applied sciences

MDPI

Article

Swin–MRDB: Pan-Sharpening Model Based on the Swin Transformer and Multi-Scale CNN

Zifan Rong, Xuesong Jiang *, Linfeng Huang and Hongping Zhou

School of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China; rzfannjfu@njfu.edu.cn (Z.R.); huanglf@njfu.edu.cn (L.H.); hpzhou@njfu.edu.cn (H.Z.)
* Correspondence: xsjiang@126.com; Tel.: +86-025-8542-7149

Abstract: Pan-sharpening aims to create high-resolution spectrum images by fusing low-resolution hyperspectral (HS) images with high-resolution panchromatic (PAN) images. Inspired by the Swin transformer used in image classification tasks, this research constructs a three-stream pan-sharpening network based on the Swin transformer and a multi-scale feature extraction module. Unlike the traditional convolutional neural network (CNN) pan-sharpening model, we use the Swin transformer to establish global connections with the image and combine it with a multi-scale feature extraction module to extract local features of different sizes. The model combines the advantages of the Swin transformer and CNN, enabling fused images to maintain good local detail and global linkage by mitigating distortion in hyperspectral images. In order to verify the effectiveness of the method, this paper evaluates fused images with subjective visual and quantitative indicators. Experimental results show that the method proposed in this paper can better preserve the spatial and spectral information of images compared to the classical and latest models.

Keywords: pan-sharpening; Swin transformer; multi-scale residuals and dense blocks; residual feature fusion block

Citation: Rong, Z.; Jiang, X.; Huang, L.; Zhou, H. Swin–MRDB: Pan-Sharpening Model Based on the Swin Transformer and Multi-Scale CNN. *Appl. Sci.* **2023**, *13*, 9022. https://doi.org/10.3390/app13159022

Academic Editors: Qizhi Xu, Jin Zheng and Feng Gao

Received: 11 July 2023
Revised: 3 August 2023
Accepted: 4 August 2023
Published: 7 August 2023

1. Introduction

Recent years have seen specific advancements in the field of remote image processing. Maccone and Ren's (2020) [1] and Marghany's (2022) [2] work contains the original theories for this purpose (2022). They also presented a brand-new theory of quantum image processing. In this view, with the development of imaging systems and remote sensing technologies, the types of sensors used in remote sensing have become diverse, allowing for the acquisition of an increasing number of types of data, most commonly hyperspectral and panchromatic images [3]. Panchromatic images have high resolution but lack spectral information, which is not conducive to the classification of ground cover. Hyperspectral images contain both spatial and spectral information, making them useful for various applications, including environmental monitoring and image classification. However, obtaining hyperspectral images with high spatial resolution is challenging due to hardware limitations. Currently, most sensors can only capture high-resolution panchromatic (PAN) images and low-resolution hyperspectral (LRHS) images. In order to make full use of the multi-source information from remote sensing images and to obtain multi-spectral images with high spatial resolution, the pan-sharpening technique was developed. Pan-sharpening fuses hyperspectral and panchromatic image data from the same scene to produce a new image with a higher spatial resolution than hyperspectral images [4]. It can help researchers acquire higher-quality hyperspectral data, thus enhancing the effectiveness of practical applications. Traditional pan-sharpening methods fall into three main categories: component replacement, multi-scale decomposition, and sparse representation.

The component replacement method is the most primitive pan-sharpening method and is generally implemented in three steps as follows: (i) first, transfer the hyperspectral image into another space to obtain the various components of the hyperspectral image;

(ii) then, replace one of the components with a panchromatic image; and (iii) finally, obtain the fused image using the corresponding inverse transformation. The main classical component replacement methods are principal component analysis (PCA) [5,6], GS transformation [7], HIS [8,9], etc. The component substitution method is only applicable when the hyperspectral image and the panchromatic image are highly correlated; otherwise, it is prone to spectral distortion, although it must be noted that many scholars have improved the algorithm to address this shortcoming. Rahmani et al. [10] extended the traditional IHS approach by proposing an adaptive IHS (AIHS) to adapt the coefficients of the IHS to preserve spectral information. Choi et al. [11] proposed partial replacement adaptive component substitution (prs), which uses partial substitution to generate synthetic component images and then injects high-frequency information based on statistical scaling.

The multi-scale decomposition-based image fusion method consists of three main steps: (i) first, multi-scale decomposition of multiple source images; (ii) then, fusion of the decomposition coefficients of different source images; and (iii) finally, multi-scale inversion on the fusion coefficients to obtain the fused image. Commonly used multi-scale decomposition methods include the Laplace pyramid [12], generalized Laplace pyramid (GLP) [13], discrete wavelet transform [14], etc. Imani et al. [15] combined a multi-scale decomposition model with a free distribution model to better preserve spectral features. Compared with the component replacement method, the multi-scale decomposition method can better preserve the spectral information of the fused image and effectively solve the problem of spectral distortion, but due to the fusion strategy, the problem of spatial information loss will occur.

Along with multi-scale decomposition and component substitution, another area of intense attention among academics is sparse representation theory-based null-spectrum fusion. The sparse representation method combines the image by using the sparseness of the image block representation in the overcomplete dictionary. Li et al. [16] proposed a remote sensing image pan-sharpening method from the perspective of compression perception as well as a remote sensing image fusion method based on local adaptive sparse representation. Yin et al. [17] proposed a pan-sharpening model based on a sparse representation injection model, which uses the ARSIS concept instead of compressed sensory reconstruction to create fused images through detail injection.

With the development of computer technology, deep learning has been widely used in pan-sharpening. Unlike traditional methods, deep learning-based spatial-spectral fusion networks use hyperspectral images and panchromatic images as input data, allowing the network to perform image data extraction and fusion autonomously, thus avoiding errors arising from manual feature extraction. Jian et al. [18] proposed a multi-scale and multi-stream fusion network that can extract spatial and spectral information at different scales. Liu et al. [19] proposed a multi-scale nonlocal-attention network that focuses on enhancing multi-scale targets in the scene, thereby improving image resolution. Iftene et al. [20] were the first to integrate deep learning with pan-sharpening, and the SRCNN model proposed by them had decent fusion but no end-to-end mapping procedure. Yang et al. [21] proposed a deep network model using a residual network structure (PanNet), which propagates spectral information directly into the reconstructed image by adding an upsampled hyperspectral image to the network output. Peng et al. [22] proposed an end-to-end pan-sharpening method based on multi-scale dense networks and designed a multi-scale dense block to extract shallow features in the network. Zheng et al. [23] proposed a residual structure-based pan-sharpening network that learns the residuals between the input and output and improves the convergence speed of the model. The above models have achieved good fusion results, but there are still the following three problems: (i) Due to the influence of the convolutional kernel reception field, the models can only extract local spatial and spectral features, resulting in poor global connectivity of the reconstructed images, and it is difficult to adjust local features according to the global spectral spatial details of the images. (ii) The size of features in remote sensing images varies, and the existing algorithm selects a single-sized convolution kernel, which cannot

achieve feature extraction for features of different sizes. (iii) The loss of feature information caused by the change in image channels in the fusion stage is not considered. To address the aforementioned issues, this paper combines the benefits of the CNN and the Swin transformer and proposes a three-stream pan-sharpening network based on a multi-scale residual dense block (MRDB) and the Swin transformer, with the goal of improving the model's ability to extract global and local features. Our contributions are as follows:

1. A pan-sharpening network with three feature extraction branches was designed to extract and combine valuable information from the various input branches.
2. The Swin transformer feature extraction module was designed to extract global features from panchromatic and hyperspectral images.
3. Multi-scale residuals and dense blocks were added to the network to extract multi-scale local features from hyperspectral images in the image.
4. The residual fusion module was used to reconstruct the retrieved features in order to decrease feature loss during fusion.

2. Methods

2.1. Existing Methods

There are various existing pan-sharpening models with fusion strategies based on different algorithms. In this section, several of the most widely used models are described in detail and selected as comparison models to evaluate the superiority of the proposed Swin–MRDB method. Table 1 briefly describes the comparison algorithms.

Among the traditional fusion methods, principal component analysis (PCA) [24] is widely used. PCA is an orthogonal linear change based on the amount of information, and the algorithm uses a linear transformation method to project the data into new spatial coordinates, transforming the massive spectral data into a small number of several components to retain the maximum information of the original data while reducing dimensionality. The pan-sharpening process of PCA is to first find the feature components of the covariance matrix between the bands of the hyperspectral image, followed by matching the histogram of the panchromatic image to the first principal component, and then replacing the first principal component of the covariance matrix with the panchromatic image. Finally, the fused result is transformed back using PCA inverse transform to obtain the final fused image. The Gram–Schmidt (GS) [25] fusion method can avoid the inconsistency of the spectral response range caused by the traditional fusion methods and maintain the consistency of the spectral information of the image before and after the fusion using the statistical analysis method for the optimal matching of each band involved in the fusion. The GS algorithm performs pan-sharpening fusion in four steps. First, a low-resolution band is replicated to create a panchromatic band. Second, a Cram–Schmidt transformation is used on both the panchromatic and multiband. Third, the high spatial resolution panchromatic band replaces the transformed first band. Lastly, the fusion map is obtained through the application of the Cram–Schmidt inverse transform. The IHS [26] algorithm is one of the widely used algorithms in the field of image fusion, which first resamples hyperspectral images and converts them from RGB to HIS (brightness, hue, saturation) space, then replaces the I component with a panchromatic image, and finally obtains the fused image via inverse conversion.

Table 1. An abbreviated description of the comparison algorithm.

