remote sensing

MDPI

Article

Unified Interpretable Deep Network for Joint Super-Resolution and Pansharpening

Dian Yu, Wei Zhang, Mingzhu Xu, Xin Tian and Hao Jiang *

Electronic Information School, Wuhan University, Wuhan 430072, China
* Correspondence: jh@whu.edu.cn

Abstract: Joint super-resolution and pansharpening (JSP) brings new insight into the spatial improvement of multispectral images. How to efficiently balance the spatial and spectral qualities in JSP is important for deep learning-based approaches. To address this problem, we propose a unified interpretable deep network for JSP, named UIJSP-Net. First, we formulate the JSP problem as an optimization problem in a specially designed physical model based on the relationship among the JSP result, the multispectral image, and the panchromatic image. In particular, two deep priors are utilized to describe latent distributions of different variables, which can improve the accuracy of the physical model. Furthermore, we adopt the alternating direction method of multipliers to solve the above optimization problem, where a series of iterative steps are generated. Finally, we design UIJSP-Net by unfolding these iterative steps into multiple corresponding stages in a unified network. Because UIJSP-Net has clear physical meanings, the spatial resolution of multispectral images can be efficiently improved while the spectral information can be kept as well. Extensive experimental results are carried out on both simulated and real datasets to demonstrate the superiority of UIJSP-Net over other state-of-the-art methods from qualitative and quantitative aspects.

Keywords: pansharpening; super-resolution; interpretable network; deep prior

Citation: Yu, D.; Zhang, W.; Xu, M.; Tian, X.; Jiang, H. Unified Interpretable Deep Network for Joint Super-Resolution and Pansharpening. *Remote Sens.* **2024**, *16*, 540. https://doi.org/10.3390/rs16030540

Academic Editor: Lionel Bombrun

Received: 18 October 2023
Revised: 21 January 2024
Accepted: 24 January 2024
Published: 31 January 2024

1. Introduction

Tremendous remote sensing images of high spatial and spectral resolution are highly demanded for many applications such as object detection, environmental protection, and land monitoring. Typically, most satellites with optical payloads capture multispectral (MS) and panchromatic (PAN) images simultaneously. Due to the difficulty in the hardware design, MS images always have multiple spectral bands but low spatial resolution. Different from MS images, PAN images have only one spectral band but high spatial resolution. Therefore, a fusion method in which a low-resolution MS image and its corresponding PAN image are employed to generate a high-resolution MS image, named pansharpening, has become a popular research topic in remote sensing [1].

Generally speaking, traditional pansharpening methods are divided into three tracks: component substitution (CS)-based, multiresolution analysis (MRA)-based, and model-based approaches [2]. The basic idea of CS based approaches is the substitution of the intensity of an MS image by a corresponding PAN image. Although CS-based approaches can generate clear fused MS images, they always suffer serious spectral distortion. To alleviate spectral distortion in CS-based approaches, MRA-based approaches attempt to solve this problem in a transformed domain of multiple resolutions. But MRA-based approaches always fail to preserve spatial resolution of MS images. By formulating models based on physical relationships between MS and PAN images, model-based approaches can provide good balances between spatial and spectral qualities [3]. Because the pansharpening problem is ill-posed, seeking appropriate priors is important in model-based approaches.

Benefiting from sophisticated non-linear feature extraction ability, deep learning (DL)-based pansharpening methods have been research hotspots in recent years [4]. In 2016,

a convolutional neural network (CNN) was applied to pansharpening (PNN), achieving superior results over traditional pansharpening methods [5]. To further improve its fusion performance, Fu et al. [6] utilized a residual CNN in the high-frequency domain. Similarly, a multiscale and multidepth CNN was adopted by Yuan et al. [7] to fuse images with multiscale features. To fully exploit spatial details at multiple spatial scales, Jin et al. [8] designed a Laplacian pyramid pansharpening network architecture based on a multi-scale loss function. Ke et al. [9] proposed a high-frequency transformer network for pansharpening based on window cross-attention, which can capture long-distance dependencies in the vision transformer. Besides the above CNN-based methods, pansharpening based on a generative adversarial network (GAN) [10] was implemented by establishing an adversarial game between the generator and the discriminator.

Although remarkable results have been achieved in DL-based pansharpening methods, the spatial resolution of the fused MS image is still limited to that of the PAN image. Thus, to obtain fusion images with better spatial resolution, combining super-resolution (SR) and pansharpening is necessary. In 2022, Chouteau et al. [11] presented a DL-based attempt to join SR and pansharpening (JSP) on Airbus's Pleiades Neo images, bringing new insight into the further spatial resolution improvement of the fused MS image, as shown in Figure 1c. However, SR and pansharpening are treated as two different steps in this attempt. Thus, the spatial improvement carried out by SR and the spectral preservation depending on pansharpening are separately controlled. As a result, such a method cannot ensure that the spatial resolution is as good as the spectral quality for the JSP results. For example, a small spatial or spectral distortion in the SR result may lead to a large error in the following pansharpening step. Therefore, it is valuable to construct a unified JSP deep network that can be trained in an end-to-end framework. However, this problem has been seldom investigated.

(**a**) (**b**) (**c**) (**d**)

Figure 1. An example of JSP. (**a**) Interpolated MS image. (**b**) Pansharpening result. (**c**) JSP result. (**d**) Groundtruth.

In order to resolve such problems, we propose a unified interpretable deep network for JSP in this study, named UIJSP-Net. To unify the SR and pansharpening in a whole framework, we start the design from the physical observations among different variables. To this end, we first construct a novel physical model based on the relationships among the JSP result, the MS image, and the PAN image. To improve the accuracy of this model, we especially utilize two deep priors to describe the distributions of the latent high-resolution MS image (JSP result). To solve the optimization problem in this model, we adopt the alternating direction method of multipliers (ADMM) to transform the solution of the optimization problem into a series of iterative steps. To construct the desired UIJSP-Net, we unfold these iterative steps into multiple corresponding stages of a unified network with the two deep priors implemented by the USRNet and DRUNet, respectively. In summary, compared with traditional deep learning-based methods, which are always trained as a black box, we first propose our JSP model with clear physical meanings. Then, we design UIJSP-Net by unfolding iterative steps of the solution to this model into multiple corresponding modules. Because each network module in UIJSP-Net relates to an iterative step with clear physical meaning, UIJSP-Net is interpretable.

The main contributions of our work are listed as follows:

- To the best of our knowledge, we are the first to build a new model to formulate the SR and pansharpening objective in a unified optimization problem, which is convenient for effectively preserving spatial as well as spectral resolution. In addition, two deep priors about the latent distributions of the latent high-resolution multispectral images are adopted to improve the accuracy of the model.
- To solve this model efficiently, we construct UIJSP-Net by utilizing the unfolding technology based on some iterative steps derived from the ADMM.
- Then we validate this method in both simulated and real datasets, proving its advantage over other state-of-the-art methods.

The rest of the study is organized as follows. The related works are summarized in Section 2. And the UIJSP-Net is introduced in Section 3. Then, experimental results are presented in Section 4, and the conclusion is made in Section 5.

2. Related Work

Considering that there are few studies focusing on JSP, we mainly introduce related work of SR and pansharpening, respectively.

2.1. SR

In this study, SR mainly refers to generating a high-resolution image out of a single low-resolution one. Typical SR methods include interpolation-based, example-based, and DL-based methods. Interpolation-based methods estimate the unknown pixels of a high-resolution image by utilizing the linear or nonlinear interpolation on their known neighbors, such as the most commonly used bi-linear and bi-cubic interpolation methods. For example, Zhu et al. [12] reconstructed a high-resolution image by interpolating pixels based on the nonlocal geometric similarities. However, these methods tend to produce visual artifacts such as aliasing and blurring. The example-based methods are reconstructed by the hypothesis that patches from low-resolution and high-resolution images have a latent relationship. Specifically, patches of low-resolution images are assumed to have a sparse representation with respect to a dictionary generated by a large exemplary dataset of high-resolution images [13,14]. Therefore, the performance of example-based methods highly correlates with the choice of the exemplary dataset. To address this problem, a self-example-based method was proposed in [15] which forms the dataset by resizing the original image into different scales, resulting in high computational complexity. With the development of DL, DL-based methods have attracted a large amount of research interest. Dong et al. [16] first establish an end-to-end mapping role between low- and high-resolution images by CNN, exhibiting superior SR performance. To explore the feature correlations of intermediate layers that can improve the ability of CNN, Dai et al. [17] proposed a second-order attention network for feature expression and correlated learning process. By integrating the advantages of exampled-based and DL-based methods, Zhang et al. [18] designed an end-to-end trainable deep network based on the unfolding optimization. Furthermore, Zhang et al. [19] adaptively modulated the convolution kernel for image SR based on the global context. Recently, Gao et al. [20] proposed a Bayesian image SR framework by modeling natural image statistics with a combination of smoothness and sparsity priors.

2.2. Pansharpening

Pansharpening is a method that fuses a low-resolution MS image and a high-resolution PAN image to generate a high-resolution MS image. Representative pansharpening methods include CS-based, MRA-based, model-based, and DL-based methods. Principal component analysis (PCA) [21], Gram–Schmidt adaptive (GSA) [22], and intensity–hue–saturation (IHS) [23] are some typical CS-based methods. They are simple in terms of implementation, but sometimes suffer serious spectral distortions. The MRA-based methods inject the spatial details of PAN images into MS images according to multiresolution decomposition, such as smoothing filter-based intensity modulation (SFIM) [24], modulation transfer

function (MTF) [25], and improved generalized Laplacian pyramid [26], thus reducing the spectral distortions of CS-based methods to a certain degree. However, the spatial qualities of MRA-based methods are always unsatisfactory. Some hybrid methods [27,28] in which CS and MRA methods are combined are also studied.

To generate images with balanced spatial and spectral quality, the model-based methods construct fusion models by making use of the relations among the fused MS, low-resolution MS, and PAN images. Considering that this fusion problem is ill-posed, seeking suitable priors is important in the model-based methods. Considering the upsampled MS image to be a blurred version of the fused MS image, Ballester et al. [29] suggested a variational model-based method, and Palsson et al. [30] further utilized the total variation regularization as a prior in the construction of the fusion model. Fu et al. [31] further constructed a local gradient prior in pansharpening for different local patches and bands of MS and PAN images. Tian et al. [32] explored priors for pansharpening from the similarities of cartoon and texture components of PAN and MS images. The difficulty in seeking accurate priors and selecting optimal model parameters is still challenging for the model-based methods.

