remote sensing

MDPI

Article

A Three-Branch Pansharpening Network Based on Spatial and Frequency Domain Interaction

Xincan Wen [1,2], Hongbing Ma [1,2,3,*] and Liangliang Li [4]

[1] School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China; 107552201337@stu.xju.edu.cn
[2] Key Laboratory of Signal Detection and Processing, Xinjiang University, Urumqi 830046, China
[3] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[4] School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China; leeliangliang@163.com
* Correspondence: hbma@tsinghua.edu.cn

Abstract: Pansharpening technology plays a crucial role in remote sensing image processing by integrating low-resolution multispectral (LRMS) images and high-resolution panchromatic (PAN) images to generate high-resolution multispectral (HRMS) images. This process addresses the limitations of satellite sensors, which cannot directly capture HRMS images. Despite significant developments achieved by deep learning-based pansharpening methods over traditional approaches, most existing techniques either fail to account for the modal differences between LRMS and PAN images, relying on direct concatenation, or use similar network structures to extract spectral and spatial information. Additionally, many methods neglect the extraction of common features between LRMS and PAN images and lack network architectures specifically designed to extract spectral features. To address these limitations, this study proposed a novel three-branch pansharpening network that leverages both spatial and frequency domain interactions, resulting in improved spectral and spatial fidelity in the fusion outputs. The proposed method was validated on three datasets, including IKONOS, WorldView-3 (WV3), and WorldView-4 (WV4). The results demonstrate that the proposed method surpasses several leading techniques, achieving superior performance in both visual quality and quantitative metrics.

Keywords: pansharpening; deep learning; three branch; interaction

check for updates

Academic Editors: Seongwook Lee and Byung-Kwan Kim

Received: 30 September 2024
Revised: 18 November 2024
Accepted: 27 November 2024
Published: 24 December 2024

Citation: Wen, X.; Ma, H.; Li, L. A Three-Branch Pansharpening Network Based on Spatial and Frequency Domain Interaction. *Remote Sens.* **2025**, *17*, 13. https://doi.org/10.3390/rs17010013

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

1. Introduction

High-resolution multispectral (HRMS) imagery has numerous applications in the domain of remote sensing image processing, such as target detection [1], military defense [2], climate monitoring [3], area classification [4], and so on. However, due to constraints associated with hardware and physical imaging capabilities, various satellite sensors are unable to directly capture HRMS data. Instead, they can only acquire paired low-resolution multispectral (LRMS) images, which contain extensive spectral information, along with panchromatic (PAN) images that offer high spatial resolution. Therefore, pansharpening technology has been developed to merge the information from these two types of images, thereby producing HRMS imagery that encompasses both rich spectral and spatial information.

Generally speaking, pansharpening methods can be divided into traditional methods and deep learning-based methods. Among them, traditional methods can be divided into three categories: the component substitution (CS) method [5], the multiresolution analysis (MRA) method [6], and the variational optimization (VO) method [7].

The CS method transforms multispectral images into a certain coordinate space, separates the spectral and spatial information, replaces the spatial information with PAN images that have been histogram-matched, and finally generates HRMS through inverse transformation. Classic algorithms include GS [8], IHS [9], and PCA [10]. This type of method will destroy the original spectral structure and produce fusion results with serious spectral distortion.

The MRA method extracts spatial detail structures by designing high-pass filters and then injects these spatial detail structures into multispectral images to obtain HRMS. Representative algorithms include GLP [11], HPF [12], and SFIM [13]. The spectral information preservation ability of this type of method is good, but during the process of injecting spatial details, spatial details are lost and introduced as shadows, resulting in poor spatial quality of the fusion result.

The VO method constructs an energy function by analyzing the relationship between LRMS, PAN, and HRMS and optimizes the model by optimizing the energy function. Typical methods include observation model-based methods [14,15] and sparse representation-based methods [16,17]. This type of method relies heavily on prior information, which often has unpredictability, resulting in a certain degree of distortion in the spectral and spatial information of the final fusion result.