Algorithm	Technology	Algorithm Type	Reference
PCA	Band replacement	traditional method	[20]
GS	Band replacement	traditional method	[21]
IHS	Spatial transformation	traditional method	[26]
PNN	Two-stage network	deep learning	[23]
TFNET	Two-stream network	deep learning	[24]
MSDCNN	Two-stream network	deep learning	[25]
SRPPNN	Characteristic injection network	deep learning	[26]

With the continuous development of deep learning, the technique has become the mainstream pan-sharpening method, and more and more scholars have proposed different feature extraction and fusion strategies to fuse panchromatic and spectral images. The convolutional neural network (CNN) is a popular deep learning model. It excels at processing multi-dimensional data by keeping the input and output in the same dimension. CNNs comprise three layers: convolutional, activation function, and pooling. Each layer has a specific function for processing data, and multiple layers can create a nonlinear transformation to map input and output data. This makes CNNs an essential tool in the field of pan-sharpening. In this paper, a few of the most typical models are selected for a detailed explanation, and these are chosen as the comparison models in this paper. PNN [27] is the first model to apply convolutional neural networks to the field of pan-sharpening, which is achieved by a simple and efficient convolutional architecture with three layers. TFNet [28] is a typical two-stream convolutional pan-sharpening network. The model consists of two feature extraction branches, which extract features of panchromatic and spectral images, respectively, and finally, the extracted features are fused by convolutional layers. MSDCNN [29] is also a typical dual-stream pan-sharpening network, differing from TFNet in that the network incorporates multi-scale feature extraction blocks for extracting feature information to different scales. SRPPNN [30] is a progressive feature injection network that injects features from panchromatic images into spectral images through high-pass residual modules to improve image spatial resolution.

2.2. The Proposed Methods

Inspired by the application of the Swin transformer and CNN in computer vision (CV), this paper designs a three-stream pan-sharpening model for fusing hyperspectral and panchromatic images. A Swin transformer module, a multi-scale residual and dense block (MRDB), and a residual feature fusion module make up the proposed network, which is shown in Figure 1. For ease of description, we use P for Pan images; HS for multi-spectral images; HS↑ for multi-spectral images after up-sampling; HRHS to represent the reconstructed image; and L for images of HS↑ and Pan concat. This paper uses the bicubic algorithm for up-sampling HS images, with this algorithm producing the best and most accurate interpolated graphics. In the feature extraction stage, the global features of L are obtained using the Swin transformer feature extraction branch. Furthermore, we utilize the multi-scale feature extraction branch of the CNN architecture to extract local features from both HS and P images. In the fusion stage, we choose the residual fusion block to fuse the features extracted from the three branches, a process discussed in detail below for the Swin–MRDB method.

2.2.1. Swin Transformer Feature Extraction Branch

This feature branch consists of four Swin transformer blocks and convolutional layers to extract deep features from the data. The following sections of this section describe the Swin transformer algorithm and its network architecture in detail.

Swin Transformer Algorithm

The CNN model, a classical deep learning architecture, effectively extracts informative features from images using a convolutional kernel. During model creation, the size of the convolution kernel can be adjusted to extract features at varying scales. Unlike CNNs, a transformer [31] discards the traditional method of convolutional feature extraction and extracts features via a self-attentive mechanism. The Swin transformer connects every pixel in an image to extract the overall features. The self-attention in the transformer is calculated as in Equation (1) and is schematically shown in Figure 2.

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{D}} + B\right)V \tag{1}$$

where $Q, K, V \in R^{M^2 \times d}$ are the query, key and value matrices, D is the query/key dimension, M^2 is number of patches in a window, and B is a positional parameter.

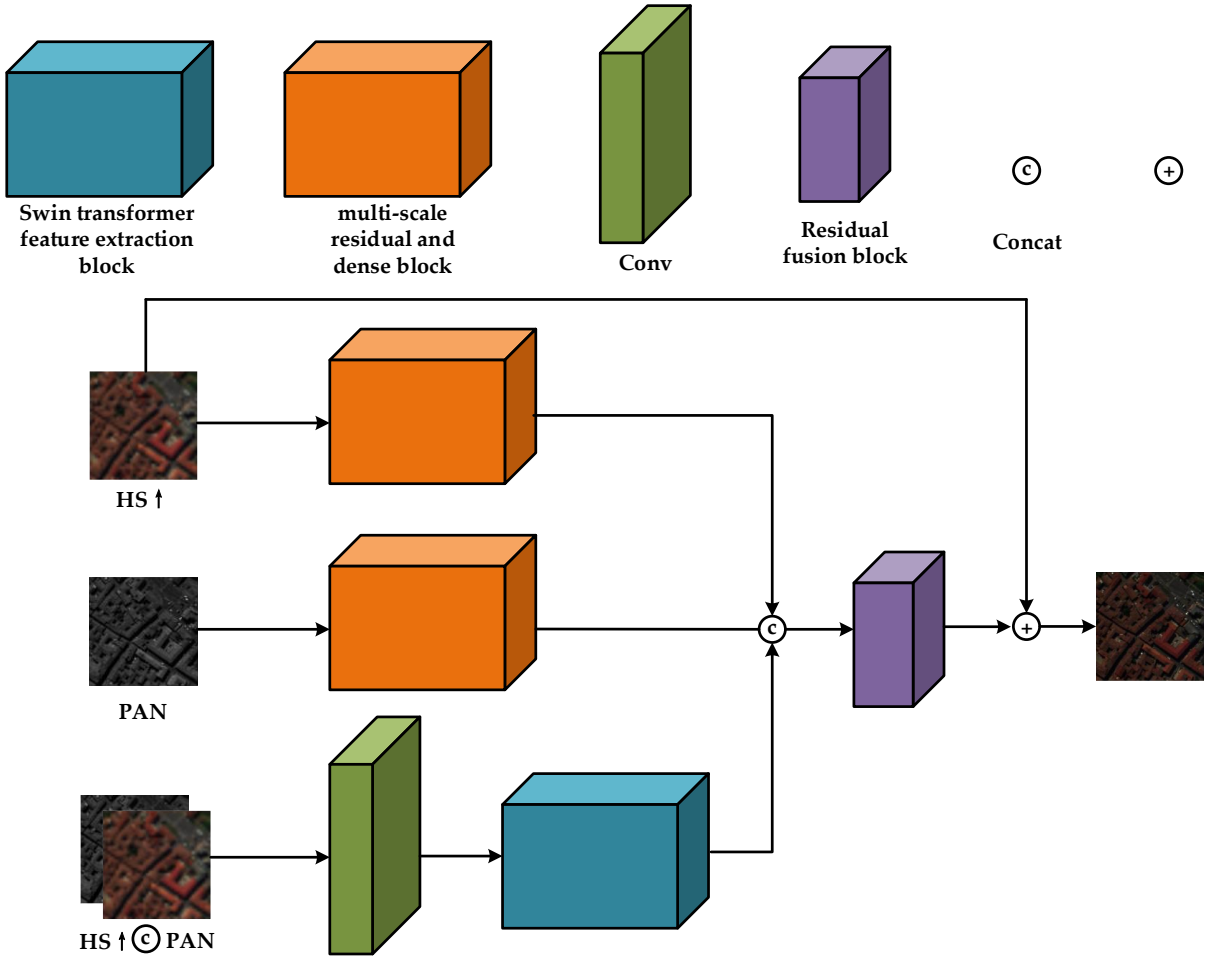

Figure 1. Overall framework diagram of Swin–MRDB. The input image is MS↑ and the PAN output image is HRHS.

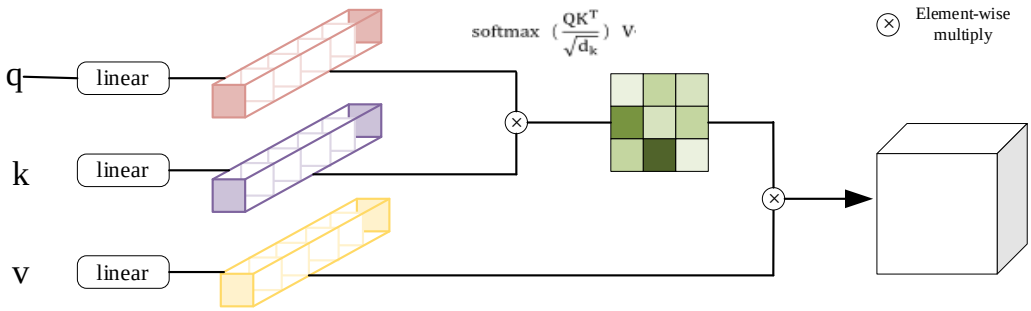

Figure 2. Self-attentive calculation process.

From the formula and the schematic, it can be seen that as long as the image size is fixed, the computational complexity of self-attention is also fixed, and the total computational complexity increases linearly with the image size, which greatly increases the computational cost. To solve this problem, the researchers proposed a vision transformer [32], a model that attempts to split the image into multiple windows and then computes the self-attention in each window. Although the vision transformer reduces the amount of computation by virtue of this operation, it also isolates the exchange of information between different windows. To address the shortcomings of the transformer, Liu et al. [33] proposed the

Swin transformer, which is a transformer model with a layered design. This model adds a moving window interaction mechanism to the vision transformer, which splits the image into several non-overlapping windows. For example, if the input image size is $H \times W \times C$, the Swin transformer will patch partition the image size to $\frac{HW}{M^2} \times M^2 \times C$, where $\frac{HW}{M^2}$ is the number of windows and M^2 is the window size. The Swin transformer will first calculate the self-attention in each window and then interact with the information of different windows through the operation of moving windows, solving the problem of not being able to interact with the information of different windows while reducing the amount of calculation.