DL-based method is a new track for pansharpening. It demonstrates powerful fusion capability because of its ability to describe data in a nonlinear way [33]. Besides PNN, DRPNN was proposed by Wei et al. [34], where the residual network is adopted to overcome the problem of the shallow network. To extract both deep and shallow features for pansharpening, Yuan et al. [7] utilized multiscale feature extraction with changeable receptive fields. Inspired by the injection process of the MRA-based methods, Zhang et al. [35] designed a bidirectional pyramid network to fuse PAN and MS images level by level. FusionNet proposed by Deng et al. [36] was designed based on a detailed injection network. Considering the necessity of constructing DL methods that have clear physical meanings in pansharpening, Tian et al. [37] designed an interpretable deep network. Similarly, Wen et al. [38] built an implicit prior for the fusion model based on the deep operators to achieve good nonlinear description ability. By constructing an adversarial game, Ma et al. [10] designed PanGan by substituting the groundtruth with the original image for unsupervised pansharpening. To effectively capture the global relationship between the MS and PAN images, Transformer was applied to pansharpening [39].

3. Proposed Method

We introduce the formulation of the proposed JSP model in this section. Then, we demonstrate an efficient solution to the optimization problem of the proposed JSP model. Lastly, we exhibit how the network is designed.

3.1. Formulation of the Proposed JSP Model

To simplify the description in this study, we adopt the following notations and definitions. Scalars and matrices are denoted as lowercase letters and bold capital letters, respectively. For example, $\mathbf{F} \in \mathbb{R}^{M \times N \times b}$ represents the high-resolution MS image of b bands, generated by JSP. $[M, N]$ is the spatial resolution of $\mathbf{F}$. $\mathbf{M} \in \mathbb{R}^{\frac{M}{sr} \times \frac{N}{sr} \times b}$ and $\mathbf{P} \in \mathbb{R}^{\frac{M}{s} \times \frac{N}{s}}$ are the low-resolution MS image and PAN image, respectively. r is the ratio of spatial resolution between the PAN image $\mathbf{P}$ and the MS image $\mathbf{M}$. s is the scale of SR for the PAN or MS image. We list the major notations of this study in Table 1.

Table 1. Major notations of this study.

Notations	Definitions
$\mathbf{F}$, $\mathbf{M}$, $\mathbf{P}$	JSP result, low-resolution MS image, and low-resolution PAN image.
$\mathbf{\Psi}_1$, $\mathbf{\Psi}_2$	Spatial blurring and down-sampling operations for SR and pansharpening, respectively.
$\Gamma(\cdot)$, $\Theta(\cdot)$	Deep operators for different modules.
$\mathbf{Z}$, $\mathbf{W}$	Auxiliary variables.
$[M, N]$	Spatial resolution of JSP result.
r	Ratio of spatial resolution between the PAN image $\mathbf{P}$ and the MS image $\mathbf{M}$.
s	Scale of SR for the PAN or MS image.

Considering that the MS image $\mathbf{M}$ can be regarded as a spatial degradation of the high-resolution MS image $\mathbf{F}$, we can utilize the following data fidelity term $D(\mathbf{F})$ to formulate their relationship as

$$D(\mathbf{F}) = \frac{1}{2}\|\mathbf{\Psi}_2\mathbf{\Psi}_1\mathbf{F} - \mathbf{M}\|_F^2, \tag{1}$$

where $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$ denote the spatial blurring and down-sampling operations for SR and pansharpening, respectively. $\|\cdot\|_F^2$ represents the Frobenius norm. Because $M \gg m$ (or $N \gg n$), Equation (1) is always ill-posed. It is necessary to seek additional priors of $\mathbf{F}$ for an accurate reconstruction.

Naturally, an image always has a sparse (smooth between pixels) property. Considering the good nonlinear feature extraction ability in DL, it is helpful to construct a latent prior $J(\mathbf{F})$ by utilizing the deep operator $\Theta(\cdot)$ to describe the inherent relationship of pixels such as the smooth characteristic in an image as

$$J(\mathbf{F}) = \Theta(\mathbf{F}). \tag{2}$$

Furthermore, the PAN image $\mathbf{P}$ can be treated as the average of b bands in the spatial degraded result $\mathbf{\Psi}_1\mathbf{F}$. Consequently, the errors between $\mathbf{P}$ and each band of $\mathbf{\Psi}_1\mathbf{F}$ should also be sparse. Based on this hypotheses, we construct a deep module $\Gamma(\cdot)$ to formulate the relationship of the sparse property of the errors between $\mathbf{P}$ and $\mathbf{\Psi}_1\mathbf{F}$ as

$$K(\mathbf{F}) = \Gamma(\mathbf{\Psi}_1\mathbf{F} - \mathbf{P}), \tag{3}$$

where $\bar{\mathbf{P}}$ denotes the duplication of $\mathbf{P}$ into b bands.

Based on the above analysis, we can formulate the desired JSP model by integrating Equations (1)–(3) as

$$\arg\min_{\mathbf{F}} \frac{1}{2}\|\mathbf{\Psi}_2\mathbf{\Psi}_1\mathbf{F} - \mathbf{M}\|_F^2 + \lambda\Gamma(\mathbf{\Psi}_1\mathbf{F} - \bar{\mathbf{P}}) + \alpha\Theta(\mathbf{F}). \tag{4}$$

λ and α are parameters to balance different terms.

3.2. Solution to the Proposed JSP Model

To solve Equation (4), we adopt the ADMM. Let $\mathbf{Z} = \mathbf{\Psi}_1\mathbf{F}$; then, the augmented Lagrange function of Equation (4) can be described as

$$\frac{1}{2}\|\mathbf{\Psi}_2\mathbf{Z} - \mathbf{M}\|_F^2 + \lambda\Gamma(\mathbf{Z} - \bar{\mathbf{P}}) + \alpha\Theta(\mathbf{F}) + \frac{\mu}{2}\|\mathbf{Z} - \mathbf{\Psi}_1\mathbf{F} - \frac{\mathbf{W}}{\mu}\|_F^2, \tag{5}$$

where $\mathbf{W}$ is an introduced auxiliary variable. μ is a parameter. Afterward, variables $\mathbf{Z}$, $\mathbf{F}$, and $\mathbf{W}$ can be solved alternatively based on the following sub-problems:

(1) **Z** sub-problem: We solve $\mathbf{Z}^{t+1}$ from

$$\arg\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{\Psi}_2\mathbf{Z}-\mathbf{M}\|_F^2+\lambda\Gamma(\mathbf{Z}-\bar{\mathbf{P}})+\frac{\mu}{2}\|\mathbf{Z}-\mathbf{\Psi}_1\mathbf{F}^t-\frac{\mathbf{W}^t}{\mu}\|_F^2. \tag{6}$$

We utilize the fast iterative shrinkage-thresholding algorithm [40] to solve Equation (6). Then, the solution to $\mathbf{Z}^{t+1}$ can be separated into the following two steps:

$$\mathbf{A}^t=\mathbf{Z}^t-\frac{\mathbf{\Psi}_2^{-1}(\mathbf{\Psi}_2\mathbf{Z}^t-\mathbf{M})+\mu(\mathbf{Z}^t-\mathbf{\Psi}_1\mathbf{F}^t-\frac{\mathbf{W}^t}{\mu})}{L}, \tag{7}$$

$$\mathbf{Z}^{t+1}=\arg\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{Z}-\mathbf{A}^t\|_F^2+\lambda\Gamma(\mathbf{Z}-\bar{\mathbf{P}}). \tag{8}$$

L is the Lipschitz constant and we set $L=1$ in the following. Let $\mathbf{G}=\mathbf{Z}-\bar{\mathbf{P}}$; then, Equation (8) can be transformed into the following form as

$$\mathbf{G}^{t+1}=\arg\min_{\mathbf{G}} \frac{1}{2}\|\mathbf{G}-(\mathbf{A}^t-\bar{\mathbf{P}})\|_F^2+\lambda\Gamma(\mathbf{G}). \tag{9}$$

The form of Equation (9) is similar to the image denoising problem with deep denoiser prior. Then, we can utilize the DRUNet [41] to solve it as

$$\mathbf{G}^{t+1}=DRUNet(\mathbf{A}^t-\bar{\mathbf{P}},\sqrt{\lambda}). \tag{10}$$

At this time, we have

$$\mathbf{Z}^{t+1}=\mathbf{G}^{t+1}+\bar{\mathbf{P}}. \tag{11}$$

(2) **F** sub-problem: We solve $\mathbf{F}^{t+1}$ from

$$\arg\min_{\mathbf{F}} \frac{\mu}{2}\|\mathbf{Z}^{t+1}-\mathbf{\Psi}_1\mathbf{F}-\frac{\mathbf{W}^t}{\mu}\|_F^2+\alpha\Theta(\mathbf{F}). \tag{12}$$

The form of Equation (12) is similar to the image SR problem with deep prior. Therefore, we adopt the USRNet [18] to solve it as

$$\mathbf{F}^{t+1}=USRNet(\mathbf{Z}^{t+1}-\frac{\mathbf{W}^t}{\mu},\sqrt{\frac{\alpha}{\mu}}). \tag{13}$$

(3) **W** sub-problem: We derive $\mathbf{W}^{t+1}$ directly from

$$\mathbf{W}^{t+1}=\mathbf{W}^t-\mu(\mathbf{Z}^{t+1}-\mathbf{\Psi}_1\mathbf{F}^{t+1}). \tag{14}$$

Therefore, the proposed JSP algorithm is summarized in Algorithm 1.

Algorithm 1: Proposed JSP algorithm.

