remote sensing

MDPI

Article

An Encoder–Decoder with a Residual Network for Fusing Hyperspectral and Panchromatic Remote Sensing Images

Rui Zhao and Shihong Du *

Institute of Remote Sensing and GIS, Peking University, Beijing 100091, China; zr_winton@pku.edu.cn
* Correspondence: dshgis@hotmail.com

Abstract: For many urban studies it is necessary to obtain remote sensing images with high hyperspectral and spatial resolution by fusing the hyperspectral and panchromatic remote sensing images. In this article, we propose a deep learning model of an encoder–decoder with a residual network (EDRN) for remote sensing image fusion. First, we combined the hyperspectral and panchromatic remote sensing images to circumvent the independence of the hyperspectral and panchromatic image features. Second, we established an encoder–decoder network for extracting representative encoded and decoded deep features. Finally, we established residual networks between the encoder network and the decoder network to enhance the extracted deep features. We evaluated the proposed method on six groups of real-world hyperspectral and panchromatic image datasets, and the experimental results confirmed the superior performance of the proposed method versus six other methods.

Keywords: image fusion; hyperspectral; panchromatic; deep learning; encoder–decoder network; residual network

check for updates

Citation: Zhao, R.; Du, S. An Encoder–Decoder with a Residual Network for Fusing Hyperspectral and Panchromatic Remote Sensing Images. *Remote Sens.* **2022**, *14*, 1981. https://doi.org/10.3390/rs14091981

Academic Editor: Edoardo Pasolli

Received: 10 March 2022
Accepted: 18 April 2022
Published: 20 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 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

Remote sensing image fusion refers to the complementary operation and processing of multisource remote sensing image data in space, time and the spectrum according to certain rules and algorithms. This obtains more accurate and richer information than single image data, and generates synthetic image data with new spatial, spectral and temporal characteristics [1,2]. In the remote sensing community, hyperspectral and panchromatic images are two important image types. Hyperspectral remote sensing images usually have high spectral resolution and can provide rich spectral information. However, due to the limited energy acquired by the remote sensing image sensor, the spatial resolution is too low to maintain images with high spectral resolutions. This means that the spatial details of ground objects cannot be reflected in hyperspectral remote sensing images [3]. Panchromatic remote sensing images usually have high spatial resolution and can provide many spatial details of the ground objects. However, the spectral resolution of panchromatic images is usually low. Thus, panchromatic images cannot provide enough spectral information [4]. As a result, fusing hyperspectral and panchromatic images can obtain images with both high spectral and high spatial resolution. Therefore, this kind of image fusion can complement the deficiency of pre-fused images, and the fused images can be used in a variety of applications such as ground object classification [5], spectral decomposition [6], and urban target detection [7], among others.

Remote sensing image fusion can be at the pixel-level, feature-level or decision-level [8]. In pixel-level fusion, each pixel of all the image data is directly fused through various algebraic operations. Then it extracts the feature information of ground objects after processing and analysis. Pixel-level fusion requires multiple sensors to be placed on the same platform to achieve both accurate spatial registration of the sensors and strict correspondence between the pixels. Pixel-level image fusion methods are generally based on the space domain and the transform domain. Feature-level fusion consists of image

spatial registration, feature extraction, feature fusion and description of the attributes according to the fusion results. When multiple image sensors report similar features at the same location, the likelihood of actual feature occurrence can be increased, and the accuracy of measured features can be improved. Feature-level image fusion is very important for target recognition and identity authentication. Decision-level fusion is the highest level of fusion. First, it carries out spatial registration of the image, and feature extraction and description of the image information's attributes is carried out by using a large-scale database and an expert decision system to simulate the process of human analysis, reasoning, recognition and decision-making. Finally, it fuses the feature information and attributes. This method mainly aims to fuse multisource information and have strong fault tolerance. The difference between decision-level and feature-level fusion lies is that the goal of feature-level fusion is to extract features from remote sensing images and directly fuse them into new features through various algorithms, while decision-level fusion aims to extract features and recognize new ground objects and then combine the ground object information into new ground objects.

Remote sensing image fusion methods can be divided into the following categories: multiresolution analysis (MRA)-based image fusion methods; component substitution (CS)-based fusion methods; matrix decomposition-based methods; Bayesian-based image fusion methods, and remote sensing image fusion methods based on deep learning (DL). MRA-based fusion methods first obtain information on spatial details by multiscale decomposition of panchromatic images, then this is fused into multispectral or hyperspectral images. MRA fusion methods mainly include undecimated wavelet transform (UWT) [9], decimated wavelet transform (DWT) [10], the indistinguishable transform method based on a curve wave [11], and the Laplacian pyramid method [12], the indistinguishable transform method based on a contour wave [13]. These extract spatial details from panchromatic images through a spatial filter, then insert the extracted spatial details into hyperspectral images.

CS-based fusion methods replace components in the multispectral or hyperspectral images with panchromatic images. CS fusion methods include the intensity–hue–saturation (IHS) method [14–16], principal component analysis (PCA) [17–19], and the Gran–Schmidt (GS) method, among others. They also rely on projection of the hyperspectral images into another spectral space to separate the spatial and spectral information, so that the transformed hyperspectral image data can be fused by replacing the spatial components of the panchromatic images. The stronger the correlation between the panchromatic images and the replaced components, the less the spectral loss in the fused images. Therefore, before replacing panchromatic images, histogram matching is usually carried out. Then the fused images are obtained through inverse spectral transformation.

Matrix decomposition-based image fusion methods assume that hyperspectral images can be decomposed into the product of the spectral primitives and the correlation coefficient matrix. The spectral primitives refer to abstract representations of the spectral information, including sparse representation [20–22] and low-rank representation [23,24]. The spectral primitive form of sparse expression has a complete dictionary and assumes that each spectrum is a linear combination of several dictionary items. The items are usually based on a complete dictionary with a low spatial resolution of the hyperspectral remote sensing images by sparse dictionary learning methods, such as K-SVD [25], online dictionaries [26], and non-negative dictionary learning. Next, sparse priors are used to regularize the coefficients; usually sparse coding algorithms to estimate the coefficients. The low-rank expression holds that spectral features can be represented by low-dimensional subspaces, and the matrix composed of spectral primitives is a low-rank matrix. The low-rank spectral elements are usually composed of vertex component analysis (VCA), simplex identification via split augmented Lagrangian (SISAL), principal component analysis, and truncated singular value decomposition (TSVD). Both sparse representation and low-rank representation methods aim to model the similarity and redundancy between spectral bands. Thus, both can maintain the spectral characteristics well. However, the low-rank

representation method can greatly reduce the dimensions of the spectral pattern and has less computational complexity than the sparse representation method.

The Bayesian method relies on the posterior distribution of hyperspectral and panchromatic images. This posterior distribution is obtained by Bayesian reasoning, and posterior distribution contains two factors: (a) a likelihood function, which is the probability density of the multispectral or hyperspectral images and panchromatic images obtained after a given target image, and (b) the prior probability density of the target image, in that the characteristics of the target image can be improved by the desired characteristics, such as segmentation smoothing. Selection of the appropriate prior information can solve the inverse ill-condition problem in the process of fusion [27]. Hyperspectral images and images with a high spatial resolution can be described in the framework of Bayesian inference. This method can intuitively explain the fusion process through the posterior distribution of the Bayesian fusion model. Since fusion problems are often ill-conditioned, Bayesian methods provide a convenient way to regularize the problem by defining an appropriate prior distribution for the scenarios of interest. According to this strategy, many scholars have designed different Bayesian estimation methods for fusing images with high spatial resolution and hyperspectral images [28,29].

DL has made achievements in many fields, such as natural language processing [30], computer vision [31,32], speech recognition [33], search engines [34] and so on. The DL fusion method is considered to be a new trend, which trains a network model to describe the mapping among the hyperspectral images, panchromatic images and target fusion images [35]. Current DL fusion methods include the Deep Residual Pansharpening Neural Network (DRPNN) [36], the Pansharpening Neural Network (PNN) [37], the Multiscale and the Multidepth Convolutional Neural Network (MSDCNN) [38]. The DIP-HyperKite method proposed in [39] defines the spatial-domain constraint as the $L1$ distance between the predicted PAN image and the actual PAN image, proposes a learnable spectral response function (SRF), and also proposes a novel over-complete network, called HyperKite, which focuses on learning high-level features by constraining the receptive from increasing in the deep layers. The RCNN method proposed in [40] utilize the network to map the differential information between the high spatial resolution panchromatic (HR-PAN) image and the low spatial resolution multispectral (LR-MS) image. Moreover, RCNN makes full use of the LR-MS image and utilizes the gradient information of the up-sampled LR-MS image (Up-LR-MS) as auxiliary data to assist the network. Furthermore, an attention module and residual blocks are incorporated in the proposed network structure. The MARB-Net proposed in [41] assigns multiple weights to each feature using multiple attention mechanism models. Then, the MARB-Net deeply mines and integrates the features using the residual network. Finally, MARB-Net performs contextual semantic integration on the deep fusion features using the Bi-LSTM network. These methods can usually obtain better spectral fidelity, but spatial enhancement is inadequate in the image fusion results. The current DL-based image fusion methods usually lack richly formalized and diversified deep features. The current DL-based image fusion methods usually lack deep feature enhancement. Meanwhile, these methods usually regard spatial and spectral features as individual units.

In the late 1980s, the invention of the back propagation algorithm for artificial neural networks brought hope to machine learning, setting off a boom in machine learning based on statistical models. In 2006, Geoffrey Hinton and Ruslan Salakhutdinov published an article in Science, a leading academic journal, that started a wave of deep learning in academia and industry. Since 2006, deep learning has been gaining momentum in academia. Today, Google, Microsoft, Baidu, and other well-known high-tech companies with large data are scrambling to invest their resources into occupation of deep learning technology.

In this study, we proposed a deep learning model of an encoder–decoder with a residual network (EDRN) for fusing hyperspectral and panchromatic images (Supplementary Materials). The advantage of the end-to-end neural network is that the model can be changed from the original input to the final output as much as possible by reducing manual pretreatment and

follow-up processing. This gives the model more space for automatic adjustment according to the data and increases the overall fit of the model. The proposed method can be divided into three parts: (1) hyperspectral and panchromatic image combination; (2) establishment of the encoder–decoder network, and (3) residual enhancement of the encoded and decoded deep features. To overcome the independence of the hyperspectral and panchromatic image features adopted in the majority of fusion methods, we first combined hyperspectral and panchromatic images by a particular means. The image features then produced interactive effects. The integration mode in our manuscript leads to a more concise combination mode for hyperspectral and panchromatic images to interact spectral-spatial information. Second, we established an encoder–decoder network for extracting the representative encoded and decoded deep features. Our model extracts richly formalized encoded and decoded deep features with different feature sizes for image fusion. Our model extracts more diversified deep features, which allows image fusion with more effective and hierarchically variable feature levels for image fusion. Finally, to solve the lack of deep feature enhancement in the current fusion methods, we established residual networks between the encoder network and the decoder network to enhance the extracted encoded and decoded deep features. Our model achieves residual enhanced encoded and decoded deep features to attain enhanced image fusion result. The establishment of the proposed method is complete. The proposed DL-based image fusion method is able to enhance the features of extracted deep features. At the same time, the proposed method is able to combine hyperspectral and panchromatic images.

In this paper, we propose a novel encoder-decoder with residual network fusion model. The main contributions and novelties of the proposed method are as follow:

(1) Spatial-spectral information interaction. We first up-sampled the hyperspectral image as the size of panchromatic image, then contacted the panchromatic and up-sampled hyperspectral images for information interaction.
(2) One-to-one encoder and decoder construction. For the process of construction, the previous encoded layers were corresponded with the latter decoded layers. Then, we constructed a one-to-one encoder-decoder network to extract encoded and decoded deep features.
(3) Encoded and decoded deep feature enhancement. We utilized the convolutional residual network from the encoded layers to corresponding decoded layers to enhance the encoded and decoded deep features. We computed the encoded deep features with a convolutional implementation and then added them to the corresponding decoded deep features to enhance the entire deep features.

The rest of this article is organized as follows: Section 2 provides a detailed description of the proposed method. Section 3 presents the experimental results. Section 4 represents a discussion. Section 5 summarizes the conclusions.

2. Proposed Method

In this article, we propose an end-to-end DL model for fusing hyperspectral and panchromatic images. "End-to-end" means that the input data of the model are the original raw data, and the output data are the result. Classical machine learning uses the raw original data as features in preprocessing, then utilizes the features in a specific application. The results of the specific application depend on the quality of the image features to a certain degree; earlier machine learning methods spent most of their time on feature design. Machine learning at that time was more appropriately named feature engineering. Later, people found that it would be better to use neural networks and let the network learn how to obtain features by itself. This led to a rise in representation learning, and this method is more flexible for data fitting. With the further deepening of the network, the multilayer concept of representation learning brought the accuracy of the algorithm to a new height and led to multilevel feature extraction as well as recognizer unified training and prediction networks. An end-to-end neural network excels in reducing manual pretreatment and follow-up processing and can be changed from the original input to the final output as

much as possible. It gives the model more space for automatic adjustment according to the data and increasing the overall fit of the model. Features can be learned by themselves, so we integrate feature extraction into the classification algorithm without human intervention. For the end-to-end DL model proposed in this research, the inputs are hyperspectral and panchromatic images, while, the output is the fusion result.

Current DL-based remote sensing image fusion methods usually regard spatial and spectral features as individual units. On the input end of the proposed end-to-end deep learning model, we first regard the hyperspectral and panchromatic images as an entirety. This operation is carried out for the sake of combining the spectral information in the hyperspectral images and the spatial information in the panchromatic images as an entirety. In addition, this spectral–spatial combination entity allows feature interaction between the spectral and spatial information in the later DL model. In this operation, the hyperspectral image is up-sampled as the spatial size of the panchromatic image. To overcome the problem that current DL-based image fusion methods usually regard spatial and spectral features as individual units, the integration of up-sampled hyperspectral and panchromatic images can provide them with spatial-spectral information interaction for image fusion. The integration mode in our manuscript leads to a more concise combination mode for hyperspectral and panchromatic images to interact spectral-spatial information. After up-sampling the hyperspectral image, we contacted the panchromatic image with the up-sampled hyperspectral image according to Equation (1):

$$X_{cat} = Cat\left(X_{HSI,up-sampled}, X_{pan}\right) \tag{1}$$

where X_{pan} is the panchromatic image and $Cat(*)$ represents the combination of the up-sampled hyperspectral and panchromatic images.

Next, we took X_{cat} as the input data for the deep learning model. After matching the panchromatic and up-sampled hyperspectral images, we used an encoder–decoder network is for elementary deep feature representation of the spectral–spatial interactive input data. An encoder–decoder is an artificial neural network used in supervised learning. The encoder is the first half of an encoder–decoder, and its function is to turn the input data into the middle layer to produce a hidden representation. This part transforms the deep representative features of input data to a hidden representation. These deep representative features are also called encoded deep features. The decoder is the back half of an encoder–decoder and its function is to refactor the hidden representation from the middle layer to the output data. This part extracts the deep representative features from the hidden representation to the output data. These deep representative features are also called decoded deep features. There are several categories of encoder–decoder: (1) ordinary encoder–decoders, which are neural networks with three layers (i.e., a neural network with a hidden layer); (2) multilayer encoder–decoders, which are extended from ordinary encoder–decoders to a encoder–decoder with multiple hidden layers, and (3) a convolutional encoder–decoder, which is extended from a fully connected network to a convolution network. We utilized the convolutional encoder–decoder in our proposed DL model. A convolutional neural network (CNN), which is a type of deep learning method, was adopted to establish a deep network fusion model for the intelligent fusion of hyperspectral and panchromatic images in this study.