Swin Transformer Extracts Branches

To generate input features for the Swin transformer, we simulate hyperspectral and panchromatic images as Q, K, and V feature vectors. Using Q and K, we calculate weight coefficients, which we then apply to V to obtain the weighted input features. As shown in Figure 3, L first goes through a convolutional layer to extract shallow features, followed by deep feature extraction through the Swin transformer feature extraction block. The Swin transformer feature extraction block is made up of four Swin transformer blocks, each of which has six Swin transformer layers and one convolutional layer. The feature received by the ith Swin transformer block is labeled $F_{i,0}$, and the features produced after the six Swin transformer layers are labeled $F_{i,1}$, $F_{i,2}$, $F_{i,3}$, $F_{i,4}$, $F_{i,5}$, and $F_{i,6}$ in the sequence stated in Equation (2).

$$F_{i,j} = H_{STL_{i,j}}(F_{i,j-1}),\ j = 1\ ,\ 2\ , \ldots , 6, \tag{2}$$

where $H_{STL_{i,j}}$ is the jth layer in the ith Swin transformer block, $F_{i,j-1}$ is the output of the j-1st layer in the ith Swin transformer block, and $F_{i,j}$ is the output of the jth layer in the ith Swin transformer block.

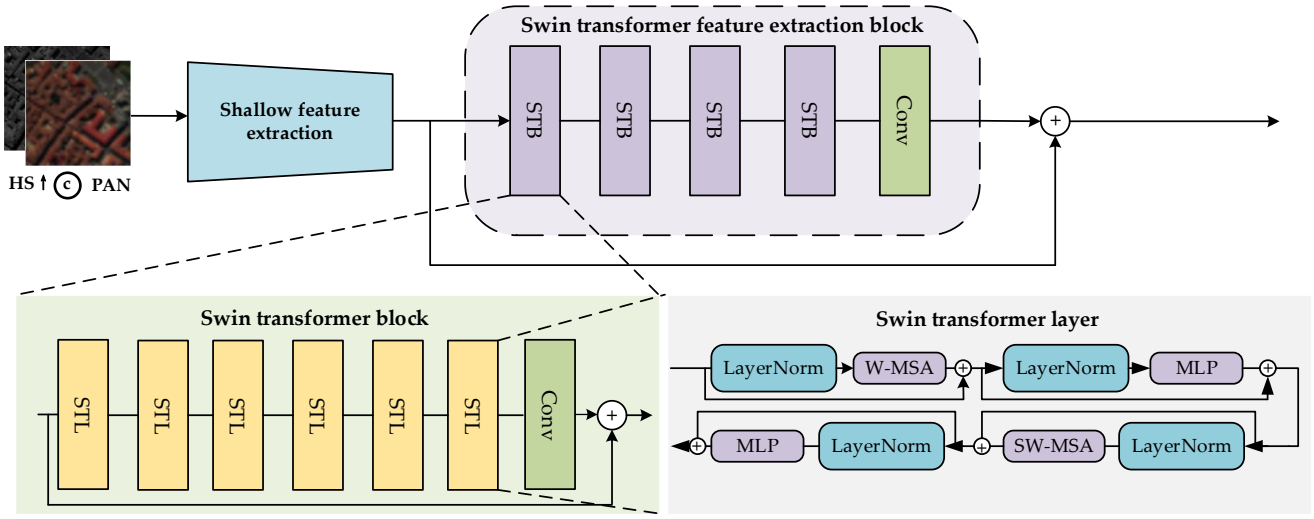

Figure 3. Overall framework diagram of the Swin transformer feature extraction branch.

Once we complete six Swin transformer layers, we obtain the extracted features $F_{i,6}$ from that section. Following this, $F_{i,6}$ is then inputted into the convolutional layer and residual structure. The model deterioration issue is resolved using the residual structure, and the translation invariance is improved using the convolutional layer. This structure is shown in Equation (3).

$$F_{i,out} = H_{conv_i}(F_{i,6}) + F_{i,0} \tag{3}$$

where $F_{i,out}$ is the output of the ith Swin transformer block, $F_{i,L}$ represents the output of the last layer in the ith Swin transformer block, H_{conv_i} represents the convolution operation of the swin transformer block, and $F_{i,0}$ represents the input of the ith layer.

Once we have gone through four Swin transformer blocks, we receive $F_{4,out}$, which we then utilize as an input for the following model. The format of this part is comparable to Equation (3). We choose a convolutional structure to process the extracted deep features and add them to the features extracted using the shallow feature module, as shown in Equation (4).

$$Swin_{ms-pan} = H_{conv}(F_{4,out}) + L_{1,0} \tag{4}$$

where H_{conv} represents the convolution operation in the Swin transformer feature extraction block, $F_{4,out}$ represents the output of the fourth swin transformer block, $L_{1,0}$ represents the output of the shallow feature extraction block, and $Swin_{ms-pan}$ represents the output of the Swin transformer branch.

2.2.2. Multi-Scale Residuals and Dense Blocks

Satellite remote sensing images contain feature compositions of different sizes, which may result in missing feature information if features are extracted using a single-scale convolution kernel. To address this problem, we refer to Guan's [34] design and introduce multi-scale convolution to extract feature information of different sizes. The multi-scale feature extraction block uses a CNN implementation that utilizes parallel connections of multiple-scale convolutional kernels. This allows for the extraction of features at different scales. As shown in Figure 4, the multi-scale residual and dense block contains four different sizes of convolutional kernels, and the underlying branch extracts small-scale features using multiple 3×3 convolutional kernels, which reduce the losses incurred in feature extraction through the residual structure. Convolutional kernels of different sizes are chosen for each of the three upper feature branches to extract feature information of different sizes. This design allows the network to retain local information of different sizes. The structure of multi-scale residuals and dense blocks is shown in Equation (5).

$$\begin{gathered} top = Conv_{9\times9}(HS/pan) \\ upper\ middle = Conv_{7\times7}(HS/PAN) \\ lower\ middle = Conv_{5\times5}(HS/pan) \\ bottom = DRB_{3\times3}(HS/PAN) \\ MRDB_{hs/pan} = Conv_{1\times1}(Concat(top, upper\ middle, lower\ middle, bottom)) \end{gathered} \tag{5}$$

where top, upper middle, and lower middle represent the top three feature extraction branches, respectively; $Conv_{9\times9}$, $Conv_{7\times7}$, $Conv_{5\times5}$, and $Conv_{1\times1}$ represent the size of the convolution kernel, respectively; $DRB_{3\times3}$ represents the dense residual block at the bottom; and $MRDB_{hs/pan}$ represents the output of the multi-scale feature extraction block.

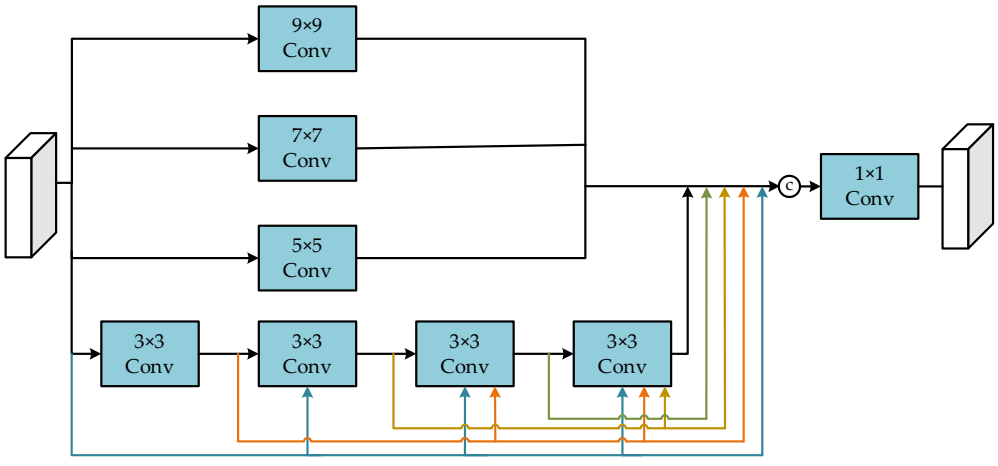

Figure 4. Overall framework diagram of multi-scale residuals and dense blocks.

2.2.3. Residual Feature Fusion Block

Once the feature extraction phase is complete, we obtain features MRDB_{hs}, MRDB_{pan}, and $swin_{hs-pan}$. To effectively combine these features, we designed a feature fusion module, which will be discussed in this section. In the feature reconstruction stage, this paper selects the residual fusion block to reconstruct the features extracted from the three branches to reduce the feature loss caused during fusion. The residual fusion block consists of a 1×1, a 3×3 convolutional, and a ReLU layer, the exact structure of which is shown in Figure 5. CNN is utilized to extract fused features, and the ReLU layer is employed to enhance nonlinearities and improve the model's representation. The fusion process starts by concatenating the features extracted from the three branches, after which the image channels are converted to be consistent with HS using a convolution layer, as shown in Equation (6).

$$F_t = H_{res}\left(concat\left(\mathrm{MRDB}_{hs}, \mathrm{MRDB}_{pan}, Swin_{hs-pan}\right)\right) \tag{6}$$

where MRDB_{hs} denotes Hs features extracted using the MRDB, MRDB_{pan} denotes PAN features extracted using the MRDB branch, $Swin_{hs-pan}$ denotes HS and PAN features extracted by the Swin transformer branch, F_t represents the output of the three-stream feature extraction module, and H_{res} denotes the residual feature fusion operation.