The pansharpening method utilizing deep learning techniques, particularly convolutional neural networks (CNNs), has garnered significant attention in recent years. The robust nonlinear feature extraction capabilities of CNNs have facilitated their extensive application in the domain of pansharpening. Drawing inspiration from super-resolution convolutional neural networks (SRCNNs) [18], Masi et al. [19] were among the first to incorporate CNNs into pansharpening networks, developing a three-layer convolutional network architecture to achieve pansharpening results. Building upon this foundation, Wei et al. [20] designed convolutional kernels of varying sizes to extract information at different scales from the source images. Recognizing that LRMS and PAN images contain different modalities of information, Liu et al. [21] proposed a dual-stream fusion network that employs independent branches to extract features from both LRMS and PAN source images. However, these independent branches, despite having identical structures, still struggle to effectively characterize spectral and spatial features. In response to this limitation, Yong et al. [22] developed separate channel and spatial attention (SA) mechanisms [23] to extract spectral and spatial features, respectively. They subsequently multiplied the channel weights by the spatial weights to derive overall weights, thereby guiding LRMS and PAN branches to enhance their feature extraction capabilities. To mitigate redundant information across different subnetworks, Wang et al. [24] introduced an adaptive feature fusion module (AFFM) that infers weight maps for various branches from the feature map, thereby increasing the network's flexibility. Other researchers reconstruct HRMS by extracting and combining common unique features. Some researchers reconstruct HRMS by extracting and combining common unique features. Xu et al. [25] used Convolutional Sparse Coding (CSC) to extract the lateral information of panchromatic images. At the same time, they split low-resolution multispectral images into panchromatic image-related feature maps and panchromatic image-independent feature maps. The former was regularized by the lateral information of the panchromatic image, and the proposed model was extended to a deep neural network using the principle of algorithm expansion techniques. Cao et al. [26] designed a CSC with two sets of filters (a common filter and a unique filter) to model PAN and MS images separately, and this model is called PanCSC. At the same time, they derived an effective optimization algorithm to optimize the PanCSC model. Inspired by the learnable iterative soft thresholding algorithm, Yin [27] proposed a coupled convolutional sparse coding-based panchromatic sharpening (PSCSC) model. The solution of PSCSC follows traditional algorithms, and PSCSC uses a deep unfolding strategy

to develop interpretable end-to-end depth panchromatic sharpening networks. Zhi et al. [28] proposed a module based on cross attention to extract common and unique features from MS and PAN images. Lin et al. [29] encodes LRMS and PAN into unique and common features, then averages the common features and combines them with the unique features in the source image to reconstruct the fused image. Deng et al. [30] developed a cross-convolutional attention network that dynamically adjusts the parameters of the convolutional kernels, thereby enhancing the interaction between the two branches to obtain complementary information. Zhou et al. [31] proposed a PAN-guided multispectral feature enhancement network that employs multi-scale convolutional blocks to aggregate features across multiple scales. Jia et al. [32] introduced an attention-based progressive network that incrementally enhances detail information to generate an enhanced version of LRMS that corresponds to the size of PAN. Concurrently, they utilized concatenated spectral data and multi-scale SA blocks to progressively extract spatial and spectral features for the reconstruction of HRMS images. Despite the promising spectral and spatial quality of the fusion results produced by existing methods, several challenges remain to be addressed as follows:

1. Many dual-branch architecture methods overlook the common features shared between LRMS and PAN images, such as those captured by [33], leading to architectural inefficiencies;
2. A lack of adaptive convolution in most current approaches weakens the network's flexibility and adaptability;
3. The majority of methods operate solely in the spatial domain, without incorporating network structures designed to process frequency domain information.