CNN is one of the representative deep learning algorithms and is a feedforward neural network containing convolution computation [42]. In the convolution layer of CNN, we only connected a neuron with the neurons of the adjacent layers, usually containing multiple feature planes. Neurons in the same feature plane share the same weights, i.e., convolution kernels. Subsampling, also known as pooling, usually takes two forms: average subsampling and maximum subsampling. Subsampling can be regarded as a special convolution process. The output result of convolution forms a layer of neurons through the activation function to form a layer characteristic map.

Figure 1 illustrates a schematic diagram of the encoder–decoder network. The encoder–decoder network adopted in our proposed DL model comprised three operations: con-

volution, pooling and up-sampling. In Figure 1, *HSI* refers to the hyperspectral image, *PAN* represents the panchromatic image, *conv* represents the convolution operation, *pool* represents the max-pooling operation and *up* represents up-sampling.

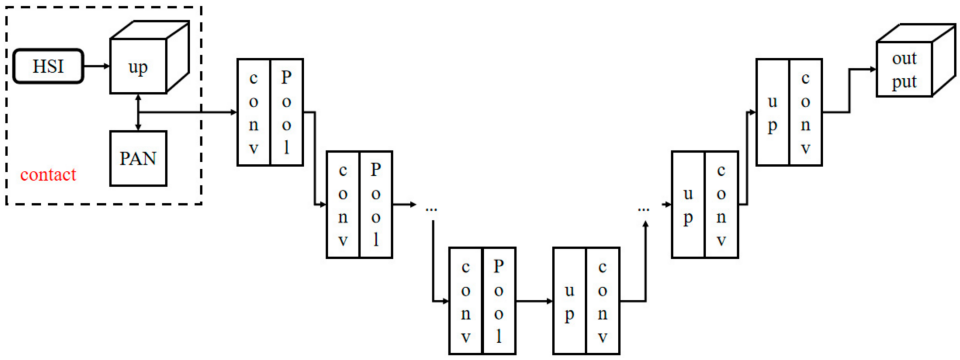


Figure 1. Schematic diagram of the encoder–decoder network.

The current DL-based image fusion methods usually lack richly formalized and diversified deep features, so we constructed an encoder-decoder network to extract richly formalized and diversified deep features. In Figure 1, there are two parts in the encoder–decoder network. The left part is the encoder network, while, the right part is the decoder network. The encoder network includes a series of convolution layers and pooling layers. The feature sizes of the convolution layers diminish gradually because of the pooling operation. The decoder network constitutes a series of up-sampling and convolution layers. The feature sizes of the convolution layers increase gradually because of up-sampling. In the encoder network, we convoluted and pooled the combined hyperspectral and panchromatic image data layer by layer to form the encoded deep features. Next, we extracted the encoded deep features from the combined hyperspectral and panchromatic image data and completed the establishment for the encoder network. Equation (2) shows the convolution operation of the encoder network:

$$\begin{aligned} X_{encoder}^{l+1} &= conv2d\left(w_{encoder}^{l+1}, X_{encoder}^{l}\right) + b_{encoder} \\ &= \left[X_{encoder}^{l} \otimes w_{encoder}^{l+1}\right](i,j) + b_{encoder} \end{aligned} \tag{2}$$

where $conv2d(*)$ and $\otimes$ is the convolution operation, $b_{encoder}$ is the deviation value and $X_{encoder}^{l}$ and $X_{encoder}^{l+1}$ represent the input and output of the $l+1$ level convolution in the encoder network, respectively. Equation (3) shows the pooling operation in the encoder network:

$$\begin{aligned} X_{encoder,k}^{l+1}(i,j) &= pool\left(X_{encoder}^{l}, s_0, p\right) \\ &= \left[\sum_{x=1}^{f}\sum_{y=1}^{f} X_{encoder,k}^{l}(s_0 i + x, s_0 j + y)^p\right]^{\frac{1}{p}} \end{aligned} \tag{3}$$

where $pool(*)$ represents the pooling function, $X_{encoder}^{l}$ and $X_{encoder}^{l+1}$ refer to the input and output of the $l+1$ level's convolution in the encoder network, the step length is s_0 and pixel (i,j) has the same meaning as the convolution layer. When $p=1$, L_p pooling takes the mean value in the pooled area, i.e., average pooling; when $p \to \infty$, L_p pooling takes the maximum value in the region, i.e., max-pooling. At the same time, in the decoder network, we subjected the middle layer between the encoder and decoder networks to the up-sampling and convolution operation layer by layer to form the decoded deep features. The decoded deep features are then extracted from the middle layer between the encoder and decoder networks to complete the establishment for the decoder network. Equation (4) shows the convolution operation of the decoder network:

$$\begin{aligned} X_{decoder}^{l+1} &= conv2d\left(X_{decoder}^{l}, w_{decoder}^{l+1}\right) + b_{decoder} \\ &= \left[X_{decoder}^{l} \otimes w_{decoder}^{l+1}\right](i,j) + b_{decoder} \end{aligned} \tag{4}$$

where $X_{decoder}^{l}$ and $X_{decoder}^{l+1}$ represent the input and output of the $l+1$ layer's convolution in the decoder network.

There are the same number of levels in the encoder and decoder networks. Each layer from the encoder network and each layer in the decoder network correspond one to one. For the encoder network, we used the pooling operation between layers to obtain encoded convolution feature blocks with different sizes and dimensions for each layer. For the decoder network, we obtained the decoded convolution feature block with the same size as the corresponding encoded convolution feature block by using up-sampling between layers. In this way, we established the encoder network and the decoder network with the corresponding feature blocks of the same size and dimension. To overcome the problem that current DL-based image fusion methods usually lack of richly formalized and diversified deep features, our model extracts richly formalized encoded and decoded deep features with different feature size for image fusion. Our model extracts more diversified deep features that makes image fusion with more effective and hierarchically variable feature levels for image fusion.

The current DL-based image fusion methods usually lack deep feature enhancement, so we utilized residual network to enhance the extracted encoded and decoded deep features. Figure 2 is a schematic diagram of residual enhancement for the encoder network and the decoder network. In Figure 2, the plus sign represents the residual block. This study used a residual network structure to adjust and enhance the encoded and decoded deep features. For the residual network, denoted as the desired underlying mapping function $H(*)$, we let the stacked nonlinear layers fit other mapping, as shown in Equation (5):

$$F(x') = H(x|x') - x \tag{5}$$

where $F(x')$ is the residual part, x is the mapping part, x' is a residual variable and x is a mapping variable. The original mapping was recast into Equation (6):

$$H(x|x') = x + F(x') \tag{6}$$

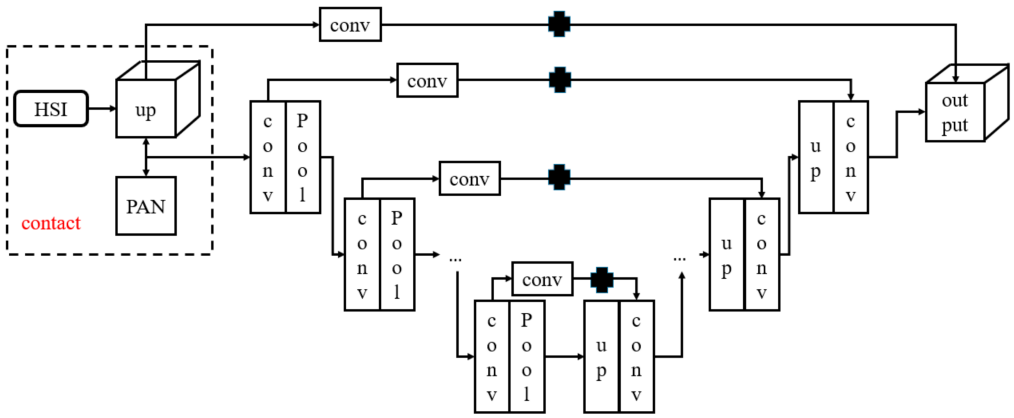


Figure 2. Schematic diagram of residual enhancement for the encoder network and the decoder network.

The formulation of $x + F(x')$ can be realized by feedforward neural networks with shortcut connections. The shortcut connections simply perform identity mapping. We also added their outputs to the outputs of the stacked layers. Identity shortcut connections add neither extra parameters nor computational complexity [43]. For the convolutional residual network with Equation (6), there is a convolutional operation in the residual part $F(x')$ and an identity mapping with the mapping part x. We adopted a convolutional residual network in the proposed EDRN method. The residuals of the network structure were

established on the decoder network at each level of convolution to join the corresponding encoder's convolution. When establishing the encoder and decoder networks, we formulated the corresponding one-to-one encoded and decoded convolutions. For establishing the residual enhancement structure between the encoder and decoder networks, we added each encoded convolution layer to the corresponding decoded convolution layer with the same convolution feature size shown in Figure 2. Before the addition operation, there was a convolution operation to enhance the residual network.

The residual network constitutes a series of residual blocks that constitute mapping and residual parts. Equation (7) shows the operation of the residual enhancement network structure:

$$X_{decoder,resi}^{L-l} = H\left(X_{encoder}^{l} \middle| X_{decoder}^{L-l}\right) = X_{decoder}^{L-l} + F(X_{encoder}^{l}) \tag{7}$$

where $X_{decoder}^{L-l}$ is the result of the convolution at the lth level in the decoder network, which is the mapping part for identity mapping in the residual block; $F(X_{encoder}^{l})$ is the residual part in the residual block according to Equation (6); $X_{decoder,resi}^{L-l}$ is the result of the enhanced residuals of the convolution of the $L-l$ layer in the decoder network, and L is the total number of convolution layers in the decoder network. For convolution of the residual network, Equation (8) represents $F(X_{encoder}^{l})$ as a convolution operation:

$$\begin{aligned} F(X_{encoder}^{l}) &= conv2d\left(X_{encoder}^{l}, w_{encoder}^{l}\right) + b_{encoder} \\ &= \left[X_{encoder}^{l} \otimes w_{encoder}^{l}\right](i,j) + b_{encoder} \end{aligned} \tag{8}$$

where $w_{encoder}^{l}$ is the convolution weight of the lth residual part and $b_{encoder}$ is the biases of the lth residual part. By substituting Equation (8) into Equation (7), Equation (9) obtains $X_{decoder,resi}^{L-l}$:

$$X_{decoder,resi}^{L-l} = X_{decoder}^{L-l} + \left[X_{encoder}^{l} \otimes w_{encoder}^{l}\right](i,j) + b_{encoder} \tag{9}$$

To overcome the problem that the current DL-based image fusion methods usually lack deep feature enhancement, our model achieves residual enhanced encoded and decoded deep features to attain enhanced image fusion result.

For the encoder and decoder network, we set each layer of the encoded and decoded layers with specific spatial size and number of layer channels. We constructed the final data cube with the final layer of the decoder network. For the final layer of the decoder network, the spatial size was the same with the spatial size of the panchromatic image, and we set the number of final layer channels as the band number of the hyperspectral image. Then, the spatial and spectral resolution of the final layer of the decoder network was the same as the ground truth image. We constructed the final data cube with the above implementation so that it could be used in loss function computed with the ground truth image.

3. Results

This section includes a series of experiments used to verify and evaluate the fusion performance of the proposed EDRN method. The components of this section are as follows.

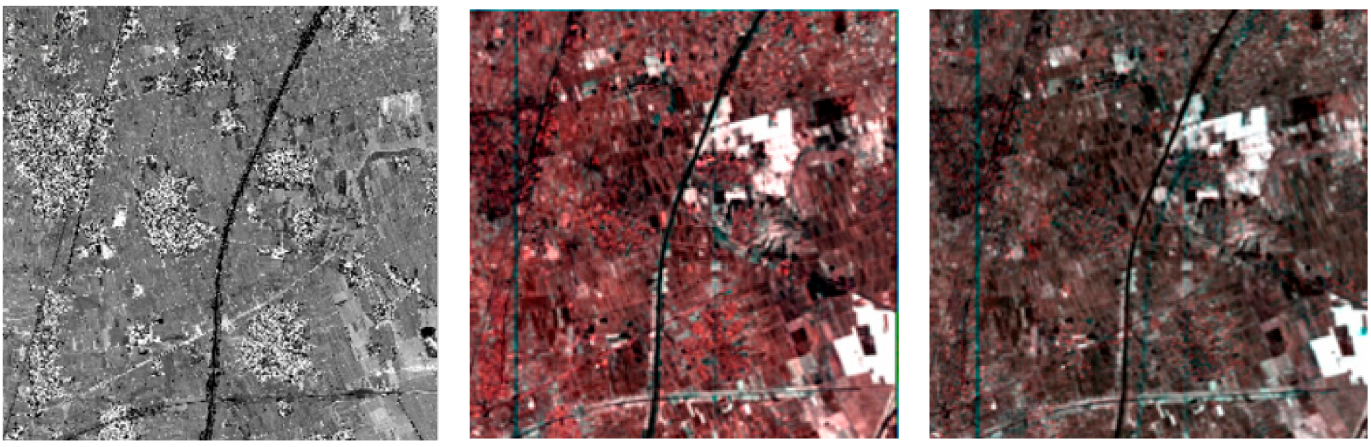
(1) Description of the hyperspectral and panchromatic datasets used for verifying and investigating the performance of the proposed EDRN method.
(2) A comparison of the experimental hyperspectral and panchromatic datasets for the proposed EDRN method versus other fusion methods.

3.1. Description of the Experimental Datasets

We utilized six groups of real-world hyperspectral and panchromatic datasets in our experiments to verify and investigate the performance of the proposed EDRN method. The three datasets had different characteristics in terms of image coverage and the distribution and clutter of ground land cover.

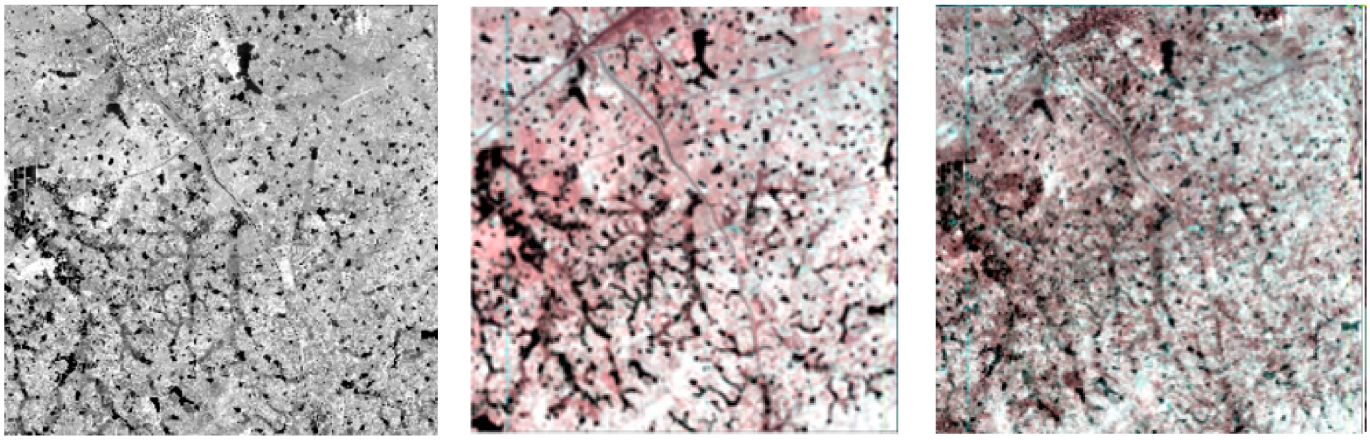
We obtained the groups of datasets using ZY-1E hyperspectral and panchromatic remote sensors. The hyperspectral sensor contains 90 spectral channels with a 30-m spatial resolution. After removing the bands of low signal-to-noise, poor quality and water absorption (22–29, 49–58 and 88–90), there were 68 bands remaining. The panchromatic remote sensor contains relatively unambiguous optical effects for the ground land cover with a 2.5 m spatial resolution. The size of all panchromatic datasets was 3600 × 3600 pixels and that of the hyperspectral datasets was 300 × 300 pixels.