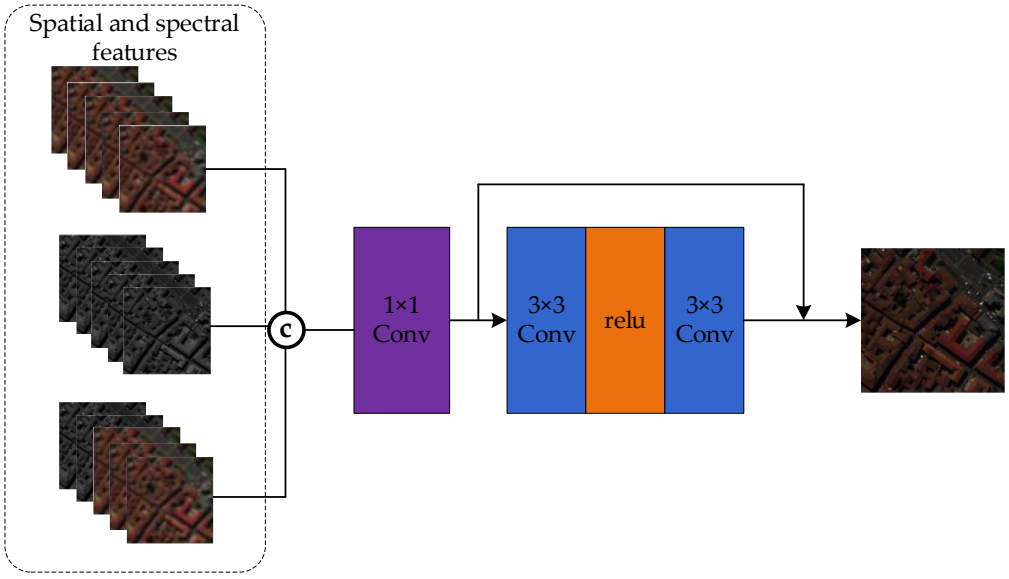

Figure 5. Overall framework diagram of the residual feature fusion block.

After the residual fusion block, we obtain a high-resolution hyperspectral image F_t that is consistent with the hyperspectral image band. To keep the original spectral information after multi-level fusion, we sum the Hs↑ and fusion characteristics, as shown in Equation (7).

$$X_t = F_t + \mathrm{Hs}\uparrow \tag{7}$$

where F_t represents the output of the residual fusion block, Hs↑ represents the up-sampled Hs, and X_t represents the final output of the network.

2.2.4. Loss Function

This paper contrasts the effects of two widely employed loss functions, L_1 and L_2, on the reconstructed images and analyzes the benefits and drawbacks of the two loss functions through ablation experiments in Section 3.5.

The L_1 loss function is the most common loss function for network training in deep learning and usually refers to the mean absolute error loss; the equation is shown in (8).

$$L_1 = \frac{1}{N}\sum_{i=1}^{N}\left\|H_{gt} - X_t\right\|_1 \tag{8}$$

The L_2 loss function is often used in regression tasks, and the equation is shown in (9):

$$L_2 = \frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(H_{gt} - X_t\right)^2 \tag{9}$$

where X_t represents HR-HS, H_{gt} represents ground truth images, and N represents the number of training sessions.

2.2.5. Algorithm Overview

Algorithm 1 gives the overall framework and training process of Swin–MRDB. Specifically, in the data preparation stage, the number of model iterations M is read first, and then hyperspectral and panchromatic images are randomly read from the training sets H_{train} and P_{train} (Line4–5). In the feature extraction phase, image features $Swin_{hs\text{-}pan}$, $MRDR_{hs}$, and $MRDB_{pan}$ are extracted based on the Swin transformer and multi-scale feature extraction block (Line7–9). In the feature fusion stage, features $Swin_{hs\text{-}pan}$, $MRDR_{hs}$, and $MRDB_{pan}$ are fused using the residual feature fusion block to obtain a high-resolution hyperspectral image D (Line11). During the parameter calculation stage, a formula is utilized to determine the disparity between G and D, resulting in the objective evaluation indicator V's calculation (Line3). Define the value of V_{best} as 0. If the V obtained by some epoch is greater than V_{best}, the value of V is used instead of the value of V_{best} (Line14–17).

Algorithm 1: Swin–MRDB

Input: hyperspectral training set H_{train}, panchromatic image training set P_{train}, hyperspectral test set H_{test}, panchromatic image test set P_{test}, ground truth image G, number of iterations M.
Output: Objective evaluation indicator V_{best}.

1 % Data preparation
2 $V_{best} \leftarrow 0$
3 Read epoch value M
4 **For** m = 1. . . . .M **do**
5 Randomly obtain image data H, P from H_{train} and P_{train}.
6 % feature extraction stage
7 The Swin Transform feature extraction block was utilized to extract the feature $Swin_{hs\text{-}pan}$ of the hyperspectral and panchromatic images.
8 The multi-scale feature extraction block was utilized to extract the feature $MRDR_{hs}$ of the hyperspectral images.
9 The multi-scale feature extraction block was utilized to extract the feature $MRDB_{pan}$ of the panchromatic images.
10 % feature fusion stage
11 The high-resolution hyperspectral image D is obtained by extracting $Swin_{hs\text{-}pan}$, $MRDB_{hs}$ and $MRDB_{pan}$ based on the residual feature fusion block.
12 % Parameter calculation stage
13 The difference between the ground truth image G and D is calculated according to the formula to obtain the objective evaluation index V
14 **IF** $V > V_{best}$ **then**
15 $V_{best} \leftarrow V$
16 **End**
17 **Return** V_{best}

3. Experiments

3.1. Datasets

We conducted experiments on three open-source hyperspectral datasets (Pavia, Botswana, and Chikusei), and since there is no ideal fusion result for labeling, this paper produces datasets according to the Wald protocol [35]. The original HS is used as the ideal fusion image, the down-sampled and then up-sampled HS is used as the HS of the network input, and the down-sampled PAN is used as the PAN of the network input, so that the relationship between the network input HS, PAN, and the ideal fusion image is consistent with the real fusion relationship. The dataset is described in detail as follows:

(1) Pavia dataset [36]: The Pavia University data are part of the hyperspectral data from the German airborne reflectance optical spectral imager image of the city of Pavia, Italy, in 2003. The spectral imager continuously images 115 bands in the wavelength range of 0.43–0.86 μm with a spatial resolution of 1.3 m. In total, 12 of these bands are removed due to noise, so the remaining 103 spectral bands are generally used. The size of the data is 610 × 340, containing a total of 2,207,400 pixels, but they contain a large number of background pixels, and only 42,776 pixels in total containing features. Nine types of features are included in these pixels, including trees, asphalt roads, bricks, meadows, etc. We partition HS into a size of 40 × 40 × 103 and Pan into a size of 160 × 160 × 1 size to make training and testing sets for training the network model.

(2) Botswana dataset [37]: The Botswana dataset was acquired using the NASA EO-1 satellite in the Okavango Delta, Botswana, in May 2001, with an image size of 1476 × 256. The wavelength range of the sensor on EO-1 is 400–2500 nm, with a spatial resolution of about 20 m. Removing 107 noise bands, the actual bands used for training are 145. We partition HS into a size of 30 × 30 × 145 and Pan into a size of 120 × 120 × 1 to make training and testing sets for training the network model.

(3) Chikusei dataset [38]: The Chikusei dataset was captured using the Headwall Hyperspec-VNIR-C sensor in Tsukishi, Japan. These data contain 128 bands in the range of 343–1018 nm, with a size of 2517 × 2335 and a spatial resolution of 2.5 m. These data were made public by Dr. Naoto Yokoya and Prof. Akira Iwasaki of the University of Tokyo. We partition HS into a size of 64 × 64 × 128 and Pan into a size of 256 × 256 × 1 to make training and testing sets for training the network model.

3.2. Evaluation Parameters

The subjective visual evaluation is to judge the fusion effect by observing the details of the image, such as sharpness, color, and contour. The subjective visual evaluation will vary from person to person and can only represent some people's viewpoints, which can be one-sided, so some objective indicators are also needed to judge the quality of the fused image. In order to evaluate the fused images more objectively, the correlation coefficient (CC) [39], spectral angle mapping (SAM) [40], root mean square error (RMSE) [41], and Errur Relative Globale Adimensionnelle Desynthese (ERAGS) [42] are chosen in this paper to evaluate the fused images objectively. These evaluation metrics have been widely used to evaluate fusion images, and they are described in detail below.

The CC reflects the linear correlation between the fused image and the reference image, which is defined as shown in Equation (10).

$$CC(X_t, H_{Gt}) = \frac{\sum_{i=1}^{m}\sum_{i=1}^{n}[X_t(i,j) - u_t][H_{GT}(i,j) - u_{Gt}]}{\sqrt{\sum_{i=1}^{m}\sum_{i=1}^{n}[X_t(i,j) - u_t]^2 \sum_{i=1}^{m}\sum_{i=1}^{n}[H_{Gt}(i,j) - u_{gt}]^2}} \tag{10}$$

where X_t represents the fused image, H_{Gt} represents the reference image, m and n represent the image sizes of X_t and H_{Gt}, u_t represents the mean value of the fused image X_t, u_t represents the mean value of the reference image H_{Gt}, represents the covariance between X_t and H_{Gt}, $\sum_{i=1}^{m}\sum_{i=1}^{n}[X_t(i,j) - u_t]^2$ represents the standard deviation of X_t, and $\sum_{i=1}^{m}\sum_{i=1}^{n}[X_t(i,j) - u_t][H(i,j) - u_{Gt}]$ represents the standard deviation of H_{Gt}. When the

CC is closer to 1, it represents a stronger linear correlation between fusion X_t and H_{Gt}, corresponding to a better fusion effect.

The RMSE reflects the degree of pixel difference between the fused image and the reference image, which is defined as shown in Equation (11).

$$\mathrm{RMSE}(X_t, H_{Gt}) = \sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{i=1}^{n}[X_t(i,j) - H_{Gt}(i,j)]^2} \tag{11}$$

where M and N represent the image sizes of x_t and G_{Gt}. RMSE reflects the degree of difference between images, so the smaller the RMSE value, the smaller the difference between the fused image and the reference image.

