*remote sensing*

MDPI

*Article*

# A Deep Unfolding Network for Multispectral and Hyperspectral Image Fusion

**Bihui Zhang [1,2], Xiangyong Cao [2,3,*] and Deyu Meng [1,4,5]**

[1] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China; zhangbihui217@gmail.com (B.Z.); dymeng@mail.xjtu.edu.cn (D.M.)
[2] The Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China
[3] School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
[4] The Peng Cheng Laboratory, Shenzhen 518066, China
[5] The Macau Institute of Systems Engineering, Macau University of Science and Technology, Taipa, Macau 999078, China
* Correspondence: caoxiangyong@mail.xjtu.edu.cn; Tel.: +86-159-9167-3997

**Abstract:** Multispectral and hyperspectral image fusion (MS/HS fusion) aims to generate a high-resolution hyperspectral (HRHS) image by fusing a high-resolution multispectral (HRMS) and a low-resolution hyperspectral (LRHS) images. The deep unfolding-based MS/HS fusion method is a representative deep learning paradigm due to its excellent performance and sufficient interpretability. However, existing deep unfolding-based MS/HS fusion methods only rely on a fixed linear degradation model, which focuses on modeling the relationships between HRHS and HRMS, as well as HRHS and LRHS. In this paper, we break free from this observation model framework and propose a new observation model. Firstly, the proposed observation model is built based on the convolutional sparse coding (CSC) technique, and then a proximal gradient algorithm is designed to solve this model. Secondly, we unfold the iterative algorithm into a deep network, dubbed as *MHF-CSCNet*, where the proximal operators are learned using convolutional neural networks. Finally, all trainable parameters can be automatically learned end-to-end from the training pairs. Experimental evaluations conducted on various benchmark datasets demonstrate the superiority of our method both quantitatively and qualitatively compared to other state-of-the-art methods.

**Keywords:** hyperspectral and multispectral image fusion; convolutional sparse coding; proximal gradient algorithm; convolutional neural networks

check for updates

**Citation:** Zhang, B.; Cao, X.; Meng, D. A Deep Unfolding Network for Multispectral and Hyperspectral Image Fusion. *Remote Sens.* **2024**, *16*, 3979. https://doi.org/10.3390/rs16213979

Academic Editor: Salah Bourennane

Received: 11 July 2024
Revised: 13 October 2024
Accepted: 15 October 2024
Published: 26 October 2024

**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Hyperspectral imagery (HSI) with high spatial resolution has the potential to greatly enhance our understanding and capabilities across a wide range of applications, including the monitoring and management of natural resources, ecosystems, biodiversity, and disasters [1]. However, due to the inherent limitations of imaging sensors, there exists a trade-off between spatial resolution and spectral resolution. Consequently, directly obtaining an HSI with high spatial resolution is a challenging task. Therefore, multispectral and hyperspectral image fusion (MS/HS fusion), i.e., generating a high-resolution hyperspectral image (HRHS) by combining a high-resolution multispectral image (HRMS) and a low-resolution hyperspectral image (LRHS), is an important task for remote sensing image processing. In recent decades, a large number of MS/HS fusion methods have been developed [1–6], and these approaches can be roughly classified into four categories, i.e., component substitution (CS) approaches [7], multiresolution analysis (MRA) methods [8,9], subspace-based methods [10–17], and deep learning (DL) approaches.

The CS-based MS/HS fusion method mainly originates from the CS-based pansharpening approaches [7,18,19], e.g., intensity–hue–saturation (IHS) [19], principal component

analysis (PCA), and Gram–Schmidt adaptive (GSA) [7]. The general CS-based scheme sharpens the low-resolution images by multiplying the difference between a high-resolution image and a synthetic intensity component by a band-wise modulation coefficient. For example, in GSA, a representative CS-based pansharpening method obtains the intensity component by the Gram–Schmidt transform [18]. A direct approach to using GSA for MS/HS fusion is to create multiple sets of images for pansharpening subproblems. Each set consists of an MS band and its associated HS bands, which are then grouped together based on correlation-based clustering. The CS-based MS/HS fusion method can enhance the spatial details well, but spectral distortion may occur, especially when there is a significant difference in the spectral range between the MS and HS images [1,7].

The MRA-based approach differs from the CS-based approach in how spatial details are obtained from high-resolution images. In the conventional pansharpening procedure, the MRA method [8,9] sharpens the entire MSI by individually sharpening the characteristics of each MS band. Specifically, it supplements the spatial details of each interpolated MS band, which is then upsampled to the same dimensions as the Pan image. Two representative methods are smoothing filtered-based intensity modulation (SFIM) [8] and generalized Laplacian pyramid (GLP) [9]. When applying such methods to MS/HS fusion task, Selva et al. proposed a novel framework called hyper-sharpening [20], where the source of high spatial resolution details is no longer a specific MS band, but rather a high-resolution image synthesized using linear regression with the MS images as input. The SFIM hyper-sharpening method (SFIM-HS) is an extension of the traditional SFIM [8] method, where a synthesized image is generated by applying a linear regression of MS bands using the least square. Similar to SFIM, GLP [9] has also been applied to MS/HS fusion effectively, i.e., GLP-HS. Compared to CS-based methods, MRA-based approaches exhibit superior spectral quality but worse spatial sharpening effects.

The subspace-based methods can be subdivided mainly into two categories, i.e., unmixing [11–13,16,17,21,22] and Bayesian-based approaches [10,14,15]. Most subspace-based methods involve decomposing the HS and MS images to obtain spectral signals of an endmember and their corresponding fractional abundances. Constrained Nonnegative Matrix Factorization (CNMF) [11] alternately unmixes the HS and MS images by NMF [23] to estimate the spectral signatures of endmembers and the high-resolution abundance maps, respectively. Similarly to CNMF, the Lanaras algorithm [13] jointly unmixes the two input images into the spectral signatures of the endmembers and the associated fractional abundances. Akhtar et al. [12] employed dictionary learning and sparse coding techniques to obtain the endmembers and high-resolution abundances, respectively. To preserve edges and smooth out noise in homogeneous regions, HySure [10] introduced total variation regularization into the subspace-based HSI super-resolution framework. The CSTF method [16] formulated the MS/HS fusion task as a tensor factorization problem. Bayesian-based methods first model the MS/HS fusion problem as a probabilistic optimization problem, and then solve it using an iterative algorithm. The MAP-SMM [14] algorithm uses the stochastic mixing model (SMM) to estimate the underlying spectral scene statistics, as well as the local statistics of the spectral mixing model. FUSE [15] utilizes a Sylvester equation to solve the maximization problem of the likelihoods obtained from the forward observation models. However, the subspace-based methods, i.e., unmixing or Bayesian-based, always require assuming subjective prior information and high computational complexity.

In recent years, deep learning-based approaches [24–28] have been extensively studied for MS/HS fusion. This type of method directly learns the mapping from HRMS and LRHS to HRHS using some network architectures. For example, Han et al. [29] presented an MS/HS fusion method using cluster-based multi-branch BP neural networks. Dian et al. [30] proposed an MS/HS fusion method based on the subspace representation and convolutional neural network (CNN) denoiser. Dian et al. [31] designed an HSRNet by restoring spatial and spectral details using a spectral preservation module and a spatial preservation module, respectively. Palsson et al. [32] introduced the deep residual network for the MS/HS fusion task. There are also some other methods to address the challenge

that multimodal remote sensing images cannot be directly compared due to their modal heterogeneity. For example, Chen et al. [33] take advantage of two types of modality-independent structural relationships in multi-modal images. In particular, they present a structural relationship graph representation learning framework for measuring the similarity of the two structural relationships. Additionally, they also designed an adaptive fusion method [34] based on frequency-decoupling to effectively fuse the local and nonlocal structural difference maps. However, these deep learning-based methods are black-box and lack interpretation. To deal with the limitations of the black-box networks, researchers proposed various model-based deep learning methods by unfolding the solution of the variational model into CNN. Xie et al. [35] first proposed a deep unfolding MS/HS fusion network based on the linear observation model, i.e., there are linear relationships between HRHS and HRMS, as well as between HRMS and LRHS. Dong et al. [36] proposed an iterative Hyperspectral Image Super-Resolution (HSISR) algorithm based on a deep HSI denoiser to leverage both domain knowledge likelihood and deep image prior and then unfold the iterative HSISR algorithm into a novel model-guided deep convolutional network. Ma et al. [37] proposed a model-based deep learning network named unfolding spatiospectral super-resolution network (US3RN) to reformulate the image degradation and incorporate the spatiospectral super-resolution (SSSR) model, which takes the observation models of SISR and SSR into consideration and then solve the model-based energy function via the alternative direction multiplier method (ADMM) technique. Sun et al. [38] propose a new image fusion network that uses deep prior information as spatial guide information. The fusion model uses spatial information as a regularization term and is solved using half splitting quadratic method. These deep unfolding methods can obtain excellent performance and sufficient interpretability by relying on a fixed linear observation model. However, the fixed linear model only considers the interaction between HRHS, HRMS, and LRHS, while the characteristics of HRHS, HRMS, and LRHS are ignored. In this paper, we continue our research along the deep unfolding method and attempt to jump out of this observation model framework by considering the image features. Specifically, we first propose a new observation model based on the convolutional sparse coding (CSC) [39,40] technique and its corresponding proximal gradient algorithm. Secondly, we unfold the iterative algorithm into a deep network (MHF-CSCNet), where the proximal operators are learned using convolutional neural networks. Finally, all trainable parameters can be learned end-to-end from the training pairs.

