.6288	0.9480
Wavelet	1.3809	1.0898	0.9593	0.7231	0.9567
HPF	1.3013	1.0562	0.9638	0.7634	0.9614
PRACS	1.2743	0.9921	0.9642	0.7705	0.9645
PNN	1.1612	0.8953	0.9586	0.8486	0.9528
Fusion-Net	1.0354	0.8778	0.9577	0.8956	0.9546
SSConv	1.0315	0.7912	0.9590	0.9012	0.9551
DS-MDNP	**1.0276**	**0.6907**	**0.9682**	**0.9240**	**0.9703**

In Table 5, the experiments show that the values are quite different from the results trained on QuickBird: The reason is that different datasets are collected in different scenarios. The spectral information collected for the same material varies depending on various factors such as the weather and temperature of the environment. In GaoFen-1, we can see that SAMs for all eight methods chosen are not very different, which indicates that the spectral losses for these fusion methods are low for GaoFen-1. In contrast, the difference of ERGAS is large, specifically, the traditional methods have an ERGAS above 0.9, while DL-based methods have an ERGAS below 0.9, and DS-MDNP ahieved the lowest value of 0.6907, which again demonstrates the advantage of DS-MDNP in terms of spectral retention. In addition, the fusion quality Q for each band shows that, band-by-band, DS-MDNP fusion results are also the best among these methods. Finally, overall quality evaluation $Q2^n$ shows that the DL-based methods generally have higher fusion quality than these traditional methods, and DS-MDNP is the highest among the DL-based methods. Similarly to QuickBird, the correlation coefficients for all these methods are not very different. In Figure 9, by looking at the fusion result plots in the figure, we can clearly see that DS-MDNP outperforms other methods in terms of texture and detail, and its fusion results are closest to Ground Truth, which fully demonstrates that DS-MDNP has also good fusion results in GaoFen-1.