To address the identified issues, this study developed a three-branch pansharpening network that leverages the interaction between spatial and frequency domains. Each branch is responsible for extracting different types of information: one branch extracts spectral information from LRMS images, another captures spatial information from PAN images, and the third branch integrates common information from both sources. The research implemented a spectral feature extraction module designed to identify nonlinear relationships among the channels in LRMS, thereby facilitating the acquisition of spectral information. Additionally, a spatial feature extraction module was created to capture spatial information from PAN images. This module employs multi-scale blocks to extract spatial features at various scales, followed by the application of SA mechanisms, as well as Sobel and Laplacian operators, to derive general spatial features and to enhance edge and texture features, respectively. Following the interaction and fusion of these two feature sets, a directional perception module is introduced to embed positional information, thereby augmenting the expressive capacity of the spatial features. For the common features derived from both LRMS and PAN, this study utilized a concatenation method that combines the bands from each image type separately and designed a dynamic convolution mechanism based on the extracted spectral and spatial features. Notably, the content of the convolution kernel is adapted according to the spectral and spatial characteristics of the input image. Subsequently, the features obtained from both the spectral and spatial feature extraction modules are processed to derive their corresponding amplitude and phase using discrete Fourier transform (DFT). These amplitude and phase features are then integrated and subjected to inverse discrete Fourier transform (IDFT) to yield frequency domain features. The branch responsible for extracting common features from LRMS and PAN utilizes these features as spatial domain inputs, which are then interacted with and fused with the frequency domain features, ultimately leading to the reconstruction of HRMS images. In summary, the contributions can be delineated as follows:

1. A three-branch pansharpening network based on spatial and frequency domain interaction was proposed, considering the distinct modal characteristics of LRMS and PAN images, as well as their common features;
2. A dynamic convolution based on spectral–spatial feature extraction was designed to improve network flexibility and adaptability;
3. Spatial–frequency domain feature fusion module (SFFIM) was developed to achieve interactive fusion of spatial and frequency domain information, leading to more comprehensive and enriched fused image features.

2. Proposed Method

Figure 1 illustrates the overall proposed network architecture. The design of the network structure is approached from two perspectives: the upsampled version of the original LRMS images, along with their unique features, and the common features shared between them, where H represents the height of the image, W represents the width of the image, and B represents the number of bands present in LRMS. Acknowledging the differences between LRMS images and PAN images, this study developed a spectral feature extraction module to extract spectral information from LRMS images and a spatial feature extraction module to derive spatial information from PAN images. Furthermore, recognizing that LRMS images and PAN images share certain common features, the research implemented a common feature extraction module. Due to the specificity of each band in LRMS [34], the study focused on each band of LRMS images with PAN images individually as input for this branch. This approach facilitates substantial interaction between PAN image and each band of LRMS. Additionally, a feature fusion module was designed to adaptively integrate information from the various branches, ultimately performing feature reconstruction to yield HRMS images.

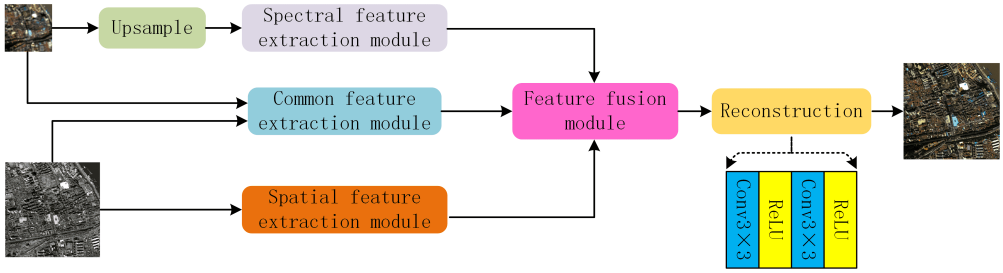


Figure 1. The architecture of the proposed method.