In summary, the contributions of our work are three-fold:

- We propose a new observation model for the MS/HS fusion task using the convolutional sparse coding (CSC) technique. Specifically, we separately model the HS and MS images using CSC, incorporating two types of features: common features and unique features. These features are derived from two key observations. First, the HS and MS images capture the same scene, indicating the presence of common features. Second, the images are generated by distinct sensors on the satellite, implying specific unique features specific to each image.
- We reformulate the observation model as an optimization problem and then place implicit prior on both the common and unique features. This allows us to leverage additional information and improve the quality of the fused image. Then, we develop a proximal gradient algorithm to solve the optimization problem efficiently, which is then unfolded into a deep network. Each module of the network corresponds to a specific operation in the iterative algorithm and, thus, the interpretability of the network can be significantly improved.
- Experimental results on benchmark datasets demonstrate the superiority of our network both quantitatively and qualitatively compared with other state-of-the-art methods.

## 2. Proposed Method

In this section, we introduce the proposed methodology. Initially, we formalize the established model as an optimization problem and decompose it into three distinct

subproblems, utilizing the proximal gradient algorithm to derive the model's solution. Next, we describe the solution steps by presenting them within a network framework and provide a clear overview of the network's structure. Finally, we explain the basic settings used during the training of the network.

### *2.1. Problem Formulation*

The MS/HS fusion model is proposed on the basis of the following observations. First, since both the LRHS image $H \in \mathbb{R}^{m\times n\times B}$ and the HRMS image $M \in \mathbb{R}^{M\times N\times b}$ are obtained from the same scenario, they should have some common features. Second, since the two images are acquired by different sensors in the satellite, thus each of them has its own unique features. Furthermore, we also assume that the information underlying both the LRHS image and the HRMS image is beneficial to estimate the HRHS image $O \in \mathbb{R}^{M\times N\times B}$. Therefore, we first extract these effective features from both images and then fuse these features to generate the HRHS image $O$. To acquire the features from both images, we adopt the CSC technique to model the LRHS and HRMS images separately.

Specifically, our new model is defined as follows:

$$M = \sum_{k=1}^{K} D_k^c \otimes C_k + \sum_{k=1}^{K} D_k^u \otimes U_k, \tag{1}$$

$$\tilde{H} = \sum_{k=1}^{K} P_k^c \otimes C_k + \sum_{k=1}^{K} P_k^v \otimes V_k, \tag{2}$$

where $\tilde{H} \in \mathbb{R}^{M\times N\times B}$ is the upsampled version of LRHS image $H$ using the polynomial kernel technique; $\otimes$ is the convolutional operation; $D_k^c \in \mathbb{R}^{s\times s\times b}$ and $P_k^c \in \mathbb{R}^{s\times s\times B}$ denote the common filter for both images; $D_k^u \in \mathbb{R}^{s\times s\times b}$ and $P_k^v \in \mathbb{R}^{s\times s\times B}$ represent the filters of both images, respectively; $C_k \in \mathbb{R}^{M\times N}$ denotes the shared common features; and $U_k \in \mathbb{R}^{M\times N}$ and $V_k \in \mathbb{R}^{M\times N}$ are unique features of HRMS image and LRHS image, respectively. Then, the HRHS image $O \in \mathbb{R}^{M\times N\times B}$ can be generated by fusing these features as follows:

$$O = \sum_{k=1}^{K} G_k^c \otimes C_k + \sum_{k=1}^{K} G_k^u \otimes U_k + \sum_{k=1}^{K} G_k^v \otimes V_k, \tag{3}$$

where $G_k^c \in \mathbb{R}^{s\times s\times B}$ represents the common filter, $G_k^u \in \mathbb{R}^{s\times s\times B}$ and $G_k^v \in \mathbb{R}^{s\times s\times B}$ denote the unique filters concerning the features of both images. To make the notations clear, we denote $C = \{C_k\}_{k=1}^K \in \mathbb{R}^{M\times N\times K}$, $U = \{U_k\}_{k=1}^K \in \mathbb{R}^{M\times N\times K}$, and $V = \{V_k\}_{k=1}^K \in \mathbb{R}^{M\times N\times K}$.

In the following, we assume that all the filters are known. Therefore, to obtain the fused HRHS image $O$, we need to calculate $C$, $U$, and $V$ as follows:

$$\begin{aligned} min_{U,V,C} & \frac{1}{2} \parallel M - \sum_{k=1}^{K} D_k^c \otimes C_k - \sum_{k=1}^{K} D_k^u \otimes U_k \parallel_F^2 \\ & + \frac{1}{2} \parallel \tilde{H} - \sum_{k=1}^{K} P_k^c \otimes C_k - \sum_{k=1}^{K} P_k^v \otimes V_k \parallel_F^2 \\ & + \lambda_1 f_1(U) + \lambda_2 f_2(V) + \lambda_3 f_3(C), \end{aligned} \tag{4}$$

where $\lambda_1, \lambda_2$, and$\lambda_3$ are the trade-off parameters. $f_1(\cdot)$, $f_2(\cdot)$, and $f_3(\cdot)$ are the regularization terms modeling the priors on $U$, $V$, and $C$, respectively.

### *2.2. Model Optimization*

To solve the problem, we adopt the alternate optimization algorithm to alternately update each variable with other variables fixed. First, the problem is broken down into three sub-problems:

$$U^{(t)} = argmin_U \parallel M - \sum_{k=1}^{K} D_k^c \otimes C_k^{(t-1)} - \sum_{k=1}^{K} D_k^u \otimes U_k \parallel_F^2 + \lambda_1 f_1(U), \tag{5}$$

$$V^{(t)} = argmin_V \parallel \tilde{H} - \sum_{k=1}^{K} P_k^c \otimes C_k^{(t-1)} - \sum_{k=1}^{K} P_k^v \otimes V_k \parallel_F^2 + \lambda_2 f_2(V), \tag{6}$$

$$\begin{aligned} C^{(t)} = argmin_C \parallel M - \sum_{k=1}^{K} D_k^c \otimes C_k - \sum_{k=1}^{K} D_k^u \otimes U_k^{(t)} \parallel_F^2 \\ + \parallel \tilde{H} - \sum_{k=1}^{K} P_k^c \otimes C_k - \sum_{k=1}^{K} P_k^v \otimes V_k^{(t)} \parallel_F^2 + \lambda_3 f_3(C). \end{aligned} \tag{7}$$

Next, we introduce the optimization for each sub-problem.

2.2.1. U-Subproblem

Instead of directly optimizing the objective (5), we first adopt the quadratic function to approximate the objective function, which is expressed as:

$$\begin{aligned} U^{(t)} = argmin_U g(U^{(t-1)}) + \frac{1}{2\eta_1} \parallel U - U^{(t-1)} \parallel_F^2 \\ + \langle \nabla g(U^{(t-1)}), U - U^{(t-1)} \rangle + \lambda_1 f_1(U), \end{aligned} \tag{8}$$

where

$$g(U^{(t-1)}) = \parallel M - \sum_{k=1}^{K} D_k^c \otimes C_k^{(t-1)} - \sum_{k=1}^{K} D_k^u \otimes U_k^{(t-1)} \parallel_F^2, \tag{9}$$

$$\nabla g(U^{(t-1)}) = D^u \otimes^{\dagger} (\sum_{k=1}^{K} D_k^c \otimes C_k^{(t-1)} + \sum_{k=1}^{K} D_k^u \otimes U_k^{(t-1)} - M). \tag{10}$$

$\nabla$ is gradient operator, $\eta_1$ is the step size, $D^u \in \mathbb{R}^{s \times s \times K \times b}$ is a 4-D tensor stacked by all $D_k^u$, and $\otimes^{\dagger}$ denotes the transposed convolution. By a simple mathematical derivation, Equation (8) can be simplified to

$$U^{(t)} = argmin_U \frac{1}{2} ||U - (U^{(t-1)} - \eta_1 \nabla g(U^{(t-1)}))||_F^2 + \eta_1 \lambda_1 f_1(U). \tag{11}$$

Equation (11) can be optimized by the general proximal operator, which is defined as

$$U^{(t)} = prox_{\eta_1 \lambda_1}(U^{(t-1)} - \eta_1 \nabla g(U^{(t-1)})), \tag{12}$$

where $prox_{\eta_1 \lambda_1}(\cdot)$ is the proximal operator related to $f_1(\cdot)$. To overcome the constraint of human-designed regularizers, we employ deep CNNs to autonomously acquire prior knowledge from the data.

2.2.2. V-Subproblem

To update V, we employ a similar approach. Firstly, we define the quadratic approximation of Equation (6) with respect to V as follows:

$$\begin{aligned} V^{(t)} = argmin_V h(V^{(t-1)}) + \frac{1}{2\eta_2} \parallel V - V^{(t-1)} \parallel_F^2 \\ + \langle \nabla h(V^{(t-1)}), V - V^{(t-1)} \rangle + \lambda_2 f_2(V), \end{aligned} \tag{13}$$

where

$$h(V^{(t-1)}) = \parallel \tilde{H} - \sum_{k=1}^{K} P_k^c \otimes C_k^{(t-1)} - \sum_{k=1}^{K} P_k^v \otimes V_k^{(t-1)} \parallel_F^2, \tag{14}$$

$$\nabla h(V^{(t-1)}) = P^v \otimes^{\dagger} (\sum_{k=1}^{K} P_k^c \otimes C_k^{(t-1)} + \sum_{k=1}^{K} P_k^v \otimes V_k^{(t-1)} - \tilde{H}). \tag{15}$$

$\nabla$ is gradient operator, $\eta_2$ is the step size, $P^v \in \mathbb{R}^{s\times s\times K\times B}$ is a 4-D tensor stacked by all $P_k^v$, and $\otimes^{\dagger}$ denotes the transposed convolution. Then, Equation (13) can be equivalently to

$$V^{(t)} = argmin_V \frac{1}{2}||V - (V^{(t-1)} - \eta_2 \nabla h(V^{(t-1)}))||_F^2 + \eta_2 \lambda_2 f_2(V). \tag{16}$$

Equation (16) can also be optimized by the general proximal operator, which is referred to as

$$V^{(t)} = prox_{\eta_2\lambda_2}(V^{(t-1)} - \eta_2 \nabla h(V^{(t-1)})), \tag{17}$$

where $prox_{\eta_2\lambda_2}(\cdot)$ is the proximal operator related to $f_2(\cdot)$.