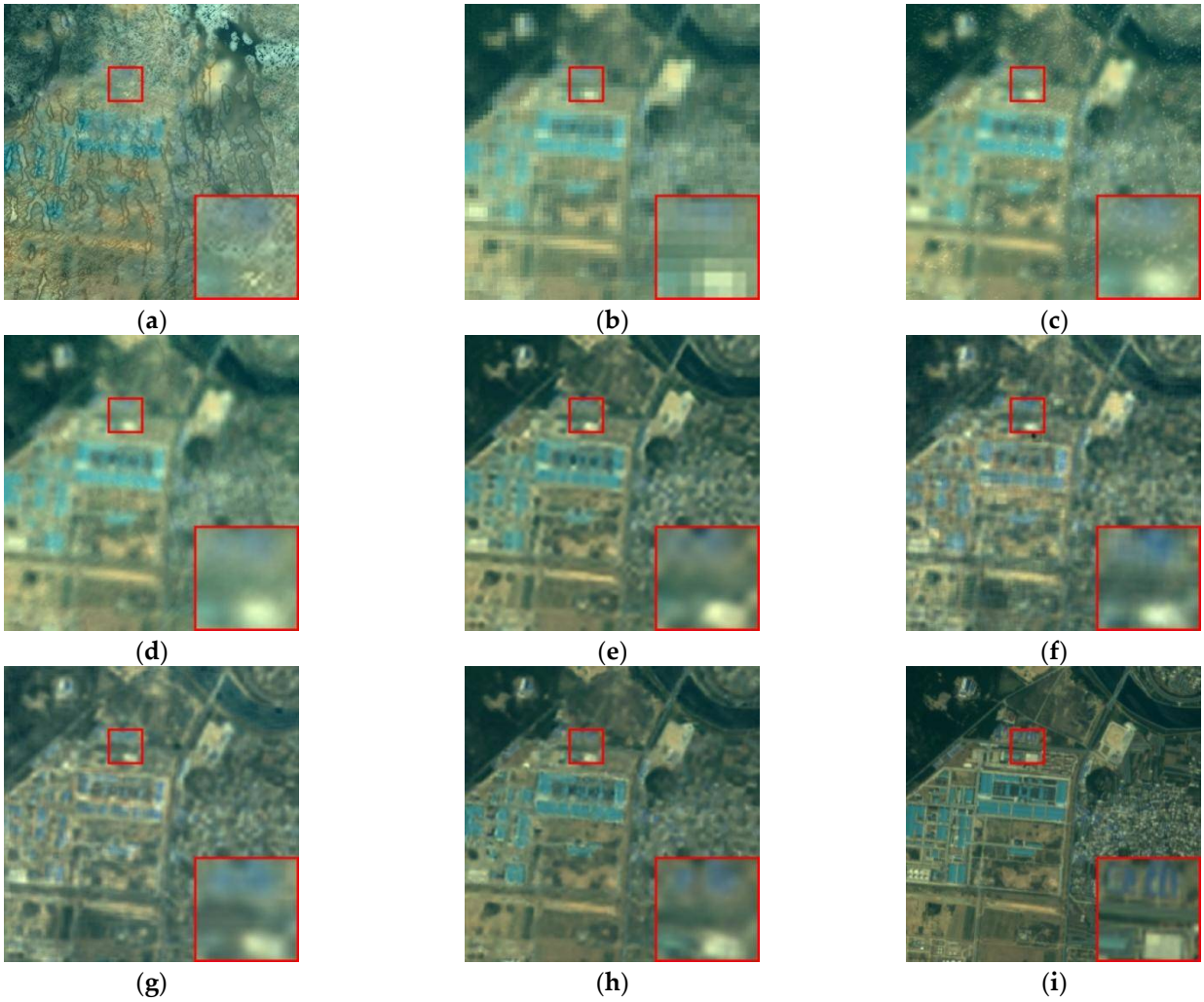

Figure 9. The fused images of GaoFen-1 where (**a**) IHS, (**b**) Wavelet, (**c**) HPF, (**d**) PRACS, (**e**) PNN, (**f**) Fusion-Net, (**g**) SSConv, (**h**) DS-MDNP, and (**i**) GT.

To verify the generalization of DS-MDNP, we also conducted a comparison experiment on WorldView-2. Compared with the above two datasets, the WorldView-2 dataset has obvious geometric structure and texture features of the image features; for example, river and channel mines and water works are clearly visible. Thus, there is more information available in this dataset for DS-MDNP. The specific experimental results are shown in Table 6, and the experimental results are shown in Figure 10.

Table 6. Quantitative evaluation comparison of fusion results on the WorldView-2 dataset. The values in bold represent the best result.

Method	SAM	ERGAS	Q	$Q2^n$	CC
IHS	4.7252	3.5953	0.9343	0.8487	0.9577
Wavelet	5.5041	3.7325	0.9211	0.8417	0.9500
HPF	4.7003	3.7872	0.9278	0.7992	0.9458
PRACS	4.6301	3.1958	0.9356	0.8706	0.9587
PNN	3.7531	2.9329	0.9563	0.9007	0.9708
Fusion-Net	2.6965	1.9571	0.9703	0.9559	0.9849
SSConv	2.5626	1.8764	0.9746	0.9580	0.9860
DS-MDNP	**2.5606**	**1.8755**	**0.9751**	**0.9624**	**0.9865**