Input $\mathbf{\Psi}_1$, $\mathbf{\Psi}_2$, $\mathbf{M}$, $\mathbf{P}$, λ, α, μ, T_{max} (maximum number of iterations), $\Gamma(\cdot)$, $\Theta(\cdot)$, $\mathbf{F}^1$, $\mathbf{W}^1$, $\mathbf{Z}^1$
For $t=1$ to T_{max} **do**
  Update $\mathbf{Z}^{t+1}$ in the closed-form from Equations (7), (10), and (11).
  Update $\mathbf{F}^{t+1}$ in the closed-form from Equation (13).
  Update $\mathbf{W}^{t+1}$ in the closed-form from Equation (14).
End
Output **F**.

3.3. Network Design

The unfolding technology [42] is utilized to construct the desired UIJSP-NET, as its efficiency is widely proven. In particular, we unfold each iteration of Algorithm 1 as one stage of UIJSP-Net. The diagram of the t-th stage of UIJSP-Net is shown in Figure 2. The inputs are $\mathbf{F}^t$, $\mathbf{Z}^t$, and $\mathbf{W}^t$, and the outputs are $\mathbf{F}^{t+1}$, $\mathbf{Z}^{t+1}$, and $\mathbf{W}^{t+1}$, respectively. Then, these outputs are further fed into the next stage. Other parameters are learned in the training process.

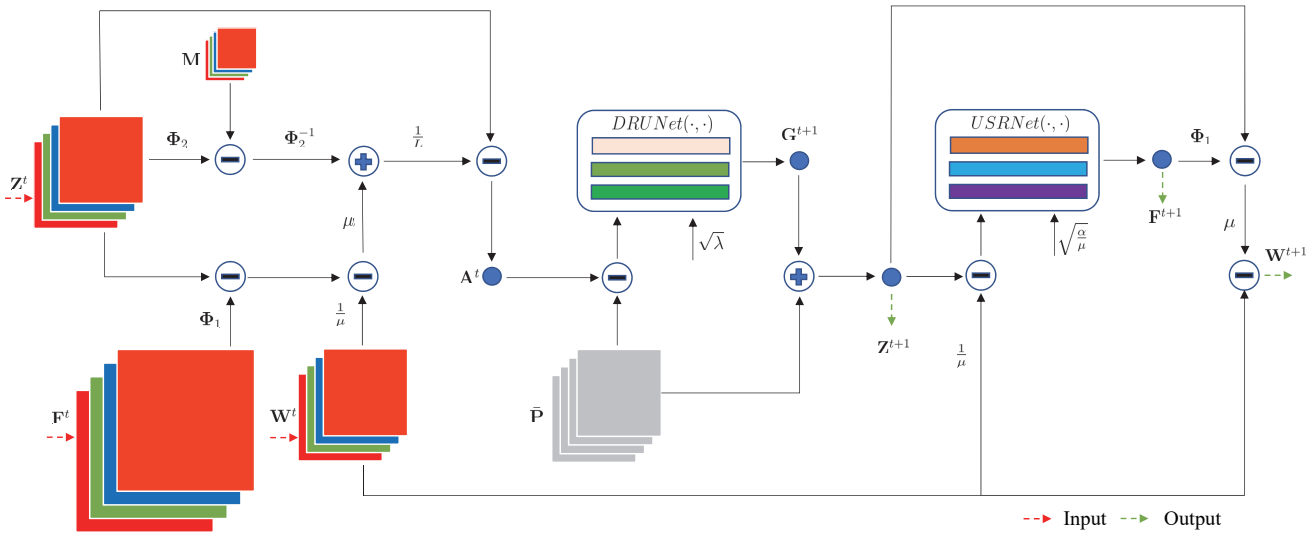

Figure 2. The t-th stage of the proposed network architecture.

The diagram of DRUNet is shown in Figure 3. It combines residual blocks into a U-Net architecture, which includes 4 scales. In each scale, 2×2 strided convolution (SConv) scalings and 2×2 transposed convolution (TConv) scalings linked by identity skip connections are utilized to implement features' downsampling and upsampling, respectively, where four successive residual blocks (ResBlocks) are included. The channel amount is 64, 128, 256, and 512 for the 4 consecutive scales, respectively. No activation functions are added to the first and the last convolution (Conv), SConv, and TConv layers. Only one ReLU activation function is provided to each residual block. The diagram of USRNet is shown in Figure 4. It includes three main modules, namely the data module $\mathcal{D}(\cdot,\cdot,\cdot,\cdot,\cdot)$, the prior module $\mathcal{P}(\cdot,\cdot)$, and the hyper-parameter module $\mathcal{H}(\cdot,\cdot)$. The data module is mainly implemented by FFT and inverse FFT operators. The backbone of the prior module is also a U-Net architecture, where identity skip connection, SConv, and TConv layers are included. The hyper-parameter module is implemented by three fully connected layers. The first and second activation functions are the ReLU layers and the last is the Softplus layer. More details about DRUNet and USRNet can be found in [18,41], respectively.

To drive the proposed UIJSP-Net, we utilize an L_2 loss function $\mathcal{L}_{JSP}$ as

$$\mathcal{L}_{JSP} = \sum_{i=1}^{b} \|\mathbf{F}_i - \bar{\mathbf{F}}_i\|_F^2. \tag{15}$$

We divide images into small blocks for training. $\mathbf{F}_i$ and $\bar{\mathbf{F}}_i$ represent the i-th training blocks of JSP results and reference high-resolution MS images, respectively. b is the number of training blocks. To better achieve balance between the computational efficiency and JSP results, we select 8 stages in our network architecture.

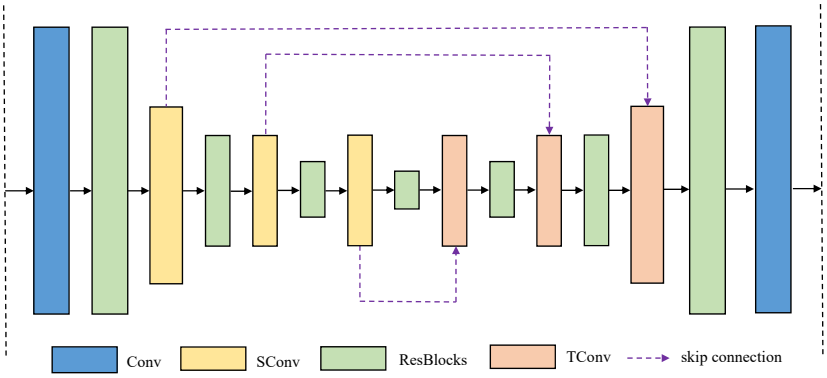

Figure 3. The diagram of DRUNet [41].

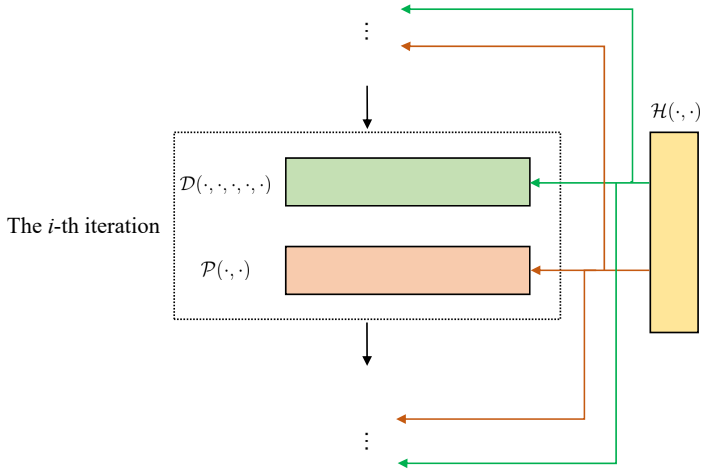

Figure 4. The diagram of USRNet [18].

4. Experimental Results

4.1. Experimental Setting

To demonstrate the efficiency of the proposed UIJSP-Net, we select three satellite datasets (GaoFen-2, QuickBird, and WorldView-3) in our experiments. GaoFen-2 is a high-resolution optical Earth observation satellite developed by China. The spatial resolution of GaoFen-2 PAN and MS images is 0.8 m and 3.2 m, respectively. QuickBird is a high-resolution commercial satellite owned by DigitalGlobe, which captures PAN and MS images with spatial resolutions of 0.61 m and 2.44 m, respectively. Similar to QuickBird, WorldView-3 is also a commercial satellite owned by DigitalGlobe. It generates PAN images with a spatial resolution of 0.31 m and MS images with eight bands and a spatial resolution of 1.24 m. We utilize blue, green, red, and near-IR bands of WorldView-3 MS images in our experiments.

Because there are no existing JSP methods, we compare traditional methods in the following ways. We first SR the input low-resolution MS and PAN images, and further fuse the SR results to generate the final results. For the convenience of a fair comparison, USRNet is adopted as the SR method in traditional methods. The pansharpening methods include two traditional pansharpening methods, SFIM and HPF, and five DL methods, PNN [5], MSDCNN [7], Hyper_DSNet [43], FusionNet [36], and DRPNN [34]. In addition, the benchmark consists of images being interpolated by polynomial kernel with 23 coefficients (EXP). For the convenience of description in the following, we denote SR (USRNet) +DL as pansharpening methods and UIJSP-Net as DL-based methods. We implement our model on PyTorch, and train all DL methods on the same dataset and on an NVIDIA GeForce RTX 3090 Ti GPU (24GB) for a fair comparison. SFIM and HPF are

implemented in Matlab. We train our model with a batch size of 40. The initialized learning rate is set to 0.0001. Finally, the model achieves stable performance in approximately 6 h.

The experiments are carried out from the following five aspects. (i) Simulation experiments. Both visual comparison and quantitative analysis are included. The diagram of the simulation process is shown in Figure 5. Because traditional methods can be regarded as a combination of SR and pansharpening, i.e., training datasets for SR and pansharpening are both required, we separate the simulation process into the following two steps. We first generate intermediate MS and PAN (IMP) blocks from input high-resolution MS and PAN (HMP) blocks by spatial blurring and downsampling. Then, IMP blocks and HMP blocks can be used to train the SR network. For simplicity, we set the resolution ratio between IMP blocks and HMP blocks as 2. Furthermore, we degrade IMP blocks by spatial blurring and downsampling based on Wald's protocol, to obtain low-resolution MS and PAN (LMP) blocks. Then, IMP blocks and LMP blocks are used for training the pansharpening network in the compared methods. For UIJSP-Net, all network modules are directly trained from HMP and LMP blocks. The quantitative metrics include the peak-signal-to-noise (PSNR), the structural similarity (SSIM), the correlation coefficient (CC), the universal image quality index (UIQI, *Q*), the erreur relative global adimensionnelle de synthese (ERGAS), and the spectral angle mapping (SAM). (ii) Real experiments. Considering there are no existing quantitative metrics for JSP in real experiments, we mainly utilize real satellite data for visual comparison. (iii) Computational efficiency. The computational speeds of different methods are compared. (iv) Ablation study. We utilize ablation studies to analyze the influence of different parts of UIJSP-Net on its performance. (v) NDVI experiments. We compare different methods on their applications in NDVI calculations, where the simulation dataset is utilized. The quantitative metrics include root mean squared error (RMSE), SSIM, and CC.