SAM reflects the similarity of the spectral data between the fused and reference images, which is defined as shown in Equation (12):

$$\mathrm{SAM}(X_t, H_{Gt}) = cos^{-1}\frac{\langle X_v, H_v\rangle}{\|X_v\|_2\|H_v\|_2} \tag{12}$$

where H_v represents the spectral vector of the reference image and x_v represents the spectral vector of the fused image. The smaller the SAM value, the smaller the spectral difference between the reference image and the fused image, indicating that the degree of spectral distortion during the fusion process is smaller.

The ERGAS reflects the degree of spectral distortion of the fused image, which is defined as shown in Equation (13):

$$\mathrm{ERGAS} = 100\frac{T_x}{T_h}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{RMSE^2(i)}{u_i^2}} \tag{13}$$

where N represents the number of bands of the image, T_x and T_h are the resolutions of the PAN image and the HS image, respectively, $RMSE^2(i)$ is the root mean square error of the ith band of the fused image and the reference image, and u_i represents the value of the ith band of the reference image. The smaller the value of ERGAS, the smaller the degree of spectral distortion.

3.3. Training Setup

This paper aims to confirm the effectiveness of the proposed method by comparing it to seven other algorithms. Three traditional algorithms (PCA, GS, and IHS) and four deep learning algorithms (PNN, TFNET, MSDCNN, and SRPPNN) were chosen for this purpose. In order to guarantee the impartiality of the experiments, all models were tested under identical conditions. The PyTorch 1.13.1 framework was employed along with a 24G RTX 3090 GPU. To train the deep learning models, we configured the batch size to 2 and set the epoch to 4000. Additionally, we utilized the Adam optimization algorithm with an initial learning rate of 0.0001. Table 2 shows the training platform parameters.

Table 2. Experimental environment setup.

Experimental Environment	Edition
Deep learning frameworks	Pytorch 1.13.1
Compilers	Python 3.8
Operating system	Window11
GPU	RTX3090
CPU	Intel(R) Xeon(R) Gold 6330 CPU

3.4. Results

To verify the validity of the models in this paper, we selected three traditional pan-sharpening methods and four more widely used deep learning pan-sharpening methods as comparison models. The traditional methods are PCA, GS, and IHS, respectively. The deep

learning-based pan-sharpening algorithms are PNN, TFNET, MSDCNN, and SRPPNN, respectively. The reconstruction results of different algorithms on the Pavia, Botswana, and Chikusei datasets are shown in Figures 6–8. In the figures, (a) denotes the ground truth reference image; (b) denotes the up-sampled MS image; and the reconstruction results of the PCA [5], GS [25], HIS [10], PNN [27], TFNet [28], MSDCNN [29], and SRPPNN [30] models are (c–g). In addition, the reconstructed images of the three datasets were evaluated objectively, and the comparison of objective evaluation metrics for each algorithm is shown in Tables 3–5. In the following sections, we analyze the reconstruction results of the Swin–MRDB method on each dataset.

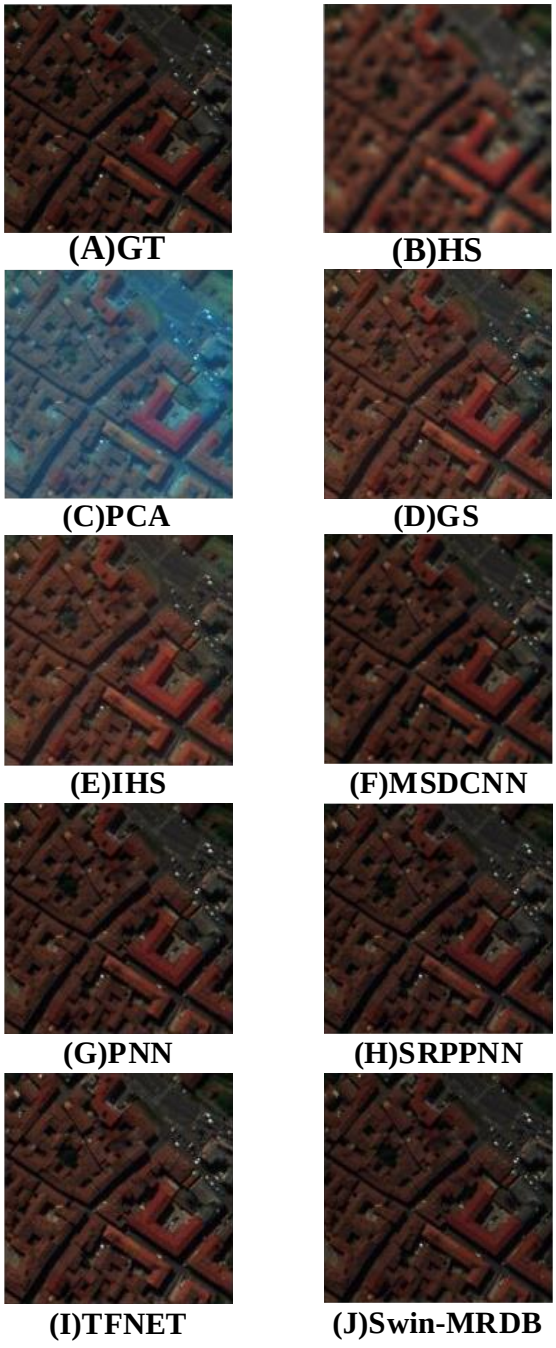

Figure 6. Visual results obtained using different methods for the Pavia Center dataset. (**A**) Ground truth; (**B**) hyperspectral image; visual results obtained using (**C**) PCA, (**D**) GS, (**E**) HIS, (**F**) MSDCNN, (**G**) PNN, (**H**) SRPPNN, (**I**) TFNET, and (**J**) Swin–MRDB (ours).

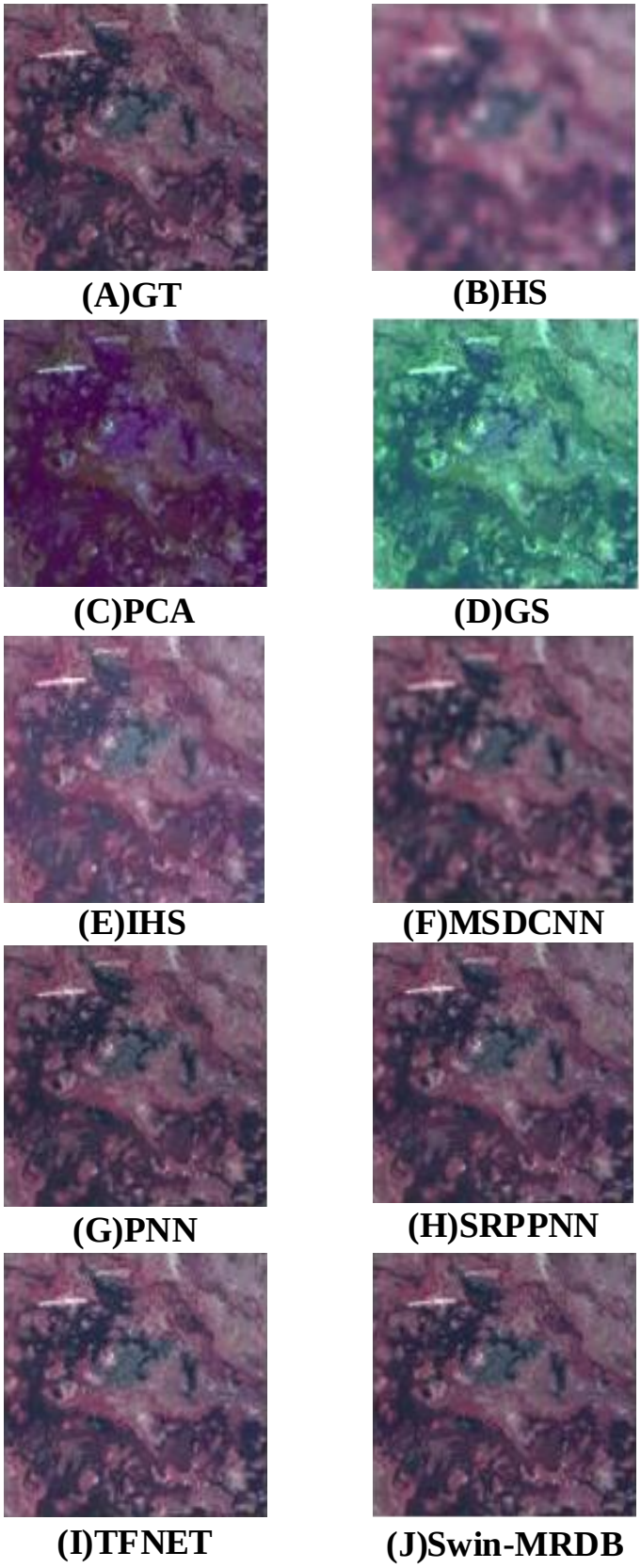

Figure 7. Pan-sharpening results for the Botswana dataset. (**A**) Ground truth; (**B**) hyperspectral image; visual results obtained using (**C**) PCA, (**D**) GS, (**E**) HIS, (**F**) MSDCNN, (**G**) PNN, (**H**) SRPPNN, (**I**) TFNET, and (**J**) Swin–MRDB (ours).