We collected the first hyperspectral and panchromatic dataset in the Baiyangdian region located at Hebei, China, which includes ground land cover of shadows, croplands, roads and buildings. The panchromatic, hyperspectral and ground truth images in this region are shown in Figure 3. We collected the second dataset from the Chaohu region in the middle range of the Yangtze River, China, which includes ground land cover of croplands, mountains, roads and water. The panchromatic, hyperspectral and ground truth images of the Chaohu region are shown in Figure 4. We collected the third dataset from Dianchi region in Kunming, China, which includes ground land cover of jungles, rivers, water and mountains. The panchromatic, hyperspectral and ground truth images of this region are shown in Figure 5. The RGB images of hyperspectral data and ground truth data are constituted with three bands for illustration. For the RGB images, we chose bands 11, 6, 2 as the red, green and blue channels for illustration. For all three datasets, the ground truth images had the same spatial resolution as respective panchromatic images, and the ground truth images had a size of 3600 × 3600 pixels. Meanwhile, the ground truth images had the same spectral resolution as the respective hyperspectral image. All the ground truth images had 68 spectral bands (after removing the bands of low signal-to-noise, poor quality and water absorption). We obtained all the ground truths of the three datasets by unmanned aerial vehicles with remote sensing image sensors. The sensor had the same retrievable spectral resolution and wave range with the hyperspectral images of the three datasets in our experiments. Meanwhile, with the low altitude flight of the unmanned aerial vehicle, the images obtained from this sensor had very high spatial resolution, which was same as the panchromatic images of the three datasets in our experiments.



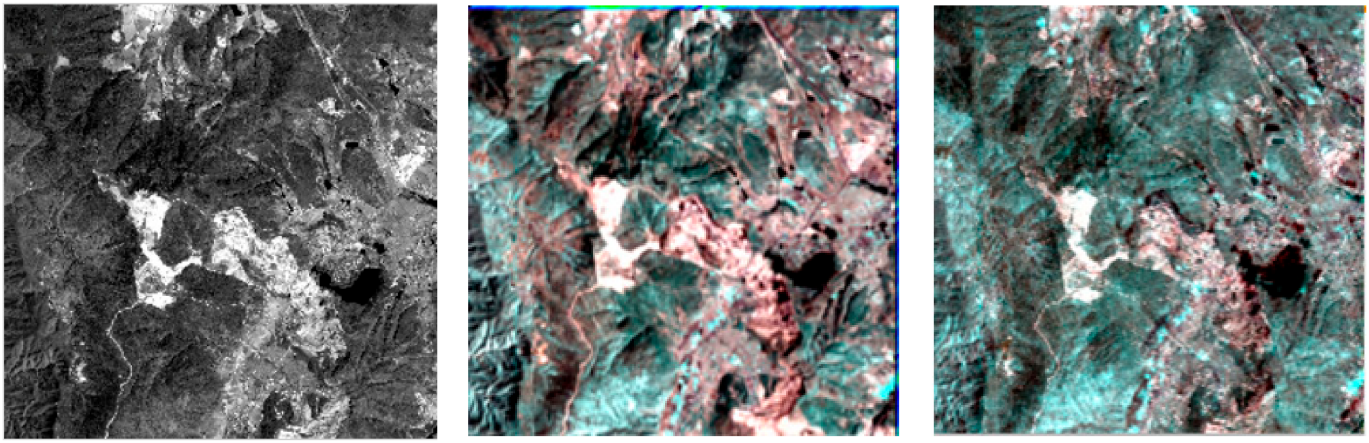
(**a**) (**b**) (**c**)

Figure 3. The Baiyangdian dataset: (**a**) Panchromatic image, (**b**) RGB image of hyperspectral data and (**c**) RGB image of ground truth data.



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

Figure 4. The Chaohu dataset: (**a**) Panchromatic image, (**b**) RGB image of hyperspectral data and (**c**) RGB image of ground truth data.



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

Figure 5. The Dianchi dataset: (**a**) Panchromatic image, (**b**) RGB image of hyperspectral data and (**c**) RGB image of ground truth data.

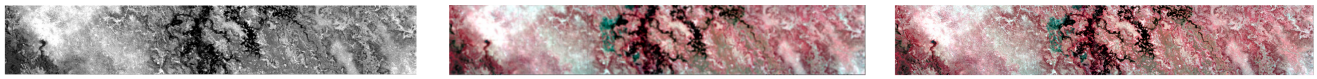
In addition, we utilized three other datasets in our experiments: (1) The Pavia Center scene was captured by the ROSIS camera. The original HSI consists of 115 spectral bands spanning from 430 to 960 nm and has 1096 × 1096 pixels with the spatial resolution of 1.3 m. Thirteen noisy bands were discarded, resulting in an HSI with 102 spectral bands spanning from 430 to 860 nm. In addition, a rectangular area with 1096 × 381 pixels with no information at the center of the original HSI was discarded, and the resulting "two-part" image with size of 1096 × 715 × 102 was used for the experiments. We also used only the top-left corner of the HSI with a size of 960 × 640 × 102, and partitioned it into 24 cubic patches of size 160 × 160 × 102 with no overlap. To generate panchromatic (PAN) images and low spatial resolution hyperspectral images (LR-HSI) corresponding to each high spatial resolution hyperspectral image (HR-HSI), we utilized Wald's protocol [44]. Following the Wald's protocol, we generated PAN images of size 160 × 160 by averaging first 61 spectral bands of HR reference HSI. To generate LR-HSIs of size 40 × 40 × 102, we spatially blurred the HR reference HSI with an 8 × 8 Gaussian filter, and then down-sampled the result with the scaling factor of 4. The panchromatic, hyperspectral and ground truth of Pavia Center dataset are shown in Figure 6. (2) The Botswana scene was acquired by the Hyperion sensor on the NASA's Earth Observing 1 (EO-1) satellite. The original Botswana HSI consisted of 242 spectral bands spanning from 400 to 2500 nm and has 1496 × 256 pixels. We removed the uncalibrated and noisy spectral bands, resulting

in an HSI with 145 spectral bands. We also used only the top-left corner of the HSI with size of 1200 × 240 × 145 and partitioned it into 20 cubic patches of size 120 × 120 with no overlapping that constituted the reference images of the Botswana dataset. To generate PAN images and the LR-HSIs corresponding to each HR reference image, we followed the Wald's protocol. We generated PAN images p of size 120 × 120 by averaging first 31 spectral bands of HR-HSI. To generate LR-HSIs, we spatially blurred the HR-HSI with an 8 × 8 window and performed down-sampling with a factor of 3. The panchromatic, hyperspectral and ground truth of Botswana dataset are shown in Figure 7. (3) The Chikusei scene was captured by a Headwall Hyperspec-VNIR-C imaging sensor over agricultural and urban areas in Chikusei, Japan. The original Chikusei HSI consisted of 128 spectral bands spanning from 363 to 1018 nm and had 2517 × 2335 pixels with a spatial resolution of 2.5 m. We used the top-left corner of the HSI with a size of 2304 × 2304 × 128. Following Wald's protocol, we generated PAN images of size 256 × 256 by averaging first 65 spectral bands of high resolution HSI. To generate LR-HSIs, we spatially blurred the HR-HSI with an 8 × 8 window and performed down-sampling. For the Chikusei dataset, we set the down-sampling factor to 4. The panchromatic, hyperspectral and ground truth of Chikusei dataset are shown in Figure 8.



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

Figure 6. The Pavia Center dataset: (**a**) Panchromatic image, (**b**) RGB image of hyperspectral data and (**c**) RGB image of ground truth data.



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

Figure 7. The Botswana dataset: (**a**) Panchromatic image, (**b**) RGB image of hyperspectral data and (**c**) RGB image of ground truth data.



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

Figure 8. The Chikusei dataset: (**a**) Panchromatic image, (**b**) RGB image of hyperspectral data and (**c**) RGB image of ground truth data.

3.2. Experimental Setup

Experiments on the three datasets were conducted to evaluate the performance of the proposed EDRN method. Six image fusion methods were chosen for the experimental comparison. They were implemented in MATLAB R2016 software and included Coupled Nonnegative Matrix Factorization (CNMF) [24], Modulation Transfer Function–Generalized Laplacian Pyramid (MTF_GLP) [45], General Intensity–Hue–Saturation (GIHS) [46], A Trous Wavelet transform-based Pan-sharpening (AWLP) [47], High Pass Filtering (HPF) [48], Smoothing Filter-based Intensity Modulation (SFIM) [48] and DIP-HyperKite [39] methods. The HPF method choses the simplest scheme achievable by using the box mask and additive injection among the possible couples of filters and coefficients. The box mask was a mask with uniform weights and implement an average. The HPF method uses a procedure to extrapolate edge information from a high-resolution band to lower spatial resolution bands. In the HPF method the higher spatial resolution data, PAN in this case, has a small high-pass spatial filter applied. The results of the small high-pass filter contained the high-frequency component/information related mostly to spatial information. The spatial filter removed most of the spectral information. On the other hand, the SFIM method employed the HPM injection scheme, named smoothing filter-based intensity modulation. The SFIM method was based on a simplified solar radiation and land surface reflection model. By using a ratio between a higher resolution image and its low pass filtered (with a smoothing filter) image, spatial details can be modulated to a co-registered lower resolution multispectral image without altering its spectral properties and contrast.

We implemented the proposed EDRN method with four convolution and pooling layers in the encoder and decoder on all three dataset and set the residual network with the previous encoded layer to the latter decoded layer. We also set the final layer of the decoder network with the number of channels as the number of hyperspectral bands to construct the final output. All the kernels of the convolution layers in both the encoder and decoder had a size of 3×3 kernels. The kernels of the pooling layers in the encoder had a 2×2 kernel size; thus, in the encoder, the output data of one convolution layer and the pooling layer shrank to a quarter of the input data size. We up-sampled all up-sampling layers in the decoder twice for the input data. For the panchromatic image, the input data was a 240×240 image patch, and the input data was a 20×20 image patch for the hyperspectral image. Then the total number of experimental image batches was 78,400. We spat the experimental image batches with 50,176 as the train data, and spat 12,544 as validation data, and spat 15,680 as the test data. For model training, when the model was initializing, we set the weight of each convolutional layer with xavier normalization, the bias of each convolutional layer was set as zero, and the batch size as 32 in our experiments. We trained our model with the Pytorch platform. We utilized a NVIDIA GeForce RTX 3080Ti graphics card to train our model. The input of the network was the contacted panchromatic and up-sampled hyperspectral images. The loss function for our network was the Root Mean Square Error (RMSE). The optimizer for our network was the Adam optimizer.

Figure 9 illustrates the qualitative details of the proposed EDRN model for which the qualitative details are mentioned above.

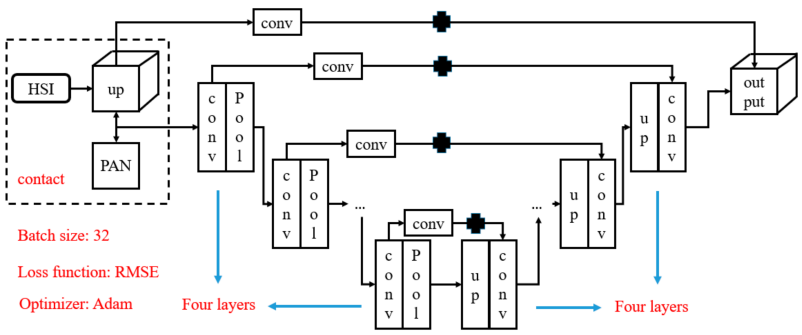


Figure 9. Qualitative details of the proposed EDRN model.

3.3. Experimental Results and Analysis

Figure 10 illustrates the RGB images of the ground truth, the different methods and the proposed EDRN method for the Baiyangdian dataset. For buildings, the AWLP, GIHS, MTF_GLP and the proposed EDRN methods achieved better fusion performance than the CNMF method. The AWLP, MTF_GLP and the proposed EDRN methods achieved better fusion performance than the CNMF and GIHS methods for roads. The AWLP and the proposed EDRN methods achieved better performance than the CNMF, GIHS and MTF_GLP methods for croplands. The GIHS and the proposed EDRN methods achieved better fusion performance than the AWLP, CNMF and MTF_GLP methods for shadows. The GIHS method did not provide reasonable spectral restoration according to the ground truth image. Therefore, the fusion result of the GIHS method achieved poor RGB contrast. The fusion results of the HPF and SFIM methods achieved good restoration for buildings and roads but poor results for croplands and shadows. The fusion result of the DIP-HyperKite method achieved good restoration for buildings, roads and shadows but poor results for croplands. To sum up, the proposed EDRN method achieved the best visual effects in terms of fusion performance among the fusion methods.

Figure 11 shows the RGB images of the ground truth, the different methods and the proposed EDRN method for Chaohu region. The fusion results of the AWLP, CNMF and the proposed EDRN methods were better than those of the GIHS, MTF_GLP methods for the ground truth images for water. For mountain and croplands, the proposed EDRN method produced better restoration than the AWLP and CNMF methods. For all classes of land cover, the fusion results of the GIHS and MTF_GLP methods had worse performance than the other methods and the proposed EDRN method. The AWLP, CNMF and the proposed EDRN methods had better fusion performance than the GIHS and MTF_GLP methods for roads. The fusion results of the GIHS method did not achieve reasonable spectral restoration according to the ground truth image and had poor RGB contrast. For roads, croplands and mountains, the MTF_GLP method had poor fusion performance. The HPF and SFIM methods had poor performance for shadows, roads and croplands but good performance for mountains. The DIP-HyperKite method had good fusion performance for roads, croplands and mountains but poor performance for shadows. In conclusion, for all classes of land cover, the proposed EDRN method had better performance and better sharpness than the other fusion methods.

Figure 12 illustrates the RGB images of ground truth, the different methods and the proposed EDRN method for Dianchi. The AWLP method, the GIHS method and the MTF_GLP method had poor restoration performance for water, while the CNMF and the proposed EDRN methods had good fusion performance. For the water class, the fusion results of the GIHS method had poor RGB contrast and the fusion results of the GIHS method did not show very good spectral performance according to the ground truth image. The CNMF and MTF_GLP methods had poor fusion performance. The AWLP, GIHS and the proposed EDRN methods had better restoration performance than the CNMF and MTF_GLP methods for jungles and mountains. The HPF and SFIM methods had good performance for mountains and water but poor performance for jungles. The DIP-HyperKite method had good performance for water and jungles but poor performance for mountains. In brief, the proposed EDRN method showed better performance for all classes of land cover than the other fusion methods. That is, the proposed EDRN method achieved better results than all the other fusion methods.

Figure 13 illustrates the RGB images of the ground truth, the different methods and the proposed EDRN method for Pavia Center. The AWLP method achieved good performance for buildings, water and vegetations but poor performance for roads. The CNMF method achieved good performance for water and vegetations but poor performance for buildings and roads. The GIHS method achieved good performance for water, buildings and roads but poor performance for vegetations. The MTF_GLP method achieved good performance for roads, water and vegetations but poor performance for buildings. The HPF method achieved good performance for water but poor performance for roads, buildings and

vegetations. The SFIM method achieved good performance for buildings and roads but poor performance for water and vegetations. The DIP-HyperKite method achieved good performance for buildings, water and vegetations but poor performance for roads. The proposed EDRN method achieved good performance for all the land covers.