2.2.3. C-Subproblem

Equation (7) can be re-written as:

$$\begin{aligned} C^{(t)} = argmin_C \parallel \hat{M} - \sum_{k=1}^{K} D_k^c \otimes C_k \parallel_F^2 \\ + \parallel \hat{H} - \sum_{k=1}^{K} P_k^c \otimes C_k \parallel_F^2 + \lambda_3 f_3(C), \end{aligned} \tag{18}$$

where $\hat{M} = M - \sum_{k=1}^{K} D_k^u \otimes U_k^{(t)}$, $\hat{H} = \tilde{H} - \sum_{k=1}^{K} P_k^v \otimes V_k^{(t)}$. Furthermore, Equation (18) can be expressed as

$$C^{(t)} = argmin||N - \sum_{k=1}^{K} L_k^c \otimes C_k||_F^2 + \lambda_3 f_3(C), \tag{19}$$

where $N \in \mathbb{R}_{M\times N\times(B+b)}$ is the concatenation of $\hat{H}$ and $\hat{M}$, and $L_k^c \in \mathbb{R}^{s\times s\times(B+b)}$ is the concatenation of $D_k^c$ and $P_k^c$. Equation (19) can then be solved by the same way as optimizing $U$ and $V$. As before, we first derive the quadratic approximation of Equation (19) with respect to $C$ as follows:

$$\begin{aligned} C^{(t)} = argmin_C l(C^{(t-1)}) + \frac{1}{2\eta_3} \parallel C - C^{(t-1)} \parallel_F^2 \\ + \langle \nabla l(C^{(t-1)}), C - C^{(t-1)} \rangle + \lambda_3 f_3(C). \end{aligned} \tag{20}$$

where

$$l(C^{(t-1)}) = \parallel N - \sum_{k=1}^{K} L_k^c \otimes C_k^{(t-1)} \parallel_F^2, \tag{21}$$

$$\nabla l(C^{(t-1)}) = L^c \otimes^{\dagger} (\sum_{k=1}^{K} L_k^c \otimes C_k^{(t-1)} - N). \tag{22}$$

$\nabla$ is gradient operator, $\eta_3$ is the step size, $L^c \in \mathbb{R}^{s\times s\times K\times B}$ is a 4-D tensor stacked by all $L_k^c$, and $\otimes^{\dagger}$ denotes the transposed convolution. Then, Equation (20) can be equivalently to

$$C^{(t)} = argmin_C \frac{1}{2}||C - (C^{(t-1)} - \eta_3 \nabla l(C^{(t-1)}))||_F^2 + \eta_3 \lambda_3 f_3(C). \tag{23}$$

Equation (23) can also be optimized by the general proximal operator, which is referred to as

$$C^{(t)} = prox_{\eta_3\lambda_3}(C^{(t-1)} - \eta_3 \nabla l(C^{(t-1)})), \tag{24}$$

where $prox_{\eta_3\lambda_3}(\cdot)$ is the proximal operator related to $f_3(\cdot)$.

### 2.3. Network Architecture

Based on the model and solution proposed in the above two sections, we construct a deep neural network, which is shown in Figure 1. This network consists of $T$ stages, which correspond to $T$ iterations of the iterative algorithm used to solve Equation (4). These stages are connected sequentially and each phase completes an update of $U$, $V$, and $C$. And finally, the output of the last stage is the ultimate result of the three features.

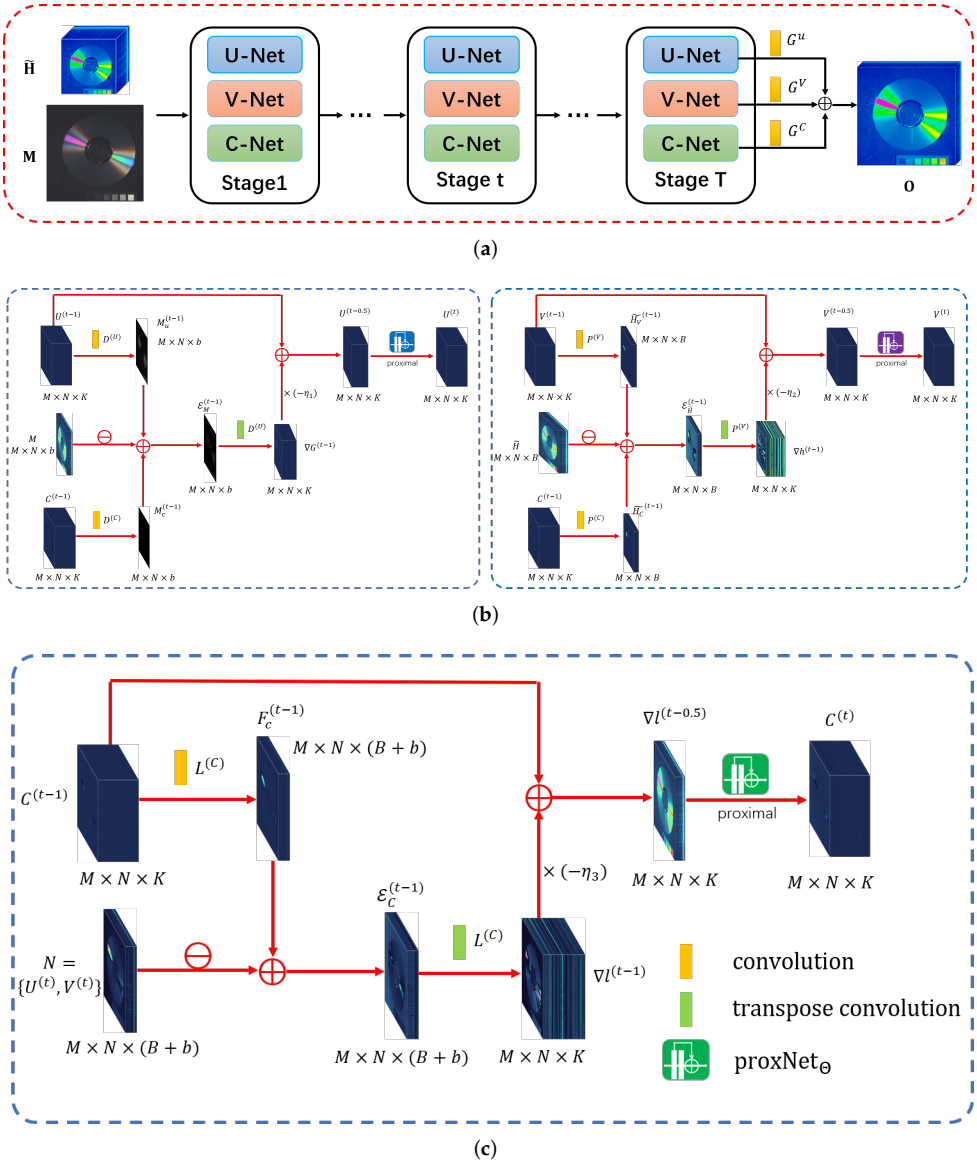



**Figure 1.** (**a**) The overall architecture of our proposed unfolding network. (**b**) The left and right sides of the figure show the structure of U-Net and V-Net, respectively. (**c**) The structure of C-Net.

As shown in Equations (12), (17), and (24), the three proximal operators $prox_{\eta_1\lambda_1}(\cdot)$, $prox_{\eta_2\lambda_2}(\cdot)$, and $prox_{\eta_2\lambda_2}(\cdot)$ play a important role in the algorithm. Usually, various hand-crafted priors are designed to make the proximal operators better and be closed form. However, due to the limited representation abilities, human-designed priors can not always achieve the expected effect. Therefore, we utilize convolutional neural networks to explore prior for $U$, $V$, and $C$, namely learning their corresponding proximal operators $prox_{\eta_1\lambda_1}(\cdot)$, $prox_{\eta_2\lambda_2}(\cdot)$, and $prox_{\eta_2\lambda_2}(\cdot)$ for updating $U^{(t)}$, $V^{(t)}$, and $C^{(t)}$ in each stage $t$. Specifically, we use ResNet to approximate the three proximal operators. Our network is designed as

several stages, and at stage $t$, $U^{(t)}$ is updated using $M$, $U^{(t-1)}$, and $C^{(t-1)}$ by the following sub-steps:

$$U-Net:\begin{cases} M_c^{(t-1)} = \sum_{k=1}^{K} D_k^c \otimes C_k^{(t-1)}, \\ M_u^{(t-1)} = \sum_{k=1}^{K} D_k^u \otimes U_k^{(t-1)}, \\ \varepsilon_M^{(t-1)} = M_c^{(t-1)} + M_u^{(t-1)} - M, \\ \nabla g^{(t-1)} = D^u \otimes^{\dagger} \varepsilon_M^{(t-1)}, \\ U^{(t-0.5)} = U^{(t-1)} - \eta_1 \nabla g^{(t-1)}, \\ U^{(t)} = proxNet_{\Theta_U^t}(U^{(t-0.5)}). \end{cases} \tag{25}$$