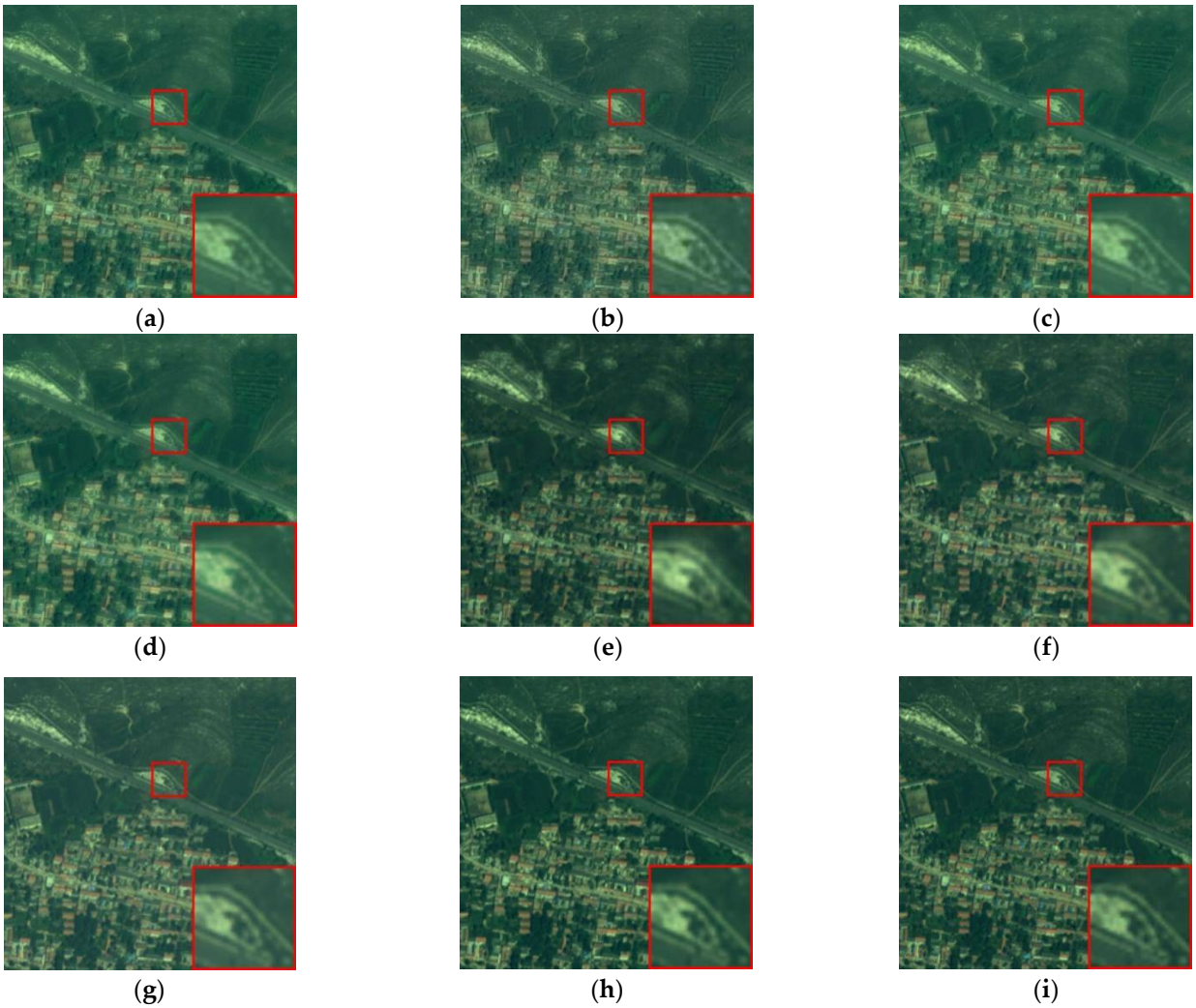

Figure 10. Fused images of WorldView-2 where (**a**) IHS, (**b**) Wavelet, (**c**) HPF, (**d**) PRACS, (**e**) PNN, (**f**) Fusion-Net, (**g**) SSConv, (**h**) DS-MDNP, and (**i**) GT.

In Table 6, we can see that the values of $Q2^n$ are better in WorldView-2 compared to QuickBird and GaoFen-1. The reason is that the dataset collected by WorldView-2 has richer texture features, which also proves that the variability of different datasets also has an impact on the training of the network. By looking at SAM and ERGAS in Table 6, it can be seen that the spectral distortion of these traditional methods is more pronounced, and the spectral quality is relatively poor. In Figure 10, we can see that DS-MDNP appears sharper in the edge texture parts of roads and looks closer to GT.

We compared the times of the selected methods for experiments on the WorldView-2 dataset, as shown in Table 7. Due to the complex design of the network structure, a large number of feature maps are generated in the process using the improved DenseNet. Although the number of feature maps is compressed at each transition layer, a considerable amount of computation is inevitably required to compress these feature maps, and each layer of the network has to output HRMS images for calculating the loss; thus, the training time of the proposed method becomes longer. However, for testing, only desired HRMS images are generated, eliminating the need to reconstruct the intermediate and underlying layers of the network; therefore, the testing time is relatively short. The superiority of DS-MDNP is further demonstrated by the fact that the proposed method maintains as much testing speed as possible while ensuring good performance.

Table 7. Comparison of training and testing times on WorldView-2.

Method	Training Time (s)	Testing Time (s)
IHS	-	1.44
Wavelet	-	1.17
HPF	-	7.98
PRACS	-	0.43
PNN	3.99	0.36
FusionNet	4.57	0.41
SS-Conv	5.52	0.30
DS-MDNP	11.14	0.38

4.3.3. Real Experiments

In practical applications, we need to fuse the original MS and PAN images. In the actual experiments, MS and PAN images, which are not degraded, are used as input to generate fused images using the parameters trained in the simulation experiments. Specifically, the fusion results of WorldView-2 are shown in Figure 11, the fusion results of QuickBird are shown in Figure 12, and the fusion results of GaoFen-1 are shown in Figure 13.