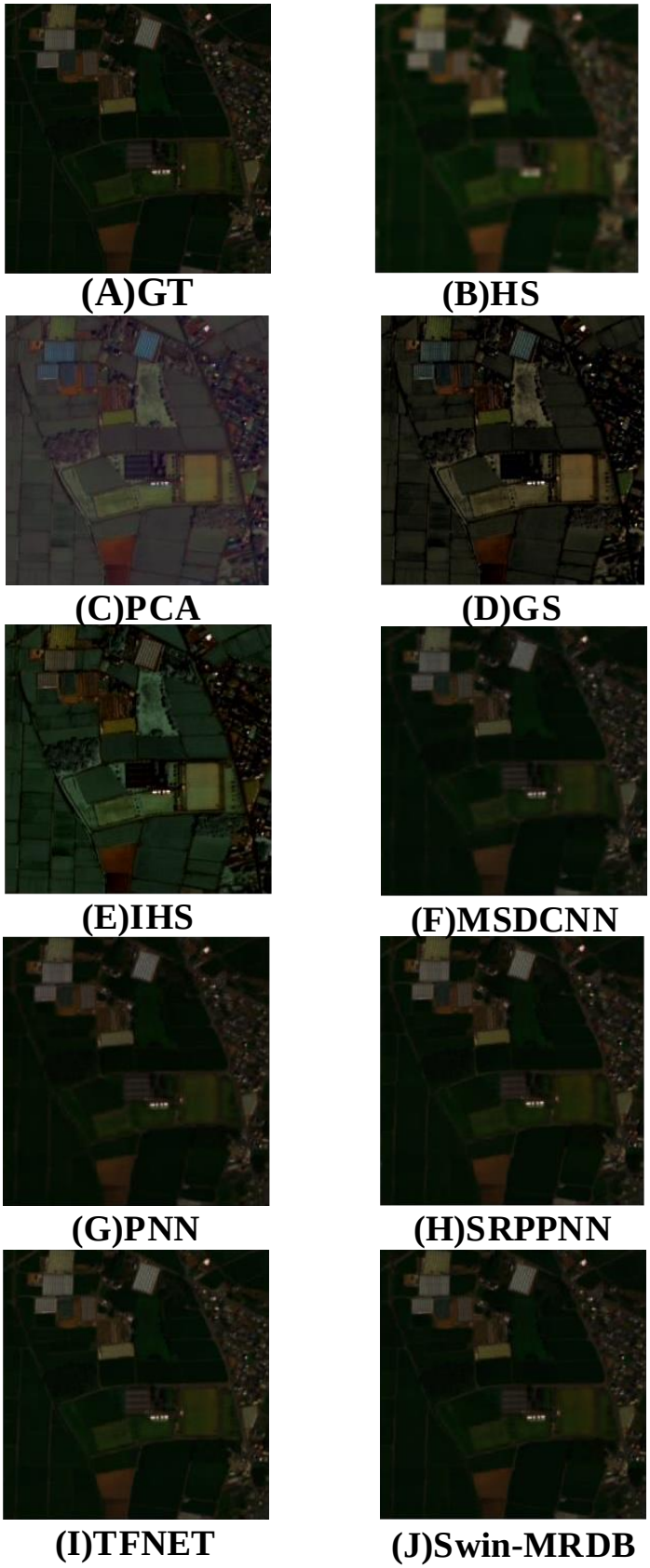

Figure 8. Pan-sharpening results for the Chikusei dataset. (**A**) Ground truth; (**B**) hyperspectral image; visual results obtained using (**C**) PCA, (**D**) GS, (**E**) HIS, (**F**) MSDCNN, (**G**) PNN, (**H**) SRPPNN, (**I**) TFNET, and (**J**) Swin–MRDB (ours).

Table 3. Quantitative evaluation of the Pavia dataset.

Models	Indicators			
	CC	**SAM**	**RMSE**	**ERGAS**
PCA [24]	0.651	16.445	6.432	9.71
GS [25]	0.964	8.924	4.432	7.56
IHS [26]	0.468	8.994	4.681	7.74
PNN [27]	0.978	5.132	1.522	3.96
TFNET [28]	0.980	4.797	1.388	3.26
MSDCNN [29]	0.980	4.867	1.361	3.12
SRPPNN [30]	0.979	5.085	1.452	3.85
ours	**0.984**	**4.394**	**1.180**	**2.64**

Higher values of CC, and lower values of SAM, RMSE, and ERGAS indicate good performance. The units of RMSE are 10^{-2}.

Table 4. Quantitative evaluation of the Botswana dataset.

Models	Indicators			
	CC	**SAM**	**RMSE**	**ERGAS**
PCA [24]	0.800	2.717	5.92	7.74
GS [21]	0.958	3.251	9.08	13.55
his [22]	0.754	1.839	6.19	8.17
PNN [23]	0.961	1.814	1.278	2.25
TFNET [24]	0.962	1.727	1.226	2.12
MSDCNN [25]	0.941	2.350	1.949	2.88
SRPPNN [26]	0.960	1.819	1.322	2.16
ours	**0.977**	**1.483**	**1.074**	**1.93**

Higher values of CC, and lower values of SAM, RMSE, and ERGAS indicate good performance. The units of RMSE are 10^{-2}.

Table 5. Quantitative evaluation of the Chikusei dataset.

Models	Indicators			
	CC	**SAM**	**RMSE**	**ERGAS**
PCA [24]	0.548	26.513	8.28	18.52
GS [21]	0.658	26.178	6.71	13.41
IHS [22]	0.672	26.566	6.75	13.64
PNN [23]	0.970	2.778	1.523	4.41
TFNET [24]	0.978	2.252	1.224	4.11
MSDCNN [25]	0.964	3.094	1.704	5.41
SRPPNN [26]	0.977	2.341	1.292	3.98
ours	**0.981**	**2.219**	**1.223**	**3.85**

Higher values of CC, and lower values of SAM, RMSE, and ERGAS indicate good performance. The units of RMSE are 10^{-2}.

3.4.1. Results for the Pavia Dataset

Urban buildings make up the majority of Pavia's characteristics; thus, we compare the pan-sharpening results mostly in terms of the architectural aspect. To demonstrate the advancement and novelty of this study, we choose PCA, GS, HIS, PNN, TFNET, MSDCNN, and SRPPNN models as comparison algorithms, respectively. The reconstruction outcomes for the various models are shown in Figure 6. To demonstrate the model's influence on reconstruction, the ground truth picture, the up-sampled MS image, and the fused image are contrasted in the visual resolution. The visual analysis shows that although each of the three machine learning techniques enhance the spatial resolution of MS pictures, they all cause some degree of spectral information loss. The PCA algorithm is the worst, resulting in extreme discoloration of the reconstructed picture; the GS and IHS algorithms have comparable visual consequences, both of which result in certain portions of the fused image being too bright. In contrast, both deep learning methods effectively preserve the

original colors in the MS image. The MSDCNN algorithm's reconstructed pictures are on the bright side, indicating a significant loss of spectral information; TFNET, SRPPNN, and PNN's reconstructed images are blurry, while effectively keeping spectral information. As can be seen from the enlarged figure, our proposed algorithm reduces the loss of spectral information while ensuring spatial resolution.

In terms of objective evaluation, a quantitative of the performance of different models on the Pavia dataset is presented in this paper, and the results are shown in Table 3. Data analysis reveals that deep learning-based pan-sharpening algorithms outperform conventional pan-sharpening methods. When the processing outcomes of the four deep learning algorithms are compared, the four deep learning methods obtain comparable results in CC metrics. The proposed model performs best compared with the other models. Specifically, the fused images produced by the Swin–MRDB method are significantly higher in SAM, RMSE, and ERGAS than other models, proving that the model can better retain spatial and spectral information. Compared with the best-performing comparative model in the objective evaluation index, the proposed model achieves an 8.4% improvement in SAM, a 13.3% improvement in RMSE, and a 15.3% improvement in ERGAS.

3.4.2. Results for the Botswana Dataset

To evaluate the reconstruction effect of our model on the Botswana dataset, we chose to compare it with seven comparison models, as shown in Table 3, respectively. The visual analysis of the reconstructed images from the Botswana dataset revealed that, among the machine learning algorithms, the IHS algorithm was the best; the PCA algorithm was the second best; and the GS algorithm reconstructed the most distorted images, losing a large amount of spectral information. Among the deep learning algorithms, the MSDCNN algorithm reconstructs images with the lowest spatial resolution, and the PNN and SRPNN reconstructed images show local detail distortion. Due to the Swin transformer's global self-attentive process, the proposed Swin–MRDB method maintains both local and global features of the reconstructed images, and no local distortion is produced.

The objective evaluation parameters for the fusion results of the Botswana dataset are shown in Table 4. With essentially the same results as with the Pavia dataset, the deep learning approach still has better reconstructions, but it is worth noting that GS maintains high CC values on both the Pavia and Botswana datasets, demonstrating that the images reconstructed using the GS algorithm have a high linear correlation with the reference image. In terms of deep learning algorithms, MSDCNN fared the poorest; PNN, SRPPNN, and TFNET achieved similar reconstruction results, while the Swin–MRDB method proposed in this paper outperformed the other algorithms in all evaluation metrics. The most significant improvement in the evaluation metric is SAM, where the results of other deep learning models are distributed between 1.72 and 2.35, while our proposed model achieves a final result of 1.483, which is a 13.7% improvement compared to the best-performing model. Next was RMSE, which received a 12.4% boost, followed by ERGAS, which received an 8.9% boost. The metric with the smallest improvement was CC, which gained 1.6%, 1.5%, 3.8%, and 1.7% over PNN, TFNET, MSDCNN, and SRPPNN, respectively.