2.1. Spectral Feature Extraction Module

LRMS images encompass extensive spectral information, which is derived from the interrelationships among various spectral bands [35]. To facilitate the extraction of this spectral information, a spectral feature extraction module was developed (Figure 2). Firstly, a 3 × 3 convolution followed by a ReLU activation function is applied sequentially to preprocess LRMS image, resulting in the output $\widetilde{\boldsymbol{L}} \in R^{H\times W\times C}$. Subsequently, global average pooling (GAP) is performed on the output $\widetilde{\boldsymbol{L}}$, which is then processed through two parallel multi-layer perceptrons (MLPs) to generate two vectors ($\boldsymbol{L}_1, \boldsymbol{L}_2 \in R^{C\times 1}$) that represent the channel features. This entire procedure is expressed through Equation (1):

$$\boldsymbol{L}_i = MLP(GAP(\widetilde{\boldsymbol{L}})) \qquad i = 1, 2 \tag{1}$$

Subsequently, a matrix multiplication operation is conducted on matrices $\boldsymbol{L}_1$ and $\boldsymbol{L}_2^T$, where $\boldsymbol{L}_2^T$ represents the matrix transposition of $\boldsymbol{L}_2$. This is followed by the application of the Sigmoid activation function $\sigma(\cdot)$ in the subsequent equation, to derive the relationship graph $\boldsymbol{M}_C \in R^{C\times C}$ among the various channels. Concurrently, the shape of matrix $\widetilde{\boldsymbol{L}}$

is transformed to yield matrix $\overline{\boldsymbol{L}} \in R^{(H\times W)\times C}$. Following this, matrix multiplication is performed on matrices $\overline{\boldsymbol{L}}$ and $\boldsymbol{M}_C$, after which their shapes are adjusted to produce matrix $\boldsymbol{L}_3 \in R^{H\times W\times C}$. This entire process can be represented in Equation (2):

$$\boldsymbol{L}_3 = \mathrm{Re}((\sigma(\boldsymbol{L}_1\boldsymbol{L}_2^T))\overline{\boldsymbol{L}}) \tag{2}$$

where Re (·) represents the undergoing shape changes. Finally, a skip connection operation is executed to retain the information from the source image, thereby yielding the output of the module $\boldsymbol{L}_{out}$.

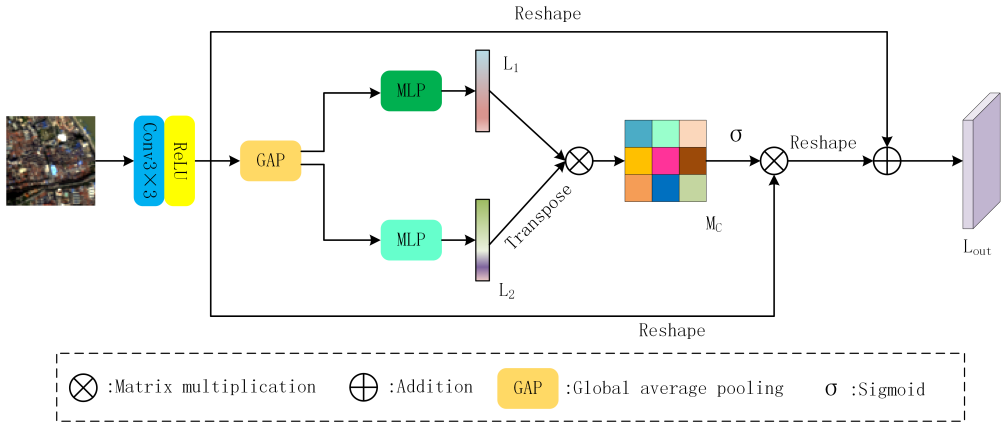


Figure 2. The architecture of the spectral feature extraction module.