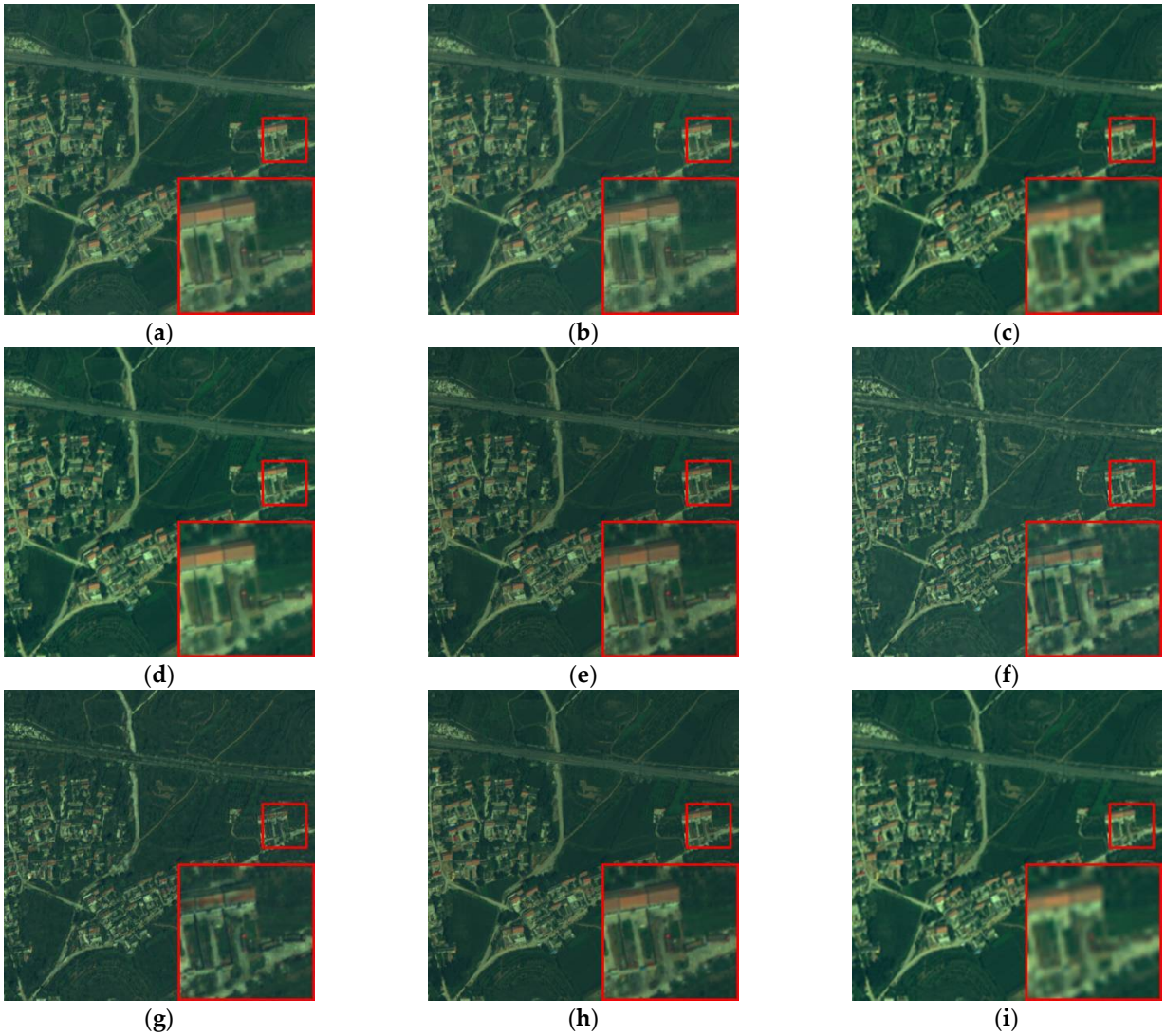

Figure 11. Fused images of WorldView-2 where (**a**) IHS, (**b**) Wavelet, (**c**) HPF, (**d**) PRACS, (**e**) PNN, (**f**) Fusion-Net, (**g**) SSConv, (**h**) DS-MDNP, and (**i**) Bicubic.