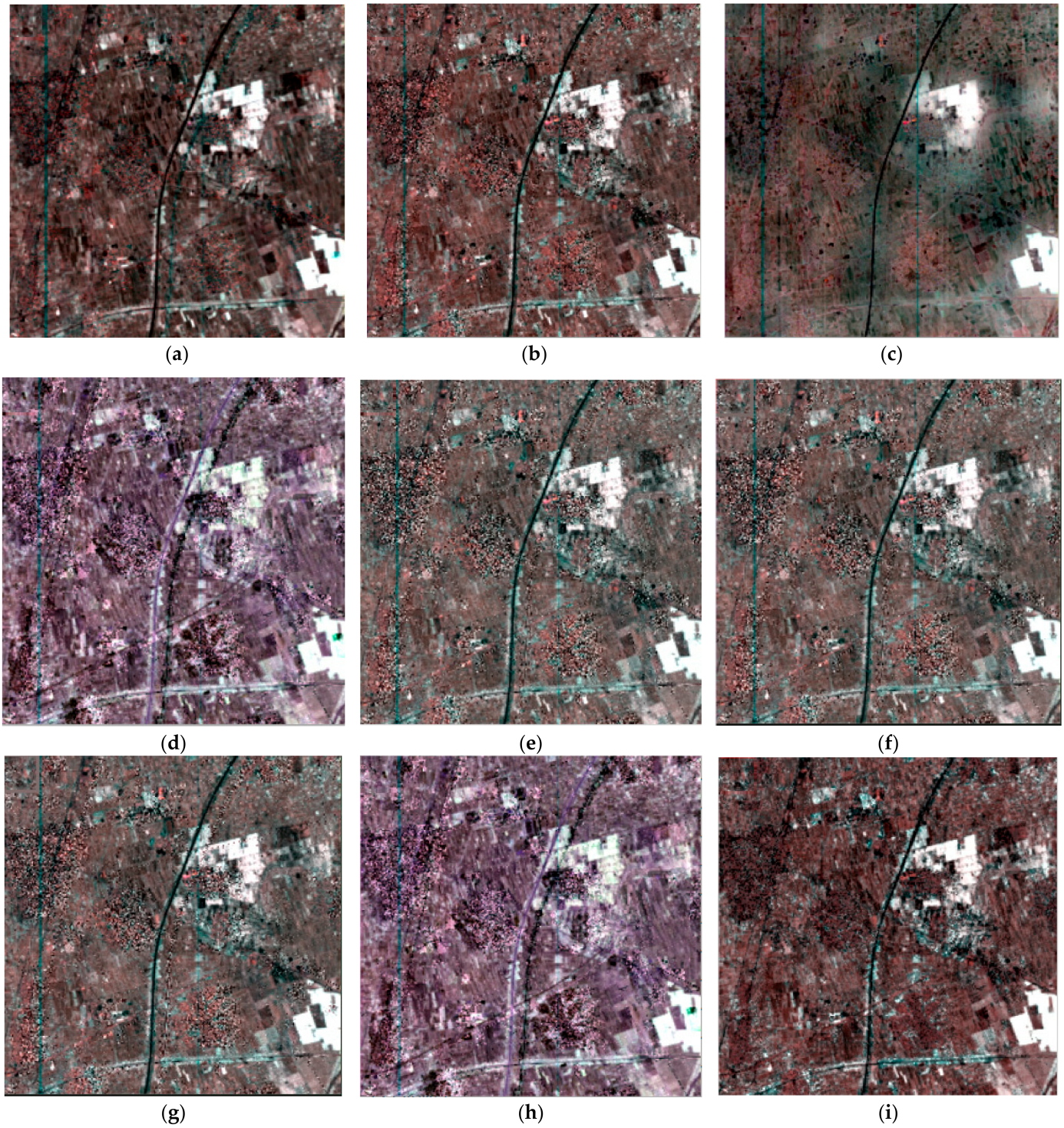


Figure 10. RGB images of ground truth, the different methods and the proposed EDRN method for the Baiyangdian dataset: (**a**) Ground truth, (**b**) AWLP, (**c**) CNMF, (**d**) GIHS, (**e**) MTF_GLP, (**f**) HPF, (**g**) SFIM, (**h**) DIP-HyperKite and (**i**) EDRN.

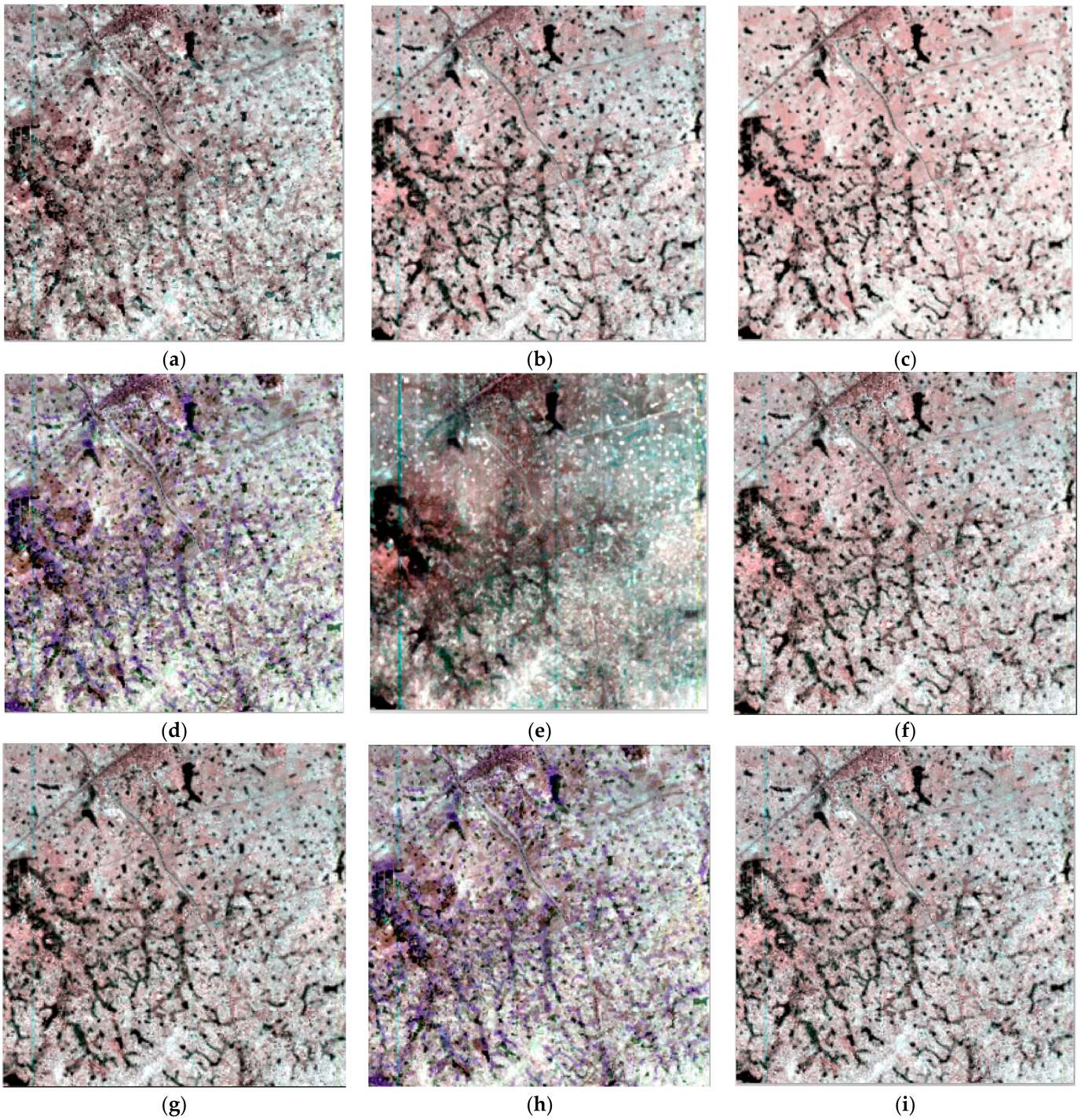


Figure 11. RGB images of the ground truth, the different methods and the proposed EDRN method on the Chaohu dataset: (**a**) Ground truth, (**b**) AWLP, (**c**) CNMF, (**d**) GIHS, (**e**) MTF_GLP, (**f**) HPF, (**g**) SFIM, (**h**) DIP-HyperKite and (**i**) EDRN.

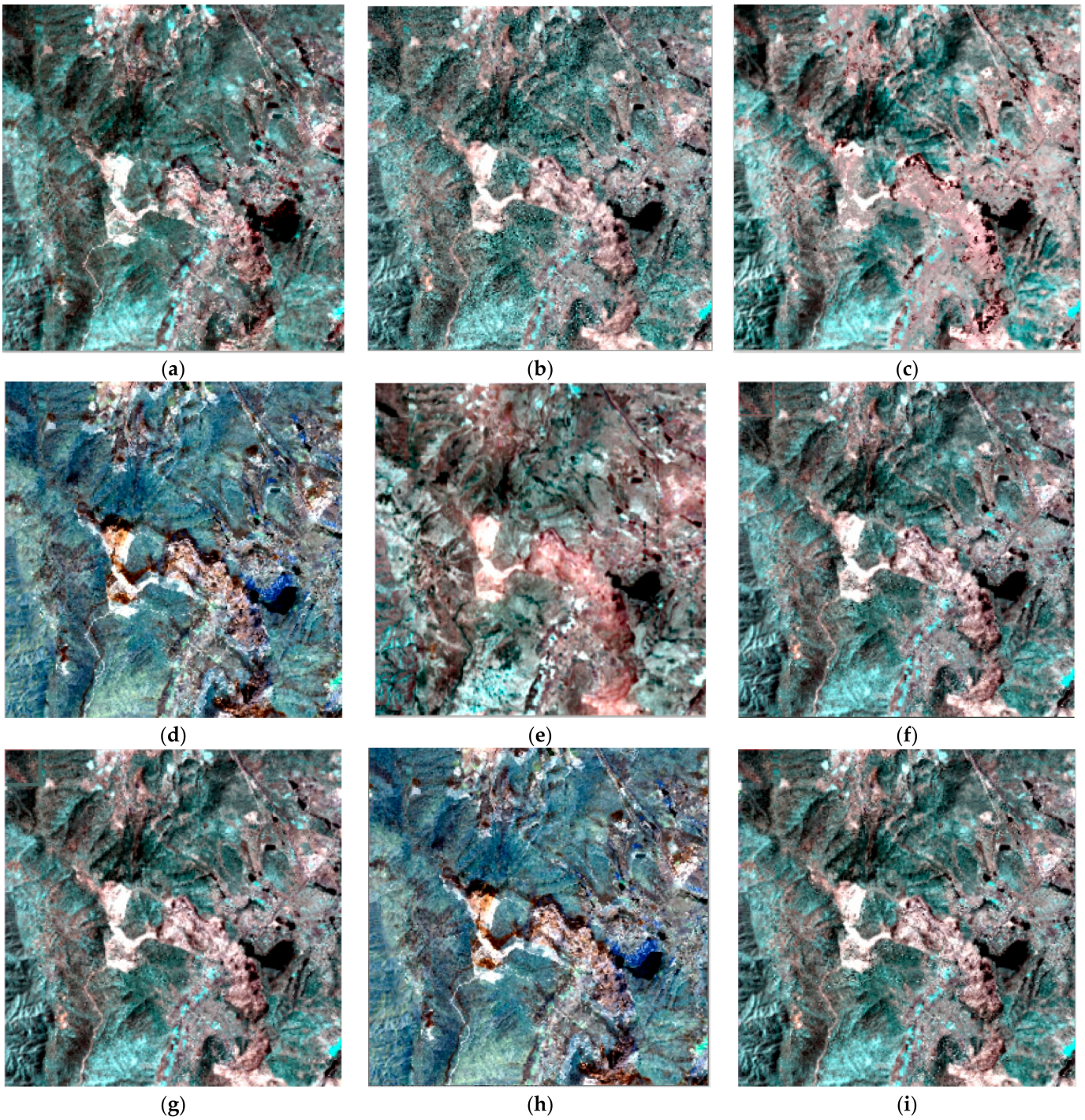


Figure 12. RGB images of ground truth, the different methods and the proposed EDRN method for the Dianchi dataset: (**a**) Ground truth, (**b**) AWLP, (**c**) CNMF, (**d**) GIHS, (**e**) MTF_GLP, (**f**) HPF, (**g**) SFIM, (**h**) DIP-HyperKite and (**i**) EDRN.

Figure 14 illustrates the RGB images of the ground truth, the different methods and the proposed EDRN method for Botswana. The AWLP method achieved good performance for mountains, water and vegetations but poor performance for jungles. The CNMF method achieved good performance for water and vegetations but poor performance for mountains and jungles. The GIHS method achieved good performance for water, mountains and jungles but poor performance for vegetations. The MTF_GLP method achieved good performance for jungles, water and vegetations but poor performance for mountains. The

HPF method achieved good performance for water but poor performance for jungles, mountains and vegetations. The SFIM method achieved good performance for mountains and jungles but poor performance for water and vegetations. The DIP-HyperKite method achieved good performance for mountains, water and vegetations but poor performance for jungles. The proposed EDRN method achieved good performance for all the land covers.

Figure 15 illustrates the RGB images of the ground truth, the different methods and the proposed EDRN method for Chikusei. The AWLP method achieved good performance for buildings, water and croplands but poor performance for bare lands. The CNMF method achieved good performance for water and croplands but poor performance for buildings and bare lands. The GIHS method achieved good performance for water, buildings and bare lands but poor performance for croplands. The MTF_GLP method achieved good performance for bare lands, water and croplands but poor performance for buildings. The HPF method achieved good performance for water but poor performance for bare lands, buildings and croplands. The SFIM method achieved poor performance for all the land covers. The DIP-HyperKite method achieved good performance for buildings, water and croplands but poor performance for bare lands. The proposed EDRN method achieved good performance for all the land covers.