Each step is unfolded into one module of the U-Net shown in Figure 1. Then, $V^{(t)}$ is updated by $\tilde{H}$, $V^{(t-1)}$, and $C^{(t-1)}$:

$$V-Net:\begin{cases} \tilde{H}_c^{(t-1)} = \sum_{k=1}^{K} P_k^c \otimes C_k^{(t-1)}, \\ \tilde{H}_u^{(t-1)} = \sum_{k=1}^{K} P_k^v \otimes V_k^{(t-1)}, \\ \varepsilon_{\tilde{H}}^{(t-1)} = \tilde{H}_c^{(t-1)} + \tilde{H}_v^{(t-1)} - \tilde{H}, \\ \nabla h^{(t-1)} = P^v \otimes^{\dagger} \varepsilon_{\tilde{H}}^{(t-1)}, \\ V^{(t-0.5)} = V^{(t-1)} - \eta_2 \nabla h^{(t-1)}, \\ V^{(t)} = proxNet_{\Theta_V^t}(V^{(t-0.5)}). \end{cases} \tag{26}$$

Similarly, each step is unfolded into one module of the V-Net as shown in Figure 1. Finally, $C^{(t)}$ is updated using $M$, $\tilde{H}$, $U^{(t)}$, and $V^{(t)}$

$$C-Net:\begin{cases} F_c^{(t-1)} = \sum_{k=1}^{K} L_k^c \otimes C_k^{(t-1)}, \\ \varepsilon_C^{(t-1)} = F_c^{(t-1)} - N, \\ \nabla l^{(t-1)} = L^c \otimes^{\dagger} \varepsilon_C^{(t-1)}, \\ \nabla l^{(t-0.5)} = C^{(t-1)} - \eta_3 \nabla l^{(t-1)}, \\ C^{(t)} = proxNet_{\Theta_C^{(t)}}(l^{(t-0.5)}). \end{cases} \tag{27}$$

Similarly, each step is unfolded into one module of the C-Net, as shown in Figure 1. In the formula above, $\Theta_U^{(t)}$, $\Theta_V^{(t)}$, and $\Theta_C^{(t)}$ are the parameters of the three ResNets at the $t^{th}$ stage. Note that all of the network parameters can be end-to-end learned from training data, including $\{\Theta_U^{(t)}, \Theta_V^{(t)}, \Theta_C^{(t)}\}_{t=1}^{T}$, $D^c = \{D_k^c\}_{k=1}^{K}$, $D^u = \{D_k^u\}_{k=1}^{K}$, $P^c = \{P_k^c\}_{k=1}^{K}$, $P^v = \{P_k^v\}_{k=1}^{K}$, $G^c = \{G_k^c\}_{k=1}^{K}$, $G^u = \{G_k^u\}_{k=1}^{K}$, $G^v = \{G_k^v\}_{k=1}^{K}$, $\eta_1, \eta_2, \eta_3$. Note that our network is interpretable. For example, to update $U^{(t)}$ by Equation (25), computing $U^{(t-0.5)}$ means the gradient descent, and the $proxNet_{\Theta_U^{(t)}}(\cdot)$ performs nonlinear operation on $U^{(t-0.5)}$.

### *2.4. Network Training*

The loss function of training process is defined as

$$L = \sum_{j=1}^{J} \| O_p^j - O_g^j \|_2^2, \tag{28}$$

where $O_p^j$ is the predicted HRHS image of the $j^{th}$ training sample, $O_g^j$ represents the reference HRHS image of the $j^{th}$ training sample, and $J$ is the number of training pairs. The proposed network is trained on Python 3.9.0 with Tensorflow 2.5.0 and Linux operating system with NVIDIA GPU GeForce GTX 2080Ti. In the network, all the convolutional kernels are set as $8 \times 8$, and the stage number of our network is set as 3. and the C-Net contains 3 ResBlocks, while the U-Net and V-Net each containing 5 ResBlocks. During optimizing the model, we chose the Adam algorithm and a decay technique to set the learning rate, i.e., decayed by 0.95 every 20,000 iterations with a fixed initial learning rate 0.00005. The epoch number is 95, and the batch size is 3.

## 3. Experimental Results

To verify the efficiency of the proposed model, we compare the proposed network with several state-of-the-art methods on three datasets. The performance assessment contains two cases. In the first case where the ground-truth HRMS is known, we utilize popular reference-based indexes. In the second case where the ground-truth HRMS is not available, we use $D_{lambda}$, $D_s$, and the quality without reference ($QNR$) criteria [41]. Specifically, the competing methods consist of GSA [7], SFIM-HS [8], GLP-HS [9], CNMF [11], S3RHSR [12], HySure [10], FUSE [15], UTV [17], CSTF [16], MHFnet [35], and HSRnet [31].

The three selected datasets are the CAVE database [42], Harvard database [43], and Chikusei database [44], respectively. Specifically, the CAVE database consists of 32 HSIs with 512 × 512 and 31 spectral bands between 400 nm to 700 nm at 10 nm steps. Meanwhile, every HSI has a corresponding RGB image with 512 × 512 and three spectral bands. The Harvard dataset is a public dataset that contains 77 HSIs of indoor and outdoor scenes covering various kinds of objects and buildings. Each HSI has a spatial size of 1392 × 1040 with 31 spectral bands acquired at an interval of 10 nm in the range of 420–720 nm. The Chikusei dataset is an airborne hyperspectral dataset, which was taken by Headwall Hyperspec-VNIR-C imaging sensor over agricultural and urban areas in Chikusei, Ibaraki, Japan. It only has one image with 128 bands in the spectral range from 363 nm to 1018 nm and the scene consists of 2517 × 2335 pixels.

We use the following six complementary and widely used quality measures for quantitative fusion assessment, i.e., the peak signal-to-noise ratio (PSNR), the spectral angle mapper (SAM [45]), the erreur relative globale asimensionnelle de synthèse (ERGAS [46]), the structure similarity (SSIM [47]), spatial correlation coefficient (SCC), and $Q2^n$ [48]. Let $x \in R^{B\times P}$ be the reference image with $B$ bands and $P$ pixels. $x = [x_1, \ldots, x_B]^T = [x_1, \ldots, x_j]$, where $x_i \in R^{P\times 1}$ is the $i$th band ($i = 1, \ldots, B$) and $x_j \in R^{B\times 1}$ is the spectral signature of the $j$th pixel ($j = 1, \ldots, P$). $\hat{x}$ denotes the estimated HS image. The PSNR measures the average qualities of spatial reconstruction between all bands of the target and of the reference image. It is the ratio of the maximum square of a signal to the square of relative residual errors. The PSNR is defined as

$$PSNR(x, \hat{x}) = \frac{1}{B}\sum_i 10 \cdot log_{10} \frac{max(x_i)^2}{\| x_i - \hat{x}_i \|_2^2 / P}. \tag{29}$$

The SAM index [45] represents the spectral similarity between reference data and reconstructed hyperspectral data. More precisely, The spectral distance is determined by calculating the angle between two vectors of the estimated and reference spectra. The SAM index is defined as

$$SAM(x, \hat{x}) = \frac{1}{P}\sum_i \arccos\left(\frac{x_j^T \hat{x}_j}{\| x_j \|_2 \| \hat{x}_j \|_2}\right). \tag{30}$$

The SSIM [47] provides a criterion to evaluate the similarity of images from three aspects: brightness, contrast, and structure. Usually, its calculation formula is described as

$$SSIM(x, \hat{x}) = \frac{1}{B}\sum_i \frac{(\mu_{x_i}\mu_{\hat{x}_i} + c_1)(2\sigma_{x_i\hat{x}_i} + c_2)}{(\mu_{\hat{x}_i}^2 + \mu_{\hat{x}_i}^2 + c_1)(\sigma_{\hat{x}_i}^2 + \sigma_{\hat{x}_i}^2 + c_2)}. \tag{31}$$

The ERGAS [46] conducted an overall evaluation of space and spectrum. ERGAS calculates the band-wise normalized root-mean-square error(RMSE) and multiplies it with the ratio of the ground sampling distance, which is defined as

$$ERGAS(x, \hat{x}) = 100d\sqrt{\frac{1}{B}\sum_{i=1}^{B} \frac{\| x_i - \hat{x}_i \|_2^2}{(\frac{1}{P}1_P^T x_i)^2}}. \tag{32}$$

The SCC is a method of assessing the similarity between the estimated and reference data. It can be achieved by calculating the correlation coefficient of the corresponding pixel values between the two images. Specifically, the calculation formula is as follows

$$SCC(x, \hat{x}) = \frac{1}{P} \sum_{j=1}^{P} \frac{cov(x_j, \hat{x}_j)}{\sigma_{x_j} \sigma_{\hat{x}_j}}. \tag{33}$$

$Q2^n$ [48] is an extension of the universal image quality index ($UIQI$ [49]) to HS image depended on hypercomplex numbers. The $UIQI$ is designed for monochromatic images to measure any distortion from three factors, i.e., loss of correlation, luminance distortion, and contrast distortion, and by the product method of the three terms. The $UIQI$ between one reference image band ($z$) and its corresponding target image band ($y$) can be written as

$$UIQI(z, y) = \frac{4\sigma_{zy}\bar{z}\bar{y}}{(\sigma_z^2 + \sigma_y^2)(\bar{z}^2 + \bar{y}^2)}. \tag{34}$$

To overcome this limitation and additionally consider the spectral distortion, $Q2^n$ extends the $UIQI$ based on modeling each pixel spectrum ($x_j$) as a hypercomplex number [48], expressed as

$$x_j = x_{j,0} + x_{j,1} i_1 + x_{j,2} i_2 + \ldots + x_{j,2^n-1} i_{2^n-1}. \tag{35}$$

The smaller ERGAS and SAM are, and the larger PSNR, SSIM, SCC and $Q2^n$ are, the better the fusion result is. The ideal values for SSIM and SCC are one.