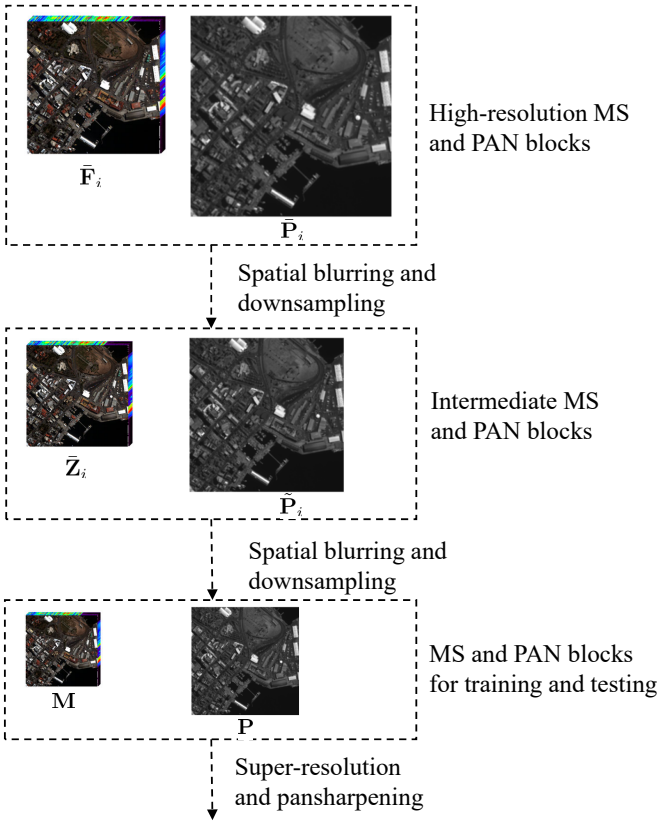

Figure 5. The diagram of the simulation process.

4.2. Simulation Experiments

For simulation (as shown in Figure 5), the spatial resolutions of MS and PAN images in LMP blocks are $64 \times 64 \times 4$ and 256×256, respectively. In addition, the spatial resolution of MS and PAN images in IMP blocks are $128 \times 128 \times 4$ and 512×512, respectively. Therefore, the spatial resolution of groundtruth is $512 \times 512 \times 4$.

4.2.1. GaoFen-2 Dataset

On the GaoFen-2 dataset, we utilize 3213 pairs of LMP blocks, IMP blocks, and groundtruth patches for training and another 235 pairs for testing. We first adopt a pair of images from the GaoFen-2 dataset to compare the performance of different methods from the aspect of visual quality (only three bands for display), as shown in Figure 6. Compared with other images in Figure 6b–i, the interpolation method EXP in Figure 6a is blurrier, indicating the necessity of JSP. Generally speaking, benefiting from the good ability of non-linear feature extraction, the clarity of DL-based methods is higher than SR +traditional pansharpening methods, such as SR+SFIM and SR+HPF. Although SR+PNN and SR+MSDCNN have clear fusion results, obvious spatial artifacts exist, degrading their fusion performance. Enlargements in Figure 6d,e show an example of such a phenomenon, where the stripes are not straight. By constructing a JSP network, UIJSP-Net in Figure 6i generates the best results, which is the closest to the groundtruth in Figure 6j. The superiority of UIJSP-Net can also be found in Figure 7, where the residual images (generated by the mean absolute errors between the fusion results and the groundtruth) are demonstrated. Due to the advantage of DL, SR+PNN generates smaller errors than SR+SFIM and SR+HPF. It is clear that UIJSP-Net produces the smallest errors among all methods. The quantitative comparison is listed in Table 2. The average values in addition to the standard deviations of all tested images are calculated. Generally speaking, UIJSP-Net and SR+DRPNN exhibit the first and second best JSP performance. For example, UIJSP-Net generates the five best metrics among all methods.

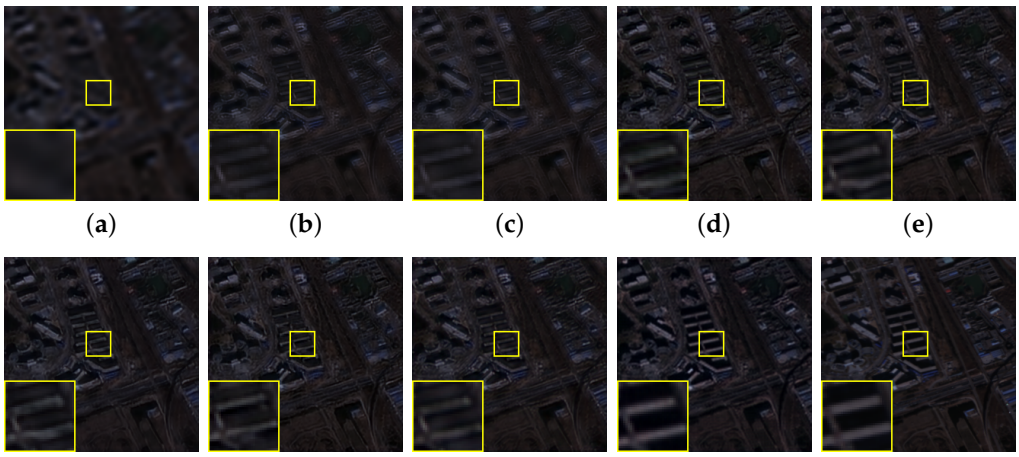

Figure 6. Visual comparison on the simulated GaoFen-2 dataset. (**a**) EXP. (**b**) SR+SFIM. (**c**) SR+HPF. (**d**) SR+PNN. (**e**) SR+MSDCNN. (**f**) SR+Hyper_DSNet. (**g**) SR+FusionNet. (**h**) SR+DRPNN. (**i**) UIJSP-Net. (**j**) Groundtruth. (Large boxes are enlargements of the small boxes).

Table 2. Quantitative comparison on the 235 test images of the simulated GaoFen-2 dataset. **BLUE**: the best, **RED**: the second-best.

Method	PSNR	SSIM	CC	*Q*	ERGAS	SAM
EXP	25.343 ± 1.342	0.688 ± 0.050	0.797 ± 0.038	0.773 ± 0.047	7.263 ± 0.953	$\mathbf{4.049 \pm 0.482}$
SR+SFIM	27.791 ± 1.425	0.847 ± 0.023	$\mathbf{0.893 \pm 0.019}$	0.884 ± 0.023	5.456 ± 0.625	4.253 ± 0.434
SR+HPF	27.769 ± 1.442	0.844 ± 0.023	0.892 ± 0.018	0.883 ± 0.022	5.469 ± 0.642	$\mathbf{4.252 \pm 0.468}$

Table 2. *Cont.*

Method	PSNR	SSIM	CC	Q	ERGAS	SAM
SR+PNN	31.004 ± 1.287	0.811 ± 0.024	0.886 ± 0.019	0.878 ± 0.020	5.847 ± 0.563	5.093 ± 0.788
SR+MSDCNN	31.238 ± 1.260	0.822 ± 0.023	0.890 ± 0.017	0.880 ± 0.021	5.683 ± 0.585	4.637 ± 0.592
SR+Hyper_DSNet	31.312 ± 1.352	0.830 ± 0.022	0.891 ± 0.018	0.886 ± 0.019	5.609 ± 0.594	4.991 ± 0.634
SR+FusionNet	29.340 ± 1.266	0.733 ± 0.037	0.857 ± 0.026	0.815 ± 0.028	7.117 ± 0.870	5.132 ± 0.774
SR+DRPNN	**31.738 ± 1.395**	**0.851 ± 0.020**	0.890 ± 0.017	**0.894 ± 0.018**	**5.375 ± 0.562**	5.238 ± 0.720
UIJSP-Net	**32.954 ± 0.904**	**0.880 ± 0.019**	**0.901 ± 0.018**	**0.929 ± 0.022**	**4.829 ± 0.667**	6.949 ± 1.387
Ideal value	$+\infty$	1	1	1	0	0

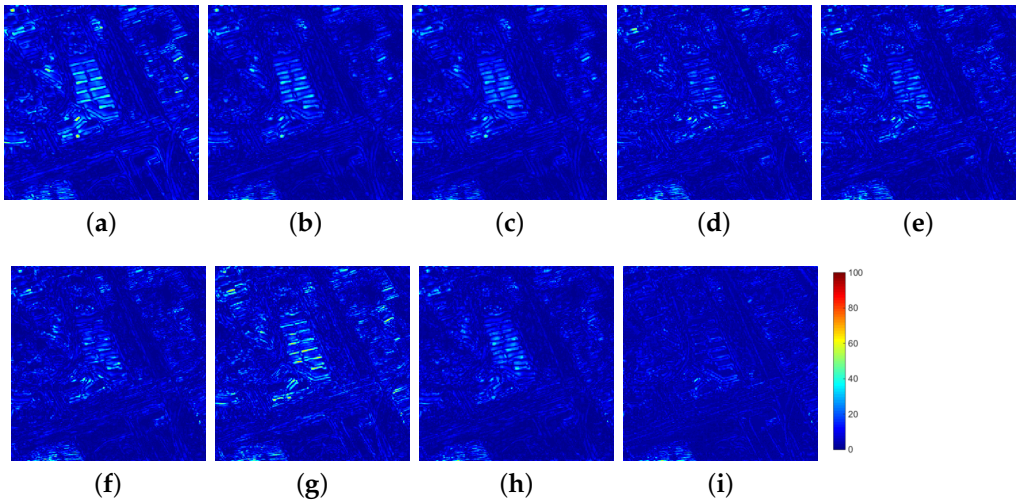

Figure 7. Residual images of the simulated GaoFen-2 dataset. (**a**) EXP. (**b**) SR+SFIM. (**c**) SR+HPF. (**d**) SR+PNN. (**e**) SR+MSDCNN. (**f**) SR+Hyper_DSNet. (**g**) SR+FusionNet. (**h**) SR+DRPNN. (**i**) UIJSP-Net.