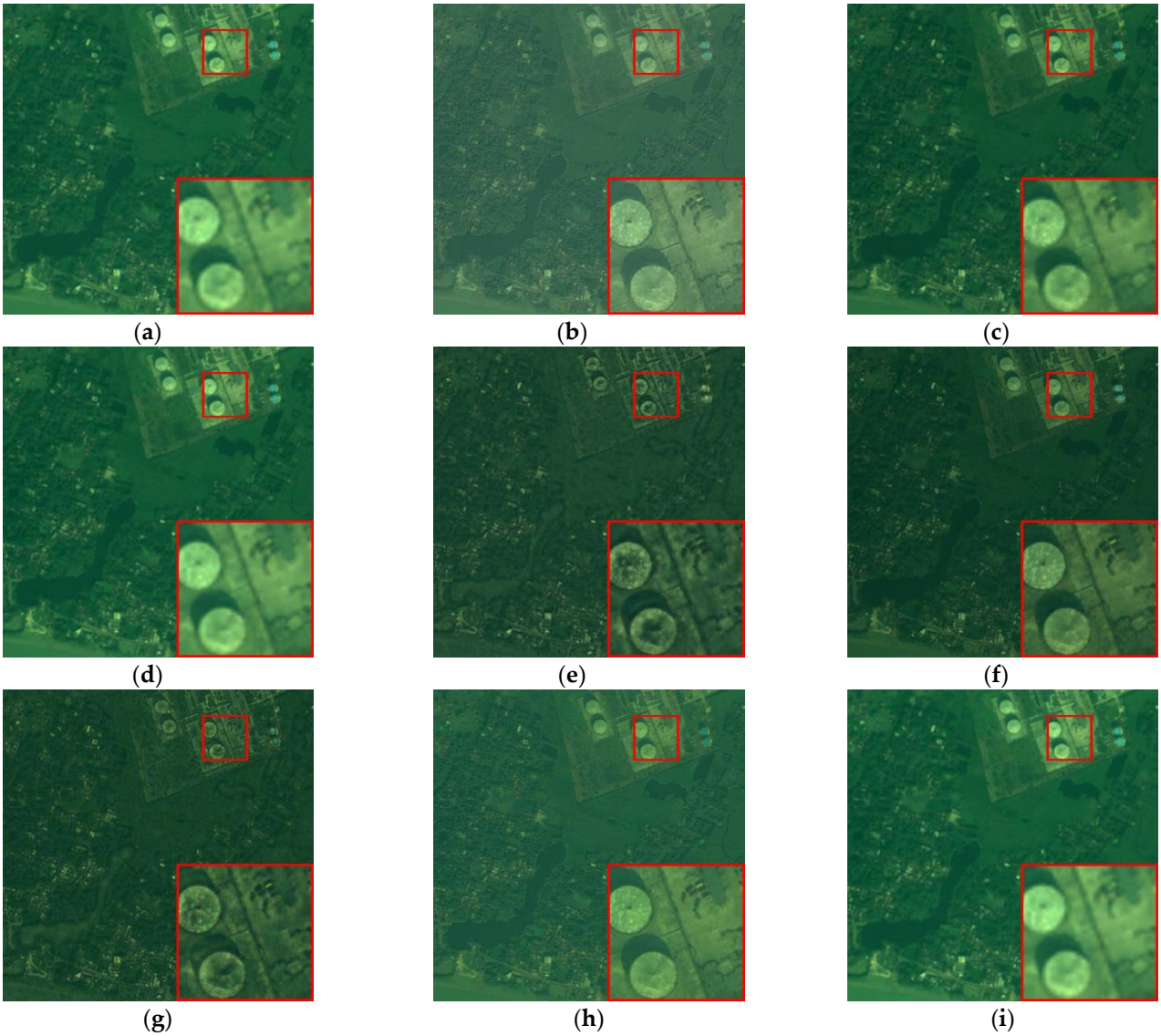

Figure 12. Fused images of QuickBird where (**a**) IHS, (**b**) Wavelet, (**c**) HPF, (**d**) PRACS, (**e**) PNN, (**f**) Fusion-Net, (**g**) SSConv, (**h**) DS-MDNP, and (**i**) Bicubic.

In Figures 11–13 we can see that the fused results were significantly enhanced compared with those obtained by Bicubic, particularly for the framed parts, and the bottom right corner shows the framed parts with 3X magnification. In detail, the framed parts in Figures 11 and 12 are roads, and the framed parts in Figure 13 are buildings. By comparing the three figures, it is noticeable that there is a significant difference in the clarity of the datasets collected by different satellites. In our selected images, it can be seen that WorldView-2 is clearer compared to QuickBird and GaoFen-1. Similarly, in the experimental results of each dataset, the comparison between DL-based methods and the traditional methods can be seen as a significant difference, and the fused results of DL-based methods have richer texture and detailed information. Specifically, in Figure 11, in the framed part of the fused result obtained by Wavelet and HPF, we can see that there is distortion, as the edges are jagged for most of the amplification. In Figure 12, the colors of the fused images obtained by these methods are different because the degree of preservation of spectral information in each band is distinctive. In addition, the fused images obtained by Wavelet and HPF have some blurring, which indicates the poor performance of these two methods in QuickBird dataset.