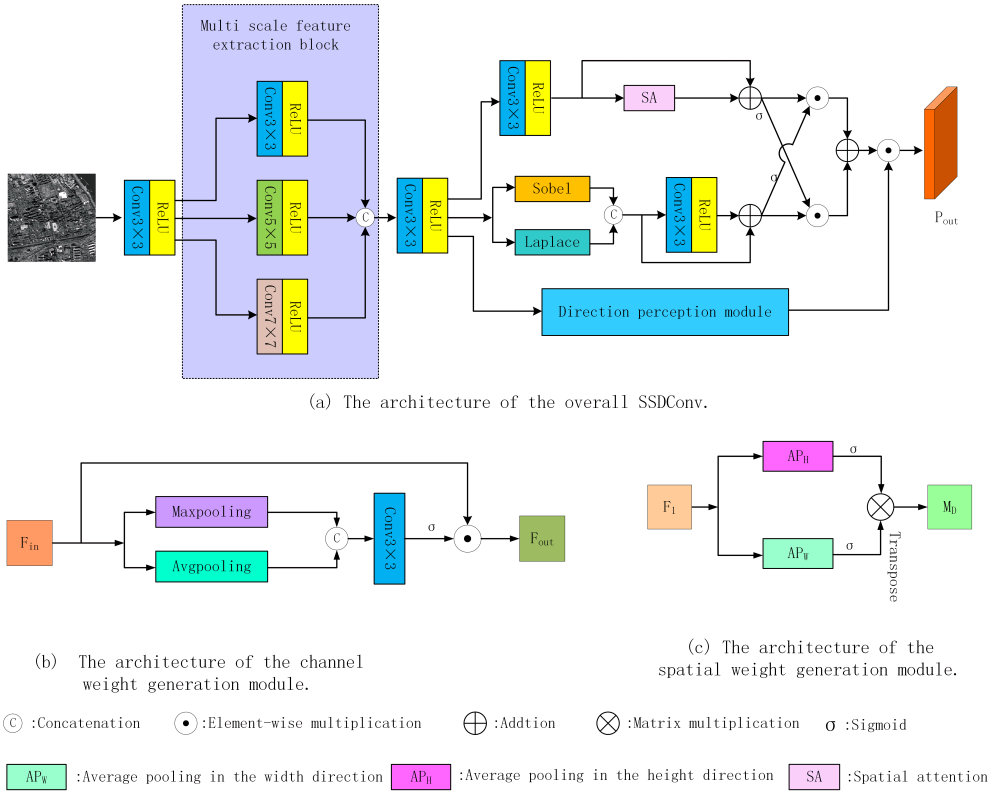
2.2. Spatial Feature Extraction Module

To acquire comprehensive spatial information in PAN imagery, this study developed a spatial feature extraction module comprising three parallel branches (Figure 3a). The subsequent sections will detail the functionality of these three branches in a top-to-bottom sequence (Figure 3).

Firstly, the PAN image was preprocessed using a 3 × 3 convolution and ReLU activation function, followed by a 3 × 3 convolution and ReLU activation function and a sequence of multi-scale blocks to preprocess the PAN image, thereby facilitating the preliminary extraction of multi-scale features. Subsequently, convolutional layers and ReLU activation functions were employed for channel dimensionality reduction, resulting in the preprocessed image $\widetilde{\boldsymbol{P}}$. In the first branch, this study utilized convolutional layers in conjunction with SA mechanisms to comprehensively extract spatial features from the PAN image while also implementing skip connection operations, which yielded the output of the branch $\boldsymbol{P}_1$.

However, addressing this issue in isolation neglects the extraction of finer textures from PAN images, which are predominantly found in details such as object edges. This observation has also been noted by several scholars. Drawing inspiration from [36,37], this study incorporated the Sobel operator and Laplacian operator in the second branch to extract fine texture details. The Sobel operator is a first-order differential operator that emphasizes pixels in proximity to the central pixel and has a high weighting on the nearest pixel to the central pixel, thereby aiding in noise reduction. It can be used to preserve strong texture features. Some researchers have also demonstrated the superiority of Sobel in extracting detailed information [38]. However, the Sobel operator is only a first-order operator, and compared with first-order operators, second-order differential operators have better edge localization ability and better pansharpening effect. The Laplace operator in second-order operators is widely used for fine feature extraction in the field of image processing, such as [37,39], which can extract weak textures of features that complement previous strong textures. Therefore, we chose the Laplace operator. Specifically, the research focused on the features extracted by these two operators, subsequently applying a

convolutional layer and executing a skip connection operation to derive the output of the branch $\boldsymbol{P}_2$.



