Article

Hierarchical Sparse Learning with Spectral-Spatial Information for Hyperspectral Imagery Denoising

Shuai Liu [1,2,*], Licheng Jiao [1,2] and Shuyuan Yang [1,2]

[1] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an 710071, China; lchjiao@mail.xidian.edu.cn (L.J.); syyang@xidian.edu.cn (S.Y.)

[2] Joint International Research Laboratory of Intelligent Perception and Computation, Xidian University, Xi'an 710071, China

[*] Correspondence: shliu_122908@yahoo.com; Tel.: +86-185-919-86967

Abstract: During the acquisition process hyperspectral images (HSI) are inevitably corrupted by various noises, which greatly influence their visual impression and subsequent applications. In this paper, a novel Bayesian approach integrating hierarchical sparse learning and spectral-spatial information is proposed for HSI denoising. Based on the structure correlations, spectral bands with similar and continuous features are segmented into the same band-subset. To exploit local similarity, each subset is then divided into overlapping cubic patches. All patches can be regarded as consisting of clean image component, Gaussian noise component and sparse noise component. The first term is depicted by a linear combination of dictionary elements, where Gaussian process with Gamma distribution is applied to impose spatial consistency on dictionary. The last two terms are utilized to fully depict the noise characteristics. Furthermore, the sparseness of the model is adaptively manifested through Beta-Bernoulli process. Calculated by Gibbs sampler, the proposed model can directly predict the noise and dictionary without priori information of the degraded HSI. The experimental results on both synthetic and real HSI demonstrate that the proposed approach can better suppress the existing various noises and preserve the structure/spectral-spatial information than the compared state-of-art approaches.

Keywords: hierarchical sparse learning; denoising; spectral-spatial information; hyperspectral images

1. Introduction

As three-dimensional (3D) data integrating both spectral and spatial information concurrently, hyperspectral imagery (HSI) can provide more reliable and accurate information about observed objects. It has numerous applications in remote sensing, diagnostic medicine and mineralogy, etc., with encouraging results [1–3]. HSI can be considered as a set of gray-scale images over wavelength, whose entries are the spectral responses. Due to the limited light, imaging system, photon effects, and calibration error, hyperspectral images are inevitably contaminated by annoying noises with different statistical properties [4,5]. The existing noises greatly reduce the quality of the HSI and increase the difficulty of subsequent processing, such as target detection, agriculture assessment, mineral exploitation and ground-cover classification. Therefore, noise reduction is an essential research topic in hyperspectral images analysis.

Over the past two decades, various approaches have been developed for the noise reduction of HSI, which aim to restore the original HSI from its noisy images as accurately as possible. The traditional denoising algorithms, in two-dimensional (2D) image domain, have been widely utilized to remove the noise for HSI [6–10], such as wavelet transform, total variation, nonlocalmeans (NLM), K-singular

value decomposition (KSVD) and block matching three-dimensional filtering (BM3D). These methods denoise the HSI band by band and destroy the latent high-dimensional structure of the HSI, which result in a great loss of spectral correlations. Especially when the spectral bands are contaminated by heavy noise, these approaches cannot effectively restore the HSI at most time.

Lam et al. [4] showed that spectral domain statistics can aid in improving the quality of the restored HSI. In this regard, some techniques treat HSI as 3D data cube rather than a 2D picture, which can fully exploit the correlations in the spectral domain. In [11,12], the kernel nonnegative Tucker decomposition and the spatial–spectral wavelet shrinkage are proposed for denoising HSI. NLM techniques based on 3D cubes are applied for the HSI recovery by exploiting the high structural sparseness/redundancy [13,14]. Total variation algorithms have been proved their effectiveness in preserving edge information for denoising HSI. In [15], the cubic total variation is developed to denoise HSI, which is achieved by exploring the 2D total variation in the spatial dimension and the 1D total variation in the spectral domain. Using spectral noise differences and spatial information differences, a spectral-spatial adaptive total variation model is proposed [16]. Nonlocal models, integrated with adaptive multidimensional total variation are shown in [17]. Noise reduction based on the tensor factorization is given in [18]. The block matching 4D (BM4D) filtering is considered in [19], which reduces the noise for the volumetric data by grouping similar cubes and exploring the collaborative filtering paradigm. By solving a Bayesian least squares optimization problem, Monte Carlo sampling has been used to estimate the posterior probability for denoising HSI [20]. In [21], an algorithm employing the hybrid conditional random field model has been introduced. To improve the denoising performance, a metric *Q*-weighted fusion algorithm is proposed to merge the denoising results of both spatial and spectral views [22].

Notably, the works based on sparse dictionary learning have proven their efficacy and popularity for image recovery. The success can be attributed to the fact that valid data in corrupted images are intrinsically sparse and the noises are uniformly spread in many domains [23]. Based on this view point, the noise can be successfully reduced. The 3D sparse coding is exploited to denoise HSI [24], which could fully explore the spectral information by extracting the different patches and obtain competitive results. In [25], the dictionary learning in a Bayesian manner is applied to recover the hyperspectral cube. Exploiting the intraband structure and the interband correlation, a joint spectral–spatial distributed sparse representation is proposed for noise reduction of HSI [26]. Sparse nonnegative matrix factorization combined with spectral–spatial information is also implemented to reduce the noise of HSI [27]. Under a new linear model exploring singular value decomposition and wavelet transform, the HSI denoising is achieved via sparse regularization [28]. In the HSI, adjacent spectral bands and spatial pixels are typically highly correlated, which exhibits the low-rank characteristics of HSI [29–32]. Using low-rank matrix recovery framework, an HSI restoration technique (LRMR) is explored to simultaneously remove various noises in HSI [31]. Zhu et al. presented an HSI mixed-noise removal method [32], which achieves the promising image quality by combining low-rank constraint and total-variation-regularized. Deep learning has been widely used in the HSI analysis [33,34]. Motivated by this, a deep dictionary learning method (DDL3+FT), consisting of the hierarchical dictionary, feature denoising and fine-turning, is developed to effectively suppress the noise in HSI [34].

However, most denoising methods are globally applied to the whole HSI or utilized on specific spectral bands obtained by partitioning the spectra with a fixed value. They obviously discard the high correlations in the spectral domain. Additionally, the real-world HSI are usually contaminated by the signal-independent noise, signal-dependent noise components or the mixed noise levels, e.g., Gaussian noise, Poisson noise, dead pixel lines and stripes. The dead pixel lines and stripes only exist in some pixels/bands, and they can be regarded as sparse noises [31]. Most studies above only focus on suppressing one or two kinds of specific image noise, such as Gaussian noise with constant variance among the bands or Poisson noise, which can not exactly depict the noise characteristics of the real-world HSI.

To address the previous limitations, this paper integrates the spectral-spatial information into the hierarchical dictionary model to efficiently suppress the existing noises in HSI. The contributions of this paper can be summarized as follows:

(1)   First, we present a spectral–spatial data extraction approach for HSI based on their high structural correlations in the spectral domain. Using this approach, the spectral bands, with similar and continuous features can be adaptively segmented into the same band-subset.

(2)   Second, a hierarchical dictionary model is organized by the prior distributions and hyper-prior distributions to deeply illustrate the noisy HSI. It depicts the noiseless data by utilizing the dictionary, in which the spatial consistency is exploited by Gaussian process. Meanwhile, by decomposing the noise term into Gaussian noise term and sparse noise term, the proposed method can well represent the intrinsic noise properties.

(3)   Last but not the least, many experiments performed under different conditions are displayed. Compared with other state-of-the-art denoising approaches, the suggested method shows superior performance on suppressing various noises, including Gaussian noise, Poisson noise, dead pixel lines and stripes.

The remainder of this paper is organized as follows: the denoising framework for 2D images based on sparse learning is presented in Section 2. The spatial-spectral data extraction and the proposed hierarchical sparse learning model, along with the Bayesian inference implemented by Gibbs sampler, are explained in detail in Section 3. Experimental results on both synthetic and real-world hyperspectral images are reported in Section 4, with comparisons between the suggested method and the state-of-the-art approaches. The concluding remarks are given in Section 5.

2. Sparse Learning for Denoising 2D Images

Let X represents a 2D noisy image of size $T_1 \times T_2$. The common restoration algorithms regard that the pixels extracted from the image are similar to the pixels extracted from their neighboring region to some degree. In this sense, the observed images are often divided into many 2D patches before restoration. By transforming the i-th patch of X into the column vector x_i, the commonly used image degradation framework can be formulated by Equation (1):

$$x_i = Da_i + N \tag{1}$$

The first term on the right side of the equation represents the clean component of the i-th patch, where D is the underlying dictionary and a_i are the sparse coefficients with respect to dictionary elements. The second part N defines the zero-mean Gaussian noise component, with the standard variance σ. Generally, the noiseless signals can be estimated by solving the l_p-minimization problem given by Equation (2):

$$\{D, \hat{a}_i\} = \arg \min_{D,a} ||x_i - Da_i||_2^2 + \lambda ||a_i||_p \tag{2}$$

In Equation (2), $||x_i - Da_i||_2^2$ is the data fidelity item, which stands for the fidelity between the observed noisy image and the original clean image, and $||a_i||_p$ is the regularization item, which gives the prior knowledge of noiseless images. $\lambda \geq 0$ is a regularization parameter, which is utilized to trade off between the fidelity item and sparseness of the coefficient. A larger λ usually implies a much sparser solution of a.

Denoising approaches aim to restore the noise-free patch $\hat{x}_i$ from degraded data x_i by calculating $\hat{x}_i = D\hat{a}_i$. The sparse prior, p is usually set to zero or one. When $p = 0$, $||a_i||_0$ represents the number of nonzero elements in a_i; when $p = 1$, $||a_i||_1$ is the sum of elements' absolute value. l_0 norm is not convex, and it cannot guarantee the uniqueness of the solution. Meanwhile, l_0-minimization is an NP-hard combinatorial search problem in general case, which is usually solved by greedy algorithms. l_1 norm is a convex function. The uniqueness of solution can be guaranteed by replacing l_0 norm with l_1 norm [35]. Traditional convex optimization methods are used to solve the problem by alternating minimization of D and a_i.

According to the types of dictionary basis, methods can be broadly classified into two groups [36]: (1) transformation-based methods. In these works, the mathematical model are developed to predict a dictionary, e.g., discrete cosine transformation (DCT), wavelet, curvelet, contourlet and shearlet; (2) learning-based methods [9,37]. The learning techniques are utilized to construct the dictionary from the training data in these studies. The learned dictionary has shown better performance in many applications. Once D and $\hat{a}_i$ are given, the idea image x_i can be obtained as $\hat{x}_i = D\hat{a}_i$. Bayesian methods have been widely used for learning the sparse dictionary [25,38], which have achieved competing restoration performance.

All these conventional methods have achieved good performance on denoising 2D images, but for more complicated pictures, e.g., hyperspectral images, the conventional 2D methods usually have less ability to explain the structural or textural details, which significantly limits the ability of representation and the range of application for the dictionary learning. In this case, the hierarchical dictionary learning methods have been introduced to deeply exploit the latent characteristics of images [37–39], with promising results.

3. HSI Denoising with Spectral-Spatial Information and Hierarchical Sparse Learning

For restoring well the HSI, we extend the denoising framework for 2D images into the hierarchical dictionary learning framework in a Bayesian manner. Let Y be a HSI with a size of $l_m \times l_n \times l_\lambda$, where $l_m \times l_n$ represents the number of spatial pixels, and l_λ defines the size of the spectral domain. Generally, spectral correlations of HSI are of more importance than spatial characteristics, which have recently been reported to enhance performance in tracking, classification and target detection, etc. The proposed method consists of two main stages: spatial-spectral data extraction and noise-free estimation based on hierarchal spare learning. Firstly, according to the structure similarity index measure (SSIM), a band-subset partition is introduced to segment the HSI into multiple/one band-subsets. Each band-subset consists of continuous bands with high structure correlation. Then local homogeneous region can be extracted for each spectral pixel. Secondly, a hierarchical sparse learning model, which is composed of the clean image term, Gaussian noise term and sparse noise term, is constructed to suppress well the various noises. To effectively capture the latent spatial information of HSI, a Gaussian process with gamma constraints is imposed to the dictionary in the clean image term. The second and the third term are used to infer the statistics characteristics of existing noises, e.g., Gaussian noise, Poisson noise, dead pixel lines, stripes or a mixture of them. By solving this framework in a Bayesian manner, the restored results for each band-subset can be obtained. The framework of the proposed method is shown in Figure 1.
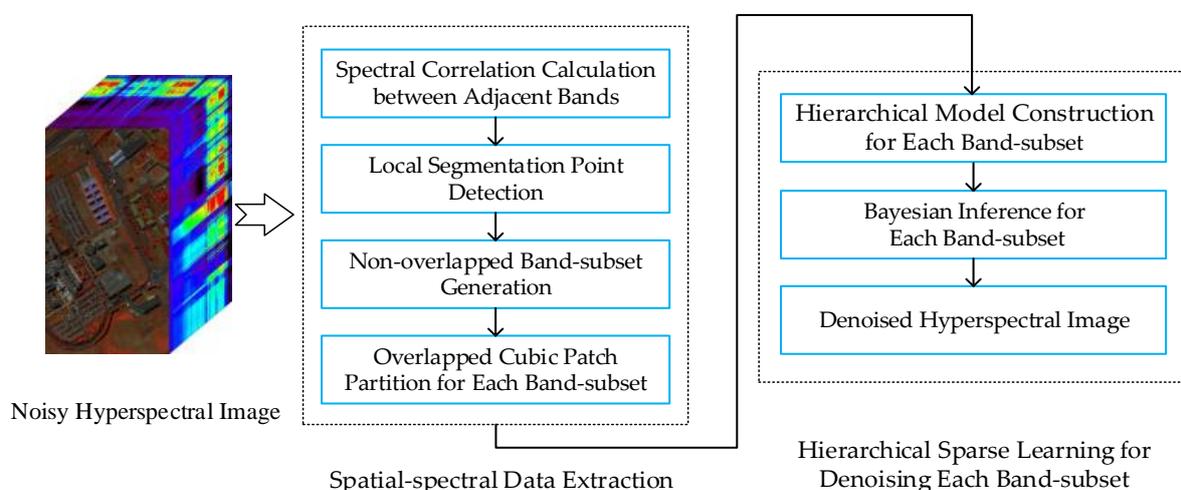

Figure 1. Framework of the proposed approach.