4.2.2. QuickBird Dataset

On the QuickBird dataset, we adopt 955 pairs of LMP blocks, IMP blocks, and groundtruth patches for training and another 224 pairs for testing. One typical pair of images from the QuickBird dataset is used for visual comparison and the result is shown in Figure 8. Although SR+SFIM and SR+HPF in Figure 8b,c can provide clearer JSP results, there are some distorted edges in the buildings. For example, the shape of the building in Figure 8c (labeled by a red ellipse) has a large deviation from that of the groundtruth in Figure 8j. Among all DL methods, UIJSP-Net generates the best JSP result. This conclusion can be verified by the comparison of enlargements in Figure 8d–i, where the building with an arrow shape in Figure 8i is closest to the groundtruth. To further verify the superiority of UIJSP-Net, we compare the residual images in Figure 9. On this dataset, SR+SFIM and SR+HPF have low JSP performance, resulting in large differences. The error of UIJSP-Net is the smallest of all methods. We list the quantitative comparison in Table 3. The DL methods generate better quantitative results than SR+SFIM and SR+HPF, especially in PSNR. UIJSP-Net has the best performance evaluated by PSNR, SSIM, CC, Q, and ERGAS, demonstrating its effectiveness in the quantitative comparison of the simulated QuickBird dataset.

Table 3. Quantitative comparison on the 224 test images of the simulated QuickBird dataset. **BLUE**: the best, **RED**: the second-best.

Method	PSNR	SSIM	CC	Q	ERGAS	SAM
EXP	27.461 ± 3.348	0.749 ± 0.137	0.828 ± 0.102	0.887 ± 0.048	8.309 ± 1.997	**5.261 ± 0.978**
SR+SFIM	27.379 ± 6.229	0.824 ± 0.094	0.842 ± 0.109	0.880 ± 0.154	17.119 ± 7.509	5.598 ± 0.709
SR+HPF	28.846 ± 3.313	0.823 ± 0.095	0.867 ± 0.097	0.922 ± 0.032	7.122 ± 1.681	5.482 ± 0.876

Table 3. *Cont.*

Method	PSNR	SSIM	CC	Q	ERGAS	SAM
SR+PNN	35.362 ± 4.855	0.914 ± 0.042	0.940 ± 0.062	0.924 ± 0.100	4.857 ± 1.176	5.463 ± 1.018
SR+MSDCNN	35.080 ± 4.872	0.910 ± 0.044	0.943 ± 0.056	0.921 ± 0.098	5.103 ± 1.224	5.460 ± 1.011
SR+Hyper_DSNet	35.478 ± 4.975	0.916 ± 0.042	0.946 ± 0.054	0.926 ± 0.096	4.794 ± 1.170	5.199 ± 0.958
SR+FusionNet	34.275 ± 5.158	0.898 ± 0.050	0.934 ± 0.057	0.911 ± 0.097	5.594 ± 1.330	5.373 ± 1.003
SR+DRPNN	34.372 ± 5.091	0.898 ± 0.051	0.932 ± 0.061	0.912 ± 0.101	5.434 ± 1.264	5.836 ± 1.081
UIJSP-Net	36.332 ± 4.011	0.926 ± 0.031	0.952 ± 0.068	0.929 ± 0.119	4.424 ± 1.181	5.683 ± 1.219
Ideal value	$+\infty$	1	1	1	0	0

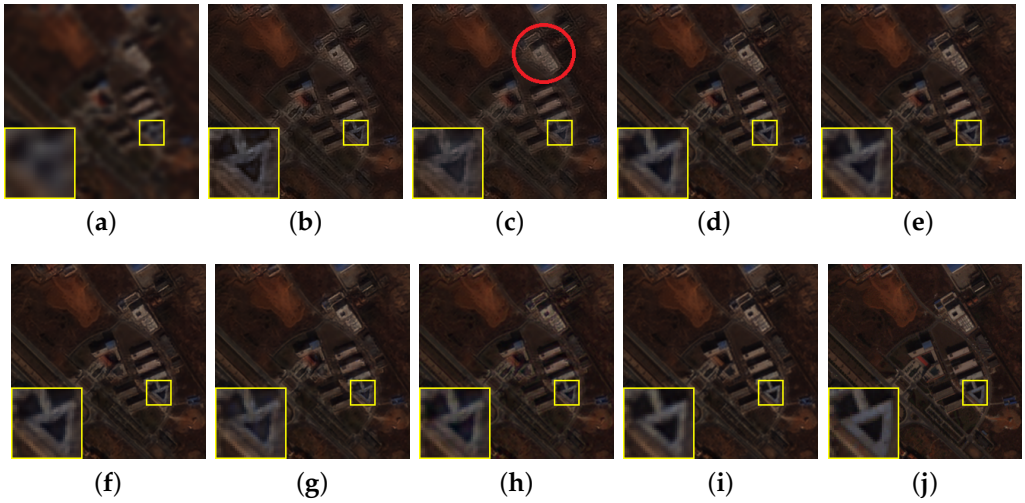

Figure 8. Visual comparison on the simulated QuickBird dataset. (**a**) EXP. (**b**) SR+SFIM. (**c**) SR+HPF. (**d**) SR+PNN. (**e**) SR+MSDCNN. (**f**) SR+Hyper_DSNet. (**g**) SR+FusionNet. (**h**) SR+DRPNN. (**i**) UIJSP-Net. (**j**) Groundtruth.

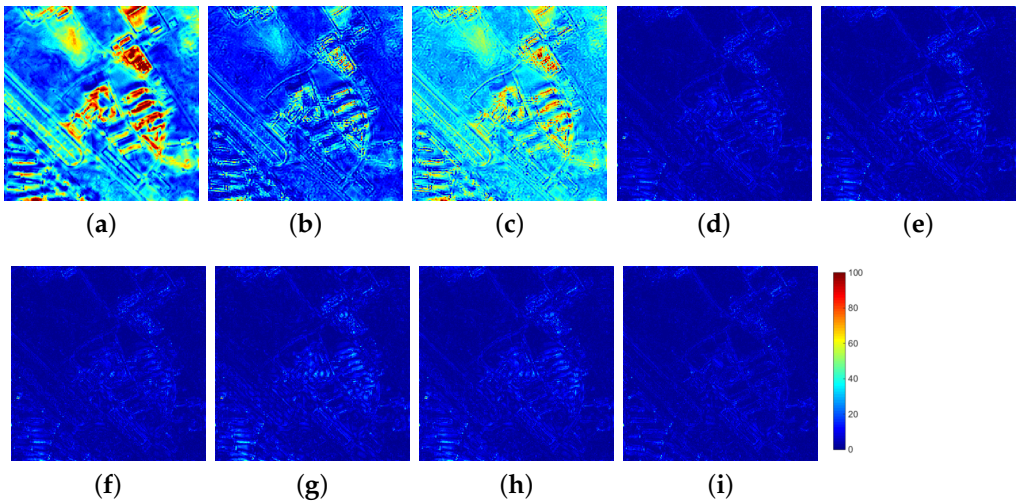

Figure 9. Residual images of the simulated QuickBird dataset. (**a**) EXP. (**b**) SR+SFIM. (**c**) SR+HPF. (**d**) SR+PNN. (**e**) SR+MSDCNN. (**f**) SR+Hyper_DSNet. (**g**) SR+FusionNet. (**h**) SR+DRPNN. (**i**) UIJSP-Net.

4.2.3. WorldView-3 Dataset

On the WorldView-3 dataset [44], we select 1027 pairs of LMP blocks, IMP blocks, and groundtruth patches for training and 26 pairs for testing. The visual comparison is shown in Figure 10. Considering that a lot of water regions (always useless for training and testing in DL due to low texture) are contained in this dataset, the performance gap

among different methods becomes small. However, the superiority of UIJSP-Net over other methods can be still clearly observed. We select the enlargements in Figure 10 as examples. The clarity of the plane in Figure 10i is higher than other methods in Figure 10b–h. The residual images of the simulated WorldView-3 dataset are shown in Figure 11. From the labeled region (by a red ellipse) in Figure 11i, it can be found that UIJSP-Net has the smallest errors. Furthermore, we compare the quantitative results in Table 4. The advantages of UIJSP-Net can be found in the four best metrics, including PSNR, CC, Q, and ERGAS, and one second best metric SAM.

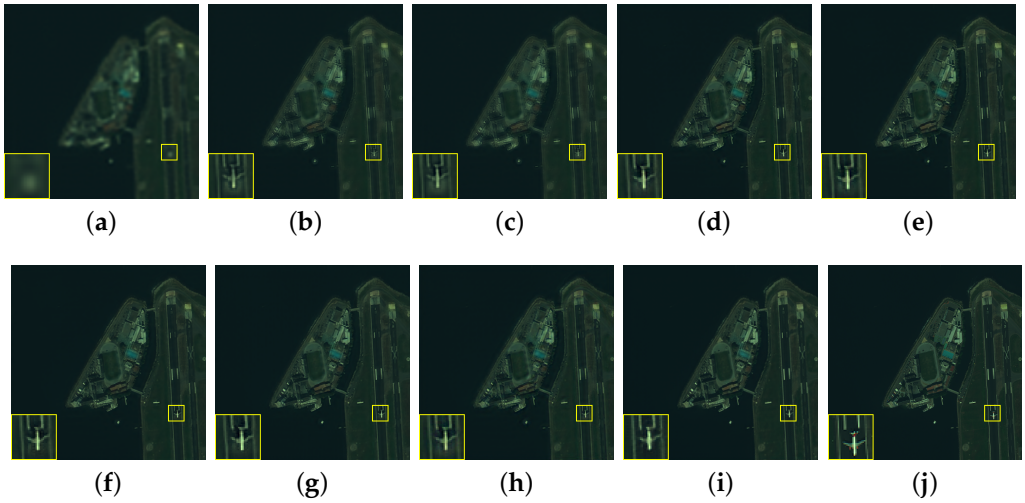

Figure 10. Visual comparison on the simulated WorldView-3 dataset. (**a**) EXP. (**b**) SR+SFIM. (**c**) SR+HPF. (**d**) SR+PNN. (**e**) SR+MSDCNN. (**f**) SR+Hyper_DSNet. (**g**) SR+FusionNet. (**h**) SR+DRPNN. (**i**) UIJSP-Net. (**j**) Groundtruth.