Additionally, we also adopt $D_\lambda, D_s$, and $QNR$ [50,51] as the indexes to measure the qualities of data with no reference at full resolutions. The spectral distortion, $D_m$, is determined by comparing the LRHS images with the fused HS images [51]. Interband $UIQI$ values are computed in two distinct sets for low and high resolutions. By subtracting the corresponding $UIQI$ values obtained at the two scales, the spectral distortion caused by the pansharpening process is derived. The spatial distortion, $D_s$, is determined by calculating the $UIQI$ values between each HS band and the MS band that has been degraded to the resolution of the HS images, as well as between the fused HS images and the full-resolution MS image. The absolute difference between the corresponding $UIQI$ values, averaged over all bands, gives the spatial distortion $D_s$ [51]. The two distortions are then combined to provide a unique quality index, referred to as $QNR \in [0, 1]$, with 1 being the best attainable value:

$$QNR = (1 - D_\lambda)^\alpha (1 - D_s)^\beta, \tag{36}$$

where usually $\alpha = \beta = 1$.

### *3.1. Evaluation at Reduced Resolution*

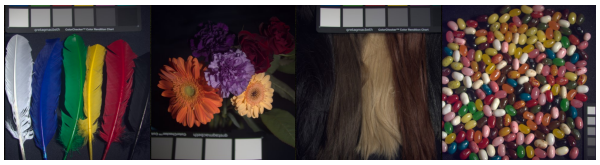
We measure the performance of our network in the reduced resolution case. We conduct experiments on two datasets, CAVE and Chikusei. In the experiment on CAVE dataset, we did not use all the data, but chose 20 images as the training set and randomly selected 10 patches of size 200 × 200 from the other 11 images for testing. The 11 CAVE test images are shown in Figure 2.

We extracted 18,000 overlapped patches with a size of 32 × 32 × 31 from the 20 images of the CAVE dataset used as GT of training dataset, thus forming the HR-HSI patches. Accordingly, the LR-HSI patches are generated starting from the HR-HSI by applying a Gaussian blur with kernel size equal to 3 × 3 and standard deviation equal to 0.5 and then downsampling the blurred patches to the size of 8 × 8, i.e., with a downsampling factor of 4. The corresponding RGB patches are generated based on the spectral response function of the Nikon D700 camera. Thus, 18,000 patches of size of 32 × 32 × 3 are available to represent the HR-MSI of training dataset. We perform the experiments on all the 10 testing reduced images. Table 1 presents the average $QIs$ on the 10 images. It can be seen that, in Table 1, the DL-based methods (i.e., HSRnet, MHFnet and our network) have better performance than other traditional methods, which may be due to the powerful learning

capabilities of DL methods. Additionally, our method performs best in terms of the PSNR, ERGAS, and SCC indexes, and is comparable to the best value on the remaining metrics. The visual comparison shown in Figure 3, implying our network performs almost the best in spatial domain, further verifies the numerical results. For better illustration, we enlarge two regions of the fused image. Meanwhile, the selected spectral vectors of the sample which is shown in Figure 3 are also presented for the fused results coming from the different fusion methods and the GT. It is worth to be remarked that the spectral vectors estimated by our method and the GT ones are very close to each other.



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



(**h**) (**i**) (**j**) (**k**)

**Figure 2.** 11 test images from the CAVE dataset. (**a**) balloons; (**b**) CD; (**c**) chart and stuffed toy; (**d**) clay; (**e**) fake and real beers; (**f**) fake and real lemon slices; (**g**) fake and real tomatoes; (**h**) feathers; (**i**) flowers; (**j**) hairs; and (**k**) jelly beans.

**Table 1.** Average *QIs* of the results on 10 patches of size 200 × 200 on the CAVE datasets. The best values are highlighted in boldface and the second place is underlined.

| Method | PSNR | SSIM | SAM | ERGAS | SCC | Q2n |
|---|---|---|---|---|---|---|
| GSA [7] | 37.578 | 0.976 | 6.621 | 3.483 | 0.992 | 0.717 |
| SFIM-HS [8] | 30.179 | 0.941 | 10.140 | 22.596 | 0.967 | 0.645 |
| GLP-HS [9] | 32.956 | 0.962 | 6.106 | 5.136 | 0.985 | 0.670 |
| CNMF [11] | 36.647 | 0.971 | 8.251 | 3.847 | 0.985 | 0.708 |
| S3RHSR [12] | 31.564 | 0.922 | 19.770 | 7.500 | 0.977 | 0.657 |
| HySure [10] | 35.879 | 0.968 | 7.274 | 4.056 | 0.989 | 0.742 |
| FUSE [15] | 27.247 | 0.934 | 5.035 | 9.248 | 0.955 | 0.601 |
| UTV [17] | 29.952 | 0.815 | 12.209 | 7.988 | 0.959 | 0.441 |
| CSTF [16] | 33.694 | 0.954 | 7.326 | 4.948 | 0.985 | 0.621 |
| HSRnet [31] | <u>42.796</u> | **0.994** | **2.814** | <u>2.147</u> | <u>0.995</u> | **0.817** |
| MHFnet [35] | 41.308 | 0.987 | 5.573 | 2.802 | 0.989 | 0.732 |
| Ours | **43.485** | <u>0.993</u> | <u>3.319</u> | **1.986** | **0.995** | <u>0.799</u> |

To demonstrate the performance of our MHF-CSCNet on remote sensing HSIs, we conducted an experiment on the Chikusei dataset, which is taken over agricultural and urban areas in Chikusei, Ibaraki, Japan. We take the original data as the GT HRHS and simulate the LRHS in the same way as the previous experiments. As for the HRMS, we use the corresponding RGB image obtained by Canon EOS 5D Mark II together with the HRHS. Then, we selected the 1000 × 2200 area in the upper left corner of the image for training, and extracted 64 × 64 overlapping patches from it as training GT HRHS patches. Moreover, the input HRMS and LRHS patches are of size 64 × 64 × 3 and 16 × 16 × 128, respectively. Additionally, we crop six nonoverlap 680 × 680 × 128 area from the remaining part of the Chikusei dataset for testing. Table 2 shows the average *QIs* for the six testing images on all methods. It can clearly be seen that our MHF-CSCNet has the best performance on almost all indices compared to all other methods. These results indicate that our method

has excellent performances on reconstruction of spatial and spectral senses. Although our approach is slightly worse than HSRnet (which directly preserves the spectrum from the hyperspectral images and then supplements it with spectral information from the high-resolution images) on SAM metrics, it is still significantly better than other approaches. It shows that our method not only has a good protection for the reconstruction of spatial details, but also for the spectrum. We also exhibit the pseudocolor images of the obtained outcomes for a visual comparison in Figure 4. Obviously, the fused results of our MHF-CSCNet either exceeds some methods in spatial feature recovery, or outperforms in spectral protection. At the same time, we show the spectral vector at a certain position in Figure 4. It can be observed from Figure 4a,b that the spectral vectors obtained by the SFIMHS, HySure, FUSE, UTV, and CSTF methods have a large degree of error, While HSRnet, MHFnet and our methods have a slight spectral distortion. Specifically, our method and HSRnet both protect the spectrum well. The effectiveness of our method is checked again.

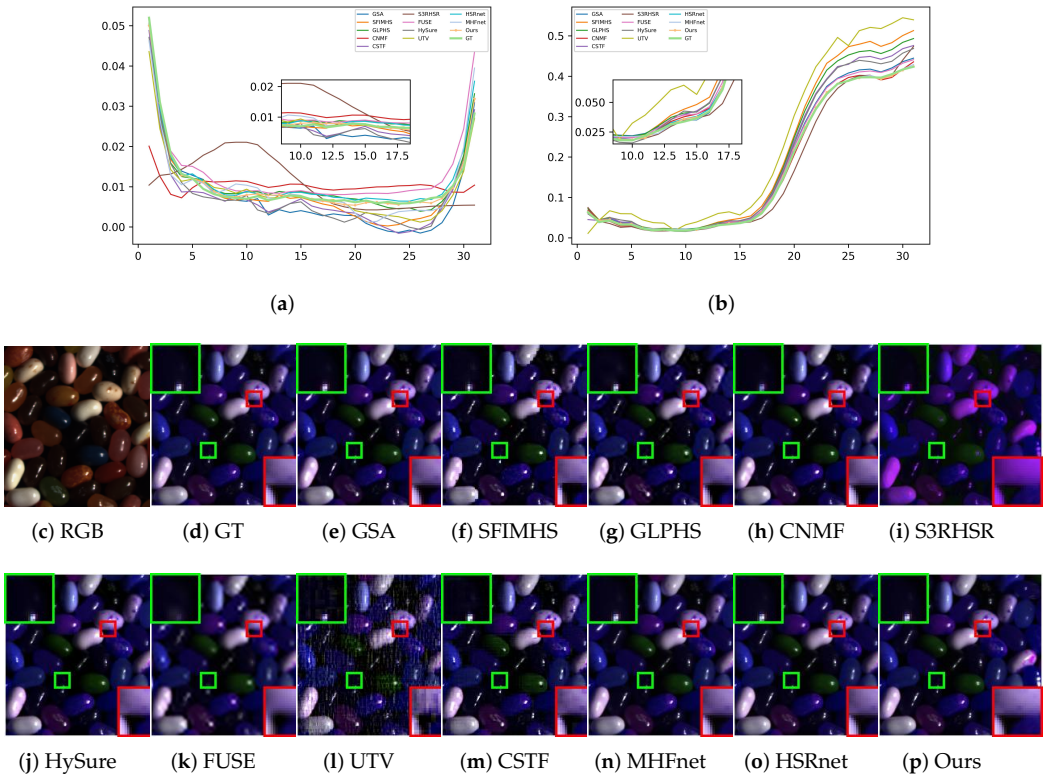



**Figure 3.** Visual representations of the fused results. Includes selected spectral vectors and the pseudocolor images for the outcomes coming from the different fusion methods on the CAVE datasets with size 200 × 200. (**a**) Spectral vectors in 2nd test image located at (31, 200). (**b**) Spectral vectors in 6th test image located at (61, 151). We show the composite image of the HS image with bands 15-8-29 as R-G-B on the final two rows. (**c**) The simulated RGB images of a test sample. (**d**) The ground-truth HRHS image. (**e**–**p**) The results obtained by 12 comparison methods and had two demarcated areas marked by red and green boxes zoomed for easy observation.