3.1. Spatial-Spectral Data Extraction

In HSI, the neighboring bands are acquired under relatively similar sensor conditions, and they are strongly correlated in the spectral domain. Based on this, SSIM is utilized to measure spectral correlations between adjacent bands [40]. Suppose B_j and B_{j+1} represent 2D images lying in the *j*-th and *j*+1-th band, respectively. Structure similarity between the *j*-th and *j*+1-th band can be defined by Equation (3):

$$SSIM(\boldsymbol{B}_j, \boldsymbol{B}_{j+1}) = \frac{\left(2\mu_{B_j}\mu_{B_{j+1}} + c_1\right)\left(2\sigma_{B_j,B_{j+1}} + c_2\right)}{\left(\mu_{B_j}^2 + \mu_{B_{j+1}}^2 + c_1\right)\left(\sigma_{B_j}^2 + \sigma_{B_{j+1}}^2 + c_2\right)} \tag{3}$$

In Equation (3), μ_{B_j} and σ_{B_j} are the mean and variance of band B_j, respectively; $\mu_{B_{j+1}}$ and $\sigma_{B_{j+1}}^2$ are the mean and variance of band B_{j+1}; The predefined constants c_1 and c_2 are applied to stabilize the division with weak denominator. Normally, the closer $SSIM(\boldsymbol{B}_j,\boldsymbol{B}_{j+1})$ is to one, the stronger the structural correlations are between the *j*-th and *j*+1-th spectral bands. Supposing $S_c(j) = SSIM(\boldsymbol{B}_j,\boldsymbol{B}_{j+1})$, the structural correlation curve $\boldsymbol{S}_c$ can be generated. Figure 2 shows the structural correlation curves of different hyperspectral images.
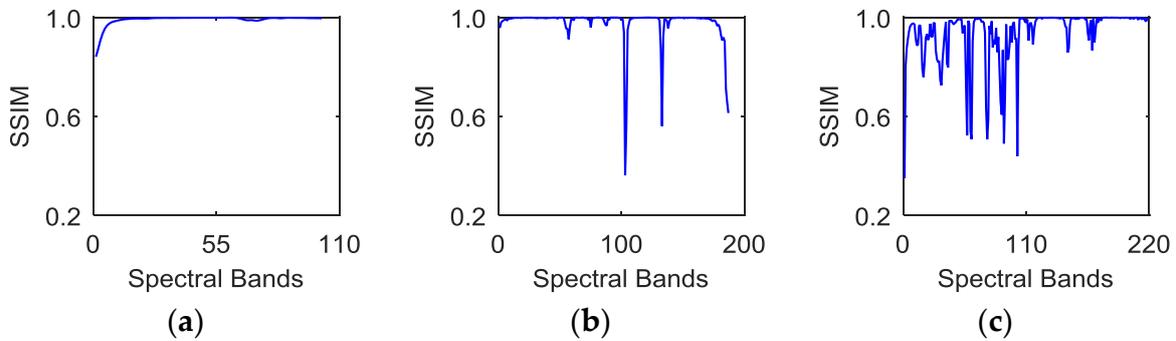

Figure 2. Structural correlation curves of different hyperspectral images: (**a**) Pavia data; (**b**) Urban data. (**c**) Indian Pines data.

Based on the curves in Figure 2, it can be found that the correlation coefficients across adjacent bands obviously vary considerably for different hyperspectral images. The curve in Figure 2a displays a relatively steady trend. Obvious drops exist in Figure 2b,c, which means that some adjacent spectral bands in Urban data and Indian Pines data have much lower structural similarity; meanwhile, continuous spectral bands between two neighboring drops show a relatively stable trend. However, most previous studies neglect this property of spectral bands. These approaches are applied to directly restore the entire bands, or to sequentially denoise the band-subset constructed by partitioning all spectral bands with a fixed value. To make the best of the correlations across neighboring bands, this paper explores the optimal partition in spectral domain by estimating the drops in curve $\boldsymbol{S}_c$. The detailed procedure of spatial-spectral data extraction is as follows: firstly, form the structural correlation curve $\boldsymbol{S}_c$ using Equation (3). Secondly, detect the local segmentation points $\boldsymbol{S}_c(j)$ in the curve $\boldsymbol{S}_c$. $\boldsymbol{S}_c(j)$ meets the condition denoted by Equation (4):

$$\begin{cases} S_c\left(j-1\right) - S_c\left(j\right) > \eta \\ S_c\left(j+1\right) - S_c\left(j\right) > \eta \end{cases} \tag{4}$$

In Equation (4), η is a predefined threshold, which is applied to avoid the local disturbance caused by noise in curve $\boldsymbol{S}_c$. The beginning and ending points in $\boldsymbol{S}_c$ can be treated as the intrinsic local partition points. To better exploit useful spectral characteristics and suppress the noise [40], a fused image is generated by using the average of the spectral bands between the neighboring segmentation points.

Based on Equations (3) and (4), it is determined whether adjacent fused images are to be merged or not. By this way, the segmentation points in curve S_c can be identified.

As shown in Figure 2a, the correlations across all neighboring bands have a relatively stable trend, and there is no need for spectral band segmentation in this case. As shown in Figure 2b,c, the HSI can be divided into the non-overlapping band-subsets based on the segmentation points. In this case, the HSI itself can be considered as a special band-subset. Let C denote the number of band-subsets. Above all, the noisy data can be reshaped as $X = \{X^1, \ldots, X^c, \ldots, X^C\}$, $c = 1, \ldots, C$ and X^c is the c-th band-subset.

Finally, to effectively preserve the local details of HSI in the spatial domain [41], we utilize the cubic patches instead of the 2D patches during the denoising process. By this, each band-subset is divided into many overlapped cubic patches. The size of each cubic patch is fixed as $l_x \times l_y \times l_c$, where $l_x \times l_y$ is the size of spatial dimensions and l_c defines the number of spectral bands in the c-th band-set. We note that there are various material categories within the same cubic patch, which increase the spectral-signature contamination (mixing/blurring) as the spatial size of cubic patches gets bigger. Therefore, the larger l_x and l_y may result in the instability of classification accuracies. If the neighborhood size is too small, it cannot well explore the spatial information. In this paper, we make a balance between the stability and the exploitation of spatial information, and we set $l_x = l_y = 4$ in the experiments.

After reshaping the cubic patches into the vectors, we can obtain the c-th band-subset $X^c = \left[x_1^c, \ldots, x_i^c, \ldots, x_M^c\right]$, where $M = (l_m - l_x + 1)(l_n - l_y + 1)$ represents the number of cubic patches and $x_i^c \in R^P$, $P = l_x \times l_y \times l_c$ is the vector generated by the i-th cubic patch of the c-th band-subset.

3.2. Hierarchical Sparse Learning for Denoising Each Band-Subset

In this subsection, the hierarchical sparse learning framework is well constructed with prior distribution and hyper-prior distribution to restore the noisy HSI [27,37,42], which integrates the Gaussian process and Gamma distribution into dictionary atoms to explore the spatial consistency of HSI simultaneously. The novel prior consisting of prior distribution and hyper-prior distribution is often called as hierarchical prior. Noticed that, the sparse learning framework with multiple hierarchical priors can be regarded as a special case of deep learning. Considering the dataset $X = \{X^1, \ldots, X^c, \ldots, X^C\}$, $c = 1, \ldots, C$, the suggested approach sequentially depicts the characteristics of each subset to recover HSI by utilizing independent draws. For the c-th band-subset $X^c = \left[x_1^c, \ldots, x_i^c, \ldots, x_M^c\right]$, the hierarchical denoising model can be denoted as Equation (5):

$$X^c = D^c A^c + N^c + Q^c \circ S^c \tag{5}$$

where the symbol ∘ represents the element-wise multiplication. The first term on the right side of the Equation (5) represents the ideally noiseless estimation for X^c, which is represented as a linear combination of dictionary atoms. $D^c = [d_1^c, \ldots, d_i^c, \ldots, d_K^c] \in R^{P \times K}$ has columns that denote the dictionary atoms to be learned, where K is the number of dictionary atoms. In the context of HSI, the signatures of x_i^c have the highly spatial consistency with samples located at its neighboring region. To better learn the HSI [43], this prior knowledge is explicitly imposed to the dictionary atom d_k^c by using the Gaussian process (GP), which is constrained by the Gamma distribution. Additionally, there is a very high probability that the spectral pixels with similar characteristics in a local region share the same dictionary atoms, which is consistent with the cognition of vision and structure. $A^c = \left[a_1^c, \ldots, a_i^c, \ldots, a_M^c\right] \in R^{K \times M}$ places a sparse restriction on the vectors to remove noise, which can be written as $A^c = W^c \circ Z^c$. The matrix $W^c = \left[w_1^c, \ldots, w_i^c, \ldots, w_M^c\right]$, with the size of $K \times M$, represents the weight of matrix A^c, where w_i^c is drawn from Gaussian distribution. The sparseness of matrix A^c is adjusted by $Z^c = \left[z_1^c, \ldots, z_i^c, \ldots, z_M^c\right] \in \{0, 1\}^{K \times M}$, which is drawn from the Beta-Bernoulli process. Let symbol $\sim$ define an i.i.d. draw from a distribution, N represents the Normal distribution, I_K is

the $K \times K$ identity matrix, Bern and Beta refer to the Bernoulli and Beta distributions, respectively. Then the noise-free data $D^c A^c$ may be represented in the following manner:

$$\begin{aligned} d_k^c &\sim N\left(0, \Sigma\right) \qquad \Sigma_{jj'} = \xi_1 \exp\left(- \left\| x_j^c - x_{j'}^c \right\|_2^2 / \xi_2\right) \qquad \xi_1 \sim \Gamma\left(\beta, \rho\right) \\ w_i^c &\sim N\left(0, I_K / \gamma_w^c\right) \quad z_{ki}^c \sim Bern\left(\pi_k^c\right) \quad \pi_k^c \sim Beta\left(a_0 / K, b_0\left(K-1\right)/K\right) \end{aligned} \tag{6}$$

where ξ_1 is drawn from the Gamma distribution and ξ_1 is the predefined constant. Both of them are reapplied to manifest the smoothness between x_j^c and $x_{j'}^c$. If $||x_j^c - x_{j'}^c||_2^2$ is small, and the corresponding components between x_j^c and $x_{j'}^c$ are regarded to have a great spatial consistency. According to this, both the spectral correlation and spatial consistency for the c-th band-subset X^c can be exploited simultaneously. z_{ki}^c denotes whether the atom d_k^c is utilized to represent x_i^c with probability π_k^c or not. When $K \rightarrow \infty$, the expectation $E(\pi_k^c) = K^{-1}a_0 / [K^{-1}a_0 + K^{-1}(K-1)b_0]$ is approximately zero. According to this, most elements in the set $\{z_{ki}^c\}_{k=1,\ldots,K}$ are equal to zero, and the sparsity is reasonably imposed to the vector a_i^c. Obviously, a_i^c consists of a few non-zero values and numerous zero values. In this sense, the constraint constructed by the Beta-Bernoulli process can be regarded as a special type of l_0 norm. Notice that, noiseless estimation for each x_j^c uses a specific subset of dictionary atoms $\{d_k^c\}_{k=1,\ldots,K}$, which is specified by the sparse vector $\{z_{ki}^c\}_{k=1,\ldots,K}$. The expression $\{z_{ki}^c\} = 1$ indicates that the atom d_k^c is employed for representing coefficient a_{ki}^c. By analyzing the number and the locations non-zero elements in the set $\{z_{ki}^c\}_{k=1,\ldots,K}$, the size of the dictionary can be learned adaptively, including atom selection and prediction. The matrix $N^c = [n_1^c, \ldots, n_i^c, \ldots, n_M^c]$, with the size of $P \times M$, defines the zero-mean Gaussian noise. n_i^c is drawn from the zero-mean Gaussian noise components, with standard variance γ_n, and it can be displayed by the expressions in Equation (7). Additionally, Poisson noise is a dependent-signal noise, which means the pixels with higher light can be very strongly interrupted. To remove the dependency of the noise variance, the variance stability transform (VST) is introduced to convert Poisson noise into Gaussian noise before implementing a denoising approach [14,44]. After recovery, a corresponding inverse transformation is utilized to obtain the final HSI restoration results.

$$n_i^c \sim N\left(0, I_P / \gamma_n^c\right) \tag{7}$$

The third term denotes the sparse noise, such as dead pixel lines, which exploits the Beta-Bernoulli process to manifest the sparseness as displayed in Equation (8). The matrix $Q^c = [q_1^c, \ldots, q_i^c, \ldots, q_M^c]$, with the size of $P \times M$, represents the intensity of the sparse noise, and $S^c = [s_1^c, \ldots, s_i^c, \ldots, s_M^c] \in R^{P \times M}$ depicts the location information where the sparse noise may exist.

$$q_i^c \sim N\left(0, I_P / \gamma_v^c\right) \quad s_{pi}^c \sim Bernoulli\left(\theta_{pi}^c\right) \quad \theta_{pi}^c \sim Beta\left(a_\theta, b_\theta\right) \tag{8}$$

When $s_{pi}^c = 1$, the m-th element of x_i^c is polluted by sparse noise with the amplitude q_{pi}^c. The expectation $E(\theta_i^c) = a_\theta / (a_\theta + b_\theta)$ could be close to zero by adjusting the shape parameters of the Beta distribution (a_0 and b_0). Each element of s_i^c is i.i.d. drawn from the Beta distribution separately, which depicts well the arbitrariness of positions in the sparse noise.