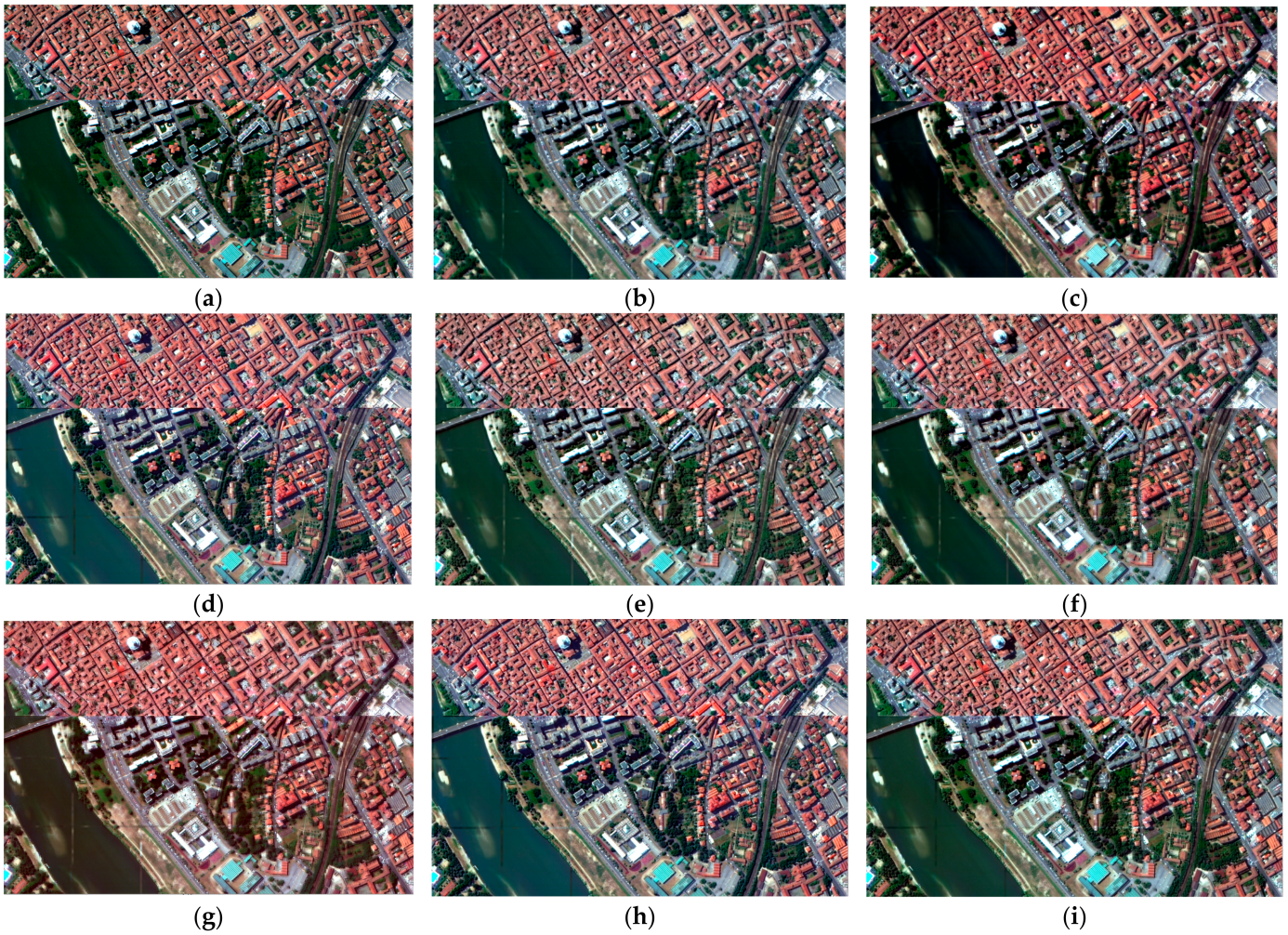


Figure 13. RGB images of ground truth, the different methods and the proposed EDRN method for the Pavia Center dataset: (**a**) Ground truth, (**b**) AWLP, (**c**) CNMF, (**d**) GIHS, (**e**) MTF_GLP, (**f**) HPF, (**g**) SFIM, (**h**) DIP-HyperKite and (**i**) EDRN.

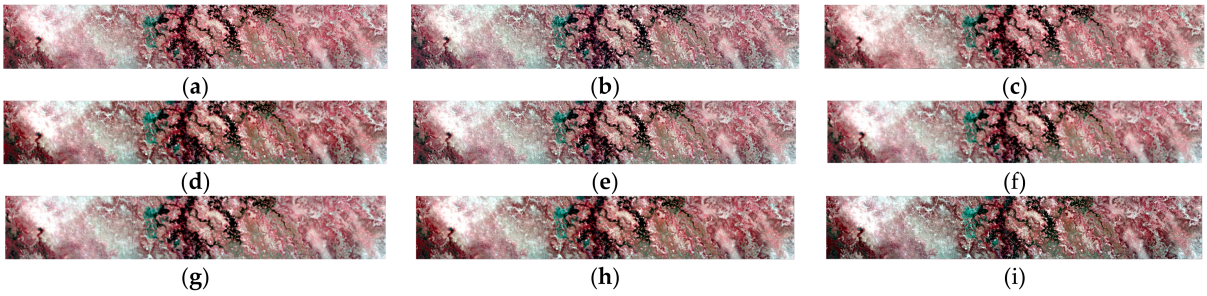


Figure 14. RGB images of ground truth, the different methods and the proposed EDRN method for the Pavia Center dataset: (a) Ground truth, (**b**) AWLP, (**c**) CNMF, (**d**) GIHS, (**e**) MTF_GLP, (**f**) HPF, (**g**) SFIM, (**h**) DIP-HyperKite and (**i**) EDRN.

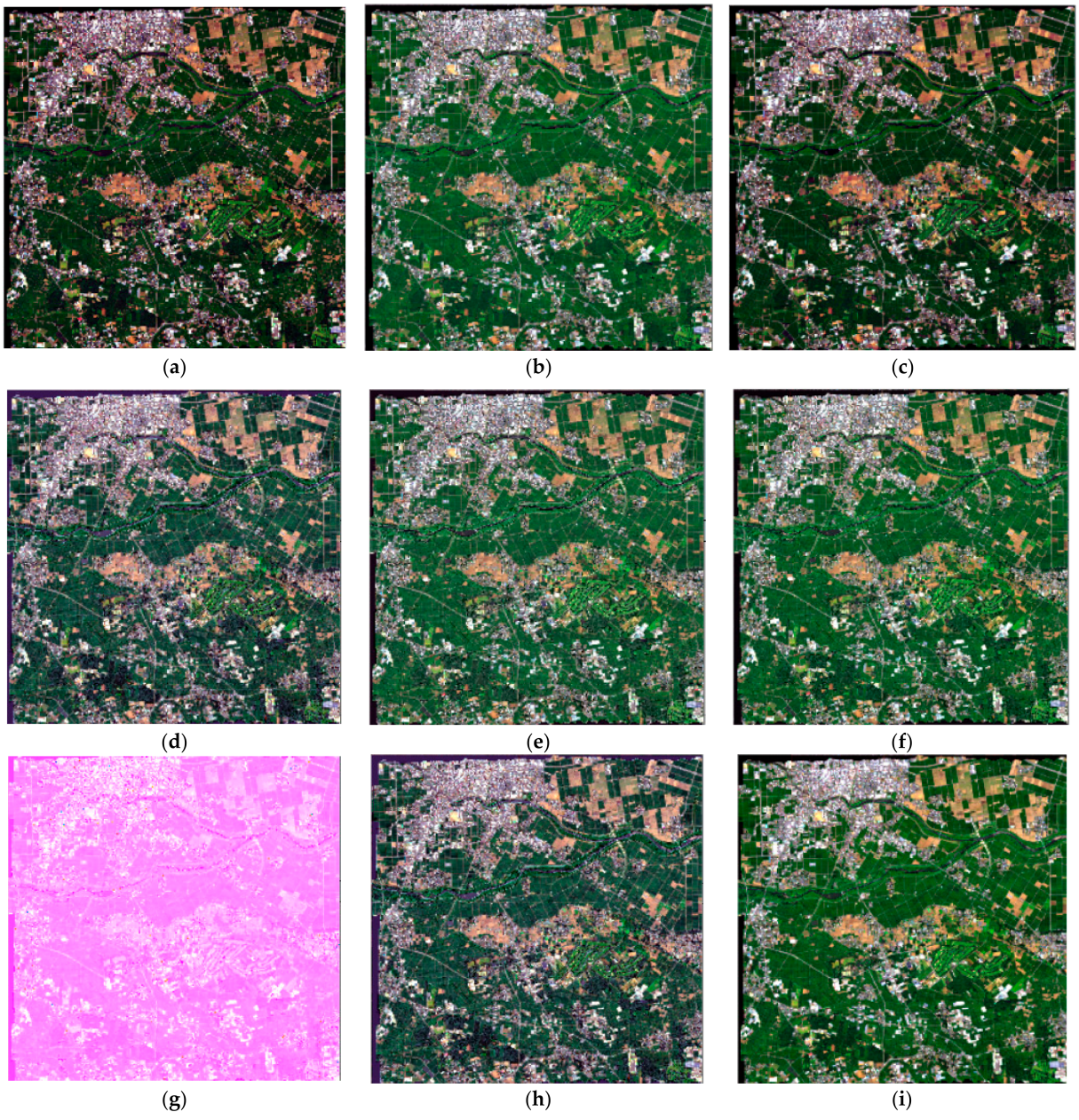


Figure 15. RGB images of ground truth, the different methods and the proposed EDRN method for the Pavia Center dataset: (**a**) Ground truth, (**b**) AWLP, (**c**) CNMF, (**d**) GIHS, (**e**) MTF_GLP, (**f**) HPF, (**g**) SFIM, (**h**) DIP-HyperKite and (**i**) EDRN.

4. Discussion

We utilized a series of image fusion performance indices to precisely verify the spectral and spatial performance of the fusion results in our experiments. We adopted eight performance indices in this study, namely the structural similarity index (SSIM), the peak-signal to noise ratio (PSNR), spectral curve comparison, spatial correlation coefficient (SCC), the spectral angle mapper (SAM), the root mean squared error (RMSE), the relative dimensionless global error in synthesis (ERGAS) and the Q metric [48,49]. We computed the SSIM index as Equation (10):

$$\mathrm{SSIM}(X, X_g) = \frac{1}{p}\sum_{i=1}^{p}\mathrm{SSIM}\left(X^i, X_g^i\right) = \frac{1}{p}\sum_{i=1}^{p}\frac{2u_{x^i}u_{x_g^i}\bullet 2\sigma_{x^i x_g^i}}{\left(u_{x^i}^2 + u_{x_g^i}^2\right)\left(\sigma_{x^i}^2 + \sigma_{x_g^i}^2\right)} \tag{10}$$

where, X is the fused image, X_g is the ground truth, $i = 1, 2, 3 \ldots, p$ with p the band number of hyperspectral image, X^i is the ith band of the fused image, X_g^i is the ith band of the ground truth, u_{x^i} is the average value of X^i, $u_{x_g^i}$ is the average value of X_g^i, $\sigma_{x^i x_g^i}$ is the covariance value between X^i and X_g^i, $\sigma_{x^i}^2$ is the variance of X^i, $\sigma_{x_g^i}^2$ is the variance of X_g^i. The SSIM index is a band-based performance index. We computed the PSNR index as in Equation (11):

$$\mathrm{PSNR}(X, X_g) = \frac{1}{p}\sum_{i=1}^{p}\mathrm{PSNR}\left(X^i, X_g^i\right) = \frac{1}{p}\sum_{i=1}^{p}10log_{10}\left(\frac{n(2^n-1)}{\sum_{j=1}^{n}\left(X_j^i - X_{g,j}^i\right)^2}\right) \tag{11}$$

where, n is the pixel number of fused image, $j = 1,2,3 \ldots, n$, X_j^i is the jth value in X^i, $X_{g,j}^i$ is the jth value in X_g^i. The PSNR index is also a band-based performance index. We computed the SCC index as in Equation (12):

$$\mathrm{SCC}\left(X^i, X_g^i\right) = \frac{n\sum_{j=1}^{n}X_j^i X_{g,j}^i - \sum_{j=1}^{n}X_j^i\sum_{j=1}^{n}X_{g,j}^i}{\sqrt{n\sum_{j=1}^{n}X_j^{i2} - \left(\sum_{j=1}^{n}X_j^i\right)^2}\sqrt{n\sum_{j=1}^{n}X_{g,j}^{i\,2} - \left(\sum_{j=1}^{n}X_{g,j}^i\right)^2}} \tag{12}$$

It is clear that the SCC index is also a band-based performance index. We computed the SAM index as in Equation (13):

$$\mathrm{SAM}(X, X_g) = \frac{1}{n}\sum_{j=1}^{n}arccos\frac{x^{jT}x_g^j}{\|x^{jT}\|_2\|x_g^j\|_2} \tag{13}$$

where, x^j is the jth pixel in fused image, x_g^j is the jth pixel in ground truth, and $\|\ \|_2$ is the L_2 norm. That the SAM index is a pixel-based performance index. We computed the RMSE index as in Equation (14):

$$\mathrm{RMSE}(X, X_g) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(x^j - x_g^j\right)^2} \tag{14}$$

The RMSE index is a pixel-based performance index. We computed the ERGAS index as in Equation (15):

$$\mathrm{ERGAS}(X, X_g) = \frac{100}{d}\sqrt{\frac{1}{p}\sum_{i=1}^{p}\left(\frac{\mathrm{RMSE}\left(X^i, X_g^i\right)}{u_{X_g^i}}\right)^2} \tag{15}$$

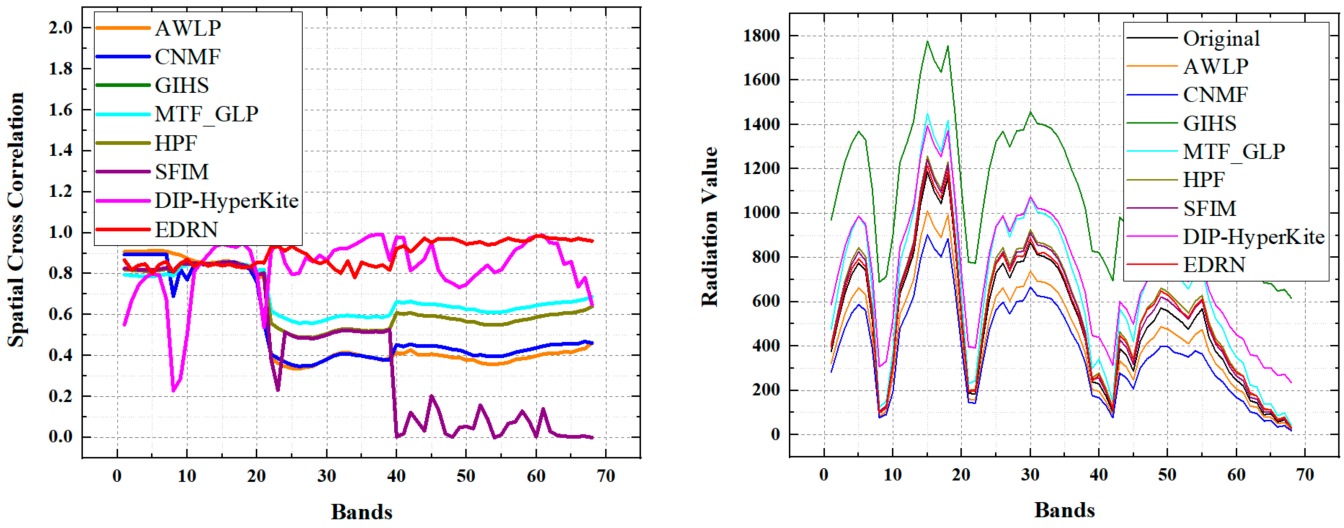
where, d is the spatial downsampling factor. The ERGAS index is a band-based performance index. We computed the Q metric index as in Equation (16):

$$\mathrm{Q}(X, X_g) = \frac{1}{n}\sum_{j=1}^{n}\frac{\sigma_{x^j x_g^j}}{\sigma_{x^j}\sigma_{x_g^j}}\bullet\frac{2u_{x^j}u_{x_g^j}}{u_{x^j}{}^2 + u_{x_g^j}{}^2}\bullet\frac{2\sigma_{x^j}\sigma_{x_g^j}}{\sigma_{x^j}{}^2 + \sigma_{x_g^j}{}^2} \tag{16}$$

The Q metric index is a pixel-based performance index.