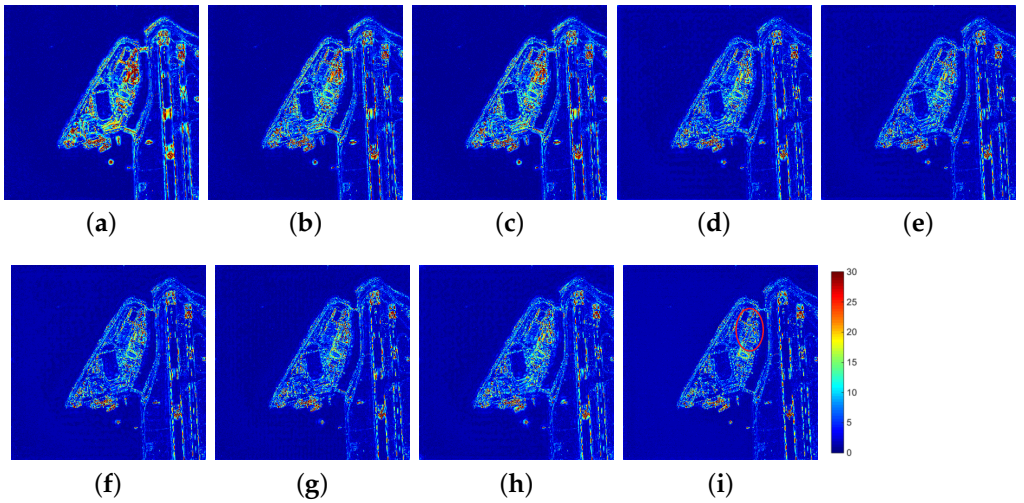

Figure 11. Residual images of the simulated WorldView-3 dataset. (**a**) EXP. (**b**) SR+SFIM. (**c**) SR+HPF. (**d**) SR+PNN. (**e**) SR+MSDCNN. (**f**) SR+Hyper_DSNet. (**g**) SR+FusionNet. (**h**) SR+DRPNN. (**i**) UIJSP-Net.

Table 4. Quantitative comparison on the 26 test images of the simulated WorldView-3 dataset. **BLUE**: the best, **RED**: the second-best.

Method	PSNR	SSIM	CC	Q	ERGAS	SAM
EXP	28.626 ± 2.326	0.880 ± 0.063	0.802 ± 0.033	0.881 ± 0.032	9.469 ± 1.184	3.969 ± 1.627
SR+SFIM	29.470 ± 2.339	$\mathbf{0.925 \pm 0.042}$	0.843 ± 0.029	0.904 ± 0.026	8.551 ± 1.004	3.754 ± 1.556
SR+HPF	29.167 ± 2.262	$\mathbf{0.918 \pm 0.043}$	0.841 ± 0.030	0.899 ± 0.026	8.845 ± 1.060	4.093 ± 1.060

Table 4. *Cont.*

Method	PSNR	SSIM	CC	*Q*	ERGAS	SAM
SR+PNN	30.454 ± 2.430	0.881 ± 0.061	0.957 ± 0.008	0.889 ± 0.024	7.742 ± 0.914	4.136 ± 1.616
SR+MSDCNN	31.061 ± 2.447	0.900 ± 0.051	0.965 ± 0.006	0.903 ± 0.019	7.201 ± 0.829	3.667 ± 1.313
SR+Hyper_DSNet	**31.164 ± 2.468**	0.903 ± 0.050	0.903 ± 0.050	0.906 ± 0.018	7.134 ± 0.814	3.511 ± 1.300
SR+FusionNet	31.154 ± 2.445	0.904 ± 0.050	**0.970 ± 0.006**	**0.908 ± 0.019**	**7.070 ± 0.832**	**3.412 ± 1.284**
SR+DRPNN	30.871 ± 2.395	0.893 ± 0.054	0.956 ± 0.008	0.896 ± 0.021	7.333 ± 0.830	3.890 ± 1.430
UIJSP-Net	**31.386 ± 2.317**	0.908 ± 0.045	**0.974 ± 0.005**	**0.913 ± 0.018**	**6.961 ± 0.844**	**3.430 ± 1.187**
Ideal value	$+\infty$	1	1	1	0	0

4.3. Real Experiment

In this experiment, the real images from the WorldView-3 dataset are used for comparison. The spatial resolutions of MS and PAN images are 256 × 256 × 4 and 1024 × 1024, respectively. In addition, the resolution of the JSP result is 2048 × 2048 × 4. We mainly compare the DL-based methods because of their superior performance in the above experiment. The visual comparison is shown in Figure 12. In a real experiment, the generalization problem in the DL-based method, such as SR+PNN, SR+MSDCNN, SR+Hyper_DSNet, SR+FusionNet, and SR+DRPNN, is more challenging than the proposed UIJSP-Net due to the following reason. Two separated DL-based processes are included in these methods and each DL-based process has its generalization problem caused by the difference between the training data and real testing data. Consequently, small artifacts in the SR process may lead to large artifacts in the following pansharpening process. Benefiting from the efficient unified network trained in one process, UIJSP-Net successfully mitigates the generalization problem. As a result, UIJSP-Net in Figure 12g demonstrates the best visual effects among all methods.

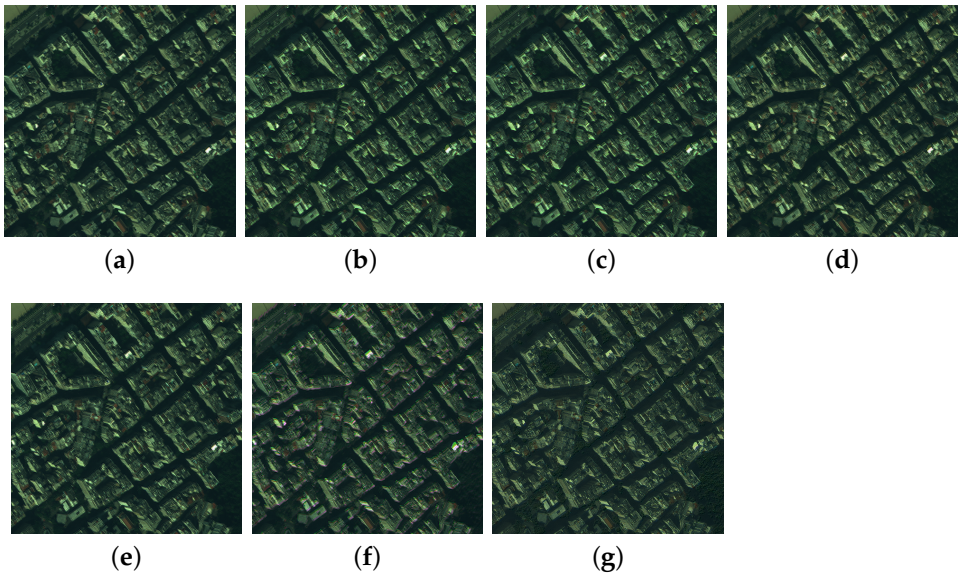

Figure 12. Visual comparison on the real WorldView-3 dataset. (**a**) EXP. (**b**) SR+PNN. (**c**) SR+MSDCNN. (**d**) SR+Hyper_DSNet. (**e**) SR+FusionNet. (**f**) SR+DRPNN. (**g**) UIJSP-Net.

4.4. Analysis of Computational Efficiency

To analyze the computational efficiency, we calculate the computational time of different methods tested on a pair of the simulated GaoFen-2 dataset. The experimental results are shown in Table 5. In particular, we mainly consider the DL-based methods, such as SR+PNN, SR+MSDCNN, SR+Hyper_DSNet, SR+FusionNet, SR+DRPNN, and UIJSP-Net, because they are all tested on the same platform (PyTorch). Generally speaking, the compu-

tational time of SR+PNN, SR+MSDCNN, and UIJSP-Net is comparable and smaller than other methods, exhibiting their computational efficiency.

Table 5. Comparison of the computational time (s) on the simulated GaoFen-2 dataset.

SR+PNN	SR+MSDCNN	SR+Hyper_DSNet	SR+FusionNet	SR+DRPNN	UIJSP-Net
0.1149	0.5295	0.5442	0.1147	0.1801	0.1190

4.5. Ablation Study

To further verify the efficiency of different modules in UIJSP-Net, we first conduct the following ablation experiment. Two models without the deep module $\Gamma(\cdot)$ and without the latent prior $\Theta(\cdot)$ trained on the simulated GaoFen-2 dataset are compared. The deep module $\Gamma(\cdot)$ that correlates to DRUNet indicates the fusion process of MS and PAN images. The latent prior $\Theta(\cdot)$ that relates to USRNet represents the SR process. The visual comparison and residual images are shown in Figure 13. The necessity of both modules in UIJSP-Net can be easily found from both the visual effect and the residuals. For example, if the SR process is not considered, then UIJSP-Net w/o $\Theta(\cdot)$ can be regarded as a DL-based pansharpening method. As a result, its clarity in Figure 13b is lower than that of UIJSP-Net in Figure 13c, thus leading to larger residuals. A similar conclusion can be observed from UIJSP-Net w/o $\Gamma(\cdot)$, which can be treated as a DL-based SR method. We list the quantitative comparison in Table 6. The module without $\Gamma(\cdot)$ has the worst performance due to lack of fusion. Compared with the module without $\Theta(\cdot)$, UIJSP-Net has a large improvement evaluated by the spatial metrics, such as PSNR, SSIM, CC, and Q, demonstrating the effectiveness of the SR process integrating with UIJSP-Net. Finally, UIJSP-Net provides the best balance between spatial and spectral qualities, having five of the best metrics, indicating the importance of both modules in UIJSP-Net.