γ_w^c, γ_n^c and γ_v^c are interpretable as non-informative hyper-parameters, which can regulate the precision of w_i^c, n_i^c and q_i^c. To flexibly solve the model with the posterior PDF, Gamma distribution is developed for these hyper-parameters as shown in Equation (9):

$$\gamma_w^c \sim \Gamma\left(c, d\right) \quad \gamma_n^c \sim \Gamma\left(e, f\right) \quad \gamma_v^c \sim \Gamma\left(g, h\right) \tag{9}$$

The negative logarithm posterior density function of the above model (utilized jointly to all data $X^c = \{x_i^c\}_{i=1,\ldots,M}$) may be represented as Equation (10):

$$
\begin{aligned}
-\log p &\left(\left\{ D^c, \{w_i^c\}, \{z_i^c\}, \{q_i^c\}, \{s_i^c\}, \{\pi_k^c\}, \{\theta_p^c\}, \xi_1^c, \gamma_w^c, \gamma_n^c, \gamma_v^c \right\} \mid \{x_i^c\} \right) = \\
&\frac{\gamma_n}{2} \sum_i \|\, x_i^c - D^c(w_i^c \circ z_i^c) - q_i^c \circ s_i^c \,\|_2^2 + \frac{1}{2} \sum_k (d_k^c)^{\mathrm{T}} \Sigma^{-1} d_k^c \\
&+ \frac{\gamma_w}{2} \sum_{i,k} \left(w_{ik}^c \right)^2 + \sum_k Beta(\pi_k^c \,|\, a_0/K, b_0\,(K-1)/K) + \sum_{i,k} \log Bernoulli \left(z_{ik}^c \,\big|\, \pi_k^c \right) \\
&+ \frac{\gamma_v}{2} \sum_{p,i} \left(q_{pi}^c \right)^2 + \sum_p Beta(\theta_p^c \,|\, a_\theta, b_\theta\,) + \sum_{p,i} \log Bernoulli \left(s_{pi}^c \,\big|\, \theta_{pi}^c \right) \\
&+ \log Gamma(\xi_1^c \,|\, \beta, \rho\,) + \log Gamma(\gamma_w^c \,|\, c, d\,) \\
&+ \log Gamma(\gamma_n^c \,|\, e, f\,) + \log Gamma(\gamma_v^c \,|\, g, h\,) + const
\end{aligned}
\tag{10}
$$

According to Equation (10), the full posterior of all parameters can be obtained rather than a point approximation, which is commonly applied by maximum a posteriori (MAP). All necessary parameters for solving the model can be learned by the prior and hyper-prior distribution, so the proposed hierarchical sparse learning framework is more robust and accurate. A graphical representation of the complete model is shown in Figure 3.
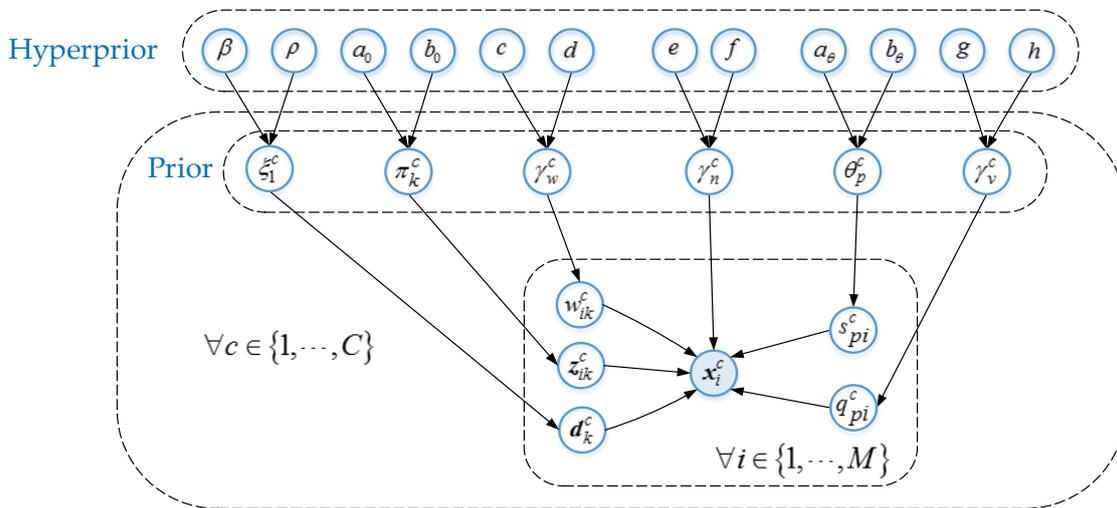

Figure 3. Graphical representation of the hierarchical sparse learning model.

Gibbs sampling is then implemented to solve the hierarchical model by employing the conjugacy of the model posterior distributions [45,46]. The update equations for each random variable are drawn from the conditional distributions on the most recent values of all other ones in the model. The details of the update equations can be found in the Appendix. After performing the Gibbs sampling, D^c and A^c can be obtained based on $\{d_k^c\}$, $\{w_i^c\}$ and $\{z_i^c\}$. Then restored images for the c-th band-subset $X^c = [x_1^c, \ldots, x_i^c, \ldots, x_M^c]$ can be inferred by calculating $D^c A^c$. Additionally, as there are several solutions for the same spectral pixel due to the usage of overlapping cubic patches, the final restored result is constructed by averaging all overlapping cubic patches. For all band-subsets $X = \{X^1, \ldots, X^c, \ldots, X^C\}$, $c = 1, \ldots, C$, the whole HSI after denoising can be obtained by sequentially performing this operation.

4. Experimental Results and Discussion

To evaluate the performance of the proposed approach, six state-of-the-art denoising methods were selected for a comparison based on both simulated and real data. The compared methods include K-SVD [9], BM3D [10], ANLM3D [13], BM4D [19], LRMR [31] and DDL3+FT [34]. The necessary parameters in the K-SVD, BM3D, ANLM3D and DDL3+FT methods, were finely tuned or selected automatically to generate the optimal experimental results as the reference suggested. In BM4D,

the noise variance is chosen from the set {0.01, 0.03, 0.04, 0.05, 0.07, 0.09, 1.1}. In LRMR, the rank of the noiseless matrix is selected from {4, 5, 6, 7}, and the cardinality of the sparse term is chosen from the set {0, 500, 1000, 1500, 2000, 3000, 4000, 5000}. For the proposed method, K is set to a relatively small value to reduce computational time, and here we choose $K = 128$. The parameters of the hyper-priors in the Gaussian process are set to $\beta = \rho = 10^{-6}$ and $\xi_2 = 200$. The iteration number of Gibbs sampling is set to 100. For setting the remaining parameters of the hyper-priors, two cases are taken into consideration: (1) The HSI is contaminated by Gaussian noise, dead pixel lines or a mixture of them. In this case the parameters are set as follows: $a_0 = b_0 = c = d = 10^{-6}$; $e = f = 10^{-6}$; $a_\theta = b_\theta = g = h = 10^{-6}$; (2) The HSI is contaminated by a mixture of Gaussian and Poisson noise or a mixture of Gaussian noise, Poisson noise and dead pixel lines. The parameters are set as follows: $a_0 = b_0 = 10^{-4}$; $c = d = 10^{-5}$; $e = f = 10^{-6}$; $a_\theta = b_\theta = 10^{-4}$; $g = h = 10^{-5}$. Once the types of noises are selected, the parameters of the hyper-priors are then set as described above and have no need to be tuned. By using these hyper parameters, the proposed framework can efficiently learn the latent information and suppress the various noises for the input HSI by sampling the infinite prior space. For one Gibbs sampling iteration, the computational complexity of the proposed method is approximate to $O(K(P + M) + PM)$. It should be pointed out that the proposed framework consumes much higher time than the six compared algorithms.

4.1. Experiment on Simulated Data

An image of Pavia University, acquired by the reflective optics system imaging spectrometer (ROSIS) over Pavia, northern Italy, is used for our experiments on simulated data. It consists of 103 spectral bands with a size of 610×340 for each band andcontains theentire spectral resolution reflectance characteristics collected from 430 nm to 860 nm in 10 nm steps. Before the simulated experiments, the gray values of the Pavia University data are scaled to the interval [0, 1].

For the sake of comparison, two kinds of evaluations are presented: (1) The visual appearances are shown to qualitatively estimate the experimental results before and after denoising, including spatial images at selected bands and spectral signatures of a few pixels; (2) The two commonly used metrics, peak signal to noise ratio (PSNR) and feature similarity index measure (FSIM) are provided to make a quantitative assessment on the denoising results. PSNR measures gray-level similarity between the restored image and reference image according to MSE. FSIM is designed to estimate the simulated result by integrating the gradient magnitude feature with the phase congruency feature from the view of human's perceptual consistency [17,47]. Better denoising performance is indicated by higher PSNR and FSIM values.

To estimate the effectiveness of the proposed approach, three types of noise are considered for the Pavia data: (1) zero-mean Gaussian noise; (2) Poisson noise. It is parameterized by the expression $X_{Poisson} = X \times peak$, where $X_{Poisson}$ refers to the noisy HSI corrupted by Poisson noise; X is the reference image; *peak* denotes the Poisson noise intensity; (3) Dead pixel lines. They are added to the same position of the selected bands in HSI, and their width varies from one line to three lines. In the simulated experiments, the noises are added for the Pavia data as the following three cases:

Case 1: The noise standard variance σ for each band of the HSI is randomly selected from 0.02 to 0.15. Figure 4 displays the PSNR and FSIM values of each band before and after recovery, which is as the quantitative evaluation for Pavia data. In Figure 4, the noisy HSI curves show big fluctuations with the spectral bands, which is caused by the changing σ with the spectral bands. Therefore, the noise information is very important to substantially improve the denoising performance. By adaptively predicting the noise and fully exploring the spectral-spatial information, the proposed method obtains the higher values of PSNR and FSIM than its competitors in most bands. KSVD and BM3D restore the HSI with the predefined fixed noise variance and do not learn σ in the simulated process. Meanwhile, they are implemented in the noisy HSI band by band and ignore the strongly spectral correlations. KSVD and BM3D show the lower values in both Figure 4a,b. ANLM3D and BM4D suppress the noise by exploiting both spatial and spectral information, which yields better simulated results compared with KSVD and BM3D. ANLM3D presents quite unstable performance, as shown in Figure 3. LRMR

takes advantage of the low-rank property in HSI, and it presents a better FSIM value by retaining well the latent features of HSI. By exploring the hierarchical deep learning and fine-tuning, DDL3+FT shows comparable PSNR values as shown in Figure 4a. LRMR and DDL3+FT transform the HSI into a 2D matrix during the recovery process, and both of them cannot effectively exploit the spatial consistencies of HSI. Obviously, the plots obtained by the suggested approach have a more stable trend than the six compared approaches. This demonstrates the effectiveness and robustness of the proposed method on reducing zero-mean Gaussian noise with the σ varying across bands.
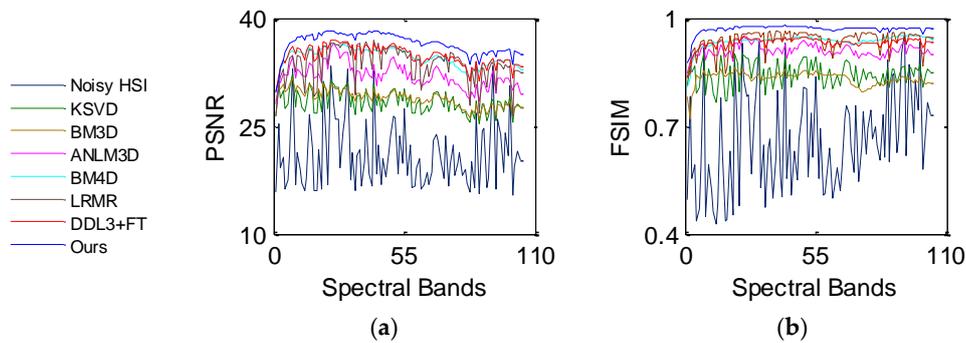
Figure 4. Quantitative evaluation results for Pavia data: (**a**) PSNR; (**b**) FSIM.

In terms of visual comparison, Figure 5 shows the denoised results of band 101 calculated by different approaches. It can be clearly seen that the proposed method works well in suppressing the noise and preserving the detailed local structure.
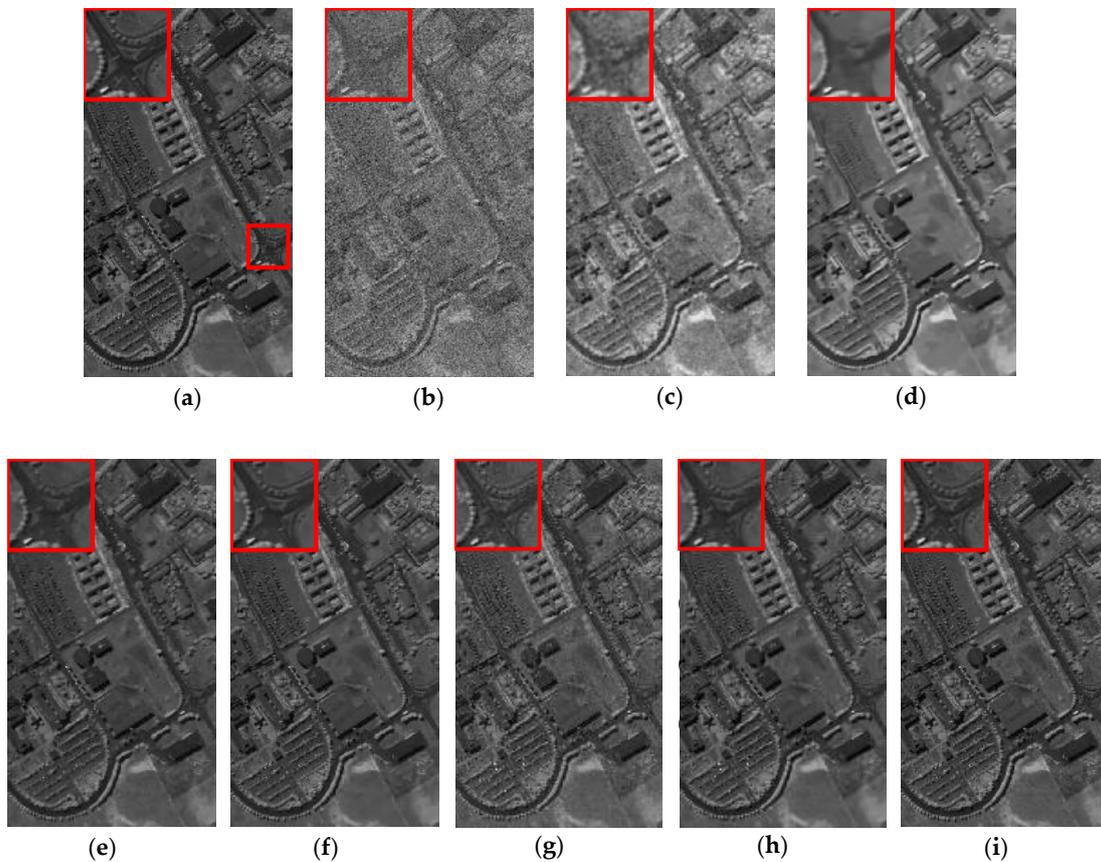
Figure 5. Restored images of band 101 corrupted with Gaussian noise: (**a**) clean HSI; (**b**) noisy HSI; (**c**) KSVD; (**d**) BM3D; (**e**) ANLM3D; (**f**) BM4D; (**g**) LRMR; (**h**) DDL3+FT; (**i**) Our method.

This is further depicted by the amplified area in the restored images of all competing methods. KSVD shows very poor performance and loses useful structural information. BM3D smooths out some important objects, which also works worse on the recovery. ANLM3D can effectively utilize the high nonlocal self-similarity to balance between smoothing and structural preservation, but it still fails to recover the profiles of local targets. The denoised result obtained by BM4D and DDL3+FT lose some fine objects. LRMR can obtain comparable results to the proposed method, but the Gaussian noise is not well reduced as shown in Figure 5g. Clearly, these visual assessment results are totally consistent with the above numerical evaluations.