**Table 2.** Average *QIs* of the results for six testing images on the Chikusei dataset. The best values are highlighted in boldface and the second place is underlined.

| Method | PSNR | SSIM | SAM | ERGAS | SCC | Q2n |
|---|---|---|---|---|---|---|
| GSA [7] | 32.519 | 0.881 | 3.745 | 5.511 | 0.698 | 0.630 |
| SFIM-HS [8] | 28.468 | 0.885 | 4.392 | 60.326 | 0.782 | 0.617 |
| GLP-HS [9] | 32.418 | 0.917 | 3.288 | 5.717 | 0.922 | 0.647 |
| CNMF [11] | 31.163 | 0.883 | 3.457 | 6.059 | 0.876 | 0.560 |
| S3RHSR [12] | 25.936 | 0.751 | 10.115 | 12.367 | 0.749 | 0.441 |
| HySure [10] | 30.464 | 0.842 | 3.716 | 7.020 | 0.606 | 0.555 |
| FUSE [15] | 31.871 | 0.906 | 3.251 | 6.170 | 0.759 | 0.627 |
| UTV | 25.085 | 0.652 | 9.130 | 12.253 | 0.712 | 0.261 |
| CSTF [16] | 29.701 | 0.829 | 4.853 | 7.376 | 0.871 | 0.460 |
| HSRnet [31] | <u>35.685</u> | <u>0.948</u> | **2.131** | <u>3.668</u> | <u>0.955</u> | **0.778** |
| MHFnet [35] | 33.611 | 0.938 | 3.587 | 5.345 | 0.936 | 0.659 |
| Ours | **36.462** | **0.955** | <u>2.242</u> | **3.174** | **0.959** | <u>0.753</u> |

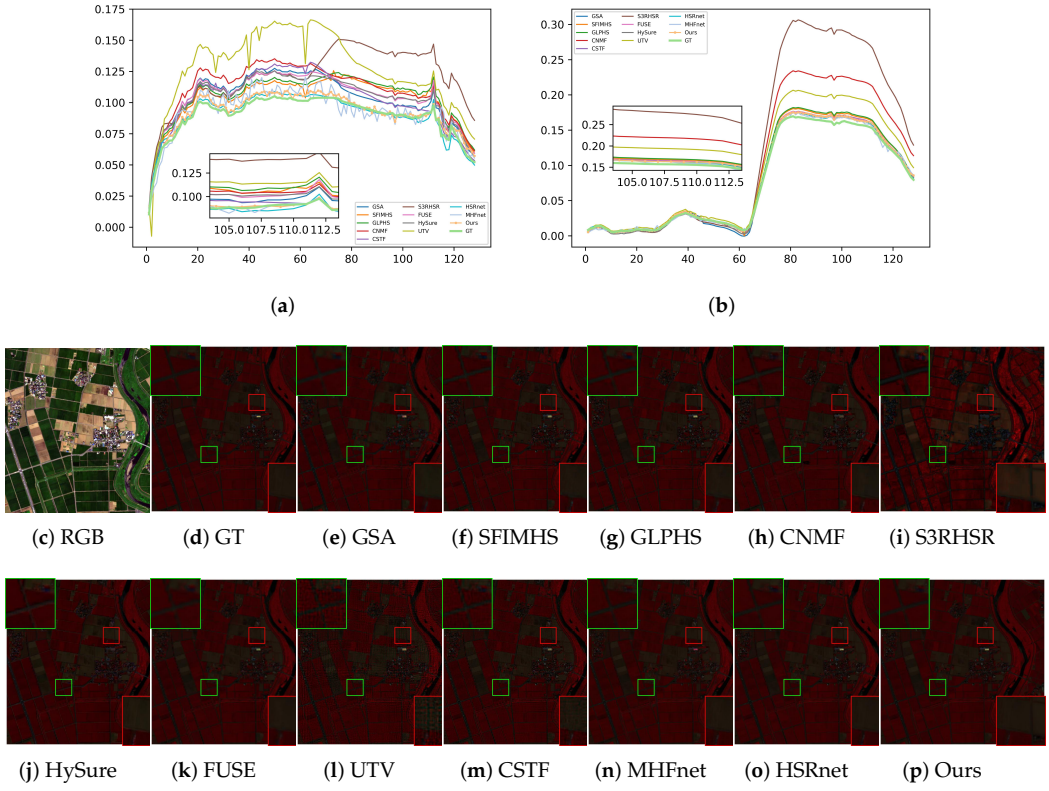
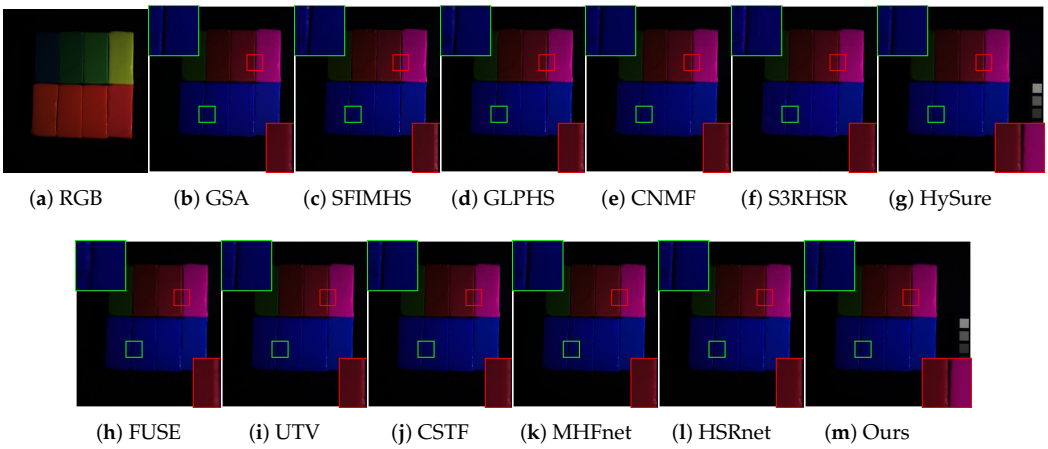



**Figure 4.** Visual representations of the fused results. Includes selected spectral vectors and the pseudocolor images for the outcomes coming from the different fusion methods on the Chikusei datasets with size 680 × 680. (**a**) Spectral vectors in 6th test image located at (250, 250). (**b**) Spectral vectors in 5th test image located at (250, 250). (**c**) The simulated RGB images of a test sample in Chikusei data set. We show the composite image of the HS image with bands 80-60-30 as R-G-B. (**d**) The ground-truth HrHS image. (**e**–**p**) The results obtained by 12 comparison methods, and two demarcated areas marked by red and green boxes zoomed for easy observation.

### *3.2. Evaluation at Full Resolution*

This section evaluates the performance of our network in the full resolution case. We conduct this experiment on one of complete test image pairs from the CAVE test set. The image pair contains one HS image (128 × 128 × 31), and one MS image (512 × 512 × 3). Since the ground-truth HRHS image is not available, we thus adopt $D_\lambda$, $D_s$, and $QNR$ as the indexes. The quantitative results of all competing methods. are reported in Table 3. As can be seen, compared with DL-based methods, our MHF-CSCNet has the best evaluation

indexes. However, some traditional methods are superior than DL-based methods on indices $D_\lambda$ and $D_s$. This phenomenon may be caused by the characteristics of different traditional methods. Figure 5 shows the composite images of a test sample obtained by the competing methods, with bands 15-8-29 as R-G-B. It is seen that the composite image obtained by MHF-net is clearer than the fused results of the other methods, while the results of other methods usually contain blurred boundary.



(**a**) RGB (**b**) GSA (**c**) SFIMHS (**d**) GLPHS (**e**) CNMF (**f**) S3RHSR (**g**) HySure

(**h**) FUSE (**i**) UTV (**j**) CSTF (**k**) MHFnet (**l**) HSRnet (**m**) Ours

**Figure 5.** (**a**) The simulated RGB images of a test sample in CAVE dataset. We show the composite image of the HS image with bands 15-8-29 as R-G-B. (**b**–**m**) The results obtained by 12 comparison methods, and had two demarcated areas marked by red and green boxes zoomed for easy observation.