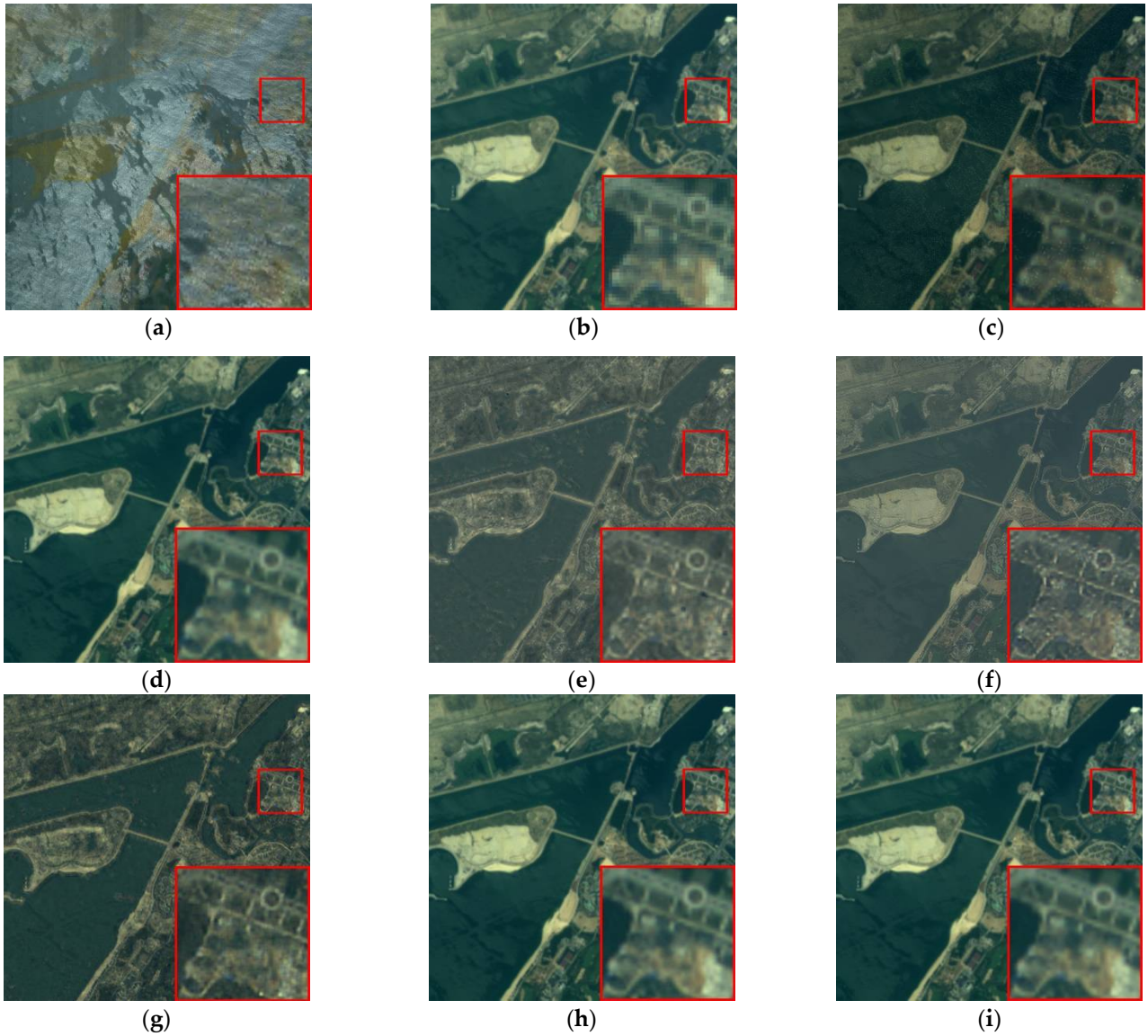

Figure 13. Fused images of GaoFen-1 where (**a**) IHS, (**b**) Wavelet, (**c**) HPF, (**d**) PRACS, (**e**) PNN, (**f**) Fusion-Net, (**g**) SSConv, (**h**) DS-MDNP, and (i) Bicubic.

In addition, the fused images obtained by PNN, Fusion-Net, and SSConv are darker in color but superior in clarity compared to traditional methods, and DS-MDNP is more similar in color and is relatively clear in resolution relative to the results obtained by Bicubic, which also deconstruct a better generalization of DS-MDNP. In Figure 13, we can see that GaoFen-1 has poorer image resolutions than the other two datasets, and the fusion results obtained by the DL-based methods perform better overall than the selected traditional methods. In these traditional methods, the fusion results obtained by IHS have obvious spectral distortion compared to PRACS, which is closer to the fusion results obtained by DL-based methods. In the enlarged image, it can be found that the fused image obtained by DS-MDNP has more obvious enhancements for texture edges; in detail, the layering between buildings is clearer and the edge parts of each building are recognizable. Finally, by evaluating the indicators, we use a non-reference evaluation metric for objectively evaluating the spectral and spatial distortion of the fusion results, as shown in Table 8.

Table 8. Quantitative evaluation comparison of real data experiments on three datasets. The values in bold represent the best result.

Method	QuickBird			GaoFen-1			WorldView-2		
	QNR	D_λ	D_S	QNR	D_λ	D_S	QNR	D_λ	D_S
IHS	0.8278	0.0899	0.0904	0.8032	0.0246	0.1753	0.8893	0.0267	0.0855
Wavelet	0.5479	0.3132	0.2023	0.9300	0.0153	0.0464	0.8257	0.1149	0.0628
HPF	0.8505	0.0416	0.1126	0.9320	0.0078	0.0607	0.8479	0.0580	0.0999
PRACS	0.8463	0.0472	0.1118	0.9284	0.0643	0.0507	0.8852	0.0326	0.0850
PNN	0.9235	0.0405	0.0375	0.9369	0.0260	0.0380	0.8587	0.0355	0.0785
Fusion-Net	0.9399	0.0268	0.0342	0.9504	0.0152	0.0347	0.9202	0.0271	0.0541
SSConv	0.9331	0.0337	0.0342	0.9673	0.0101	0.0227	0.9162	0.0290	0.0563
DS-MDNP	**0.9455**	**0.0239**	**0.0312**	**0.9841**	**0.0044**	**0.0115**	**0.9206**	**0.0240**	**0.0501**

In Table 8, we can see on the one hand that the traditional method has more spectral and spatial distortion on the QuickBird and GaoFen-1, while the DL-based method has an advantage in spectral and spatial preservation. On the other hand, DS-MDNP has the best spectral and spatial retention values among these methods in terms of both spectral retention and spatial retention. The spectral distortion of the DL-based methods is worse on WorldView-2 than QuickBird and GaoFen-1, but in this case, DS-MDNP still ensures the best spectral retention, which also proves that DS-MDNP has a good generalization property.

5. Discussion

Extensive experiments have been designed to verify the effectiveness of MS-MDNP. We conducted ablation experiments using various strategies to determine the final network architecture. The ablation experiments demonstrated that the best performance is obtained when the input uses the differential strategy, the feature extraction network uses the improved DenseNet, and loss uses hybrid loss strategy, which draws on the hybrid attention-based residual network [41]. To validate the effectiveness of DS-MDNP, we chose four traditional methods including IHS [3], Wavelet [8], HPF [18], and PRACS [6] for comparison and three DL-based methods including PNN [22], Fusion-Net [25], and SSConv [28] for comparison. By comparing the results of the comparison experiments on the three datasets, in Figures 8–10 it is clear that DS-MDNP is more advantageous in terms of spectral preservation and spatial detail enhancement, and it is closer to the GT image in subjective visualization. In addition, by comparing the objective evaluation metrics [42] on the three datasets, in Tables 4–6 DS-MDNP has optimal results both in terms of spatial and spectral aspects. Finally, in real experiments, our method still has good robustness, the best quality evaluation, the clearest subjective visual effects, possesses good spectral fidelity, and DS-MDNP can have relatively fast testing speed while maintaining good accuracy; overall, DS-MDNP has the best performance.