Case 2: Mixed noise consisting of zero-mean Gaussian noise with standard variance $\sigma = 0.2$ and dead pixel lines. For Case 2, the restored results at band 95 calculated by different methods, including KSVD, BM3D, ANLM3D, BM4D, LRMR, DDL3+FT and the proposed approach, are shown in Figure 6. One area of interest is amplified in the recovery images of all methods. As shown in Figure 6a–h, we can find that the result obtained by the suggested method is much closer to the clean HSI in visual performance. The dead pixel lines in Figure 6c–h are still obvious, which means that the six compared methods fail to suppress the dead pixel lines. What's more, according to the amplified area it can be easily observed that our restoration method can effectively restore the homogenous region while preserving the edges. KSVD can partly ameliorate the noisy image quality, but it destroys the structure of sparse coefficients in dictionary learning process, which results in the loss of edges and other structural details as shown in Figure 6c. BM3D reduces the noises by utilizing the statistics of the similar neighboring patches and achieves a much better visual impression than KSVD, but it smooths out the structural information, and its result is blurred.
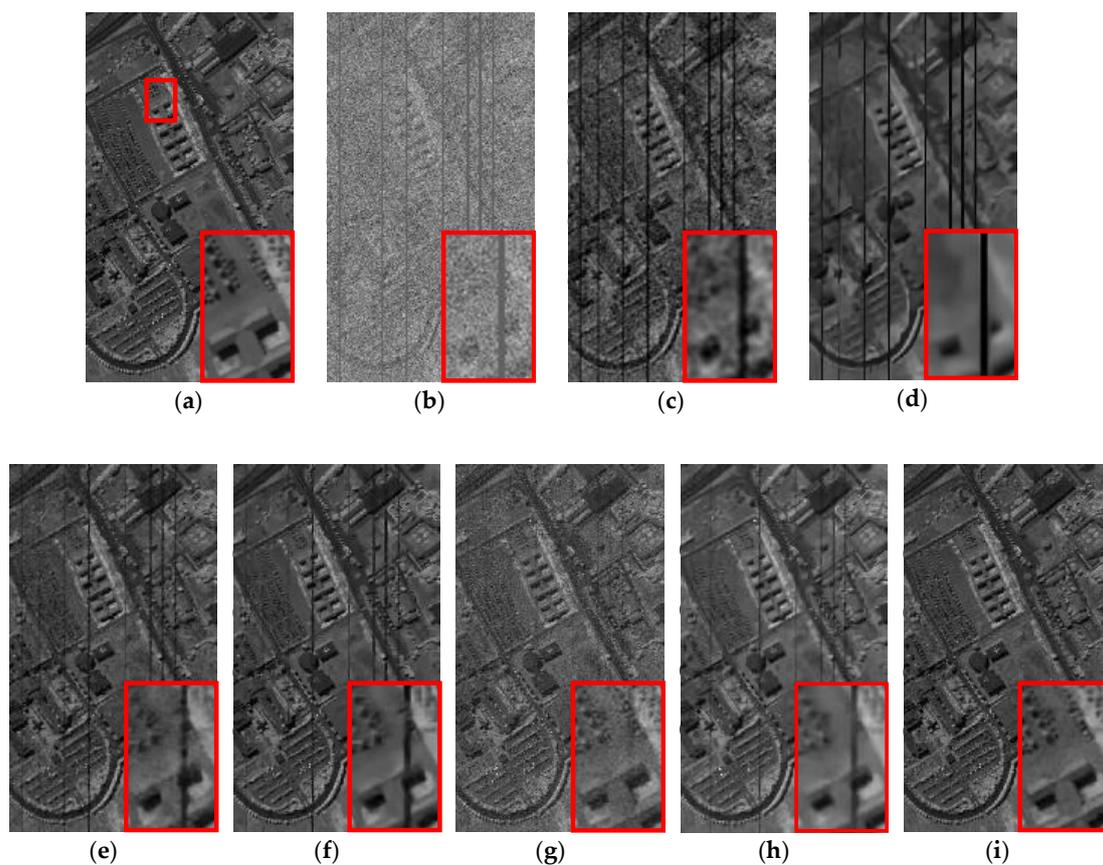

Figure 6. Restored results of band 95 corrupted with a mixture of Gaussian noise and dead pixel lines: (**a**) clean HSI; (**b**) noisy HSI; (**c**) KSVD; (**d**) BM3D; (**e**) ANLM3D; (**f**) BM4D; (**g**) LRMR; (**h**) DDL3+FT; (**i**) Our method.

ANLM3D can effectively utilize the high nonlocal self-similarity to achieve better balance smoothing and structural preserving. It fails however to preserve the edges and the details. BM4D can achieve visual improvements by utilizing the 3D nonlocal self-similarity of data cube, however, it removes some of the fine objects and over smooths the HSI. As shown in Figure 6g, the blurred black lines mean that LRMR only removes part of the dead pixels lines; meanwhile, the blurry white dots indicate that LRMR cannot efficiently remove the heavy Gaussian noise. Clearly, the DDL3+FT method fails to restore the mixed noisy image, which can be seen from the blurry edges and obvious dead pixel lines as shown in Figure 6h.

Figures 7 and 8 present the vertical and horizontal profiles of band 95 at pixel (559, 150) in the simulated experiment of Case 2, respectively. Clearly, the results are visually different in both shape and amplitude, where rapid fluctuations are introduced by the existence of dead pixel lines.
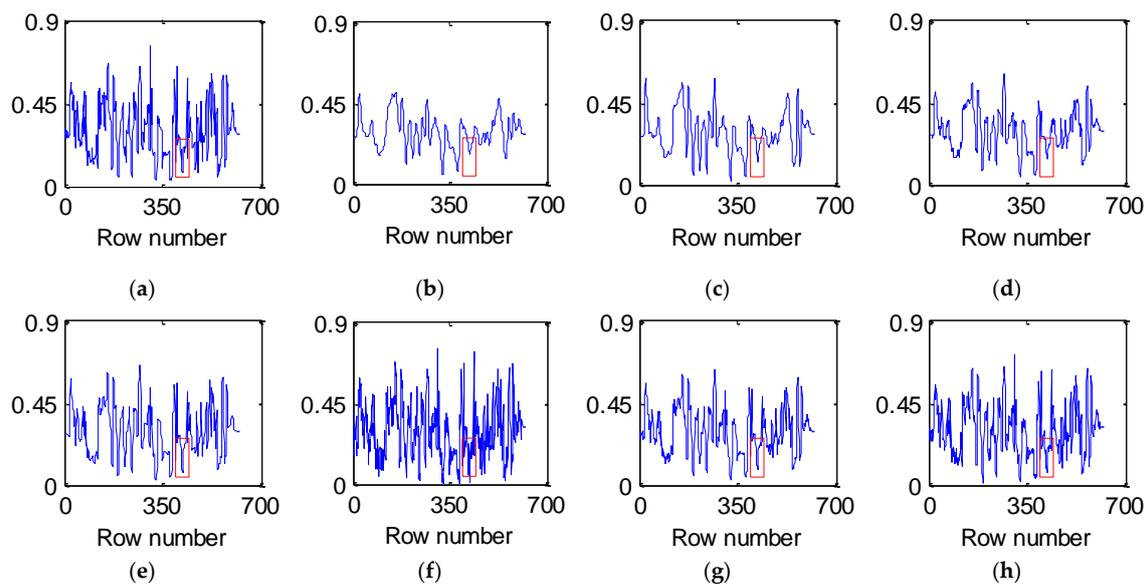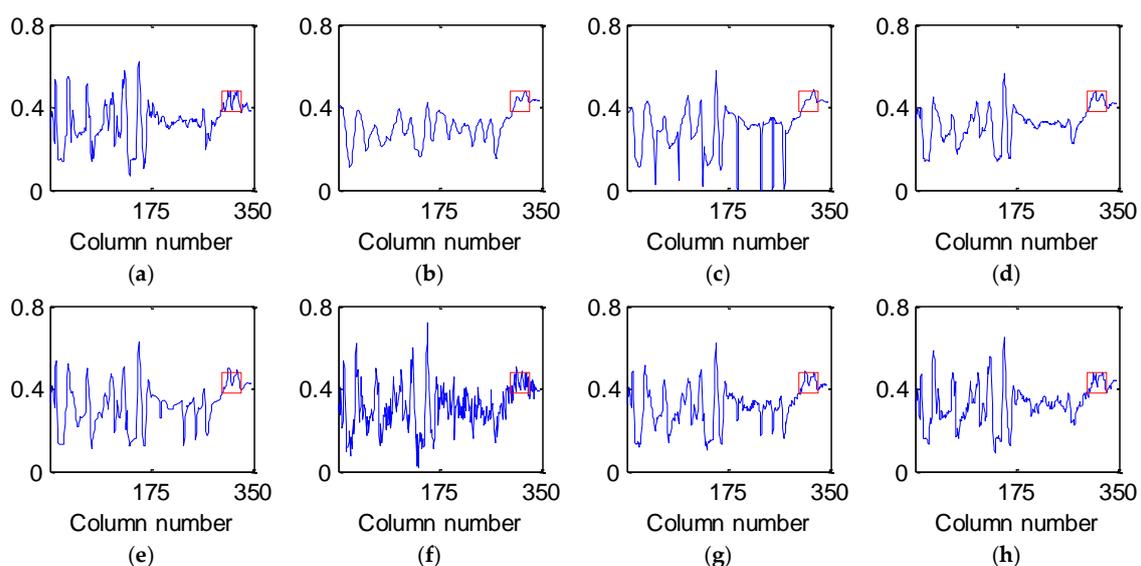

Figure 7. Vertical profiles of band 95 at pixel (559, 150) before and after denoising: (**a**) clean HSI; (**b**) KSVD; (**c**) BM3D; (**d**) ANLM3D; (**e**) BM4D; (**f**) LRMR; (**g**) DDL3+FT; (**h**) Our method.

Figure 8. Horizontal profiles of band 95 at pixel (559, 150) before and after denoising: (**a**) clean HSI; (**b**) KSVD; (**c**) BM3D; (**d**) ANLM3D; (**e**) BM4D; (**f**) LRMR; (**g**) DDL3+FT; (**h**) Our method.

From Figures 7 and 8, it can be observed that the profiles produced by the proposed method are closest to those of the original HSI and give the best removal of Gaussian noise and dead pixel lines. The curves in Figures 7b–g and 8b–g are not ideally consistent with those in Figures 7a and 8a and therefore lead to the limited ability for denoising HSI, which greatly supports the analysis above.

Spectral signatures of the clean image and the restored images at pixel (559, 150) are displayed in Figure 9, which are very critical for the classification and identification of the HSI. KSVD and BM3D denoise the HSI band by band and destroy the spectral–spatial correlation. As shown in Figure 9b,c, there are recognizable artifacts in the restored signatures obtained by KSVD and BM3D. Using ANLM3D and BM4D, the restored signatures are much nearer to the initial spectral reflectance curve due to the exploitation of the spectral–spatial information, but they appear to fluctuate strongly compared with the clean spectral signature. The restored signatures LRMR and DDL3+FT have the similar trends and shapes, but the detailed information cannot be well remained. Obviously, spectral signature calculated by the proposed method is the optimal, which demonstrates the superiority of the proposed method in reducing the Gaussian noise and dead pixel lines.
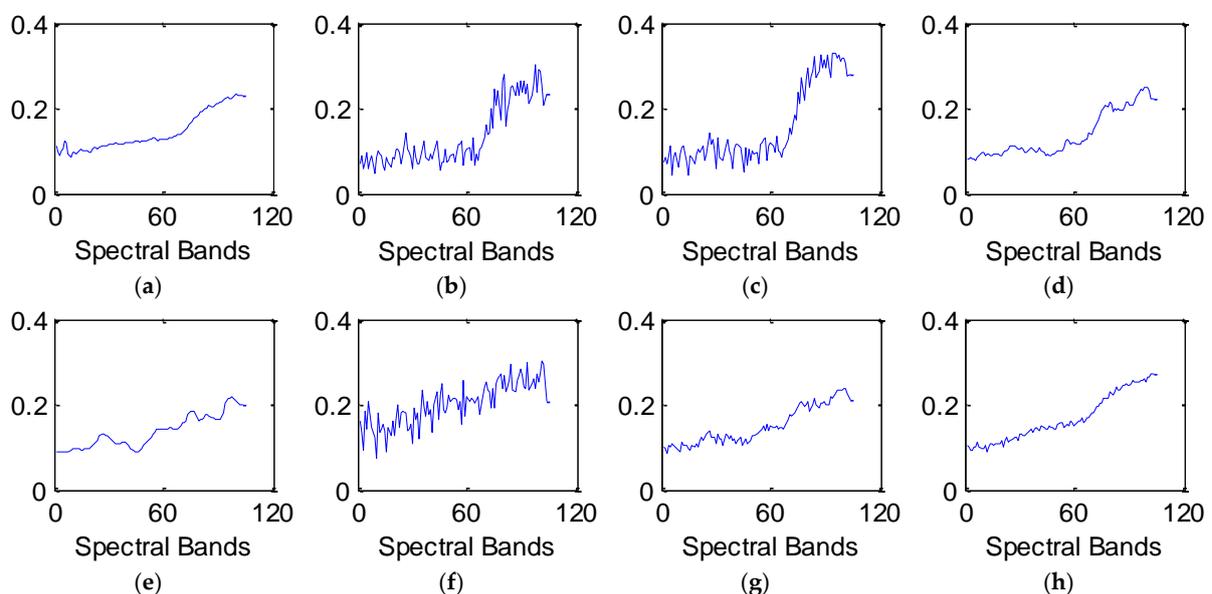

Figure 9. Spectral reflectance curves at pixel (559, 150) before and after denoising: (**a**) clean HSI; (**b**) KSVD; (**c**) BM3D; (**d**) ANLM3D; (**e**) BM4D; (**f**) LRMR; (**g**) DDL3+FT; (**h**) Our method.

Case 3: Mixed noise consisting of zero-mean Gaussian noise, Poisson noise and dead pixel lines, with σ = 0.15 and *peak* = 30. For Case 3, band 90 of the initial HSI and denoised results are presented in Figure 10. One area of all listed images is magnified for making a clear comparison. KSVD shows poor performance as shown in Figure 10c, and BM3D is oversmoothing and loses some useful objects. Also, there are obvious dead pixel lines in Figure 10c,d, which indicates that the KSVD and BM3D methods fail to remove the dead pixel lines. The ANLM3D and BM4D algorithms can only reduce part of the dead pixel lines, as presented in Figure 10d,f. Both of them do not work well in preserving the fine objects. The restored image of LRMR still has dead pixels lines and some Gaussian noise. From Figure 10h, it can be found that DDL3+FT fails in suppressing the dead pixels lines. As presented in Figure 10i, our approach can effectively remove the Gaussian noise, Poisson noise and dead pixel lines while preserving the local details such as edges and textures. Obviously, it performs the best compared to the six compared methods.