(a) The architecture of the overall SSDConv.

(b) The architecture of the channel weight generation module.

(c) The architecture of the spatial weight generation module.

Figure 3. The architecture of the overall and partial spatial feature extraction module.

After extracting the features from the two aforementioned branches, a methodical approach was employed to integrate the information from these branches, ensuring that they mutually inform one another and that the details are examined. Specifically, for each branch, the Sigmoid activation function is utilized to derive the weight maps $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ corresponding to the feature maps of each branch. Subsequently, $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are cross-multiplied with $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$. Ultimately, the information from the two branches, following their interaction, is combined to yield $\boldsymbol{P}_{12}$ (Equation (3)):

$$\boldsymbol{P}_{12} = \boldsymbol{P}_1 \cdot \sigma(\boldsymbol{P}_2) + \boldsymbol{P}_2 \cdot \sigma(\boldsymbol{P}_1) \tag{3}$$

In addition to providing rich spatial information, including details and contours, PAN images also encompass positional information. Specifically, the rows and columns within PAN images exhibit interrelatedness. The first two branches do not distinguish between rows and columns of the image but use a unified operation, which means a lack of perception of positional information. To effectively capture this location information, this study developed a spatial perception module within the third branch (Figure 3c). Following the pooling, $\widetilde{\boldsymbol{P}}$, along the width and height dimensions of the feature map, feature representation in the height dimension $\boldsymbol{P}_{3h} \in R^{1\times W\times C}$ and feature representation of the width dimension $\boldsymbol{P}_{3w} \in R^{H\times 1\times C}$ were derived. Subsequently, a multiplication of $\boldsymbol{P}_{3h}$ and $\boldsymbol{P}_{3w}$ was performed along the channel dimension, applying the Sigmoid activation function to generate the direction-aware map $\boldsymbol{M}_D$. Ultimately, the multiplication of variable $\boldsymbol{M}_D$ with $\boldsymbol{P}_{12}$, conducted channel by channel and element by element, makes each element in $\boldsymbol{P}_{12}$ obtain a weight map with directional information, serves to embed the positional information into $\boldsymbol{P}_{12}$, thereby yielding the output of the spatial feature extraction module $\boldsymbol{P}_{out}$. This process is articulated in Equation (4) as follows:

$$\boldsymbol{P}_{out} = \sigma(\boldsymbol{P}_{3w} * \boldsymbol{P}_{3h}) \odot \boldsymbol{P}_{12} \tag{4}$$

where $*$ represents the matrix multiplication along the channel direction and $\odot$ represents the multiplication channel by channel and element by element.

2.3. Common Feature Extraction Module

The integration of spatial information from LRMS images and spectral information from PAN images reveals a shared informational basis between the two modalities. To effectively harness this common information, this study developed a dedicated feature extraction module, the architecture of which is depicted in Figure 4. Given the unique characteristics of each spectral band in LRMS images, the research implemented a dynamic convolution kernel, termed spectral spatial feature extraction (SSDConv), tailored for the corresponding branch of each band, rather than employing conventional convolution with uniform parameters. This approach not only takes into account the different properties of each band but also enhances the model's adaptability. The detailed structure of SSDConv is presented in Figure 5a.

The fundamental aspect of generating dynamic convolution kernels involves acquiring the weight matrix of local blocks $\boldsymbol{W} \in R^{k \times k \times C_{in}}$, where the local blocks present the patch segments created by concatenating the PAN image with each band of LRMS data, followed by the segmentation of the feature map through preprocessing in the convolutional layer; k and C_{in} represent the dimensions and the number of channels of the input local blocks, respectively. Given that the methodology for generating dynamic convolution kernels for each branch corresponding to a specific band is consistent, a particular band was used as a representative example and select a local block $\boldsymbol{F}_P \in R^{k \times k \times C_{in}}$ to elucidate the process of generating dynamic convolution kernels.