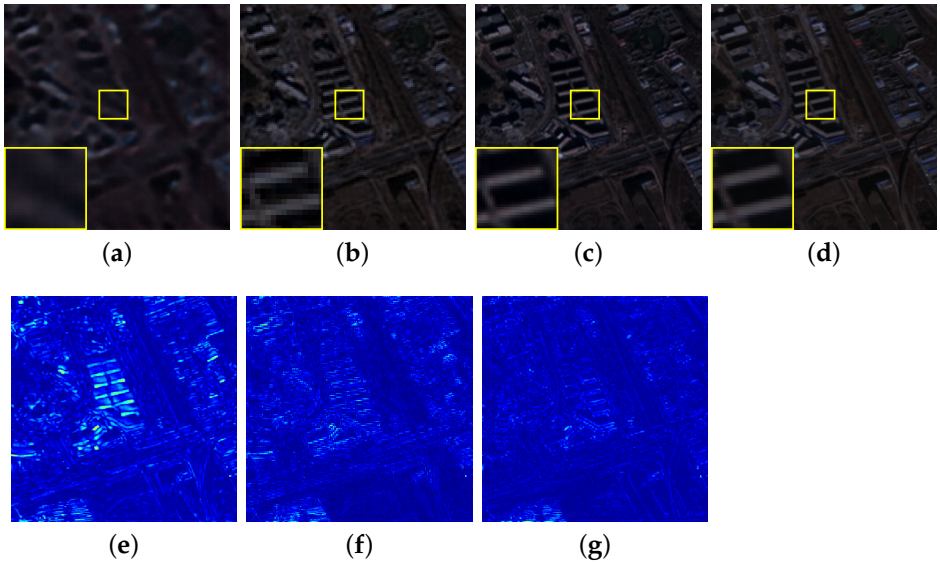

Figure 13. Visual results of ablation experiments on the simulated GaoFen-2 dataset. (**a**) w/o $\Gamma(\cdot)$. (**b**) w/o $\Theta(\cdot)$. (**c**) UIJSP-Net. (**d**) Groundtruth. (**e**) Residual of (**a**). (**f**) Residual of (**b**). (**g**) Residual of (**c**).

Table 6. Quantitative results of ablation experiments for different modules on the simulated GaoFen-2 dataset. **Bold**: the best.

Models	$\Gamma(\cdot)$	$\Theta(\cdot)$	PSNR	SSIM	CC	Q	ERGAS	SAM
w/o $\Gamma(\cdot)$	×	√	28.862 ± 1.384	0.680 ± 0.045	0.834 ± 0.031	0.766 ± 0.047	7.689 ± 0.973	6.432 ± 1.099
w/o $\Theta(\cdot)$	√	×	31.254 ± 1.270	0.824 ± 0.024	0.889 ± 0.023	0.895 ± 0.017	5.620 ± 0.672	$\mathbf{5.594 \pm 1.024}$
UIJSP-Net	√	√	$\mathbf{32.954 \pm 0.904}$	$\mathbf{0.880 \pm 0.019}$	$\mathbf{0.901 \pm 0.018}$	$\mathbf{0.929 \pm 0.022}$	$\mathbf{4.829 \pm 0.667}$	6.949 ± 1.387

Furthermore, to demonstrate the effectiveness of the proposed interpretable deep network, we also compare its performance with another unified JSP deep network USR-Net+Hyper_DSNet (U-USRNet+Hyper_DSNet) due to the good performance of SR+Hyper_DSNet in the simulated experiments. At this time, U-USRNet+Hyper_DSNet is directly trained from HMP and LMP blocks. In this experiment, we select the simulated GaoFen-2 dataset for verification. We show the visual comparison in Figure 14. Due to the special design of the interpretable deep network from a new physical model, UIJSP-Net generates clearer results and smaller errors than U-USRNet+Hyper_DSNet. The quantitative comparison is shown in Table 7, where UIJSP-Net outperforms U-USRNet+Hyper_DSNet when evaluated by five metrics, that is, PSNR, SSIM, CC, Q, and ERGAS.

(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 14. Visual results of ablation experiments on the simulated GaoFen-2 dataset. (**a**) U-USRNet+Hyper_DSNet. (**b**) UIJSP-Net. (**c**) Groundtruth. (**d**) Residual of (**a**). (**e**) Residual of (**b**).

Table 7. Quantitative results of ablation experiments about U-USRNet+Hyper_DSNet and UIJSP-Net on the simulated GaoFen-2 dataset. **Bold**: the best.

Method	PSNR	SSIM	CC	Q	ERGAS	SAM
U-USRNet+Hyper_DSNet	31.734 ± 1.110	0.837 ± 0.026	0.878 ± 0.023	0.897 ± 0.021	5.509 ± 0.666	$\mathbf{6.702 \pm 0.968}$
UIJSP-Net	$\mathbf{32.954 \pm 0.904}$	$\mathbf{0.880 \pm 0.019}$	$\mathbf{0.901 \pm 0.018}$	$\mathbf{0.929 \pm 0.022}$	$\mathbf{4.829 \pm 0.667}$	6.949 ± 1.387

4.6. NDVI Experiment

To further analyze the application of UIJSP-Net, we utilize an experiment based on normalized difference vegetation index (NDVI), which can be calculated as

$$I_{\text{NDVI}} = \frac{I_{NIR} - I_R}{I_{NIR} + I_R}.$$

I_{NIR} and I_R are the near-infrared and red bands of an MS image, respectively. Therefore, we can use the JSP result to analyze their performance on the NDVI application. In this experiment, we select the simulated WorldView-3 dataset as an example. The visual comparison is shown in Figure 15. The NDVI of EXP shown in Figure 15a is blurry according to the low spatial resolution of the MS image. By introducing JSP, the spatial clarity of NDVI can be improved. However, there are some large errors in SR+HPF, SR+PNN, SR+MSDCNN, and SR+DRPNN due to large spectral distortion in the JSP results. UIJSP-Net produces the best result over others, being closest to the groundtruth. Furthermore, the residual images of NDVI are calculated in Figure 16. Red and blue indicate large and small residuals, respectively. The quantitative comparison on NDVI is listed in Table 8, where UIJSP-Net achieves the best results evaluated by all metrics, demonstrating its prospect in the NDVI application.

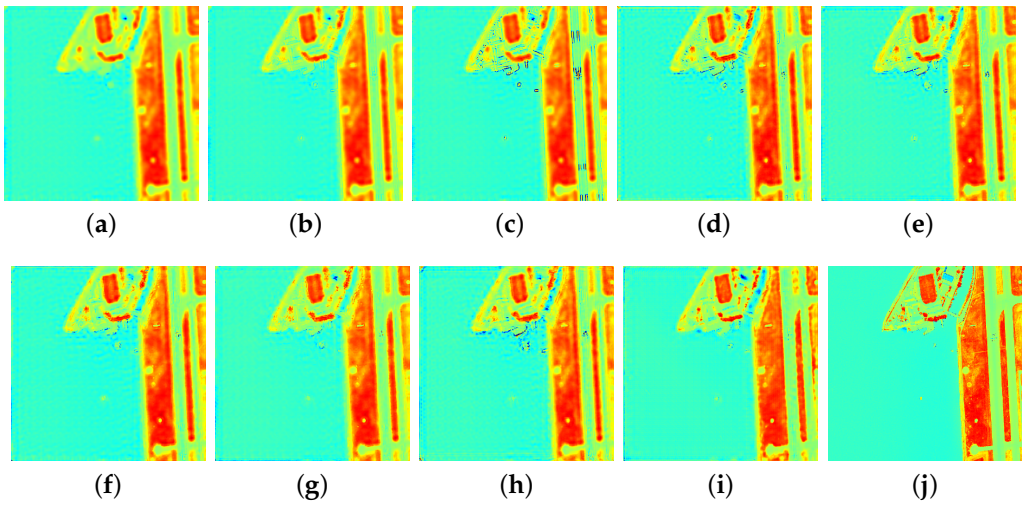

Figure 15. Visual comparison on NDVI. (**a**) EXP. (**b**) SR+SFIM. (**c**) SR+HPF. (**d**) SR+PNN. (**e**) SR+MSDCNN. (**f**) SR+Hyper_DSNet. (**g**) SR+FusionNet. (**h**) SR+DRPNN. (**i**) UIJSP-Net. (**j**) Groundtruth.

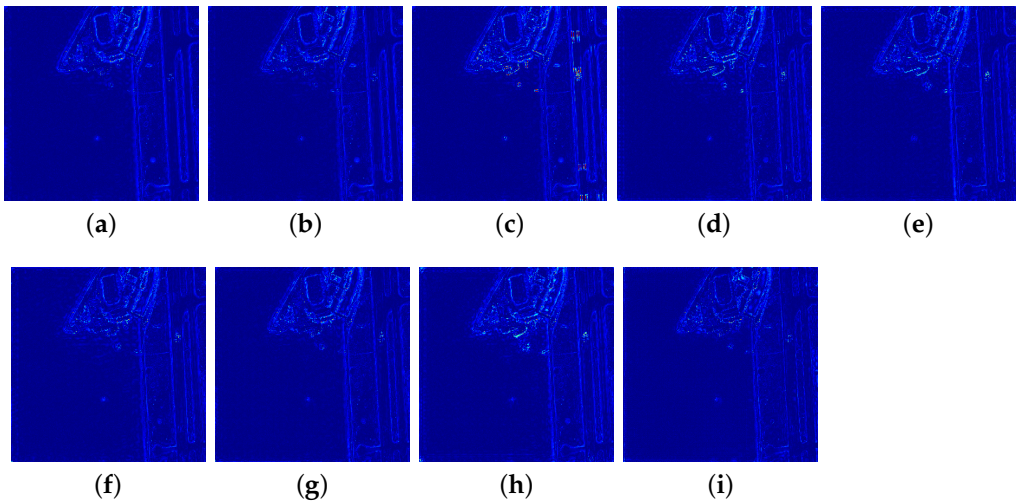

Figure 16. Residual images of NDVI. (**a**) EXP. (**b**) SR+SFIM. (**c**) SR+HPF. (**d**) SR+PNN. (**e**) SR+MSDCNN. (**f**) SR+Hyper_DSNet. (**g**) SR+FusionNet. (**h**) SR+DRPNN. (**i**) UIJSP-Net.