Figures 11 and 12 display the vertical and horizontal profiles of band 90 at pixel (399, 290) in simulated experiment of Case 3, separately. The spectral reflectance curves of all competing approaches at location (399, 290) before and after denoising are shown in Figure 13.
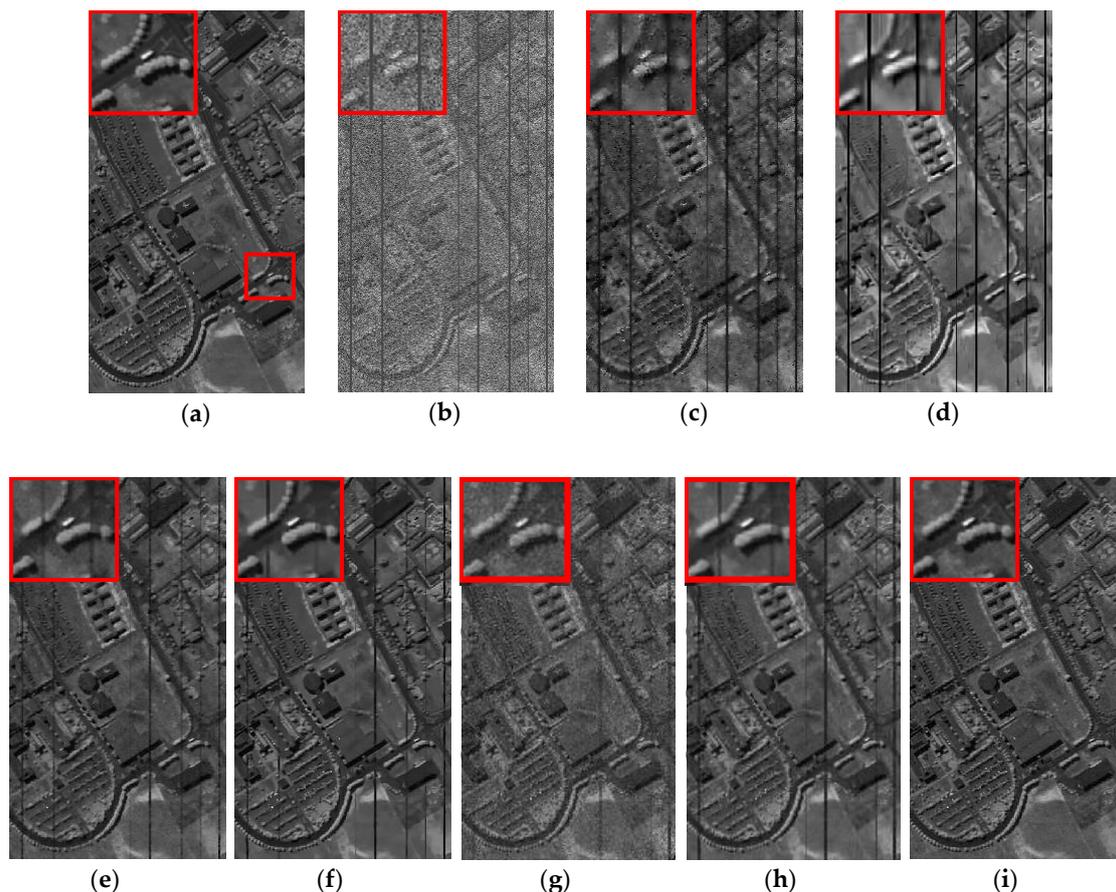
Figure 10. Restored results of band 90 corrupted with a mixture of Gaussian noise, Poisson noise and dead pixel lines:(**a**) clean HSI; (**b**) noisy HSI; (**c**) KSVD; (**d**) BM3D; (**e**) ANLM3D; (**f**) BM4D; (**g**) LRMR; (**h**) DDL3+FT; (**i**) Our method.
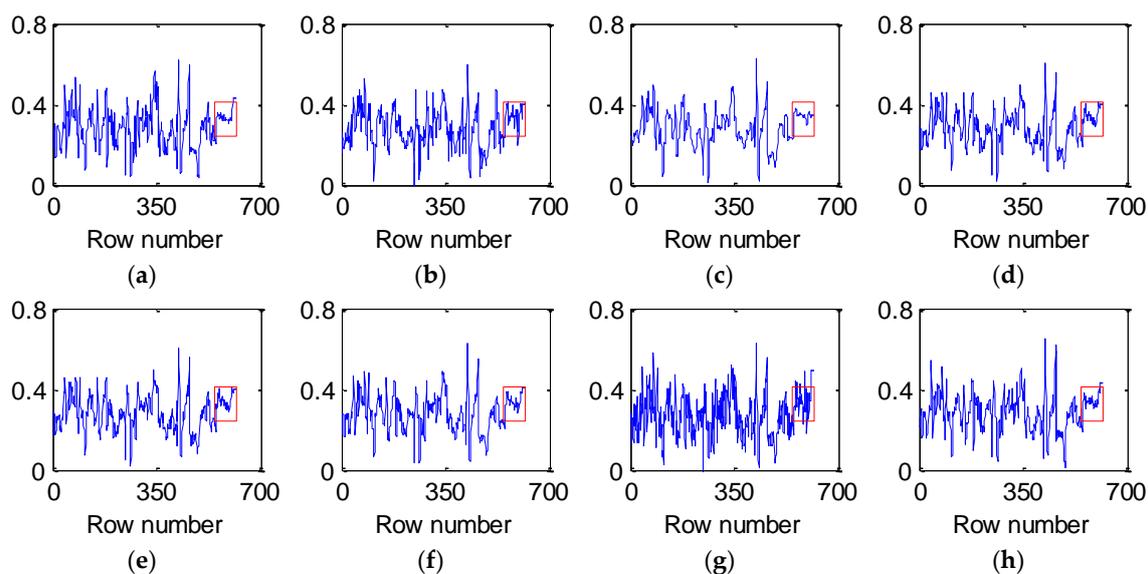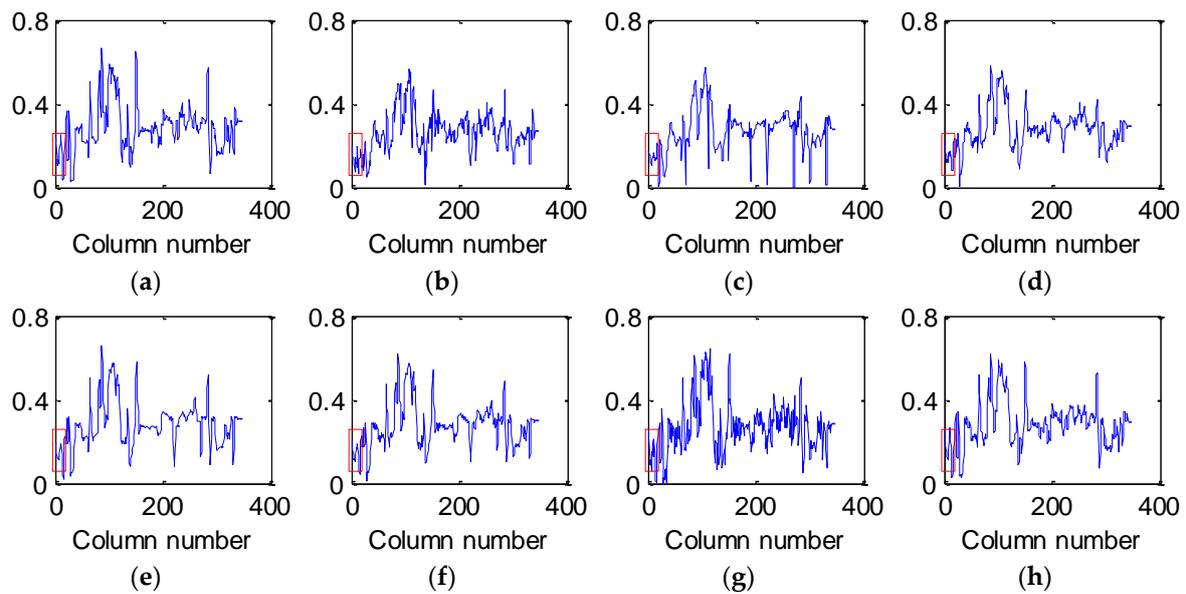

Figure 11. Vertical profiles of band 90 at pixel (399, 290) before and after denoising: (**a**) clean HSI; (**b**) KSVD; (**c**) BM3D; (**d**) ANLM3D; (**e**) BM4D; (**f**) LRMR; (**g**) DDL3+FT; (**h**) Our method.

Figure 12. Horizontal profiles of band 90 at pixel (399, 290) before and after denoising:(**a**) clean HSI; (**b**) KSVD; (**c**) BM3D; (**d**) ANLM3D; (**e**) BM4D; (**f**) LRMR; (**g**) DDL3+FT; (**h**) Our method.
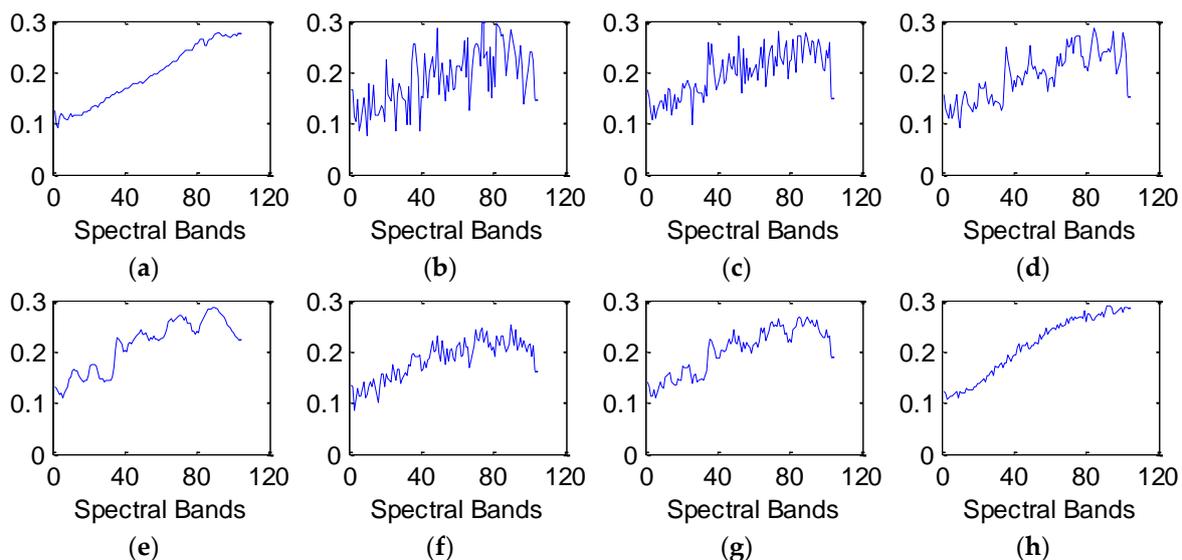

Figure 13. Spectral reflectance curves before and after denoising at pixel (399, 290): (**a**) clean HSI; (**b**) KSVD; (**c**) BM3D; (**d**) ANLM3D; (**e**) BM4D; (**f**) LRMR; (**g**) DDL3+FT; (**h**) Our method.

The rapid fluctuations of results in Figure 10 are caused by the dead pixel lines. A visual comparison is made based on the difference in the shape and amplitude of Figures 11–13. KSVD is inferior to suppress the mixed noise and preserve the spectral information, which can be clearly seen in Figures 11b, 12b and 13b. According to Figures 11c, 12c and 13c, BM3D is over smoothing while preserving the structure and introduces some recovery artifacts. From Figures 11d, 12d and 13d, it can be found that ANLM3D can partly reduce the mixed noise, which can be easily seen by the reduction of rapid fluctuations as shown in Figure 12d. As shown in Figures 11e,f, 12e,f and 13e,f, the curves obtained by BM4D and LRMR are much near to the initial curves compared to KSVD, BM3D, ANLM3D and DDL3+FT, but they fail in the detail preservation. By comparing the region marked by red rectangle, it is easily obtained that our method achieves the best approximation to the intrinsic patterns of the clean HSI, which is fully consistent with the analysis above.

4.2. Experiment on Real Data

Two well-known real data sets are adopted for evaluating the proposed method, including Urban data and Indian Pines data. The visual impressions are presented to qualitatively estimate the experimental results before and after restoration. For real HSI, there is no reference image to implement the numerical evaluation, e.g., PSNR and FSIM. Hence, the classification accuracies on Indian Pines data are utilized to quantitatively estimate the denoising performance.

4.2.1. Denoising for Urban Data

Urban data, with the original size of $307 \times 307 \times 210$, is acquired by the HYDICE sensor. Due to the detector-to-detector difference, it has different intensity strips and mixed noises versus bands. After removing the bands 104–108, 139–151, and 207–210 of Urban data, polluted by the atmosphere and water absorption, a subset with the size of $150 \times 150 \times 188$ is used in the following experiments.

Figure 14 presents the recovered images of band 186 obtained by different methods. To facilitate the visual comparison, the yellow arrows are utilized to mark the obvious stripes in Figure 14. Meanwhile, Figure 15 gives the enlarged details in the red rectangle of Figure 14. KSVD shows poor performance on the reduction of stripes and the preservation of structures as shown in Figures 14b and 15b. Obviously, BM3D and BM4D smooth out the important texture and fine targets, and both of them fail in removing the stripes. According to Figures 14d,g and 15d,g, it can be found that ANLM3D and DDL3+FT can partly suppress the stripes, and DDL3+FT has a better ability for preserving the texture and edge information than ANLM3D. LRMR shows better recovery performance than the other five compared methods on the noise reduction and structure preservation. From Figure 14f,h, it can be clearly observed that LRMR is inferior to the proposed method in removing the stripes. Generally, our proposed restoration method can efficiently restore the Urban data and convincingly outperform the six compared denoising approaches.
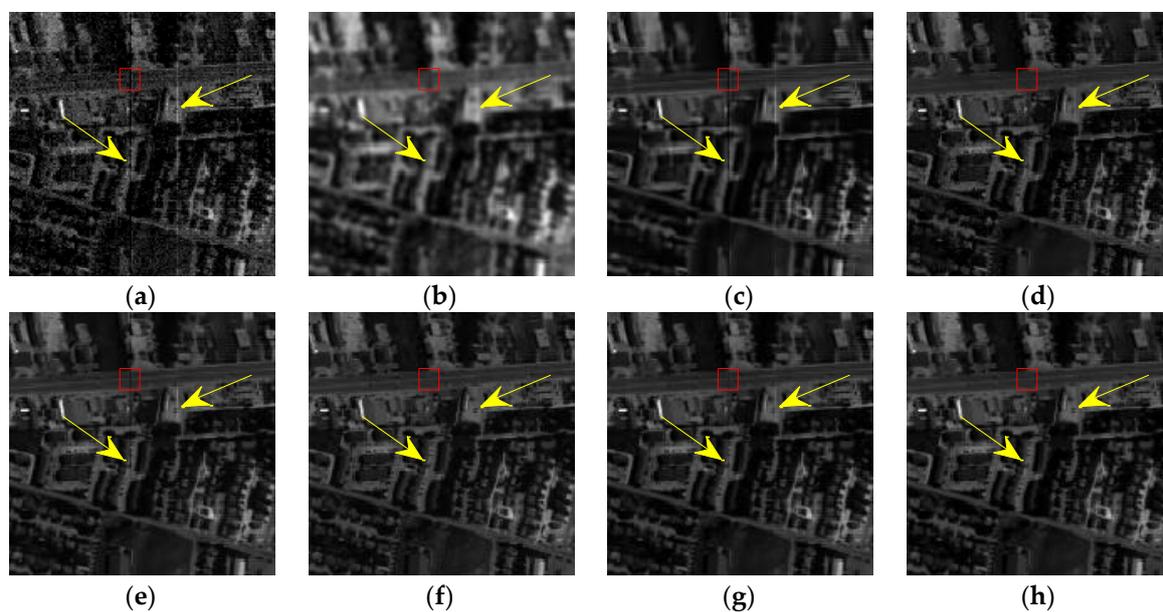

Figure 14. Restored results in Urban image: (**a**) Original band 186; (**b**) KSVD; (**c**) BM3D; (**d**) ANLM3D; (**e**) BM4D; (**f**) LRMR; (**g**) DDL3+FT; (**h**) Our method.