Among the performance indices, the SCC and SSIM are spatial quality metrics, and SAM and the spectral curve are spectral quality metrics. The RMSE, PSNR, ERGAS and the Q metric are comprehensive spatial–spectral quality metrics. We utilized the SAM, spectral curve comparison, RMSE, PSNR, ERGAS and Q metric to test the spectral information enhancement in high-resolution images. We also utilized the SCC, SSIM, RMSE, PSNR, ERGAS and Q metric to test the spatial information enhancement in hyperspectral images. The spectral curve comparison compares the spectral curves of a pixel in the results of a fusion method with the corresponding spectral curve of the pixel in the original hyperspectral image. In our experiments, we compared the spectral curve of the (360, 360) pixel in the fusion results of all the compared and proposed methods with the spectral curve of the (30, 30) pixel in the corresponding hyperspectral image for Baiyangdian, Chaohu and Dianchi datasets. For Pavia Center dataset, we compared the spectral curve of the (120, 120) pixel in the fusion results of all the compared and proposed methods with the spectral curve of the (30, 30) pixel in the corresponding hyperspectral image. For the Botswana dataset, we compared the spectral curve of the (90, 90) pixel in the fusion results of all the compared and proposed methods with the spectral curve of the (30, 30) pixel in the corresponding hyperspectral image. For the Chikusei dataset, we compared the spectral curve of the (120, 120) pixel in the fusion results of all the compared and proposed methods with the spectral curve of the (30, 30) pixel in the corresponding hyperspectral image. It is important to emphasize that PSNR, SSIM and the Q metric are better when they are larger, and RMSE, SAM and ERGAS are better when they are smaller. For SCC, the performance is better when most of the SCC values of the bands are bigger than those of the other fusion methods. For spectral curve comparison, the performance is better when the spectral curve is near to the original spectral curve in the hyperspectral images.

Figure 16 shows the quality of the compared and proposed fusion methods for the Baiyangdian dataset in terms of SCC and spectral curve comparison, while Table 1 illustrates the quality of the compared and proposed fusion methods for the Baiyangdian dataset in terms of RMSE, SAM PSNR, SSIM, ERGAS and the Q metric. In Table 1, the best performance for each of all the indices is shown in bold font. The AWLP, MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than the CNMF and GIHS methods to a great degree, while the proposed EDRN method achieved the best RMSE performance, with a RMSE lower than 100. The AWLP, CNMF, MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than the GIHS method to a great degree, while the EDRN method achieved the best performance for the SAM index. The GIHS and the proposed EDRN methods achieved better performance than the other methods, while the proposed EDRN method achieved the best SCC performance in most of the spectral bands. The proposed EDRN method also achieved the best performance for the spectral curve comparison. AWLP, MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than CNMF and GIHS, while the proposed EDRN method achieved the best performance for the PSNR index. GIHS, MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than AWLP and CNMF, while the proposed EDRN method achieved the best performance for the SSIM index. AWLP, MTF_GLP, HPF, DIP-HyperKite and the proposed EDRN methods achieved better performance than the CNMF, GIHS and SFIM methods, while the proposed EDRN method achieved the best performance for the ERGAS index. MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than the AWLP, CNMF, GIHS, HPF and SFIM methods, while the proposed EDRN method had the best performance for the Q metric index. Hence, the proposed EDRN method achieved the best performance for all eight evaluation indices.



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

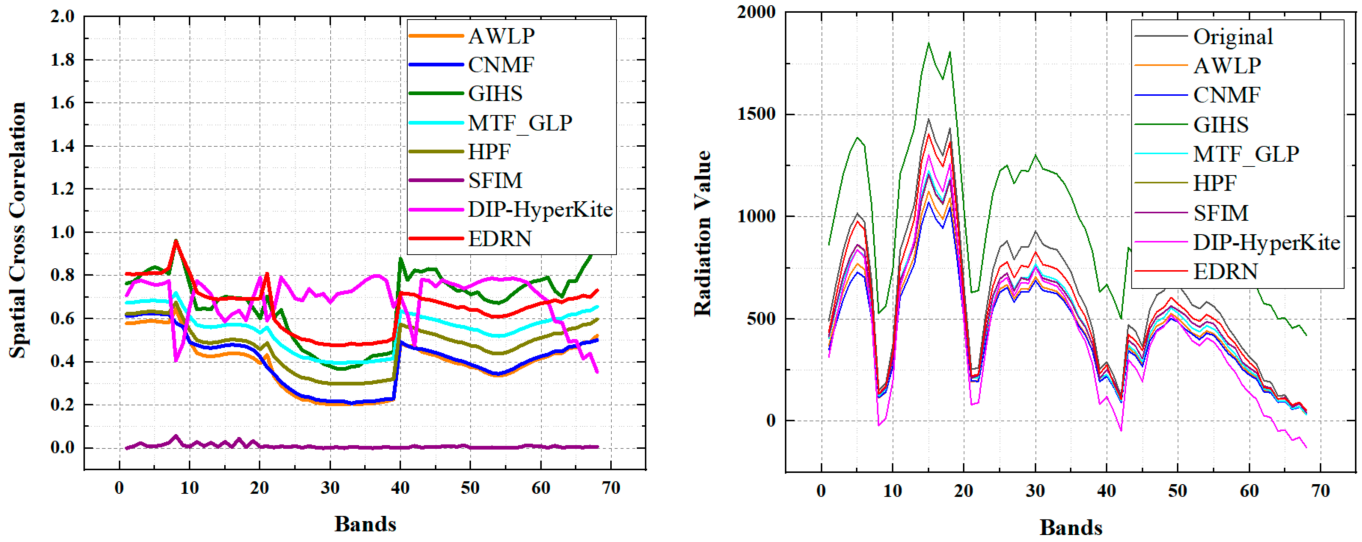
Figure 16. Quality of the compared and proposed fusion methods on the Baiyangdian dataset: (**a**) SCC; (**b**) spectral curve comparison.

Table 1. Quality of the compared and proposed fusion methods on the Baiyangdian dataset for RMSE, SAM, PSNR, SSIM, ERGAS and the Q metric.

Method	RMSE	SAM	PSNR	SSIM	ERGAS	Q Metric
AWLP	111.34	5.1192	24.7554	0.5366	2.1064	0.4092
CNMF	621.19	5.0474	9.6141	0.1542	4.0826	0.4739
GIHS	388.22	12.0829	10.7795	0.6366	4.1850	0.5552
MTF_GLP	102.23	4.7519	24.7507	0.6690	1.8436	0.6234
HPF	105.27	4.8796	24.6616	0.6341	1.9423	0.5736
SFIM	622.49	4.8642	10.7677	0.3848	4.3758	0.5593
DIP-HyperKite	102.21	4.7465	25.1498	0.6835	1.7569	0.6832
EDRN	**95.00**	**4.6514**	**26.9133**	**0.7060**	**1.6884**	**0.7142**

Figure 17 shows the quality of the compared and proposed methods for the Chaohu dataset in terms of SCC and spectral curve comparison, while Table 2 illustrates the quality of the compared and proposed fusion methods for the Chaohu dataset in terms of RMSE, SAM PSNR, SSIM, ERGAS and the Q metric. In Table 2, the best performance of each of the indices is shown in bold font. The AWLP, MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than the CNMF and GIHS methods to a great degree, while the proposed EDRN method achieved the best RMSE performance. The AWLP, CNMF, MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than the GIHS method to a great degree, while the EDRN method achieved the best performance for the SAM index. The proposed EDRN method achieved better performance than all the compared methods in most of the spectral bands for the SCC index. The proposed EDRN method also achieved the best performance for the spectral curve comparison. AWLP, MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than CNMF and GIHS, and the proposed EDRN method achieved the best performance for the PSNR index. GIHS, MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than AWLP and CNMF, and the proposed EDRN method achieved the best performance for the SSIM index. AWLP, CNMF, MTF_GLP, HPF, DIP-HyperKite and the proposed EDRN methods achieved better performance than the GIHS and SFIM methods, while the proposed EDRN method had the best performance for the ERGAS index. GIHS, MTF_GLP, HPF, SFIM, DIP-HyperKite and the proposed EDRN methods achieved better performance than the AWLP and CNMF

methods, while the proposed EDRN method achieved the best performance for the Q metric index. Hence, the proposed EDRN method achieved the best performance for all eight evaluation indices.



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

Figure 17. Quality of the compared and proposed fusion methods on the Chaohu dataset: (**a**) SCC; (**b**) spectral curve comparison.

Table 2. Quality of the compared and proposed fusion methods on the Chaohu dataset for RMSE, SAM, PSNR, SSIM, ERGAS and the Q metric.

Method	RMSE	SAM	PSNR	SSIM	ERGAS	Q Metric
AWLP	168.62	3.6677	21.1784	0.3925	2.7310	0.3818
CNMF	613.49	4.0911	10.6790	0.4393	3.4038	0.3366
GIHS	554.96	15.6016	8.2922	0.6044	4.9276	0.5739
MTF_GLP	154.44	3.5090	21.9589	0.5454	2.4534	0.6327
HPF	162.02	3.6128	21.5446	0.4735	2.6103	0.5465
SFIM	404.52	4.2470	18.7483	0.5868	4.3799	0.5367
DIP-HyperKite	135.21	3.4879	21.9635	0.6263	1.9651	0.6754
EDRN	**119.88**	**3.3733**	**22.2971**	**0.6562**	**1.8949**	**0.7245**

Figure 18 shows the quality of the compared and proposed methods for the Dianchi dataset in terms of the SCC and spectral curve comparison. Table 3 illustrates the quality of the compared and proposed fusion methods for the Dianchi dataset in terms of RMSE, SAM PSNR, SSIM, ERGAS and the Q metric. In Table 3, the best performance for each of all the indices is shown in bold font. The AWLP, MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than the CNMF and GIHS methods to a great degree, and the proposed EDRN method achieved the best RMSE performance. The AWLP, CNMF, MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than the GIHS method to a great degree, and the EDRN method achieved the best performance for the SAM index. The proposed EDRN method achieved better performance than all the other methods in most of the spectral bands for the SCC index. AWLP and the proposed EDRN methods achieved better performance than the CNMF, GIHS and MTF_GLP methods for the spectral curve comparison. The proposed EDRN method achieved the best performance for the spectral curve comparison. AWLP, MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than CNMF and GIHS, and the proposed EDRN method achieved the best performance for the PSNR index. GIHS, MTF_GLP, DIP-HyperKite and the proposed EDRN methods

achieved better performance than AWLP and CNMF, and the proposed EDRN method achieved the best performance for the SSIM index. MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than the AWLP, CNMF, GIHS, HPF and SFIM methods, while the proposed EDRN method achieved the best performance for the ERGAS index. GIHS, MTF_GLP, HPF, SFIM, DIP-HyperKite and the proposed EDRN methods achieved better performance than the AWLP and CNMF methods, while the proposed EDRN method achieved the best performance for the Q metric index.

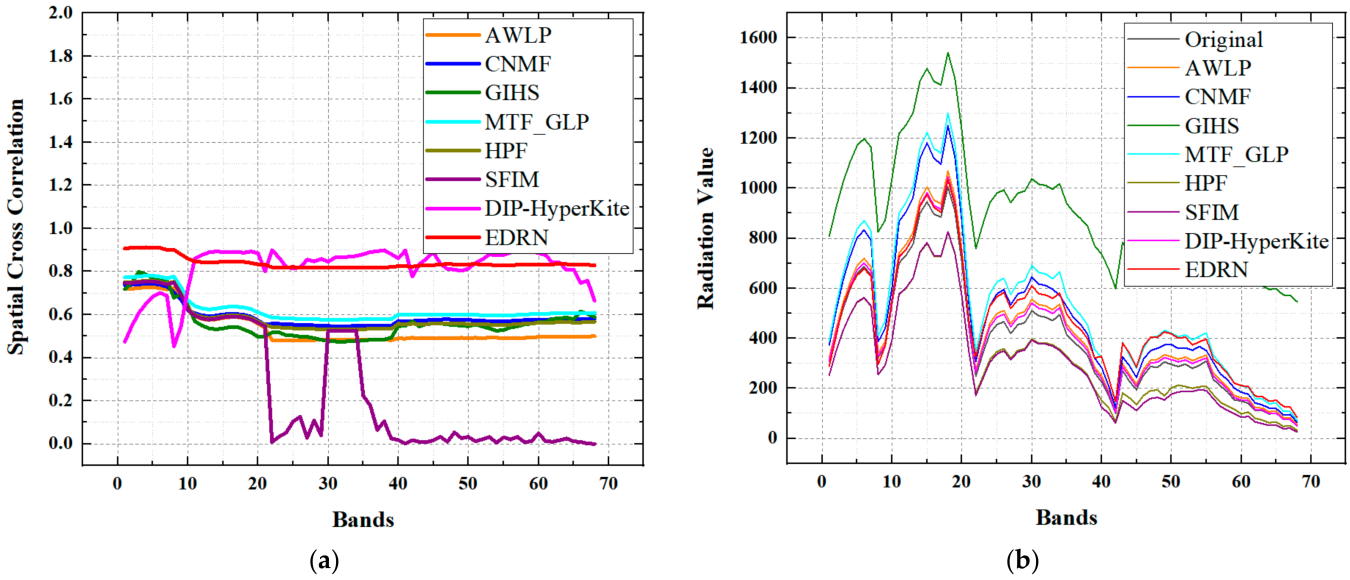


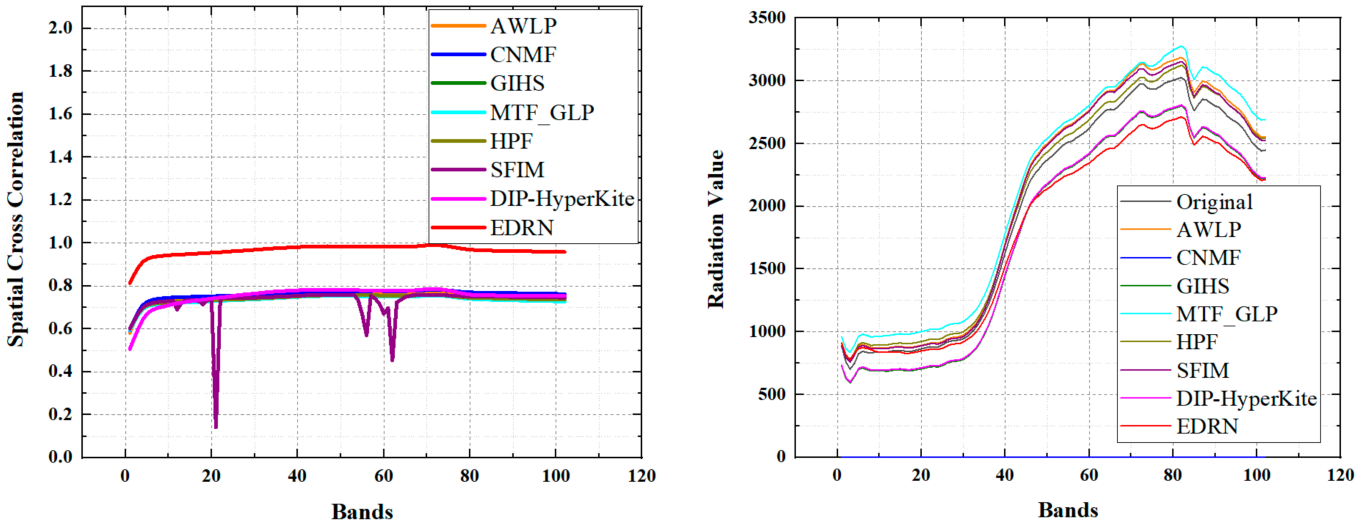
Figure 18. Quality of the compared and proposed fusion methods for the Dianchi dataset: (**a**) SCC; (**b**) spectral curve comparison.

Table 3. Quality of the compared and proposed fusion methods for the Dianchi dataset for RMSE, SAM, PSNR, SSIM, ERGAS and the Q metric.