3.4.3. Results for the Chikusei Dataset

The Chikusei dataset contains both urban and rural areas, with arable land making up the majority. Since we have already evaluated the reconstruction of urban images on the Pavia dataset, we evaluated the reconstruction of cultivated land images by using the Swin–MRDB on the Chikusei dataset. To assess the reconstruction effect in a more objective manner, we selected the following models for comparison: PCA, GS, HIS, PNN, TFNET, MSDCNN, and SRPPNN. As shown in Figure 8, the images fused using the PCA, GS, and IHS algorithms are overall brighter, resulting in a significant loss of spectral information. All of the deep learning techniques perform well in terms of preserving spectral information; however, they all exhibit variable degrees of texture discontinuity,

most notably in MSDCNN reconstructed pictures. It is clear from Figure 8 that our Swin–MRDB reconstruction results are most similar to the ground truth reference image.

This section also shows the pan-sharpening effect of our model on the Chikusei dataset using objective metrics. As can be seen in Table 5, our Swin–MRDB method achieves the best results for CC, SAM, RMSE, and ERGAS, indicating that our method retains more spatial detail and spectral information. Among them, ERGAS showed significant improvement, with 12.6%, 6.3%, 28%, and 3.2% improvement compared to PNN, TFNet, MSDCNN, and SRPPNN, respectively. The performance on SAM is also outstanding, with 20%, 1.4%, 28%, and 15.2% improvement compared to PNN, TFNet, MSDCNN, and SRPPNN, respectively. For the RMSE, compared to PNN, TFNet, MSDCNN, and SRPPNN, the improvement is 19.6%, 0.1%, 28%, and 5.3%, respectively. For the CC, compared to PNN, TFNet, MSDCNN, and SRPPNN, the improvement is 1.1%, 0.3%, 1.7%, and 0.4%, respectively.

3.5. Ablation Experiments

To investigate the influence of alternative loss functions on the model, the ADAM optimizer is used to train the Swin–MRDB model with L_1 and L_2 loss functions on the Botswana dataset. In this experiment, the CC, SAM, RMSE, and ERGAS were selected for objective evaluation of the results, and the detailed results are shown in Table 6. The comparison shows that the model based on the L_1 loss function achieves better results, preserving the spatial details of the fused images and reducing the loss of spectral information. Compared with the L_2 loss function, when the model is trained with the L_1 loss function, all metrics are improved significantly, including the CC (by 9.5%), SAM (by 65%), RMSE (by 44%), and ERGAS (by 75%), so we finally choose to train the model with the L_1 loss function.

Table 6. Ablation test results.

Model	L_1	L_2	RMDB-pan	RMDB-MS	CC	SAM	RMSE	ERGAS
Swin–MRDB	√	×	√	√	0.977	1.483	1.074	1.93
	×	√	√	√	0.892	2.461	1.942	3.38
	√	×	√	×	0.965	1.533	1.214	2.19
	√	×	×	√	0.962	1.634	1.245	2.25

The units of RMSE are 10^{-2}.

In addition, this section compares the effects of different multi-scale residuals and dense feature extraction branches on the model. The addition of a multi-scale feature extraction branch to the Swin transformer can effectively improve the performance of the model and help reconstruct higher-quality remote sensing images. As can be seen from Table 6, the addition of two multi-scale feature extraction branches significantly improves the objective evaluation metrics. When adding multi-scale feature extraction blocks for PAN images, the CC, SAM, RMSE, and ERGAS are improved by 1.5%, 9.2%, 13.7%, and 14.2%, respectively. When adding multi-scale feature extraction blocks for HS images, the CC, SAM, RMSE, and ERGAS are improved by 1.2%, 3.2%, 11.5%, and 14%, respectively.

In order to evaluate the impact of the number of residual blocks in the reconstruction module on the Swin–MRDB model, this paper selects Botswana as the experimental dataset and compares the impact of different numbers of residual blocks on objective evaluation metrics. The detailed results are shown in Figure 9. The horizontal coordinate is the number of residual blocks, and it can be seen that increasing the number of residual blocks results in better performance for the Swin–MRDB model, but there is no significant change in model ability as the number of residual blocks increases. Considering that more residual fusion blocks will affect the speed of the model, we finally choose one residual block to fuse the extracted features.

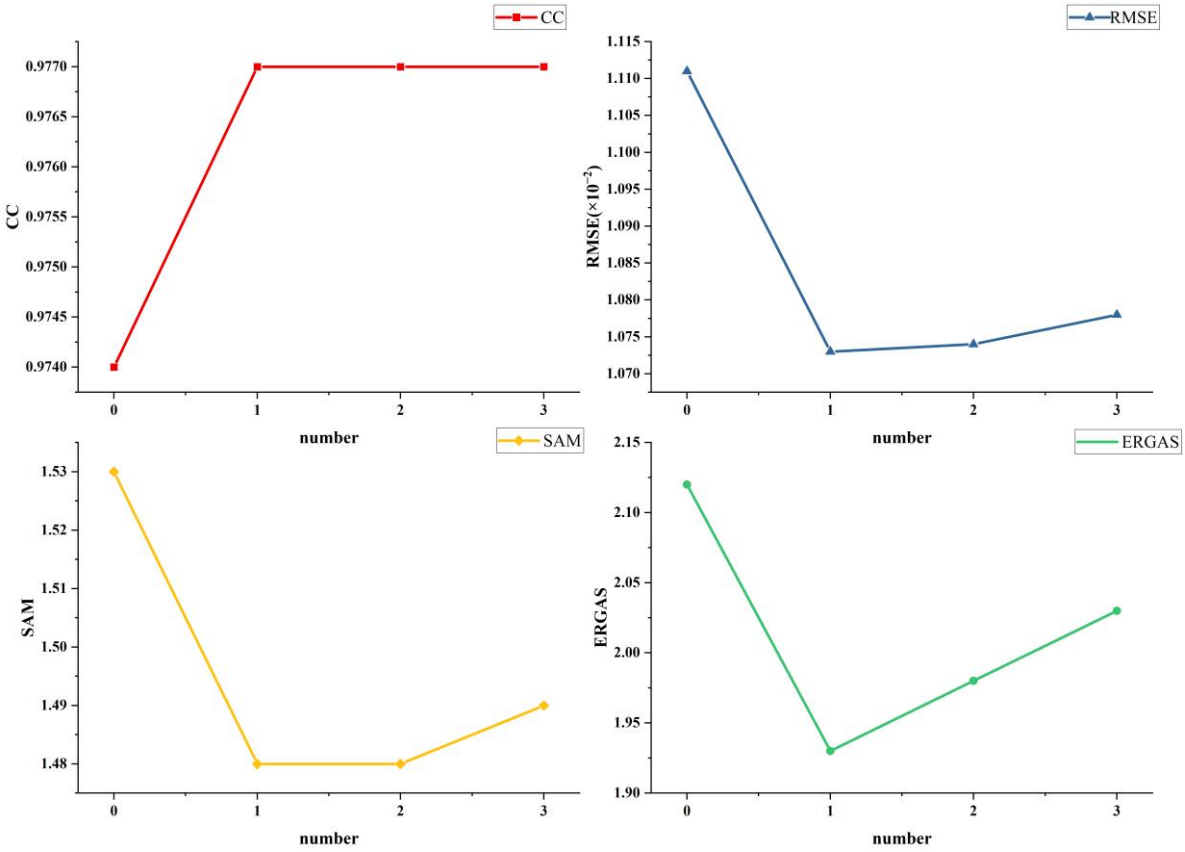

Figure 9. Effect of different numbers of residual blocks on the indicator.

4. Discussion

Most of the current pan-sharpening tasks are implemented based on CNN architecture. CNN can extract features containing more local details, but the global connectivity of these features is poor [42,42]. In recent years, the Swin transformer has attracted the attention of many scholars due to its excellent performance in the field of machine vision. The Swin transformer can establish the connection between individual pixel points of an image and possesses strong global feature extraction, but the model is poor in perceiving the details of an image [43]. Therefore, we propose a Swin–MRDB pan-sharpening model, which employs CNN and the Swin transformer to extract shallow feature information and deep feature information of the image, respectively. Then, the features are fused by residual blocks to obtain a hyperspectral image with high resolution.

In this section, we further discuss the effectiveness of the Swin–MRDB model. First, in order to demonstrate the advantages of the model, experiments are designed to compare the proposed model with the most current state-of-the-art models (PCA, GS, IHS, PNN, TFNet, SRPPNN, and MSDCNN), and three widely used datasets are selected for simulation experiments to reduce the experimental errors caused by the datasets. In order to show the effect of fusion, the fused images generated by different models are put into the text in this paper to better show the gap between the fused images and the ground truth images, and the effect is shown in Figures 6–8. Compared with traditional methods, deep learning methods can preserve spatial and spectral information better, and our proposed method is the closest to the ground truth image. In addition, to reduce the bias caused by subjective evaluation, we choose the four most commonly used image evaluation metrics (CC, SAM, RMSE, and ERGAS) to evaluate the fusion results objectively. As shown in Tables 3–5, the proposed model achieved the best results in all three datasets, with the most significant improvement in the Pavia dataset, where SAM, RMSE, and ERGAS obtained 8.4%, 13.3%, and 15.3% improvement, respectively.

In addition to the objective and subjective evaluation of the fused images, we performed ablation experiments to evaluate the effect of different modules on the fusion results. Specifically, the CC, SAM, RMSE, and ERGAS obtained 1.5%, 9.2%, 13.7%, and 14.2% improvement, respectively, after adding the Pan image multi-scale feature extraction branch. With the addition of the Hs image multi-scale feature extraction branch, the above metrics were improved by 1.2%, 3.2%, 11.5%, and 14%, respectively.