In addition, Figure 16 presents the results at band 104 before and after denoising. As shown in Figure 16a, there are many heavy stripes within the initial band 104. The false-color images of Urban data before and after restoration are presented in Figure 17, which consist of the *1*st, *104*th and *135*th bands. KSVD updates the dictionary atoms one by one, and destroys the structure of sparse

coefficients. As shown in Figures 16b and 17b, it blurs the image structures and shows very poor performance. By exploiting the block matching and 3D collaborative filter, BM3D greatly enhances the image quality of the initial ones, but there are still obvious stripes in Figures 16c and 17c. According to the regions marked by the red ellipse, it can be found that ANLM3D, BM4D, LRMR and DDL3+FT can partly remove the stripes, and LRMR works worse in effectively restoring the images polluted by serious noises. Obviously, the proposed method shows the superior performance on denoising the flat region with the better edge preservation. Especially, by compared the regions marked by a blue rectangle in Figure 16, it can be easily found that the suggested method can restore well the targets and remain the edges while efficiently reducing the stripes and mixed noises. Above all, the restored images obtained by the proposed method present the best visual result than KSVD, BM3D, ANLM3D, BM4D, LRMR and DDL3+FT, with the satisfying detail preservation.
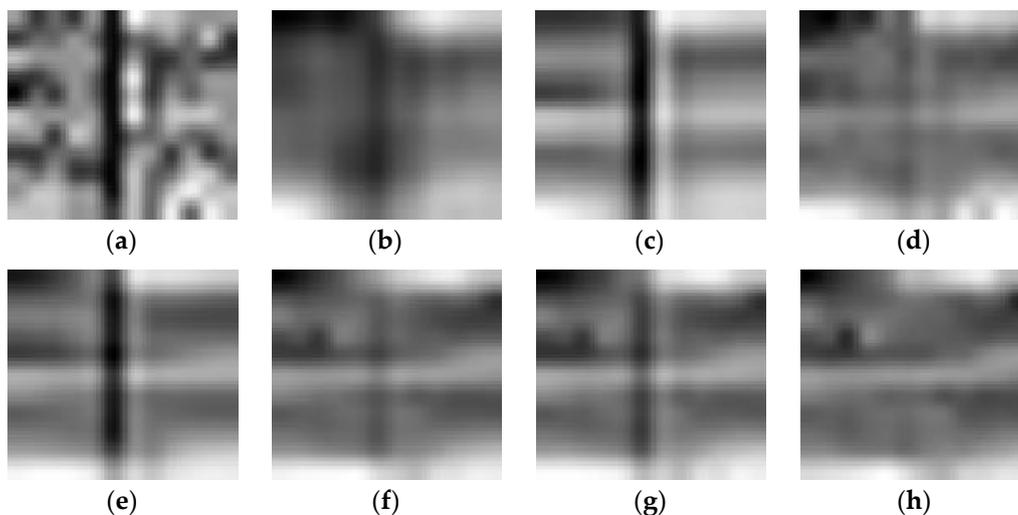

Figure 15. ×4 magnified results of the various approaches in the red rectangle of Figure 14: (**a**) Original band 186; (**b**) KSVD; (**c**) BM3D; (**d**) ANLM3D; (**e**) BM4D; (**f**) LRMR; (**g**) DDL3+FT; (**h**) Our method.
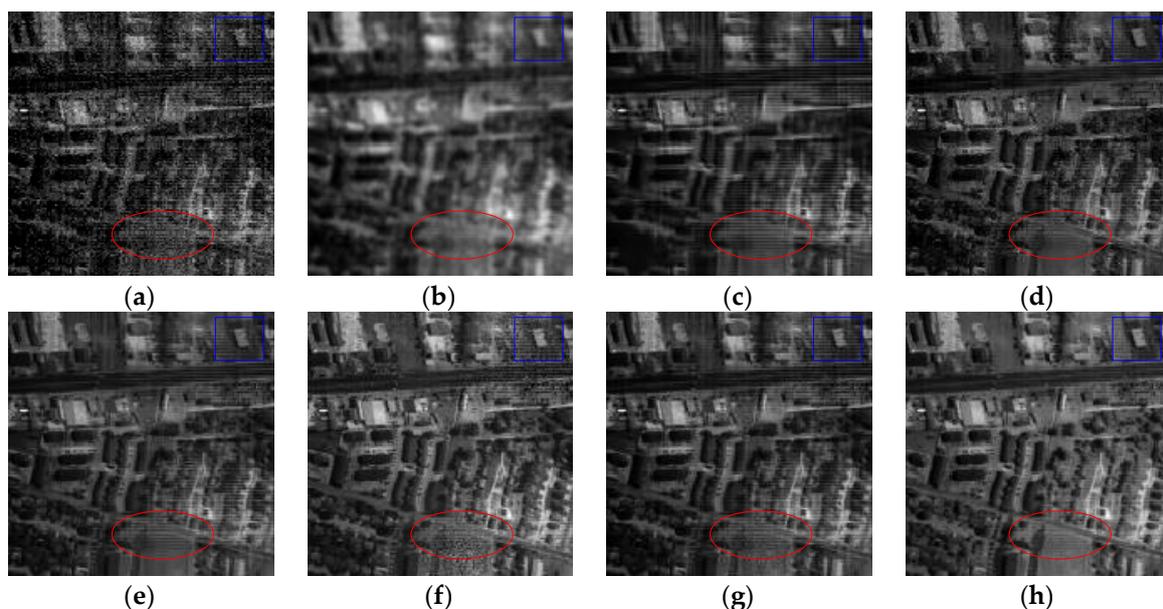
Figure 16. Restored results in Urban image: (**a**) Original band 104; (**b**) KSVD; (**c**) BM3D; (**d**) ANLM3D; (**e**) BM4D; (**f**) LRMR; (**g**) DDL3+FT; (**h**) Our method.
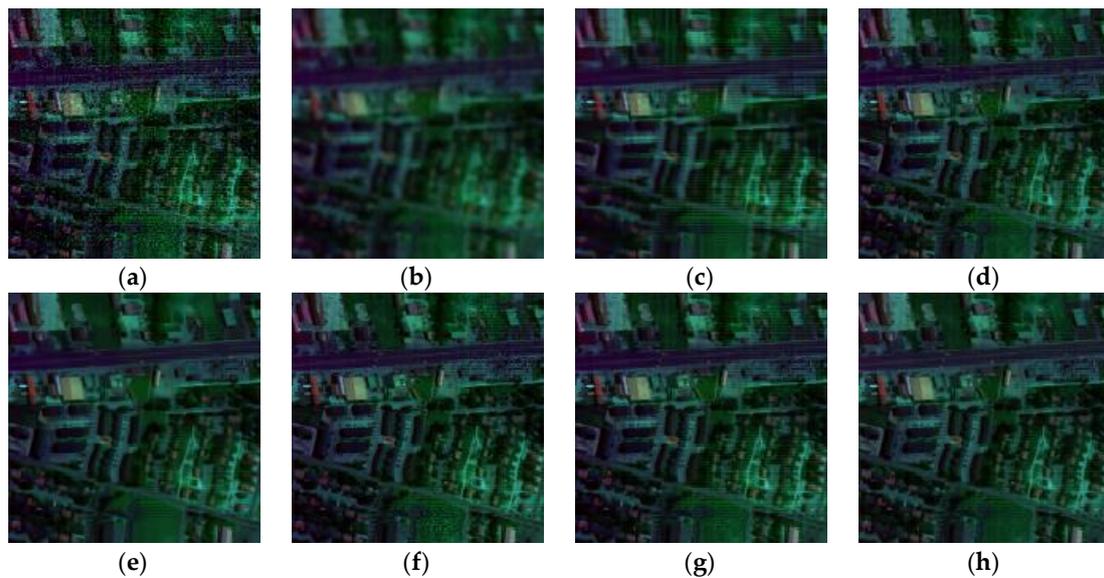
Figure 17. Restored results of Urban image: (**a**) original false-color image (R: 1, G: 104, and B: 135); (**b**) KSVD; (**c**) BM3D; (**d**) ANLM3D; (**e**) BM4D; (**f**) LRMR; (**g**) DDL3+FT; (**h**) Our method.

4.2.2. Experimental Results on Indian Pines Data

The second data set is named Indian Pines, which is recorded by the NASA AVIRIS sensor over the Indian Pines region in 1992, and this dataset contains much random noise in some bands during the acquiring process. It comprises 220 spectral bands and the spatial dimension of each spectral band is 145 × 145 pixels. For Indian Pines data, the ground truth has 16 land cover classes and a total of 10,366 labeled pixels.

Figures 18–20 display the visual comparisons of different bands polluted by different noises. Obviously, the proposed method achieves better visual quality than the compared ones. From Figures 18b, 19b and 20b, KSVD shows poorer denoising capability than other methods. BM3D significantly blurs the images and loses the texture information and edges.
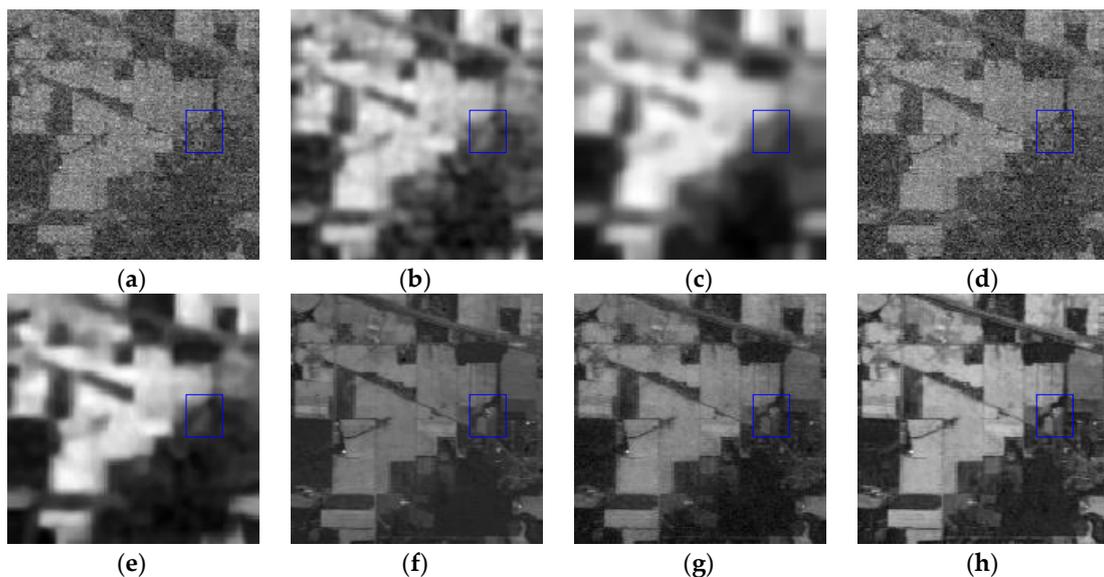
Figure 18. Restored results in Indian Pines image: (**a**) Original band 164; (**b**) KSVD; (**c**) BM3D; (**d**) ANLM3D; (**e**) BM4D; (**f**) LRMR; (**g**) DDL3+FT; (**h**) Our method.
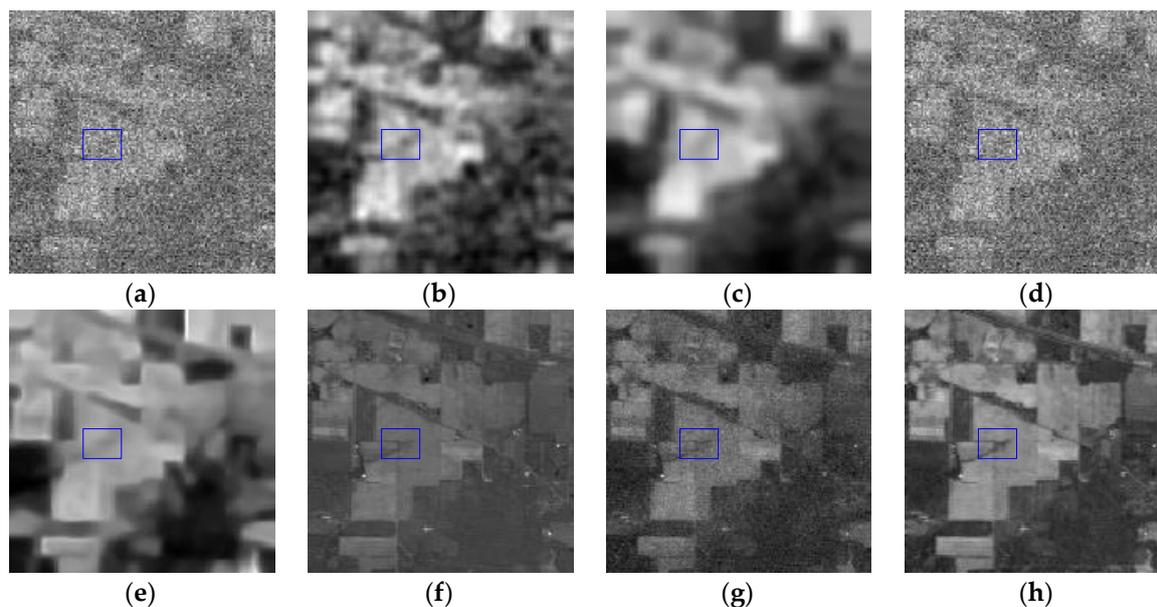
Figure 19. Restored results in Indian Pines image: (**a**) Original band 220; (**b**) KSVD; (**c**) BM3D; (**d**) ANLM3D; (**e**) BM4D; (**f**) LRMR; (**g**) DDL3+FT; (**h**) Our method.
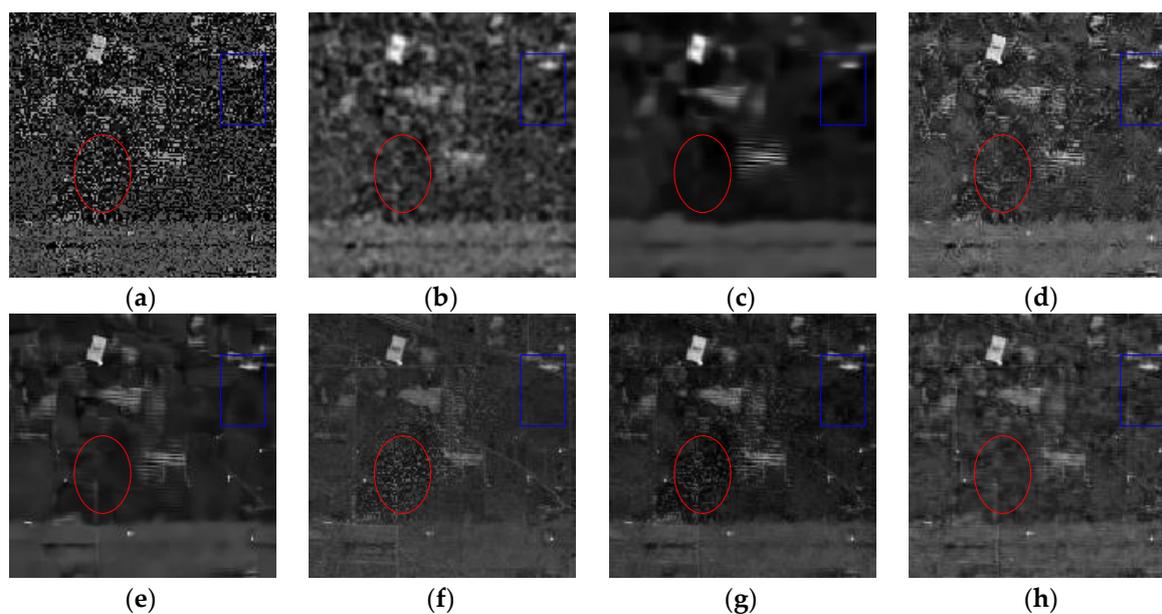
Figure 20. Restored results in Indian Pines image: (**a**) Original band 1; (**b**) KSVD; (**c**) BM3D; (**d**) ANLM3D; (**e**) BM4D; (**f**) LRMR; (**g**) DDL3+FT; (**h**) Our method.