Method	RMSE	SAM	PSNR	SSIM	ERGAS	Q Metric
AWLP	216.96	7.8959	22.1282	0.5327	4.1298	0.4136
CNMF	614.75	6.9455	12.9519	0.3517	4.5741	0.3810
GIHS	504.64	14.6475	13.0600	0.6715	4.4962	0.6300
MTF_GLP	201.17	7.1726	22.8580	0.6203	3.7628	0.6929
HPF	208.43	7.3917	22.5630	0.5810	3.9549	0.6165
SFIM	204.03	7.2389	14.3158	0.4463	4.1040	0.5839
DIP-HyperKite	195.21	7.0265	22.9435	0.6853	3.3542	0.7236
EDRN	**178.97**	**6.7030**	**22.4281**	**0.7896**	**3.3295**	**0.7418**

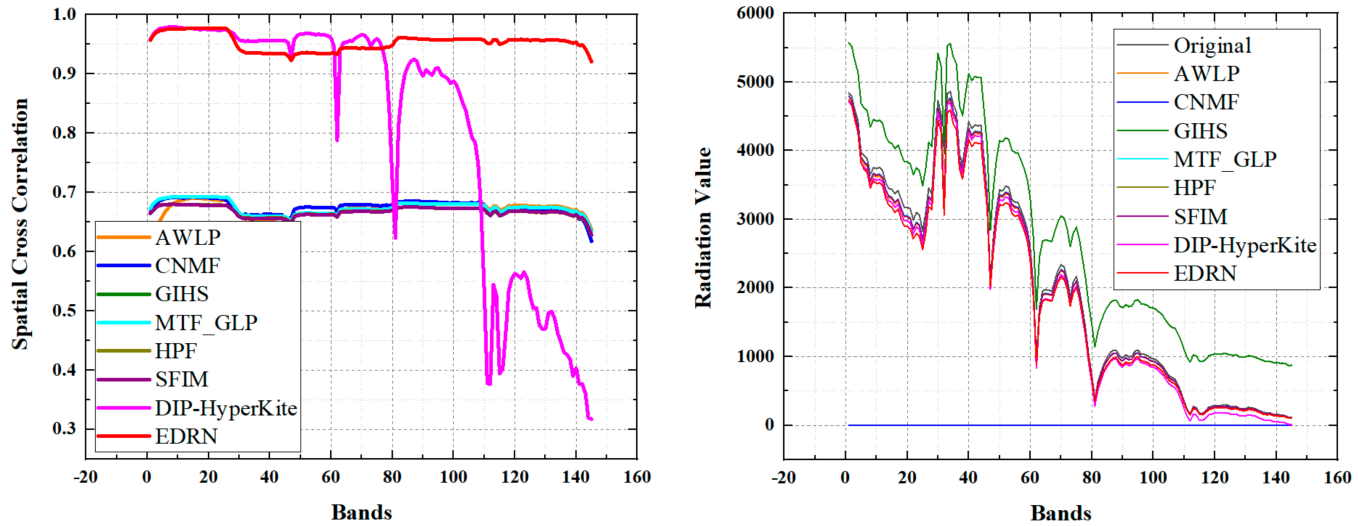
Figures 19–21 show the quality of the compared and proposed methods for the Pavia Center, Botswana and Chikusei dataset in terms of SCC and spectral curve comparison, while Tables 4–6 illustrate the quality of the compared and proposed fusion methods for the Pavia Center, Botswana and Chikusei dataset in terms of RMSE, SAM PSNR, SSIM, ERGAS and the Q metric. In Tables 4–6, the best performance corresponded by each index is shown in bold font. The AWLP, MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than the CNMF and GIHS methods to a great degree, while the proposed EDRN method achieved the best RMSE performance. The AWLP, CNMF, MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than the GIHS method to a great degree, while the EDRN method achieved the best performance for the SAM index. The proposed EDRN method achieved better performance than all the compared methods in most of the spectral bands for the SCC index. The proposed EDRN method also achieved the best performance for the spectral curve comparison. AWLP, MTF_GLP, DIP-HyperKite and the proposed EDRN methods

achieved better performance than CNMF and GIHS, and the proposed EDRN method achieved the best performance for the PSNR index. GIHS, MTF_GLP, DIP-HyperKite and the proposed EDRN methods achieved better performance than AWLP and CNMF, and the proposed EDRN method achieved the best performance for the SSIM index. AWLP, CNMF, MTF_GLP, HPF, DIP-HyperKite and the proposed EDRN methods achieved better performance than the GIHS and SFIM methods, while the proposed EDRN method had the best performance for the ERGAS index. GIHS, MTF_GLP, HPF, SFIM, DIP-HyperKite and the proposed EDRN methods achieved better performance than the AWLP and CNMF methods, while the proposed EDRN method achieved the best performance for the Q metric index. Hence, the proposed EDRN method achieved the best performance for all eight evaluation indices.



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

Figure 19. Quality of the compared and proposed fusion methods for the Pavia Center dataset: (**a**) SCC; (**b**) spectral curve comparison.



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

Figure 20. Quality of the compared and proposed fusion methods for the Botswana dataset: (**a**) SCC; (**b**) spectral curve comparison.

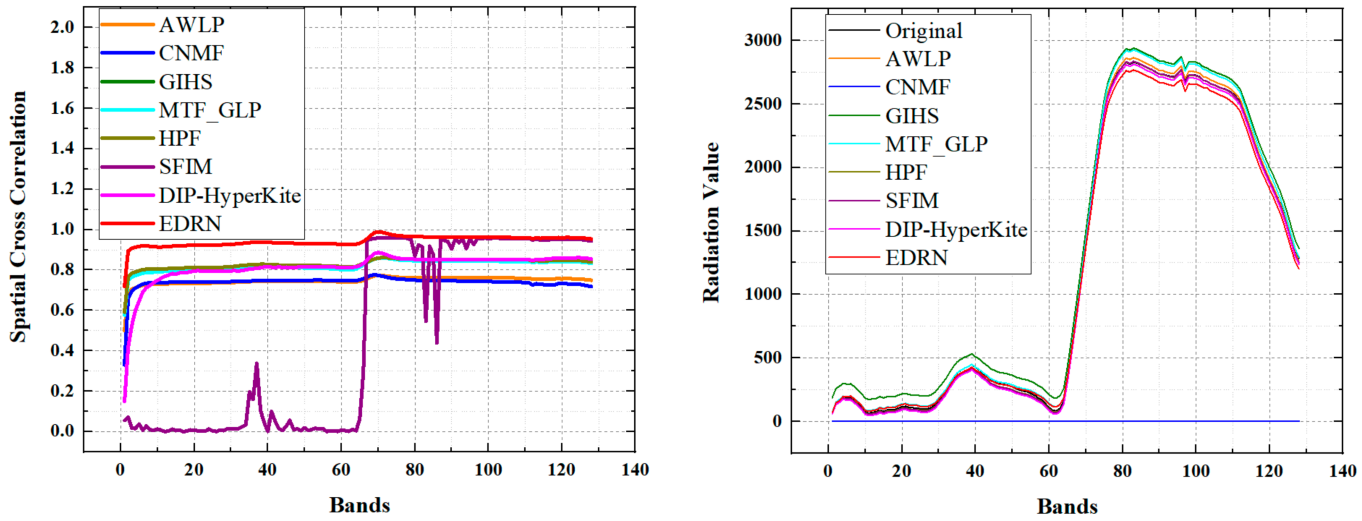


Figure 21. Quality of the compared and proposed fusion methods for the Chikusei dataset: (**a**) SCC; (**b**) spectral curve comparison.

Table 4. Quality of the compared and proposed fusion methods for the Pavia Center dataset for RMSE, SAM, PSNR, SSIM, ERGAS and the Q metric.

Method	RMSE	SAM	PSNR	SSIM	ERGAS	Q Metric
AWLP	237.06	7.6978	25.4689	0.7564	3.5698	0.8765
CNMF	141.46	8.0866	28.5679	0.7231	2.9587	0.1024
GIHS	247.34	9.0839	26.4587	0.7432	3.2569	0.8575
MTF_GLP	275.86	9.6245	27.2685	0.7863	3.5478	0.8583
HPF	263.78	7.8827	35.2654	0.7695	3.6921	0.8563
SFIM	309.36	7.3746	32.5482	0.8154	2.9874	0.8619
DIP-HyperKite	129.02	5.6112	38.6532	0.8532	2.8515	0.8571
EDRN	**125.41**	**4.5655**	**40.5698**	**0.8765**	**2.7451**	**0.9076**

Table 5. Quality of the compared and proposed fusion methods for the Botswana dataset for RMSE, SAM, PSNR, SSIM, ERGAS and the Q metric.

Method	RMSE	SAM	PSNR	SSIM	ERGAS	Q Metric
AWLP	134.68	2.2227	25.4687	0.7632	2.6985	0.9137
CNMF	228.71	2.0591	26.4698	0.7321	2.5487	0.5684
GIHS	376.24	11.2962	27.3654	0.7465	2.1265	0.6054
MTF_GLP	117.01	2.0283	30.6854	0.7795	2.2548	0.9176
HPF	131.11	2.1127	31.6287	0.7832	1.9875	0.8920
SFIM	132.46	2.1454	29.5874	0.7562	2.2031	0.8910
DIP-HyperKite	96.24	1.6827	32.1221	0.7932	1.8924	0.9114
EDRN	**90.79**	**1.4608**	**34.1512**	**0.8165**	**1.7548**	**0.9215**

In terms of the quality evaluation for the compared and proposed EDRN methods on the six datasets, the proposed EDRN method achieved better evaluation performance than all the other fusion methods.

We counted the time cost for taking entire hyperspectral image as up-sampling input and taking hyperspectral image patch as up-sampling input. The time costs are shown in Table 4 as follow.

We set the size of image patches for the hyperspectral image as 20 × 20, and the size of hyperspectral image as 300 × 300. Then, we segmented the hyperspectral image with a total of 255 image patches. As shown in Table 7, taking the entire hyperspectral image as up-sampling input cost a little more time, at roughly 37 s. Then, we computed the time

cost for taking one image patch as up-sampling input, which cost roughly 0.11 s. We also computed the total time cost for taking all image patches as up-sampling input which is that time cost for one image patch multiplied by 255. Taking image patches as up-sampling input saves more time than taking the entire hyperspectral image as up-sampling input.

Table 6. Quality of the compared and proposed fusion methods for the Chikusei dataset for RMSE, SAM, PSNR, SSIM, ERGAS and the Q metric.

Method	RMSE	SAM	PSNR	SSIM	ERGAS	Q Metric
AWLP	205.47	3.6548	35.6254	0.7932	4.6598	0.8457
CNMF	163.21	4.5987	36.5487	0.7521	4.2154	0.6254
GIHS	249.38	3.5874	32.6985	0.7832	4.0215	0.7451
MTF_GLP	219.93	2.9847	40.5632	0.8125	3.8624	0.8945
HPF	199.89	3.0215	39.2541	0.8265	3.7541	0.8865
SFIM	262.82	3.1256	38.3652	0.8032	3.8365	0.8798
DIP-HyperKite	103.32	2.8512	43.5312	0.8763	3.6231	0.9024
EDRN	**100.94**	**2.7546**	**45.1236**	**0.8863**	**3.5362**	**0.9123**

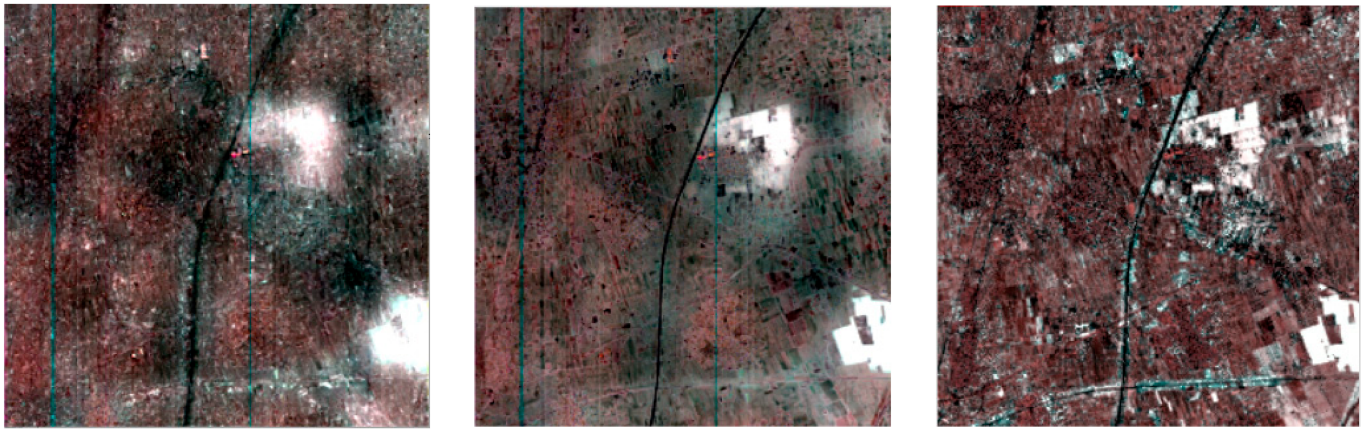
Table 7. Time cost (second) for taking different input for up-sampling.

Time Cost (Second)	Baiyangdian	Chaohu	Dianchi
Entire HSI	37.61	36.79	37.23
One image patch	0.109	0.115	0.114
Image patch for total	27.795	29.325	29.07

We set a series of experiments for comparing up-sampling with nearest, bilinear and bicubic modes that affect the downstream task. The fused results of up-sampling with nearest, bilinear and bicubic modes on Baiyangdian, Chaohu and Dianchi dataset are shown in Figures 22–24.

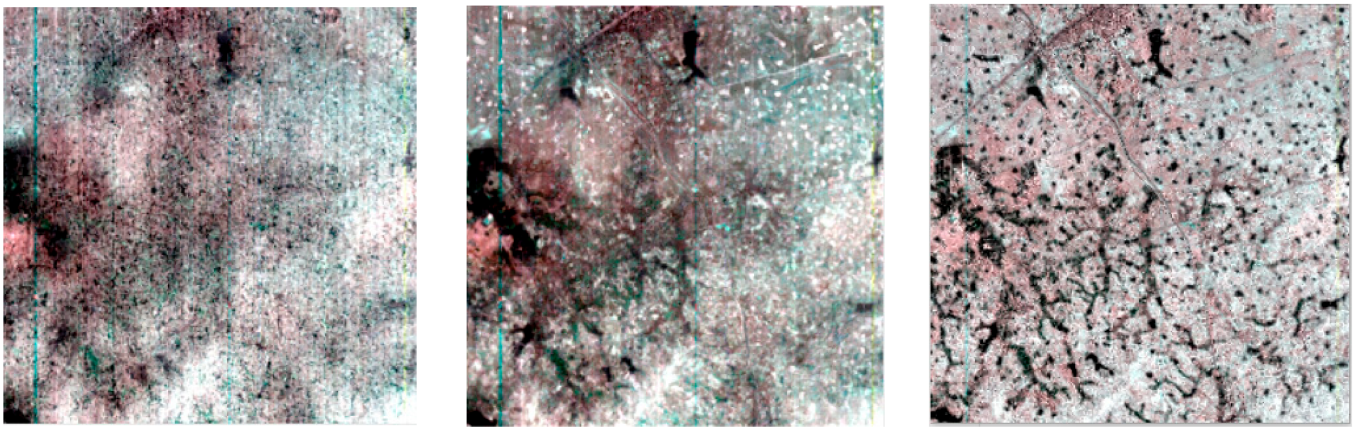
As shown in Figures 22–24, fusion results with up-sampling of the nearest mode have a very fuzzy visual effect, the results with up-sampling of bilinear mode have a little clearer visual effect than that with up-sampling of nearest mode, while the results with up-sampling of bicubic mode have the best visual effect. As shown in Figure 22a, the fusion result with up-sampling of nearest mode achieved fuzzy fusion performance on the land covers of croplands, roads, buildings and shadows in the Baiyangdian region dataset. As shown in Figure 22b, the fusion result with up-sampling of bilinear mode achieved fuzzy fusion performance on the land covers of cropland, buildings and shadows but a little clearer fusion performance on the land cover of road on the Baiyangdian region dataset. As shown in Figure 22c, the fusion result with up-sampling of bicubic achieved clear fusion performance on all of the land covers in the Baiyangdian region dataset. As shown in Figure 23a, the fusion result with up-sampling of nearest mode achieved fuzzy fusion performance on the land covers of croplands, mountains, roads and water in the Chaohu region dataset. As shown in Figure 23b, the fusion result with up-sampling of bilinear mode achieved a fuzzy fusion performance on the land covers of croplands, mountains and roads but a little clearer fusion performance on the land cover of water in the Chaohu region dataset. As shown in Figure 23c, the fusion result with up-sampling of bicubic mode achieved clear fusion performance on all land covers in the Chaohu region dataset. As shown in Figure 24a, the fusion result with up-sampling of nearest mode achieved fuzzy fusion performance on the land covers of rivers, water and mountains in the Dianchi region dataset. As shown in Figure 24b, the fusion result with up-sampling of bilinear mode achieved fuzzy fusion performance on the land covers of mountains but a little clearer fusion performance on the land covers of jungles, rivers and water in the Dianchi region dataset. As shown in Figure 24c, the fusion result with up-sampling of bicubic mode achieved clear fusion performance for all of the land covers in the Dianchi region dataset. It can be concluded that up-sampling with nearest and bilinear modes affected

the downstream task with a comparatively fuzzy up-sampling result. Up-sampling with the bicubic mode reached the sweet spot for the up-sampling step. If the image was up-sampled beyond this point, the performance of the downstream task began to diminish.