**Table 3.** Average *QIs* of the results on one image (clay) of CAVE test set. The best values are highlighted in boldface and the second place is underlined.

| Method | $D_\lambda$ | $D_s$ | $QNR$ |
|---|---|---|---|
| GSA [7] | 0.1096 | 0.0948 | 0.806 |
| SFIM-HS [8] | 0.1712 | <u>0.0602</u> | 0.7789 |
| GLP-HS [9] | 0.0778 | 0.0891 | 0.8401 |
| CNMF [11] | 0.0704 | 0.0786 | 0.8565 |
| S3RHSR [12] | 0.1512 | 0.1484 | 0.7229 |
| HySure [10] | 30.0789 | 0.0885 | 0.8396 |
| FUSE [15] | 0.0223 | 0.1022 | 0.8778 |
| UTV [17] | 0.1046 | **0.0545** | 0.8466 |
| CSTF [16] | 30.0741 | 0.0631 | 0.8675 |
| HSRnet [31] | **0.0171** | 0.1027 | <u>0.882</u> |
| MHFnet [35] | 0.1284 | 0.1124 | 0.7736 |
| Ours | <u>0.0222</u> | 0.0967 | **0.8832** |

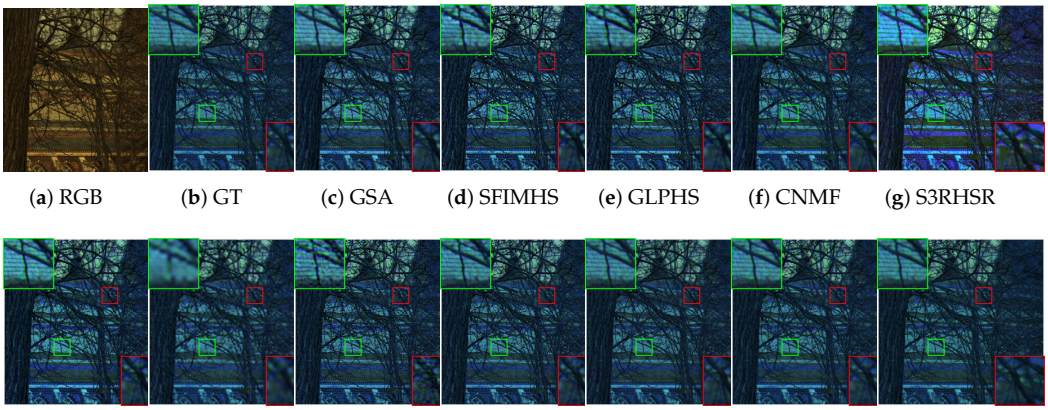
### *3.3. Generalization to New Satellites*

In this section, we randomly choose the 512 × 512 size in the lower right part of the image, then twenty images are randomly determined for testing from the Harvard dataset. As in the previous settings, the original data are regarded as the GT HRHS. The LRHS are simulated by the same method as in the CAVE dataset. The HRMS is also obtained by applying the spectral response of the Nikon D700 camera. In the experiment on the Harvard dataset, we would like to remark that our method is trained on the CAVE dataset, and directly test them without any retraining or fine-tuning. Thus, the performance on the Harvard dataset of these methods could reflect the model's generalization abilities. Table 4 records the average *QIs* over 20 images of all competing methods. In the four indicators, PSNR, SSIM, ERGAS and Q2n, the method based on deep learning is significantly better than the traditional method as shown in Table 4. Figure 6 shows the pseudo-graph composed of bands 8-15-29 on all the completing methods. It is easy to observe that the proposed method performs better than other competing ones, in the better recovery of both

finer-grained textures and spectral distribution. Additionally, we plot the spectral vectors for two exemplary cases in Figure 7. It is worth to be remarked that the spectral vectors estimated by our method and the GT ones are very close to each other. Similar to the results in the CAVE dataset, we can observe the superiority of our method on almost all evaluation measures, which implies that our network can generalize well to new satellites.

**Table 4.** Average *QIs* of the results for twenty testing images on the Harvard dataset. The best values are highlighted in boldface and the second place is underlined.

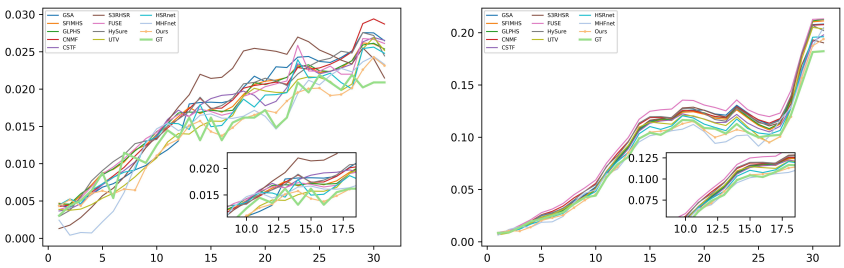
| Method | PSNR | SSIM | SAM | ERGAS | SCC | Q2n |
|---|---|---|---|---|---|---|
| GSA [7] | 31.926 | 0.952 | 3.350 | 5.796 | **0.988** | 0.738 |
| SFIM-HS [8] | 28.701 | 0.916 | 3.446 | 7.200 | 0.971 | 0.665 |
| GLP-HS [9] | 29.558 | 0.928 | 3.630 | 6.671 | 0.977 | 0.678 |
| CNMF [11] | 31.440 | 0.949 | **3.065** | 6.027 | 0.984 | 0.734 |
| S3RHSR [12] | 30.420 | 0.920 | 5.778 | 7.001 | 0.977 | 0.683 |
| HySure [10] | 31.240 | 0.948 | 3.341 | 6.141 | 0.986 | 0.730 |
| FUSE [15] | 27.803 | 0.876 | 3.242 | 7.636 | 0.952 | 0.516 |
| UTV [17] | 33.170 | 0.931 | 4.722 | 4.351 | 0.975 | 0.647 |
| CSTF [16] | 33.927 | 0.930 | 3.660 | 4.042 | 0.978 | 0.654 |
| HSRnet [31] | <u>38.634</u> | <u>0.974</u> | <u>3.159</u> | **2.828** | 0.985 | <u>0.752</u> |
| MHFnet [35] | 37.820 | 0.969 | 4.412 | 4.747 | 0.978 | 0.736 |
| Ours | **38.930** | **0.975** | 3.405 | <u>3.585</u> | <u>0.986</u> | **0.755** |



(**a**) RGB (**b**) GT (**c**) GSA (**d**) SFIMHS (**e**) GLPHS (**f**) CNMF (**g**) S3RHSR

(**h**) HySure (**i**) FUSE (**j**) UTV (**k**) CSTF (**l**) MHFnet (**m**) HSRnet (**n**) Ours

**Figure 6.** (**a**) The simulated RGB images of a test sample in Harvard data set. We show the composite image of the HS image with bands 8-15-29 as R-G-B. (**b**) The ground-truth HRHS image. (**c**–**n**) The results obtained by 12 comparison methods, and had two demarcated areas marked by red and green boxes zoomed for easy observation.



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

**Figure 7.** Visual representations of the fused results. Including selected spectral vectors for the outcomes coming from the different fusion methods on the Harvard datasets with size 512 × 512. (**a**) Spectral vectors in 5th test image located at (101, 101). (**b**) Spectral vectors in 18th test image located at (51, 51).

## 4. Parameters and Time Comparisons

Table 5 shows the parameter number and floating-point operations per second (FLOPs) of our network with the other two DL-based methods. All the methods used 20 images of Harvard dataset as same as the test data the last subsection for inference with NVIDIA GPU GeForce GTX 2080Ti. As Table 5 shows the parameter number, running time and FLOPs of our model are larger than other methods, although our method has better results, and we will try to improve this problem in our future work.

**Table 5.** Comparison of running time, parameter number, and FLOPs.

| Method | Image Size | HSRnet | MHFnet | Our |
|---|---|---|---|---|
| Parameters ($\times 10^6$) | $128 \times 128 \times 31{,}512 \times 512$ | 0.63 | 1.21 | 5.52 |
| FLOPs ($\times 10^{11}$) | $128 \times 128 \times 31{,}512 \times 512$ | 2.57 | 5.10 | 52.33 |
| Running time (s) | $128 \times 128 \times 31{,}512 \times 512$ | 0.47 | 12.91 | 23.24 |

## 5. Conclusions

This paper proposes a novel model-based deep neural network for the MS/HS fusion task. Firstly, a new MS/HS fusion observation model based on convolutional sparse coding (CSC) is proposed. Then, we design a proximal gradient algorithm to solve this model, and the algorithm is unfolded into a network. Our network has a clear physical interpretation because each module corresponds to a specific operation of the algorithm. Experimental results conducted on some benchmark datasets demonstrate that our network can obtain comparable or better quantitative and qualitative performance compared to other methods.

**Author Contributions:** Methodology, X.C.; Validation, B.Z.; Formal Analysis, B.Z.; Resources, X.C.; Writing—Original Draft B.Z.; Writing—Review and Editing, X.C.; Funding Acquisition, X.C. and D.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0112902 and in part by the China NSFC Projects under Contract 62272375 and Contract 12226004.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors on request.