Table 8. Quantitative comparison on NDVI. **BLUE**: the best, **RED**: the second-best.

Method	RMSE	SSIM	CC
EXP	11.702	0.945	0.828
SR+SFIM	**11.476**	**0.946**	0.832
SR+HPF	14.626	0.914	0.795
SR+PNN	12.991	0.933	0.784
SR+MSDCNN	12.545	0.937	0.804
SR+Hyper_DSNet	11.818	0.945	0.813
SR+FusionNet	11.529	**0.946**	**0.837**
SR+DRPNN	13.496	0.927	0.770
UIJSP-Net	**11.154**	**0.949**	**0.864**
Ideal Value	0	1	1

5. Conclusions

In this paper, we propose a novel unified interpretable deep network for JSP, named UIJSP-Net. To this end, we first formulate the desired problem as a unified optimization

model with two deep priors. This strategy is helpful to balance the spatial and spectral qualities well. Furthermore, we utilize the unfolding technology to form UIJSP-Net by mapping a series of iterative steps derived from the alternating direction method of multipliers into several network stages. Finally, we carry out extensive experiments based on datasets from GaoFen-2, QuickBird, and WorldView-3 satellites to exhibit the advantage of UIJSP-Net compared with other state-of-the-art methods evaluated by both visual comparison and quantitative analysis. An experiment based on NDVI is further utilized to indicate the application aspect of UIJSP-Net in remote sensing. An ablation study is also provided to analyze the importance of different modules in UIJSP-Net. For simplicity, we only employ the SR network implemented by USRNet in our deep prior. More efficient SR networks that can achieve SR in arbitrary resolution will be considered to handle higher-resolution images in the future.

Author Contributions: Conceptualization, D.Y.; methodology, D.Y.; software, D.Y. and W.Z.; validation, M.X.; formal analysis, X.T.; investigation, D.Y.; resources, X.T.; data curation, W.Z.; writing—original draft preparation, D.Y.; writing—review and editing, X.T. and H.J.; visualization, M.X.; supervision, H.J.; project administration, X.T.; funding acquisition, X.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (61971315).

Data Availability Statement: Data will be made available on request.

Acknowledgments: The authors would like to thank the editors and reviewers for their very competent comments and helpful suggestions to improve this paper.

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

References

1. Zhang, T.J.; Deng, L.J.; Huang, T.Z.; Chanussot, J.; Vivone, G. A Triple-Double Convolutional Neural Network for Panchromatic Sharpening. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 9088–9101. [CrossRef]
2. Zhang, M.; Li, S.; Yu, F.; Tian, X. Image fusion employing adaptive spectral-spatial gradient sparse regularization in UAV remote sensing. *Signal Process.* **2020**, *170*, 107434. [CrossRef]
3. Tian, X.; Zhang, W.; Yu, D.; Ma, J. Sparse Tensor Prior for Hyperspectral, Multispectral, and Panchromatic Image Fusion. *IEEE/CAA J. Autom. Sin.* **2023**, *10*, 284–286. [CrossRef]
4. Wang, W.; Zhou, Z.; Zhang, X.; Lv, T.; Liu, H.; Liang, L. DiTBN: Detail Injection-Based Two-Branch Network for Pansharpening of Remote Sensing Images. *Remote Sens.* **2022**, *14*, 6120. [CrossRef]
5. Masi, G.; Cozzolino, D.; Verdoliva, L.; Scarpa, G. Pansharpening by convolutional neural networks. *Remote Sens.* **2016**, *8*, 594. [CrossRef]
6. 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.
7. 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]
8. Jin, C.; Deng, L.J.; Huang, T.Z.; Vivone, G. Laplacian pyramid networks: A new approach for multispectral pansharpening. *Inf. Fusion* **2022**, *78*, 158–170. [CrossRef]
9. Ke, C.; Liang, H.; Li, D.; Tian, X. High-Frequency Transformer Network Based on Window Cross-Attention for Pansharpening. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
10. Ma, J.; Yu, W.; Chen, C.; Liang, P.; Guo, X.; Jiang, J. Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion. *Inf. Fusion* **2020**, *62*, 110–120. [CrossRef]
11. Chouteau, F.; Gabet, L.; Fraisse, R.; Bonfort, T.; Harnoufi, B.; Greiner, V.; Le Goff, M.; Ortner, M.; Paveau, V. Joint Super-Resolution and Image Restoration for PLÉIADES NEO Imagery. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *43*, 9–15. [CrossRef]
12. Zhu, S.; Zeng, B.; Zeng, L.; Gabbouj, M. Image interpolation based on non-local geometric similarities and directional gradients. *IEEE Trans. Multimed.* **2016**, *18*, 1707–1719. [CrossRef]
13. Yang, J.; Wright, J.; Huang, T.; Ma, Y. Image super-resolution as sparse representation of raw image patches. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
14. Zhang, K.; Gao, X.; Tao, D.; Li, X. Multi-scale dictionary for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1114–1121.
15. Freedman, G.; Fattal, R. Image and video upscaling from local self-examples. *ACM Trans. Graph. (TOG)* **2011**, *30*, 1–11. [CrossRef]

16. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a deep convolutional network for image super-resolution. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 184–199.
17. Dai, T.; Cai, J.; Zhang, Y.; Xia, S.T.; Zhang, L. Second-order attention network for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11065–11074.
18. Zhang, K.; Gool, L.V.; Timofte, R. Deep unfolding network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 3217–3226.
19. Zhang, Y.; Wei, D.; Qin, C.; Wang, H.; Pfister, H.; Fu, Y. Context reasoning attention network for image super-resolution. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 4278–4287.
20. Gao, S.; Zhuang, X. Bayesian image super-resolution with deep modeling of image statistics. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1405–1423. [CrossRef]
21. Zhou, Z.; Ma, N.; Li, Y.; Yang, P.; Zhang, P.; Li, Y. Variational PCA fusion for Pan-sharpening very high resolution imagery. *Sci. China Inf. Sci.* **2014**, *57*, 1–10. [CrossRef]
22. 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]
23. Rahmani, S.; Strait, M.; Merkurjev, D.; Moeller, M.; Wittman, T. An adaptive IHS pan-sharpening method. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 746–750. [CrossRef]
24. Liu, J. Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. *Int. J. Remote Sens.* **2000**, *21*, 3461–3472. [CrossRef]
25. Vivone, G.; Restaino, R.; Dalla Mura, M.; Licciardi, G.; Chanussot, J. Contrast and error-based fusion schemes for multispectral image pansharpening. *IEEE Geosci. Remote Sens. Lett.* **2013**, *11*, 930–934. [CrossRef]
26. Addesso, P.; Restaino, R.; Vivone, G. An Improved Version of the Generalized Laplacian Pyramid Algorithm for Pansharpening. *Remote Sens.* **2021**, *13*, 3386. [CrossRef]
27. Alparone, L.; Garzelli, A.; Vivone, G. Intersensor statistical matching for pansharpening: Theoretical issues and practical solutions. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4682–4695. [CrossRef]
28. Liu, J.; Liang, S. Pan-sharpening using a guided filter. *Int. J. Remote Sens.* **2016**, *37*, 1777–1800. [CrossRef]
29. Ballester, C.; Caselles, V.; Igual, L.; Verdera, J.; Rougé, B. A variational model for P+ XS image fusion. *Int. J. Comput. Vis.* **2006**, *69*, 43–58. [CrossRef]
30. Palsson, F.; Sveinsson, J.R.; Ulfarsson, M.O. A new pansharpening algorithm based on total variation. *IEEE Geosci. Remote Sens. Lett.* **2013**, *11*, 318–322. [CrossRef]
31. Fu, X.; Lin, Z.; Huang, Y.; Ding, X. A variational pan-sharpening with local gradient constraints. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10265–10274.
32. Tian, X.; Chen, Y.; Yang, C.; Ma, J. Variational pansharpening by exploiting cartoon-texture similarities. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
33. Xu, H.; Le, Z.; Huang, J.; Ma, J. A Cross-Direction and Progressive Network for Pan-Sharpening. *Remote Sens.* **2021**, *13*, 3045. [CrossRef]
34. Wei, Y.; Yuan, Q.; Shen, H.; Zhang, L. Boosting the accuracy of multispectral image pansharpening by learning a deep residual network. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1795–1799. [CrossRef]
35. 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]
36. Deng, L.J.; Vivone, G.; Jin, C.; Chanussot, J. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6995–7010. [CrossRef]
37. Tian, X.; Li, K.; Wang, Z.; Ma, J. VP-Net: An interpretable deep network for variational pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
38. Wen, R.; Deng, L.J.; Wu, Z.C.; Wu, X.; Vivone, G. A novel spatial fidelity with learnable nonlinear mapping for panchromatic sharpening. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5401915. [CrossRef]
39. Zhou, M.; Huang, J.; Fang, Y.; Fu, X.; Liu, A. Pan-sharpening with customized transformer and invertible neural network. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2022; Volume 36, pp. 3553–3561.
40. Beck, A.; Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2009**, *2*, 183–202. [CrossRef]
41. Zhang, K.; Li, Y.; Zuo, W.; Zhang, L.; Van Gool, L.; Timofte, R. Plug-and-Play Image Restoration With Deep Denoiser Prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 6360–6376. [CrossRef] [PubMed]
42. Yang, D.; Sun, J. Proximal dehaze-net: A prior learning-based deep network for single image dehazing. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 702–717.

43. Zhuo, Y.W.; Zhang, T.J.; Hu, J.F.; Dou, H.X.; Huang, T.Z.; Deng, L.J. A Deep-Shallow Fusion Network With Multidetail Extractor and Spectral Attention for Hyperspectral Pansharpening. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 7539–7555. [CrossRef]
44. Zhou, H.; Liu, Q.; Wang, Y. PGMAN: An Unsupervised Generative Multiadversarial Network for Pansharpening. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 6316–6327. [CrossRef]