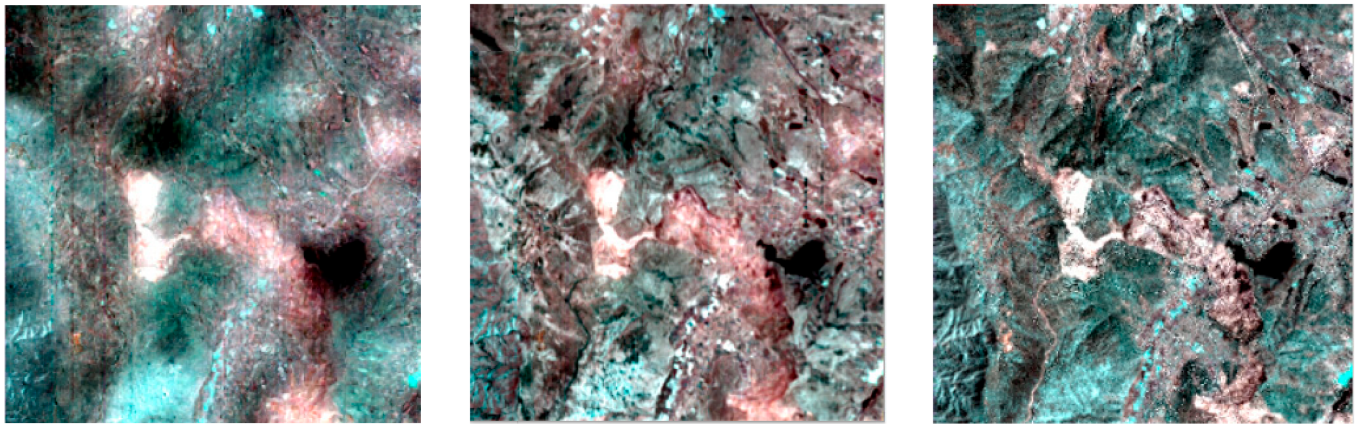
(**a**) (**b**) (**c**)

Figure 22. Different mode of up-sampling for the Baiyangdian dataset. (**a**) Nearest, (**b**) bilinear, (**c**) bicubic.



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

Figure 23. Different mode of up-sampling for the Chaohu dataset. (**a**) Nearest (**b**) bilinear (**c**) bicubic.



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

Figure 24. Different mode of up-sampling for the Dianchi dataset. (**a**) Nearest (**b**) bilinear (**c**) bicubic.

5. Conclusions

In this research, we established a deep network fusion model for the process of fusing hyperspectral and panchromatic images. This method first matches the panchromatic images and up-sampled hyperspectral images to combine the original spectral and spatial information. Second, the method establishes an encoder–decoder network structure to extract representative deep features from the spectral–spatial information combined as input. Finally, the method establishes residual networks for a one-to-one encoder layer and a decoder layer. The latter two operations aim at adjusting the deep features learned from the deep learning network to make the deep features more representational. We compared the proposed method with the AWLP, CNMF, GIHS, MTF_GLP, HPF, SFIM and DIP-HyperKite methods. The experimental results suggest that the proposed method can achieve competitive spatial quality compared with the existing methods and recover most of the spectral information that the corresponding sensor would observe with the highest spatial resolution. For our model, we treated the input as contacted panchromatic and up-sampled hyperspectral images. This limited the deep feature extracted from the data with entire spatial and spectral features. In the future, we will pay attention to the deep features extracted from respective panchromatic and hyperspectral images to achieve deeper insight of effective spatial and spectral deep features.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/rs14091981/s1. Data Organizing and Code Running.

Author Contributions: Methodology, R.Z.; supervision, S.D. All authors have read and agreed to the published version of the manuscript.

Funding: National Key Research and Development Program of China: 2021YFE0117100.

Data Availability Statement: Data supported by Capital Normal University.

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

References

1. Ghassemian, H. A review of remote sensing image fusion methods. *Inf. Fusion* **2016**, 75–89. [CrossRef]
2. Li, C.; Xu, H. Spectral Fidelity in High-resolution Remote Sensing Image Fusion. *Geo-Inf. Sci.* **2008**, *10*, 520–526.
3. Shaw, G.; Manolakis, D. Signal processing for hyperspectral image exploitation. *IEEE Signal Process. Mag.* **2002**, *19*, 12–16. [CrossRef]
4. Choi, M.; Kim, R.Y.; Nam, M.R.; Kim, H.O. Fusion of multispectral and panchromatic Satellite images using the curvelet transform. *IEEE Geosci. Remote Sens. Lett.* **2005**, *2*, 136–140. [CrossRef]
5. Benediktsson, J.A.; Pesaresi, M.; Arnason, K. Classification and Feature Extraction for Remote Sensing Images from Urban Areas based on Morphological Transformations. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 1940–1949. [CrossRef]
6. Huete, A.R.; Escadafal, R. Assessment of Biophysical Soil Properties through Spectral Decomposition Techniques. *Remote Sens. Environ.* **1991**, *35*, 149–159. [CrossRef]
7. Churchill, S.; Randell, C.; Power, D.; Gill, E. Data fusion: Remote sensing for target detection and tracking. In Proceedings of the 2004 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2004), Anchorage, AK, USA, 20–24 September 2004.
8. Federico, C. A Review of Data Fusion Techniques. *Sci. World J.* **2013**, *2013*, 704504.
9. Sridhar, V.; Reddy, P.R.; Saiteja, B. Image Enhancement through Discrete and Stationary Wavelet Transforms. *Int. J. Eng. Sci. Res. Technol.* **2014**, *3*, 137–147.
10. Mallat, S.G. A Theory of Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 674–693. [CrossRef]
11. Starck, J.L.; Fadili, J.M.; Murtagh, F. The Undecimated Wavelet Decomposition and its Reconstruction. *IEEE Trans. Image Process.* **2007**, *16*, 297–309. [CrossRef] [PubMed]
12. Burt, P.J.; Adelson, E.H. The Laplacian Pyramid as a Compact Image Code. *IEEE Trans. Commun.* **2003**, *31*, 532–540. [CrossRef]
13. Do, M.N.; Vetterli, M. The Contourlet Transform: An Efficient Directional Multiresolution Image Representation. *IEEE Trans. Image Process.* **2005**, *14*, 2091–2106. [CrossRef] [PubMed]
14. Carper, W.J.; Lillesand, T.M.; Kiefer, P.W. The Use of Intensity-Hue-Saturation Transformations for Merging SPOT Panchromatic and Multispectral Image Data. *Photogramm. Eng. Remote Sens.* **1990**, *56*, 459–467.
15. Tu, T.M.; Su, S.C.; Shyu, H.C.; Huang, P.S. A New Look at IHS-Like Image Fusion Methods. *Inf. Fusion* **2001**, *2*, 177–186. [CrossRef]
16. Chavez, P.S.; Sides, S.C.; Anderson, J.A. Comparison of Three Different Methods to Merge Multiresolution and Multispectral Data: Landsat TM and SPOT Panchromatic. *Photogramm. Eng. Remote Sens.* **1991**, *57*, 265–303.

17. Kwarteng, P.; Chavez, A. Extracting spectral contrast in Landsat Thematic Mapper image data using selective principal component analysis. *Photogramm. Eng. Remote Sens.* **1989**, *55*, 339–348.
18. Shettigara, V.K. A Generalized Component Substitution Technique for Spatial Enhancement of Multispectral Images Using a Higher Resolution Data Set. *Photogram Eng. Remote Sens.* **1992**, *58*, 561–567.
19. Shah, V.P.; Younan, N.H.; King, R.L. An Efficient Pan-Sharpening Method via a Combined Adaptive PCA Approach and Contourlets. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1323–1335. [CrossRef]
20. Huang, B.; Song, H.; Cui, H.; Peng, J.; Xu, Z. Spatial and Spectral Image Fusion Using Sparse Matrix Factorization. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 1693–1704. [CrossRef]
21. Dong, W.; Fu, F.; Shi, G.; Cao, X.; Wu, J.; Li, G.; Li, X. Hyperspectral Image Super-Resolution via Non-Negative Structured Sparse Representation. *IEEE Trans. Image Process.* **2016**, *25*, 2337–2352. [CrossRef] [PubMed]
22. Wei, Q.; Bioucas-Dias, J.; Dobigeon, N.; Tourneret, J.Y. Hyperspectral and Multispectral Image Fusion based on a Sparse Representation. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3658–3668. [CrossRef]
23. Simoes, M.; Bioucas-Dias, J.; Almeida, L.B.; Chanussot, J. A Convex Formulation for Hyperspectral Image Superresolution via Subspace-Based Regularization. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3373–3388. [CrossRef]
24. Yokoya, N.; Yairi, T.; Iwasaki, A. Coupled Nonnegative Matrix Factorization Unmixing for Hyperspectral and Multispectral Data Fusion. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 528–537. [CrossRef]
25. 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]
26. Nascimento, J.M.P.; Dias, J.M.B. Vertex Component Analysis: A Fast Algorithm to Unmix Hyperspectral Data. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 898–910. [CrossRef]
27. Palsson, F.; Sveinsson, J.R.; Ulfarsson, M.O. A New Pansharpening Algorithm Based on Total Variation. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 318–322. [CrossRef]
28. Zhang, Y.; De, B.S.; Scheunders, P. Noise-Resistant Wavelet-Based Bayesian Fusion of Multispectral and Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 3834–3843. [CrossRef]
29. Joshi, M.; Jalobeanu, A. MAP Estimation for Multiresolution Fusion in Remotely Sensed Images Using an IGMRF Prior Model. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 1245–1255. [CrossRef]
30. Li, H. Deep Learning for Natural Language Processing: Advantages and Challenges. *Natl. Sci. Rev.* **2018**, *5*, 22–24. [CrossRef]
31. Soysa, S.D.; Manawadu, S.; Sendanayake, S.; Athipola, U. Computer Vision, Deep Learning and IOT Based Enhanced Early Warning system for the safety of Rail Transportation. In Proceedings of the International Conference on Advances in Computing and Technology, Virtual Online, 28 November 2020.
32. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep Learning for Computer Vision: A Brief Review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef]
33. Silva, L.G.D.; Guedes, L.L.V.; Colcher, S. Using Deep Learning to Recognize People by Face and Voice. In *Anais Estendidos do XXV Simpósio Brasileiro de Sistemas Multimídia e Web*; SBC: Porto Alegre, Brazil, 2019.
34. Ahmad, F.; Abbasi, A.; Kitchens, B.; Adjeroh, D.A.; Zeng, D. Deep Learning for Adverse Event Detection from Web Search. *IEEE Trans. Knowl. Data Eng.* **2020**, *99*, 1. [CrossRef]
35. Scarpa, G.; Vitale, S.; Cozzolino, D. Target-adaptive CNN-based pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5443–5457. [CrossRef]
36. Wei, Y.; Yuan, Q.; Shen, H.; Zhang, L. Boosting the Accuracy of Multispectral Image Pansharpening by Learning a Deep Residual Network. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1795–1799. [CrossRef]
37. Masi, G.; Cozzolino, D.; Verdoliva, L.; Scarpa, G. Pansharpening by Convolutional Neural Networks. *Remote Sens.* **2016**, *8*, 594. [CrossRef]
38. Yuan, Q.; Wei, Y.; Meng, X. A Multiscale and Multidepth Convolutional Neural Network for Remote Sensing Imagery Pan-sharpening. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 978–989. [CrossRef]
39. Bandara, W.G.C.; Valanarasu, J.M.J.; Patel, V.M. Hyperspectral Pansharpening Based on Improved Deep Image Prior and Residual Reconstruction. *arXiv* **2021**, arXiv:2107.02630. [CrossRef]
40. Jiang, M.; Shen, H.; Li, J.; Yuan, Q.; Zhang, L. A differential information residual convolutional neural network for pansharpening. *ISPRS J. Photogramm. Remote Sens.* **2020**, *163*, 257–271. [CrossRef]
41. Cai, W.; Wei, Z.; Liu, R.; Zhuang, Y.; Wang, Y.; Ning, X. Remote sensing image recognition based on multi-attention residual fusion networks. *ASP Trans. Pattern Recognit. Intell. Syst.* **2021**, *1*, 1–8. [CrossRef]
42. Lawrence, S.; Giles, C.L.; Tsoi, A.C.; Back, A.D. Face Recognition: A Convolutional Neural Network Approach. *IEEE Trans. Neural Netw.* **1997**, *8*, 98–113. [CrossRef] [PubMed]
43. Pandey, P.C.; Tate, N.J.; Balzter, H. Mapping Tree Species in Coastal Portugal Using Statistically Segmented Principal Component Analysis and Other Methods. *Sens. J. IEEE* **2014**, *14*, 4434–4441. [CrossRef]
44. Zeng, Y.; Huang, W.; Liu, M.; Zhang, H.; Zou, B. Fusion of satellite images in urban area: Assessing the quality of resulting images. In Proceedings of the IEEE 2010 18th International Conference on Geoinformatics, Beijing, China, 18–20 June 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 1–4.
45. Aiazzi, B.; Alparone, L.; Baronti, S.; Garzelli, A. Context-Driven Fusion of High Spatial and Spectral Resolution Images based on Oversampled Multiresolution Analysis. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 300–312. [CrossRef]

46. Zhou, X.; Liu, J.; Liu, S.; Cao, L.; Zhou, Q.; Huang, H. A GIHS-based Spectral Preservation Fusion Method for Remote Sensing Images Using Edge Restored Spectral Modulation. *ISPRS J. Photogramm. Remote Sens.* **2014**, *88*, 16–27. [CrossRef]
47. Otazu, X.; González-Audícana, M.; Fors, O.; Núñez, J. Introduction of sensor spectral response into image fusion methods. Application of wavelet-based methods. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 2376–2385. [CrossRef]
48. Vivone, G.; Alparone, L.; Chanussot, J.; Dalla Mura, M.; Garzelli, A.; Licciardi, G.A.; Restaino, R.; Wald, L. A Critical Comparison Among Pansharpening Algorithms. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 2565–2586. [CrossRef]
49. Dian, R.; Li, S.; Guo, A.; Fang, L. Deep hyperspectral image sharpening. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 5345–5355. [CrossRef] [PubMed]