6. Conclusions

Deep learning techniques are increasingly applied in more fields and have achieved impressive results. In this paper, we proposed a pansharpening method named DS-MDNP, which obtains MS images at different levels by up-sampling operations on original MS images. A PAN image is obtained at different levels by performing down-sampling operations on the original PAN image. Using the difference strategy, we can map the spatial information of PAN image to each band of the MS image in different levels. These different-level MS and PAN images can be used to generate differential images of different levels as the input of DS-MDNP. The different layers of DS-MDNP use the same feature extraction and reconstruction network, which is the improved DenseNet. The features at different levels obtained by DenseNet are correlated by feedback connections, while the features extracted from the bottom layer network are fed back to the upper layer network to make full use of the features at different levels. The training process of different layers is controlled by

a hybrid loss strategy to obtain more expected spatial information; finally, the spectral information of the MS images are injected into the extracted information to obtain a fused image with high spatial resolution and high spectral information.

In future work, a feasible solution to improve the performance of the network could be to combine DS-MDNP with traditional methods. Additionally, a spectral feature extraction network could be proposed for the preservation of spectral information rather than simply injecting MS images' spectral information. In addition, our scheme can be improved by applying multi-level features to the fusion of hyperspectral and MS images or the fusion of hyperspectral and PAN images.

Author Contributions: Methodology, software, and conceptualization, J.Y. (Junru Yin) and J.Q.; modification and writing—review and editing, Q.C. and J.Y. (Jun Yu); validation, M.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Henan Province Science and Technology Breakthrough Project, grant number 212102210102 and 212102210105.

Data Availability Statement: The data presented in this study are available in article.

Acknowledgments: The authors would like to thank the editors and reviewers for their advice.

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

References

1. Sun, L.; Zhao, G.; Zheng, Y.; Wu, Z. Spectral-Spatial Feature Tokenization Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
2. He, C.; Sun, L.; Huang, W.; Zhang, J.; Zheng, Y.; Jeon, B. TSLRLN: Tensor subspace low-rank learning with non-local prior for hyperspectral image mixed denoising. *Signal Process.* **2021**, *184*, 108060. [CrossRef]
3. Tu, T.M.; Su, S.C.; Shyu, H.C.; Huang, P.S. A new look at IHS-like image fusion methods. *Inf. Fusion* **2001**, *2*, 177–186. [CrossRef]
4. Kwarteng, P.; Chavez, A. Extracting spectral contrast in Landsat Thematic Mapper image data using selective principal component analysis. *Photogramm. Eng. Remote Sens.* **1989**, *55*, 339–348.
5. Laben, C.A.; Brower, B.V. Process for Enhancing the Spatial Resolution of Multispectral Imagery Using Pan-Sharpening. U.S. Patent 6,011,875, 4 January 2000.
6. 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]
7. Zhou, J.; Civco, D.L.; Silander, J.A. A wavelet transform method to merge Landsat TM and SPOT panchromatic data. *Int. J. Remote Sens.* **1998**, *19*, 743–757. [CrossRef]
8. Nunez, J.; Otazu, X.; Fors, O.; Prades, A.; Pala, V.; Arbiol, R. Multiresolution-based image fusion with additive waveletdecomposition. *IEEE Trans. Geosci. Remote Sens.* **1999**, *37*, 1204–1211. [CrossRef]
9. Burt, P.J.; Adelson, E.H. The Laplacian pyramid as a compact image code. *IEEE Trans. Commun.* **1983**, *31*, 532–540. [CrossRef]
10. Shah, V.P.; Younan, N.H.; King, R.L. An Efficient Pan-Sharpening Method via a Combined Adaptive PCA Approach and Contourlets. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1323–1335. [CrossRef]
11. Ghahremani, M.; Ghassemian, H. Remote-sensing image fusion based on Curvelets and ICA. *Int. J. Remote Sens.* **2015**, *36*, 4131–4143. [CrossRef]
12. Ballester, C.; Caselles, V.; Igual, L.; Verdera, J.; Rouge, B. A variational model for P + XS image fusion. *Int. J. Comput. Vis.* **2006**, *69*, 43–58. [CrossRef]
13. Vivone, G.; Simoes, M.; Dalla Mura, M.; Restaino, R.; Bioucas-Dias, J.M.; Licciardi, G.A.; Chanussot, J. Pansharpening based on semiblind deconvolution. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 1997–2010. [CrossRef]
14. Liu, Y.; Wang, Z. A practical pan-sharpening method with wavelet transform and sparse representation. In Proceedings of the IEEE International Conference on Imaging Systems and Techniques (IST), Santorini Island, Greece, 14–17 October 2014.
15. Zeng, D.; Hu, Y.; Huang, Y.; Xu, Z.; Ding, X. Pan-sharpening with structural consistency and $\ell 1/2$ gradient prior. *Remote Sens. Lett.* **2016**, *7*, 1170–1179. [CrossRef]
16. Fu, X.; Lin, Z.; Huang, Y.; Ding, X. A variational pan-sharpening with local gradient constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
17. Gillespie, A.; Kahle, A.B.; Walker, R.E. Color enhancement of highly correlated images-II. Channel ration and "Chromaticity" Transform techniques. *Remote Sens. Environ.* **1987**, *22*, 343–365. [CrossRef]
18. Vivone, G.; Alparone, L.; Chanussot, J. A Critical Comparison Among Pansharpening Algorithms. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2565–2586. [CrossRef]
19. Tian, X.; Chen, Y.; Yang, C. A variational pansharpening method based on gradient sparse representation. *IEEE Signal Process. Lett.* **2020**, *27*, 1180–1184. [CrossRef]