5. Conclusions

This paper proposes a three-stream feature fusion network called Swin–MRDB for the pan-sharpening of remote sensing images. In the feature extraction phase, the proposed Swin–MRDB model contains three separate feature extraction branches to extract PAN and HS features, where the Swin transformer branch is used to capture the long-range dependencies between PAN and HS, and the multi-scale residual feature extraction branch is used to extract local features at different scales. In the feature fusion section, we propose residual fusion blocks to reduce feature loss in order to retain more spatial and spectral information. Compared to the existing methods, the model proposed in this paper shows greater performance.

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

Funding: This research was funded by the Jiangsu Province Science and Technology Program Special Funds (Key R&D Program Modern Agriculture) Project (BE2022374), Modern Agricultural Equipment and Technology Demonstration and Promotion Project in Jiangsu Province (NJ2022-12).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Acknowledgments: The authors gratefully acknowledge the Space Application Laboratory, Department of Advanced Interdisciplinary Studies, the University of Tokyo, for providing the hyperspectral data.

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

References

1. Marghany, M. *Remote Sensing and Image Processing in Mineralogy*; CRC Press: Boca Raton, FL, USA, 2021; ISBN 978-1-00-303377-6.
2. Maccone, L.; Ren, C. Quantum Radar. *Phys. Rev. Lett.* **2020**, *124*, 200503. [CrossRef]
3. Ghassemian, H. A Review of Remote Sensing Image Fusion Methods. *Inf. Fusion* **2016**, *32*, 75–89. [CrossRef]
4. Li, J.; Hong, D.; Gao, L.; Yao, J.; Zheng, K.; Zhang, B.; Chanussot, J. Deep Learning in Multimodal Remote Sensing Data Fusion: A Comprehensive Review. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102926. [CrossRef]
5. Kwarteng, P.; Chavez, A. Extracting Spectral Contrast in Landsat Thematic Mapper Image Data Using Selective Principal Component Analysis. In Proceedings of the 6th Thematic Conference on Remote Sensing for Exploration Geology, London, UK, 11–15 September 1988.
6. Shandoosti, H.R.; Ghassemian, H. Combining the Spectral PCA and Spatial PCA Fusion Methods by an Optimal Filter. *Inf. Fusion* **2016**, *27*, 150–160. [CrossRef]
7. Laben, C.A.; Brower, B.V. Process for Enhancing the Spatial Resolution of Multispectral Imagery Using Pan-Sharpening. U.S. Patent No. 6,011,875, 4 January 2000.
8. Chavez, P.; Sides, S.; Anderson, J. Comparison of Three Different Methods to Merge Multiresolution and Multispectral Data: Landsat TM and SPOT Panchromatic. *Photogramm. Eng. Remote Sens.* **1991**, *57*, 295–303. [CrossRef]
9. Huang, P.S.; Tu, T.-M. A New Look at IHS-like Image Fusion Methods (Vol 2, Pg 177, 2001). *Inf. Fusion* **2007**, *8*, 217–218. [CrossRef]
10. Rahmani, S.; Strait, M.; Merkurjev, D.; Moeller, M.; Wittman, T. An Adaptive IHS Pan-Sharpening Method. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 746–750. [CrossRef]
11. Choi, J.; Yu, K.; Kim, Y. A New Adaptive Component-Substitution-Based Satellite Image Fusion by Using Partial Replacement. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 295–309. [CrossRef]

12. 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]
13. Vivone, G.; Restaino, R.; Chanussot, J. Full Scale Regression-Based Injection Coefficients for Panchromatic Sharpening. *IEEE Trans. Image Process.* **2018**, *27*, 3418–3431. [CrossRef]
14. Khan, M.M.; Chanussot, J.; Condat, L.; Montanvert, A. Indusion: Fusion of Multispectral and Panchromatic Images Using the Induction Scaling Technique. *IEEE Geosci. Remote Sens. Lett.* **2008**, *5*, 98–102. [CrossRef]
15. Imani, M. Adaptive Signal Representation and Multi-Scale Decomposition for Panchromatic and Multispectral Image Fusion. *Future Gener. Comput. Syst. Int. J. Escience* **2019**, *99*, 410–424. [CrossRef]
16. Li, S.; Yang, B. A New Pan-Sharpening Method Using a Compressed Sensing Technique. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 738–746. [CrossRef]
17. Yin, H. Sparse Representation Based Pansharpening with Details Injection Model. *Signal Process.* **2015**, *113*, 218–227. [CrossRef]
18. Jian, L.; Wu, S.; Chen, L.; Vivone, G.; Rayhana, R.; Zhang, D. Multi-Scale and Multi-Stream Fusion Network for Pansharpening. *Remote Sens.* **2023**, *15*, 1666. [CrossRef]
19. Liu, N.; Li, W.; Sun, X.; Tao, R.; Chanussot, J. Remote Sensing Image Fusion With Task-Inspired Multiscale Nonlocal-Attention Network. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
20. Iftene, M.; Arabi, M.E.A.; Karoui, M.S. Transfering super resolution convolutional neural network for remote. Sensing data sharpen-ing. In Proceedings of the Name of IEEE Transfering Super Resolution Convolutional Neural Network for Remote. Sensing Data Sharpening, Amsterdam, The Netherlands, 23–26 September 2018.
21. Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; Paisley, J. PanNet: A Deep Network Architecture for Pan-Sharpening. In Proceedings of the Name of 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1753–1761.
22. Peng, J.; Liu, L.; Wang, J.; Zhang, E.; Zhu, X.; Zhang, Y.; Feng, J.; Jiao, L. PSMD-Net: A Novel Pan-Sharpening Method Based on a Multiscale Dense Network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4957–4971. [CrossRef]
23. Zheng, Y.; Li, J.; Li, Y.; Cao, K.; Wang, K. Deep Residual Learning for Boosting the Accuracy of Hyperspectral Pansharpening. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1435–1439. [CrossRef]
24. Aiazzi, B.; Baronti, S.; Selva, M. Improving Component Substitution Pansharpening through Multivariate Regression of MS plus Pan Data. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 3230–3239. [CrossRef]
25. Masi, G.; Cozzolino, D.; Verdoliva, L.; Scarpa, G. Pansharpening by Convolutional Neural Networks. *Remote Sens.* **2016**, *8*, 594. [CrossRef]
26. Liu, X.; Liu, Q.; Wang, Y. Remote Sensing Image Fusion Based on Two-Stream Fusion Network. *Inf. Fusion* **2020**, *55*, 1–15. [CrossRef]
27. Yuan, Q.; Wei, Y.; Meng, X.; Shen, H.; Zhang, L. A Multiscale and Multidepth Convolutional Neural Network for Remote Sensing Imagery Pan-Sharpening. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 978–989. [CrossRef]
28. Cai, J.; Huang, B. Super-Resolution-Guided Progressive Pansharpening Based on a Deep Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5206–5220. [CrossRef]
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Name of Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
30. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 87–110. [CrossRef]
31. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the Name of 2021 IEEE/CVF International Conference on Computer Vision (ICCV 2021), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.
32. Guan, P.; Lam, E.Y. Multistage Dual-Attention Guided Fusion Network for Hyperspectral Pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
33. Wald, L.; Ranchin, T.; Mangolini, M. Fusion of Satellite Images of Different Spatial Resolutions: Assessing the Quality of Resulting Images. *Photogramm. Eng. Remote Sens.* **1997**, *63*, 691–699. [CrossRef]
34. Plaza, A.; Benediktsson, J.A.; Boardman, J.W.; Brazile, J.; Bruzzone, L.; Camps-Valls, G.; Chanussot, J.; Fauvel, M.; Gamba, P.; Gualtieri, A.; et al. Recent Advances in Techniques for Hyperspectral Image Processing. *Remote Sens. Environ.* **2009**, *113*, S110–S122. [CrossRef]
35. Ungar, S.; Pearlman, J.; Mendenhall, J.; Reuter, D. Overview of the Earth Observing One (EO-1) Mission. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 1149–1159. [CrossRef]
36. 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.
37. Alparone, L.; Wald, L.; Chanussot, J.; Thomas, C.; Gamba, P.; Bruce, L.M. Comparison of Pansharpening Algorithms: Outcome of the 2006 GRS-S Data-Fusion Contest. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 3012–3021. [CrossRef]
38. Yuhas, R.H.; Goetz, A.F.H. Discrimination among Semi-Arid Landscape Endmembers Using the Spectral Angle Mapper (SAM) Algorithm. In Proceedings of the Name of Summaries of the Third Annual JPL Airborne Geoscience Workshop, Pasadena, CA, USA, 7 June 1992.

39. Karunasingha, D.S.K. Root Mean Square Error or Mean Absolute Error? Use Their Ratio as Well. *Inf. Sci.* **2022**, *585*, 609–629. [CrossRef]
40. Wald, L. Data Fusion. Definitions and Architectures—Fusion of Images of Different Spatial Resolutions. In Proceedings of the Name of Fusion of Earth Data: Merging Point Measurements, Raster Maps, and Remotely Sensed image, Washington, DC, USA, 11 October 2002.
41. Nan, Y.; Zhang, H.; Zeng, Y.; Zheng, J.; Ge, Y. Intelligent Detection of Multi-Class Pitaya Fruits in Target Picking Row Based on WGB-YOLO Network. *Comput. Electron. Agric.* **2023**, *208*. [CrossRef]
42. Zhou, J.; Zhang, Y.; Wang, J. RDE-YOLOv7: An Improved Model Based on YOLOv7 for Better Performance in Detecting Dragon Fruits. *Agronomy* **2023**, *13*, 1042. [CrossRef]
43. Li, A.; Zhao, Y.; Zheng, Z. Novel Recursive BiFPN Combining with Swin Transformer for Wildland Fire Smoke Detection. *FORESTS* **2022**, *13*, 2032. [CrossRef]