As shown in Figures 18d and 19d, the image quality improvements obtained by ANLM3D appear very small and can be neglected. BM4D is oversmoothing and loses texture information. By observing the regions marked by the blue rectangle in Figures 18–20, our denoising method has the better ability on the edge and structure preservation than LRMR and DDL3+FT; meanwhile, LRMR and DDL3+FT work worse in efficiently removing the random noises as shown in the regions marked by the red ellipse in Figures 18–20. According to Figures 18–20, our algorithm is much more superior than the six compared methods for seriously corrupted images, which is consistent with the above analysis, so our algorithm can perform best in the removal of random noises from the Indian Pines data, while effectively improving the quality of the noisy HSI and retoring the texture and structure details.

For classification-based evaluation, two cases are taken into consideration according to the testing data: (1) classifying the 20 heavily corrupted bands of Indian Pines data, including bands 104–108, 150–163 and 220; (2) classifying the Indian Pines data with 20 heavily corrupted bands removed. Similar with the setting in [48], the number of training samples for small classes "alfalfa", "grass/pasture mowed", and "oats" is set as 15 samples per class, and the number for the remaining classes is set to 50. Support vector machine (SVM) is utilized as the classification method. As usual, the commonly used overall accuracy (OA) and the kappa coefficient are selected as evaluation metrics and the map of the results is used for visual estimation.

Table 1 lists the overall accuracy (OA) and kappa coefficient (κ) of the results for the heavily corrupted bands. Table 2 shows the OA and kappa coefficient of the results for the remaining bands. After achieving the recovery for testing data, the values of OA and kappa coefficient are obviously enhanced as shown in Tables 1 and 2, which demonstrate the necessity of HSI denoising before implementing the classification. Compared with other algorithms, the proposed method obtains the best OA and kappa coefficients in both Tables 1 and 2, which means that our denoising method can greatly restore the structure information (which is essential for classification) in the seriously polluted bands or the remaining 200 bands. We note that ANLM3D method obtains an OA value of 25.17% and κ of 0.2135 in Table 1, which are just 9.43% and 0.1223 higher than initial HSI. This lower classification accuracy and kappa coefficient are totally in line with the poor denoising performance as displayed in Figure 19d.

Table 1. SVM classification accuracies on heavily corrupted bands (104–108, 150–163, 220) of the Indian Pines data.

	Initial HSI	KSVD	BM3D	ANLM3D	BM4D	LRMR	DDL3+FT	Ours
OA	15.74%	57.96%	81.2%	25.17%	69.27%	48.38%	49.12%	83.76%
κ	0.0912	0.5368	0.7803	0.2135	0.679	0.3664	0.4284	0.8109

Table 2. SVM classification accuracies on the Indian Pines data (bands 104–108, 150–163, 220 removed).

	Initial HSI	KSVD	BM3D	ANLM3D	BM4D	LRMR	DDL3+FT	Ours
OA	74.39%	87.96%	0.9215%	81.06%	87.59%	87.18%	85.92%	90.35%
κ	0.7183	0.8531	0.8673	0.775	0.8568	0.8548	0.8442	0.8726

The classification maps of different algorithms are displayed in Figure 21, where the first row is the results of the 20 heavily corrupted bands and the second row presents the results of the remaining 200 bands before and after restoration. According to Figure 21, it can be easily observed that the result of suggested method presents the better visual effect than the six compared algorithms.
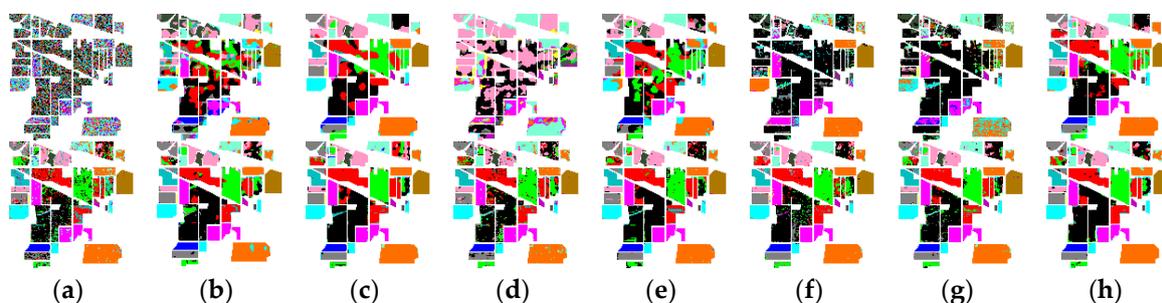

(**a**)   (**b**)   (**c**)   (**d**)   (**e**)   (**f**)   (**g**)   (**h**)

Figure 21. Classification results for Indian Pines data before and after denoising: the first row is the results of the 20 heavily corrupted bands; the second row presents the results of the remaining 200 bands: (**a**) Original HSI; (**b**) KSVD; (**c**) BM3D; (**d**) ANLM3D; (**e**) BM4D; (**f**) LRMR; (**g**) DDL3+FT; (**h**) Our method.

4.3. Discussion

4.3.1. Threshold Parameter η

The value of the parameter η in Equation (4) is selected from 0 to 1. Quantitative evaluations of different η values were implemented by comparing the OA metric on Indian Pines data. Noticing that for the data consisting of the 20 heavily polluted bands, the SSIM values of adjacent bands are very similar with each other as shown in Figure 2c, we just analyzed the remaining 200 bands here. Figure 22 presents the relative SVM classification accuracies. It can be seen that the highest classification accuracies occur in $\eta \in (0.4,0.5)$ and the lowest values is in $\eta \in (0,0.1)$. In particular, the case when $\eta = 0$ means that the segmentation points are selected in line with the local dropping points, which result in the absence of sample information for some band-subsets and decrease the classification accuracy. When $\eta = 1$, there is no need to divide the bands. Observing the curve in Figure 22, it can be found that the appropriate η may well enhance the classification accuracies, which also demonstrates the necessary for the band-subset segmentation.
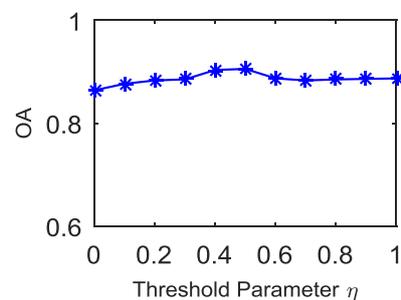

Figure 22. SVM classification accuracies of Indian Pines data (104–108, 150–163, 220 removed).

The number of the band subsets, *C*, is only dominated by the parameter η. For the HSI with *C* band subsets, its computational complexity is approximately equal to $\mathrm{O}(\sum_{c=1}^{C} \left(l_x l_y l_c \left(K_c + M \right) + (K_c M) \right)$ for one iteration. The values of l_x, l_y and M are fixed in our experiments. Hence, the computational complexity of the proposed method is determined by the parameters l_c and K_c. To simplify the analysis, we consider the special case that the number of dictionary atoms, K_c, is set as the same value K' for all band-subsets. Then the computational complexity can be rewritten as $\mathrm{O}\left(l_x l_y \left(K' + M \right) \left(l_1 + \ldots + l_C \right) + C K' M \right)$. It is noted that the value of the expression $(l_1 + \ldots + l_C)$ is the number of bands in HSI and *M* is much larger than K', so the computational complexity is mainly dominated by the expression $K'M$. Due to the strong spectral correlations, *C* is usually smaller than 10, as shown in Figure 2. When *C* changes from 1 to 10, the empirical value of K' shows a rapid decrease, and the relative value of $CK'M$ shows a downwards trend, which implies a lower computational complexity.

4.3.2. The Sparse Noise Term

To evaluate the effectiveness of the sparse noise term, we performed a comprehensive comparison between the restored images obtained by the proposed method with the sparse noise term disabled and enabled. The Urban dataset is applied for the simulation, and the method with disabled sparse noise term is named as SNT-DIS. Figure 23 shows the visual impression of band 133 obtained by SNT-DIS and the proposed method. Compared with the corrupted image in Figure 23a, it can be easily found that both algorithms can greatly improve the image quality as displayed in Figure 23b,c, but the SNT-DIS can only reduce a small part of the stripes, as presented in Figure 23b, and it also fails in preserving the fine objects. As presented in Figure 23c, the suggested method can effectively remove heavy stripes and mixed noises while preserving the local details such as edges and textures.

The obvious superiority of the proposed method demonstrates that the denoising performance can be greatly improved by introducing the sparse noise term.
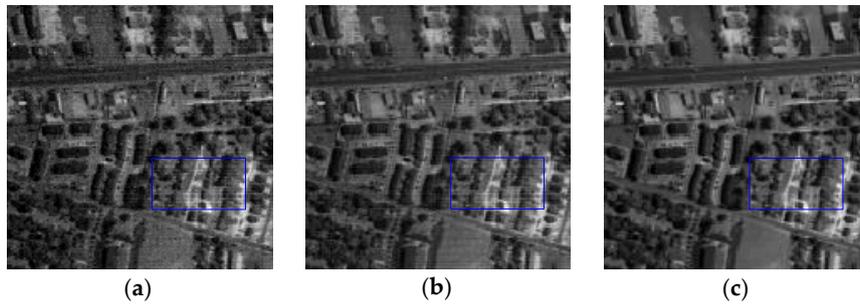

(**a**)	(**b**)	(**c**)

Figure 23. Restored results in Urban data: (**a**) Original band 133; (**b**) SNT-DIS; (**c**) Our method.

5. Conclusions

In this paper, a novel hierarchical framework combining structural correlations in spectral domain and sparse learning is developed for HSI denoising. First, with the suggested band-subset partition, spectral-spatial cubic patches can be effectively extracted, where each patch has strong structure correlations between adjacent bands and high spatial similarities within the cubic patch. Second, noise-free estimations for each band-subset are generated by solving a sparse problem constructed by a Beta-Bernoulli process. The Gaussian process with Gamma distribution is regarded as the precursor of dictionary elements to fully exploit the spatial consistency in HSI. By utilizing the spectral-spatial information and Gibbs sampling, the suggesting framework can effectively learn the latent detail and structure information in HSI. It has the advantage of adaptively predicting the noise in a data-driven manner, which enables the framework to better depict the existing noises in HSI. Meanwhile, it can automatically infer the size and the atoms of dictionary. Compared with KSVD, BM3D, ANLM3D, BM4D, LRMR and DDL3+FT, the proposed framework achieves superior performance both numerically and visually, which has been verified on the simulated and real HSI. It can suppress various noises in HSI, such as Gaussian noise, Poisson noise, dead pixel lines, stripes or a mixture of them, with better image quality improvement and structure characteristics preservation. Further research will be directed toward exploiting the appropriately hidden structure and reducing the computational time for the analysis of the HSI.

Author Contributions: Shuai Liu is the corresponding author who theoretically proposed the original method and wrote this paper; Licheng Jiao and Shuyuan Yang gave some valuable suggestions.

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

Appendix A

Supposing that symbol—represents "the remaining variables except the one being sampled". The update equations for the *c*-th band-subset are listed asfollows:

Sampling d_k^c:

$$p\left(d_k^c \mid -\right) \sim N\left(d_k^c \mid \mu_{d_k^c}, \Omega_{d_k^c}\right) \tag{A1}$$

where the covariance $\Omega_{d_k^c}$ and mean $\mu_{d_k^c}$ are expressed as:

$$\Omega_{d_k^c} = \left(\Sigma^{-1} + \gamma_n^c \sum_i \left(w_{ik}^c\right)^2 \left(z_{ik}^c\right)^2\right)^{-1} \tag{A2}$$

$$\boldsymbol{\mu}_{\boldsymbol{d}_k^c} = \gamma_n^c \boldsymbol{\Omega}_{\boldsymbol{d}_k^c} \sum_i w_{ik}^c z_{ik}^c \boldsymbol{x}_{(i,-k)}^c \tag{A3}$$

where $\boldsymbol{x}_{(i,-k)}^c = \boldsymbol{x}_i^c - \boldsymbol{D}^c(\boldsymbol{w}_i^c \odot \boldsymbol{z}_i^c) - \boldsymbol{q}_i^c \odot \boldsymbol{s}_i^c + (w_{ik}^c \odot z_{ik}^c)\boldsymbol{d}_k^c$.