20. Fang, F.; Li, F.; Shen, C. A variational approach for pan-sharpening. *IEEE Trans. Image Process.* **2013**, *22*, 2822–2834. [CrossRef]
21. Dong, C.; Loy, C.C.; He, K. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [CrossRef]
22. Masi, G.; Cozzolino, D.; Verdoliva, L. Pansharpening by convolutional neural networks. *Remote Sens.* **2016**, *8*, 594. [CrossRef]
23. Wei, Y.; Yuan, Q.; Shen, H. Boosting the accuracy of multispectral image pansharpening by learning a deep residual network. *IEEE Geosci. Remote. Sens. Lett.* **2017**, *14*, 1795–1799. [CrossRef]
24. Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; Paisley, J. PanNet: A deep network architecture for pan-sharpening. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5449–5457.
25. Deng, L.J.; Vivone, G.; Jin, C. Detail Injection-Based Deep Convolutional Neural Networks for Pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6995–7010. [CrossRef]
26. Wang, W.; Zhou, Z.; Liu, H. MSDRN: Pansharpening of Multispectral Images via Multi-Scale Deep Residual Network. *Remote Sens.* **2021**, *13*, 1200. [CrossRef]
27. Wang, D.; Li, Y.; Ma, L. Going deeper with densely connected convolutional neural networks for multispectral pansharpening. *Remote Sens.* **2019**, *11*, 2608. [CrossRef]
28. Wang, Y.; Deng, L.J.; Zhang, T.J. SSconv: Explicit Spectral-to-Spatial Convolution for Pansharpening. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 4472–4480.
29. Wang, W.; Liu, H. An Efficient Detail Extraction Algorithm for Improving Haze-Corrected CS Pansharpening. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
30. Maneshi, M.; Ghassemian, H.; Imani, M. Sparse Representation of Injected Details for MRA-Based Pansharpening. In Proceedings of the 2020 IEEE India Geoscience and Remote Sensing Symposium, Gujarat, India, 2–4 December 2020; pp. 86–89.
31. Li, W.; Xiang, M.; Liang, X. MDCwFB: A Multilevel Dense Connection Network with Feedback Connections for Pansharpening. *Remote Sens.* **2021**, *11*, 2218. [CrossRef]
32. Xiao, S.; Jin, C.; Zhang, T.; Ran, R.; Deng, L. Progressive Band-Separated Convolutional Neural Network for Multispectral Pansharpening. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium, Brussels, Belgium, 11–16 July 2021; pp. 4464–4467.
33. Zhang, T.; Deng, L.; Huang, T.; Chanussot, J.; Vivone, G. A Triple-Double Convolutional Neural Network for Panchromatic Sharpening. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–14, *accepted*. [CrossRef]
34. Yuhas, R.H.; Goetz, A.F.; Boardman, J.W. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In Proceedings of the Summaries 3rd Annual JPL Airborne Geoscience Workshop, Pasadena, CA, USA, 1–5 June 1992; pp. 147–149.
35. Khademi, G.; Ghassemian, H. A multi-objective component-substitution-based pansharpening. In Proceedings of the 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA), Shahrekord, Iran, 19–20 April 2017; pp. 248–252.
36. Liu, X.; Liu, Q.; Wang, Y. Remote sensing image fusion based on two-stream fusion network. *Inf. Fusion* **2020**, *55*, 1–15. [CrossRef]
37. Wang, Z.; Bovik, A.C. A universal image quality index. *IEEE Signal Process. Lett.* **2002**, *9*, 81–84. [CrossRef]
38. Alparone, L.; Baronti, S.; Garzelli, A.; Nencini, F. A global quality measurement of pan-sharpened multispectral imagery. *IEEE Geosci. Remote Sens. Lett.* **2004**, *1*, 313–317. [CrossRef]
39. 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]
40. He, K.; Zhang, X.; Ren, S. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
41. Liu, Q.; Han, L.; Tan, R. Hybrid Attention Based Residual Network for Pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2021**, *13*, 1962. [CrossRef]
42. Zhu, R.; Zhou, F.; Xue, J.H. MvSSIM: A quality assessment index for hyperspectral images. *Neurocomputing* **2018**, *272*, 250–257. [CrossRef]