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

## References

1. Yokoya, N.; Grohnfeldt, C.; Chanussot, J. Hyperspectral and Multispectral Data Fusion: A comparative review of the recent literature. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 29–56. [CrossRef]
2. Borsoi, R.; Imbiriba, T.; Bermudez, J. Super-Resolution for Hyperspectral and Multispectral Image Fusion Accounting for Seasonal Spectral Variability. *IEEE Trans. Image Process.* **2020**, *29*, 116–127. [CrossRef] [PubMed]
3. Hong, D.; Yao, J.; Li, C.; Meng, D.; Yokoya, N.; Chanussot, J. Decoupled-and-Coupled Networks: Self-Supervised Hyperspectral Image Super-Resolution with Subpixel Fusion. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5527812.
4. Liu, N.; Li, W.; Tao, R. Geometric Low-Rank Tensor Approximation for Remotely Sensed Hyperspectral And Multispectral Imagery Fusion. In Proceedings of the ICASSP 2022—2022 IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP), Virtual, 7–13 May 2022; pp. 2819–2823.
5. Wang, K.; Wang, Y.; Zhao, X.; Chan, J.; Xu, Z.; Meng, D. Hyperspectral and Multispectral Image Fusion via Nonlocal Low-Rank Tensor Decomposition and Spectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7654–7671. [CrossRef]
6. Jin, W.; Wang, M.; Wang, W.; Yang, G. FS-Net: Four-Stream Network With Spatial–Spectral Representation Learning for Hyperspectral and Multispecral Image Fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 8845–8857. [CrossRef]
7. 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]
8. 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]
9. Aiazzi, B.; Alparone, L.; Baronti, S.; Garzelli, A.; Selva, M. MTF-tailored multiscale fusion of high-resolution MS and Pan imagery. *Photogramm. Eng. Remote Sens.* **2006**, *72*, 591–596. [CrossRef]

10. Simoes, M.; Bioucas-Dias, J.; Almeida, L.; Chanussot, J. A convex formulation for hyperspectral image superresolution via subspace-based regularization. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 3373–3388. [CrossRef]
11. Yokoya, N.; Yairi, T.; Iwasaki, A. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Trans. Geosci. Remote Sens.* **2011**, *50*, 528–537. [CrossRef]
12. Akhtar, N.; Shafait, F.; Mian, A. Sparse spatio-spectral representation for hyperspectral image super-resolution. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part VII 13, pp. 63–78.
13. Lanaras, C.; Baltsavias, E.; Schindler, K. Hyperspectral super-resolution by coupled spectral unmixing. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3586–3594.
14. Eismann, M. Resolution Enhancement of Hyperspectral Imagery Using Maximum a Posteriori Estimation with a Stochastic Mixing Model. Ph.D. Thesis, University of Dayton, Dayton, OH, USA, 2004.
15. Wei, Q.; Dobigeon, N.; Tourneret, J. Fast fusion of multi-band images based on solving a Sylvester equation. *IEEE Trans. Image Process.* **2015**, *24*, 4109–4121. [PubMed]
16. Li, S.; Dian, R.; Fang, L.; Bioucas-Dias, J. Fusing hyperspectral and multispectral images via coupled sparse tensor factorization. *IEEE Trans. Image Process.* **2018**, *27*, 4118–4130. [CrossRef] [PubMed]
17. Xu, T.; Huang, T.; Deng, L.; Zhao, X.; Huang, J. Hyperspectral image superresolution using unidirectional total variation with tucker decomposition. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4381–4398. [CrossRef]
18. Laben, C.; Brower, B. Process for Enhancing the Spatial Resolution of Multispectral Imagery Using Pan-Sharpening. U.S. Patent 6,011,875, 4 January 2000.
19. Carper, W.; Lillesand, T.; Kiefer, R. Others The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data. *Photogramm. Eng. Remote Sens.* **1990**, *56*, 459–467.
20. Selva, M.; Aiazzi, B.; Butera, F.; Chiarantini, L.; Baronti, S. Hyper-sharpening: A first approach on SIM-GA data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3008–3024. [CrossRef]
21. Nascimento, J.; Dias, J. Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 898–910. [CrossRef]
22. Hong, D.; Yokoya, N.; Chanussot, J.; Zhu, X. CoSpace: Common Subspace Learning From Hyperspectral-Multispectral Correspondences. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 4349–4359. [CrossRef]
23. Lee, D.; Seung, H. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. [CrossRef]
24. Yang, J.; Zhao, Y.; Chan, J. Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network. *Remote Sens.* **2018**, *10*, 800. [CrossRef]
25. Li, Y.; Hu, J.; Zhao, X.; Xie, W.; Li, J. Hyperspectral image super-resolution using deep convolutional neural network. *Neurocomputing* **2017**, *266*, 29–41. [CrossRef]
26. Wang, Q.; Li, Q.; Li, X. Hyperspectral image superresolution using spectrum and feature context. *IEEE Trans. Ind. Electron.* **2020**, *68*, 11276–11285. [CrossRef]
27. Wei, W.; Nie, J.; Li, Y.; Zhang, L.; Zhang, Y. Deep recursive network for hyperspectral image super-resolution. *IEEE Trans. Comput. Imaging* **2020**, *6*, 1233–1244. [CrossRef]
28. Hu, J.; Huang, T.; Deng, L.; Dou, H.; Hong, D.; Vivone, G. Fusformer: A transformer-based fusion network for hyperspectral image super-resolution. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6012305. [CrossRef]
29. Han, X.; Yu, J.; Luo, J.; Sun, W. Hyperspectral and multispectral image fusion using cluster-based multi-branch BP neural networks. *Remote Sens.* **2019**, *11*, 1173. [CrossRef]
30. Dian, R.; Li, S.; Kang, X. Regularizing hyperspectral and multispectral image fusion by CNN denoiser. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 1124–1135. [CrossRef]
31. Hu, J.; Huang, T.; Deng, L.; Jiang, T.; Vivone, G.; Chanussot, J. Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 7251–7265. [CrossRef]
32. Palsson, F.; Sveinsson, J.; Ulfarsson, M. Sentinel-2 image fusion using a deep residual network. *Remote Sens.* **2018**, *10*, 1290. [CrossRef]
33. Chen, H.; Yokoya, N.; Wu, C.; Du, B. Unsupervised Multimodal Change Detection Based on Structural Relationship Graph Representation Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5635318. [CrossRef]
34. Chen, H.; Yokoya, N.; Chini, M. Fourier domain structural relationship analysis for unsupervised multimodal change detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *198*, 99–114. [CrossRef]
35. Xie, Q.; Zhou, M.; Zhao, Q.; Meng, D.; Zuo, W.; Xu, Z. Multispectral and hyperspectral image fusion by MS/HS fusion net. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1585–1594.
36. Dong, W.; Zhou, C.; Wu, F.; Wu, J.; Shi, G.; Li, X. Model-guided deep hyperspectral image super-resolution. *IEEE Trans. Image Process.* **2021**, *30*, 5754–5768. [CrossRef]
37. Ma, Q.; Jiang, J.; Liu, X.; Ma, J. Deep Unfolding Network for Spatiospectral Image Super-Resolution. *IEEE Trans. Comput. Imaging* **2022**, *8*, 28–40. [CrossRef]
38. Sun, Y.; Liu, J.; Yang, J.; Xiao, Z.; Wu, Z. A deep image prior-based interpretable network for hyperspectral image fusion. *Remote Sens. Lett.* **2021**, *12*, 1250–1259.

39. Deng, X.; Dragotti, P. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3333–3348. [CrossRef] [PubMed]
40. Li, M.; Cao, X.; Zhao, Q.; Zhang, L.; Meng, D. Online rain/snow removal from surveillance videos. *IEEE Trans. Image Process.* **2021**, *30*, 2029–2044. [CrossRef]
41. Vivone, G.; Garzelli, A.; Xu, Y.; Liao, W.; Chanussot, J. Panchromatic and Hyperspectral Image Fusion: Outcome of the 2022 WHISPERS Hyperspectral Pansharpening Challenge. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 166–179.
42. Yasuma, F.; Mitsunaga, T.; Iso, D.; Nayar, S. Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum. *IEEE Trans. Image Process.* **2010**, *19*, 2241–2253. [CrossRef]
43. Chakrabarti, A.; Zickler, T. Statistics of real-world hyperspectral images. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 193–200.
44. Yokoya, N.; Iwasaki, A. *Airborne Hyperspectral Data over Chikusei*; Technical Report SAL-2016-05-27; Space Application Laboratory, the University of Tokyo: Tokyo, Japan, 2016; Volume 5, p. 5.
45. Yuhas, R.; Boardman, J.; Goetz, A. Determination of semi-arid landscape endmembers and seasonal trends using convex geometry spectral unmixing techniques. In *Summaries of the 4th Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*; JPL: Pasadena, CA, USA, 1993.
46. Wald, L. Quality of high resolution synthesised images: Is there a simple criterion? In Proceedings of the Third Conference "Fusion of Earth Data: Merging Point Measurements, Raster Maps and Remotely Sensed Images", Sophia Antipolis, France, 26–28 January 2000; pp. 99–103.
47. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612.
48. Garzelli, A.; Nencini, F. Hypercomplex quality assessment of multi/hyperspectral images. *IEEE Geosci. Remote. Sens. Lett.* **2009**, *6*, 662–665. [CrossRef]
49. Wang, Z.; Bovik, A. A universal image quality index. *IEEE Signal Process. Lett.* **2002**, *9*, 81–84.
50. Vivone, G.; Alparone, L.; Chanussot, J.; Dalla Mura, M.; Garzelli, A.; Licciardi, G.; Restaino, R.; Wald, L. A Critical Comparison Among Pansharpening Algorithms. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2565–2586.
51. Alparone, L.; Aiazzi, B.; Baronti, S.; Garzelli, A.; Nencini, F.; Selva, M. Multispectral and panchromatic data fusion assessment without reference. *Photogramm. Eng. Remote Sens.* **2008**, *74*, 193–200. [CrossRef]


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