Sampling z_{ik}^c:

$$p\left(z_{ik}^c \mid -\right) \sim \mathrm{Bern}(\frac{p_1}{p_1 + p_0}) \tag{A4}$$

The posterior probability of $z_{ik}^c = 1$ is proportional to:

$$p_1 = \pi_k^c \exp(-\frac{\gamma_n^c}{2}((w_{ik}^c)^2 (\boldsymbol{d}_k^c)^{\mathrm{T}} \boldsymbol{d}_k^c - 2w_{ik}^c (\boldsymbol{d}_k^c)^{\mathrm{T}} \boldsymbol{x}_{(i,-k)}^c)) \tag{A5}$$

The posterior probability of $z_{ik}^c = 0$ is proportional to:

$$p_0 = 1 - \pi_k^c \tag{A6}$$

Sampling $\boldsymbol{w}_i^c$:

$$p\left(w_{ik}^c \mid -\right) \sim N(w_{ik}^c \mid \mu_{wik}^c, \Omega_{wik}^c) \tag{A7}$$

$$\Omega_{wik}^c = (\gamma_w^c + \gamma_n^c (z_{ik}^c)^2 (\boldsymbol{d}_k^c)^{\mathrm{T}} \boldsymbol{d}_k^c)^{-1} \tag{A8}$$

$$\mu_{wik}^c = \gamma_n^c \Omega_{wik}^c (\boldsymbol{d}_k^c)^{\mathrm{T}} \boldsymbol{x}_{(i,-k)}^c \tag{A9}$$

Sampling π_k^c:

$$p\left(\pi_k^c \mid -\right) \sim \mathrm{Beta}(a_0/K + \sum_i z_{ik}^c, b_0 (K-1)/K + M - \sum_i z_{ik}^c) \tag{A10}$$

Sampling ξ_1^c:

$$p\left(\xi_1^c \mid -\right) \sim \Gamma(\beta + PK/2, \rho + \sum_k (\boldsymbol{d}_k^c)^{\mathrm{T}} \Sigma^{-1} \boldsymbol{d}_k^c/2) \tag{A11}$$

Sampling γ_w^c:

$$p\left(\gamma_w^c \mid -\right) \sim \Gamma(c + MK/2, d + \sum_i (w_i^c)^{\mathrm{T}} w_i^c/2) \tag{A12}$$

Sampling γ_n^c:

$$p\left(\gamma_n^c \mid -\right) \sim \Gamma(e + PM/2, f + \sum_i \|x_i^c - \boldsymbol{D}^c(\boldsymbol{w}_i^c \circ \boldsymbol{z}_i^c) - \boldsymbol{q}_i^c \circ \boldsymbol{s}_i^c\|_2^2) \tag{A13}$$

Sampling s_{pi}^c:

$$p\left(s_{pi}^c \mid -\right) \sim \mathrm{Bern}(\frac{v_1}{v_1 + v_0}) \tag{A14}$$

$$v_1 = \theta_p^c \exp(-\frac{\gamma_v}{2}((q_{pi}^c)^2 - 2q_{pi}^c x_{(i,s)}^c)) \tag{A15}$$

$$v_0 = 1 - \theta_p^c \tag{A16}$$

where $x_{(i,s)}^c = x_i^c - \boldsymbol{D}^c(\boldsymbol{w}_i^c \circ \boldsymbol{z}_i^c)$. The posterior probability that $s_{pi}^c = 1$ is proportional to v_1, and the posterior probability that $s_{pi}^c = 0$ is proportional to v_0.

Sampling $\boldsymbol{q}_{pi}^c$:

$$p\left(q_{pi}^c \mid -\right) \sim N(\mu_{qpi}, \Omega_{qpi}) \tag{A17}$$

$$\Omega_{qpi} = \left(\gamma_v^c + \gamma_n^c \left(s_{pi}^c\right)^2\right)^{-1} \tag{A18}$$

$$\mu_{qpi} = \gamma_n^c \Omega_{qpi} s_{pi}^c x_{(i,s)}^c \tag{A19}$$

Sampling θ_p^c:

$$p\left(\theta_p^c\mid-\right)\sim\text{Beta}(a_\theta+\sum_i s_{pi}^c, b_\theta+M-\sum_i s_{pi}^c)\tag{A20}$$

Sampling γ_v^c:

$$p\left(\gamma_v^c\mid-\right)\sim\Gamma(g+PM/2, h+\sum_i (q_i^c)^{\mathrm{T}} q_i^c/2)\tag{A21}$$

References

1. Van der Meer, F.D.; van der Werff, H.M.; van Ruitenbeek, F.J.; Hecker, C.A.; Bakker, W.H.; Noomen, M.F.; van der Meijde, M.; Carranza, E.J.M.; de Smeth, J.B.; Woldai, T. Multi-and Hyperspectral geologic remote sensing: A review. *Int. J. Appl. Earth Obs. Geoinf.* **2012**, *14*, 112–128. [CrossRef]
2. Lu, G.; Fei, B. Medical hyperspectral imaging: A review. *J. Biomed. Opt.* **2014**, *19*, 010901. [CrossRef] [PubMed]
3. Gmur, S.; Vogt, D.; Zabowski, D.; Moskal, L.M. Hyperspectral analysis of soil nitrogen, carbon, carbonate, and organic matter using regression trees. *Sensors* **2012**, *12*, 10639–10658. [CrossRef] [PubMed]
4. Lam, A.; Sato, I.; Sato, Y. Denoising hyperspectral images using spectral domain statistics. In Proceedings of the IEEE International Conference on Pattern Recognition (ICPR'12), Tsukuba, Japan, 11–15 November 2012; pp. 477–480.
5. Fu, P.; Li, C.Y.; Xia, Y.; Ji, Z.X.; Sun, Q.S.; Cai, W.D.; Feng, D.D. Adaptive noise estimation from highly textured hyperspectral images. *Appl. Opt.* **2014**, *53*, 7059–7071. [CrossRef] [PubMed]
6. Chen, G.Y.; Qian, S.E. Denoising and dimensionality reduction of hyper-spectral imagery using wavelet packets, neighbour shrinking and principal component analysis. *Int. J. Remote Sens.* **2009**, *30*, 4889–4895. [CrossRef]
7. Beck, A.; Teboulle, M. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Trans. Image Process.* **2009**, *18*, 2419–2434. [CrossRef] [PubMed]
8. Buades, A.; Coll, B.; Morel, J.M. A non-local algorithm for image denoising. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 7–12 June 2005; pp. 60–65.
9. Aharon, M.; Elad, M.; Bruckstein, A. K-SVD: An algorithm for designing overcomplete dictionaries for Sparse Representation. *IEEE Trans. Signal Process.* **2006**, *54*, 4311–4322. [CrossRef]
10. Dabov, K.; Foi, A.; Karkovnik, V. Image denoising by sparse 3Dtransform-domain collaborative filtering. *IEEE Trans. Image Process.* **2007**, *16*, 2080–2094. [CrossRef] [PubMed]
11. Karami, A.; Yazdi, M.; Asli, A.Z. Noise reduction of hyperspectral images using kernel non-negative Tucker decomposition. *IEEEJ. Sel. Top. Signal Process.* **2011**, *5*, 487–493. [CrossRef]
12. Chen, G.Y.; Qian, S.E. Denoising of hyperspectral imagery using principal component analysis and wavelet shrinkage. *IEEE Trans. Geosci.Remote Sens.* **2011**, *49*, 973–980. [CrossRef]
13. Manjón, J.V.; Pierrick, C.; Luis, M.B.; Collins, D.L.; Robles, M. Adaptive non-local means denoising of MR images with spatially varying noise levels. *J. Magn. Reson. Imaging* **2010**, *31*, 192–203. [CrossRef] [PubMed]
14. Qian, Y.T.; Ye, M.C. Hyperspectral imagery restoration using nonlocal spectral-spatial structured sparse representation with noise estimation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 499–515. [CrossRef]
15. Aggarwal, H.K.; Majumdar, A. Hyperspectral Image Denoising Using Spatio-Spectral Total Variation. *IEEE Trans. Geosci. Remote Sens.* **2016**, *13*, 442–446. [CrossRef]
16. Yuan, Q.Q.; Zhang, L.P.; Shen, H.F. Hyperspectral image denoising employing a spectral–spatial adaptive total variation model. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 3660–3677. [CrossRef]
17. Li, J.; Yuan, Q.Q.; Shen, H.F.; Zhang, L.P. Hyperspectral image recovery employing a multidimensional nonlocal total variation model. *Signal Process.* **2015**, *111*, 230–446. [CrossRef]
18. Liao, D.P.; Ye, M.C.; Jia, S.; Qian, Y.T. Noise reduction of hyperspectral imagery based on nonlocal tensor factorization. In Proceedings of the 2013 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Melbourne, Australia, 21–26 July 2013; pp. 1083–1086.

19. Maggioni, M.; Katkovnik, V.; Egiazarian, K.; Foi, A. Nonlocal transform-domain filter for volumetric data denoising and reconstruction. *IEEE Trans. Image Process.* **2013**, *22*, 119–133. [CrossRef] [PubMed]

20. Xu, L.L.; Li, F.; Wong, A.; Clausi, D.A. Hyperspectral Image Denoising Using a Spatial–Spectral Monte Carlo Sampling Approach. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3025–3038. [CrossRef]

21. Zhong, P.; Wang, R.S. Jointly learning the hybrid CRF and MLR model for simultaneous denoising and classification of hyperspectral imagery. *IEEE Trans. Neural Netw. Learn. Syst.* **2014**, *25*, 1319–1334. [CrossRef]

22. Yuan, Q.Q.; Zhang, L.P.; Shen, H.F. Hyperspectral image denoising with a spatial–spectral view fusion strategy. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 2314–2325. [CrossRef]

23. Mairal, J.; Bach, F.; Ponce, J. Sparse Modeling for Image and Vision Processing. *Found. Trends Comput. Graph. Vis.* **2014**, *8*, 85–283. [CrossRef]

24. Wu, D.; Zhang, Y.; Chen, Y. 3D sparse coding based denoising of hyperspectral images. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium, Milan, Italy, 26–31 July 2015; pp. 3115–3118.

25. Xing, Z.; Zhou, M.; Castrodad, A.; Sapiro, G.; Carin, L. Dictionary learning for noisy and incomplete hyperspectral images. *SIAM J. Imaging Sci.* **2012**, *5*, 33–56. [CrossRef]

26. Li, J.; Yuan, Q.Q.; Shen, H.F.; Zhang, L.P. Noise Removal from Hyperspectral Image with Joint Spectral-Spatial Distributed Sparse Representation. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 5425–5439. [CrossRef]

27. Ye, M.C.; Qian, Y.T.; Zhou, J. Multitask sparse nonnegative matrix factorization for joint spectral–spatial hyperspectral imagery denoising. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2621–2639. [CrossRef]

28. Rasti, B.; Sveinsson, J.R.; Ulfarsson, M.O.; Benediktsson, J.A. Hyperspectral image denoising using a new linear model and Sparse Regularization. In Proceedings of the 2013 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Melbourne, Australia, 21–26 July 2013; pp. 457–460.

29. Yang, J.X.; Zhao, Y.Q.; Chan, J.C.W.; Kong, S.G. Coupled Sparse Denoising and Unmixing with Low-Rank Constraint for Hyperspectral Image. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1818–1833. [CrossRef]

30. Zhu, R.; Dong, M.Z.; Xue, J.H. Spectral nonlocal restoration of hyperspectral images with low-rank property. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3062–3067.

31. Zhang, H.Y.; He, W.; Zhang, L.P.; Shen, H.F.; Yuan, Q.Q. Hyperspectral image restoration using low-rank matrix recovery. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 4729–4743. [CrossRef]

32. He, W.; Zhang, H.Y.; Zhang, L.P.; Shen, H.F. Total-variation-regularized low-rank matrix factorization for hyperspectral image restoration. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 178–188. [CrossRef]

33. Makantasis, K.; Karantzalos, K.; Doulamis, A.; Doulamis, N. Deep supervised learning for hyperspectral data classification through convolutional neural networks. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 4959–4962.

34. Huo, L.; Feng, X.; Huo, C.; Pan, C. Learning deep dictionary for hyperspectral image denoising. *IEICE Trans. Inf. Syst.* **2015**, *7*, 1401–1404. [CrossRef]

35. Donoho, D.L. For most large underdetermined systems of linear equations the minimal -norm solution is also the sparsest solution. *Commun. Pure Appl. Math.* **2006**, *59*, 797–829. [CrossRef]

36. Mairal, J.; Bach, F.; Ponce, J. Task-driven dictionary learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 791–804. [CrossRef] [PubMed]

37. Shah, A.; David, K.; Ghahramani, Z.B. An empirical study of stochastic variational inference algorithms for the beta Bernoulli process. In Proceedings of the 32nd International Conference on Machine Learning (ICML'15), Lille, France, 6–11 July 2015.

38. Thiagarajan, J.J.; Ramamurthy, K.N.; Spanias, A. Learning stable multilevel dictionaries for sparse representations. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *26*, 1913–1926. [CrossRef] [PubMed]

39. Yan, R.M.; Shao, L.; Liu, Y. Nonlocal hierarchical dictionary learning using wavelets for image denoising. *IEEE Trans. Image Process.* **2013**, *22*, 4689–4698. [CrossRef] [PubMed]

40. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]

41. Papyan, V.; Elad, M. Multi-scale patch-based image restoration. *IEEE Trans. Image Process.* **2016**, *25*, 249–261. [CrossRef] [PubMed]

42. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2014.

43. Rasmussen, C.; Williams, C. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.

44. Makitalo, M.; Foi, A. Noise parameter mismatch in variance stabilization, with an application to Poisson-Gaussian noise estimation. *IEEE Trans. Image Process.* **2014**, *23*, 5349–5359. [CrossRef] [PubMed]

45. Casella, G.; George, E.I. Explaining the Gibbs sampler. *Am. Stat.* **1992**, *46*, 167–174.

46. Rodriguez, Y.G.; Davis, R.; Scharf, L. *Efficient Gibbs Sampling of Truncated Multivariate Normal with Application to Constrained Linear Regression*; Columbia University: New York, NY, USA, 2004.

47. Zhang, L.; Zhang, L.; Mou, X.Q.; Zhang, D. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Trans. Image Process.* **2011**, *20*, 2378–2386. [CrossRef] [PubMed]

48. Xu, Y.; Wu, Z.; Wei, Z. Spectral–Spatial Classification of Hyperspectral Image Based on Low-Rank Decomposition. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2370–2380. [CrossRef]