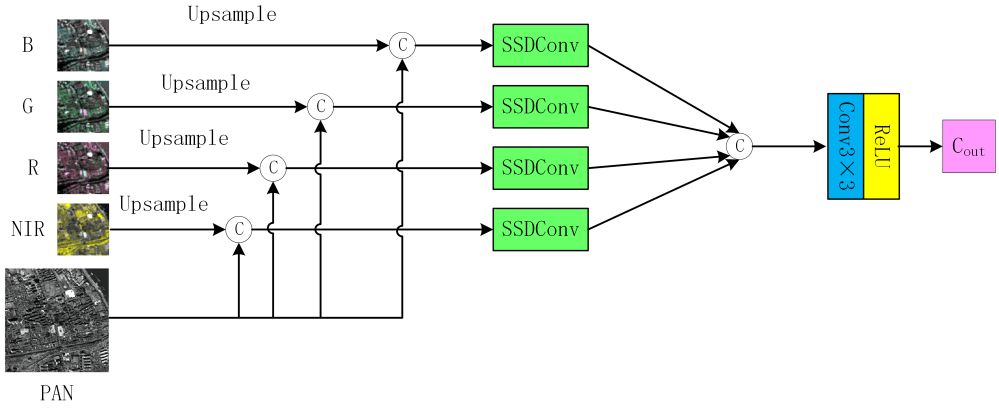


Figure 4. The architecture of the overall common feature extraction module.

$\boldsymbol{F}_P$ comprises information derived from a specific band of LRMS and PAN imagery. Consequently, numerous studies examined the channel and spatial characteristics of the feature map, leading to the development of three-branch feature extraction architectures that address both channel (one-dimensional) and spatial (two-dimensional) dimensions. However, the network structures of these three branches are identical and do not adequately account for the inherent differences between spatial and spectral features. To address this limitation, this study proposed a dynamic convolution approach focused on spectral–spatial feature extraction. For channel features, the research employed the "squeeze excitation" operation [40] (SE(·) in Equation (5)) to derive the weight for the channel dimension $\boldsymbol{W}_C \in R^{1 \times 1 \times C}$. In terms of spatial features, this study utilized the SA mechanism $SA_W(\cdot)$ and $SA_H(\cdot)$ in Equation (5), to generate the corresponding weight maps $\boldsymbol{W}_w \in R^{H \times 1 \times C}$

and $\boldsymbol{W}_h \in R^{1\times W\times C}$ for the width and height dimensions of the feature map, respectively. Ultimately, these three weights were combined to obtain $\boldsymbol{W}$ (Equation (5)):

$$\boldsymbol{W} = SE(\boldsymbol{F}_P) \odot (SA_W(\boldsymbol{F}_P) \otimes SA_H(\boldsymbol{F}_P)) \tag{5}$$

where $\odot$ represents the channel-wise multiplication and $\otimes$ represents the matrix multiplication along the channel direction. Following this, an inner product operation is conducted between W and a set of candidate kernels, $K \in R^{C_{out}\times k\times k\times C_{in}}$, to derive a dynamic convolution kernel. Subsequently, this dynamic convolution kernel is employed to execute convolution operations on the input local block, resulting in the corresponding features of the local block $\boldsymbol{F'}_P$ (Equation (6)):

$$\boldsymbol{F'}_P = (W \odot K) * \boldsymbol{F}_P \tag{6}$$

where $\odot$ represents the inner product operation and $*$ represents the convolution operation.

At the conclusion of the common feature extraction module, this study focused on features from various bands and subsequently processed them through convolutional layers followed by ReLU activation functions to derive the output of the module $\boldsymbol{C}_{out}$.

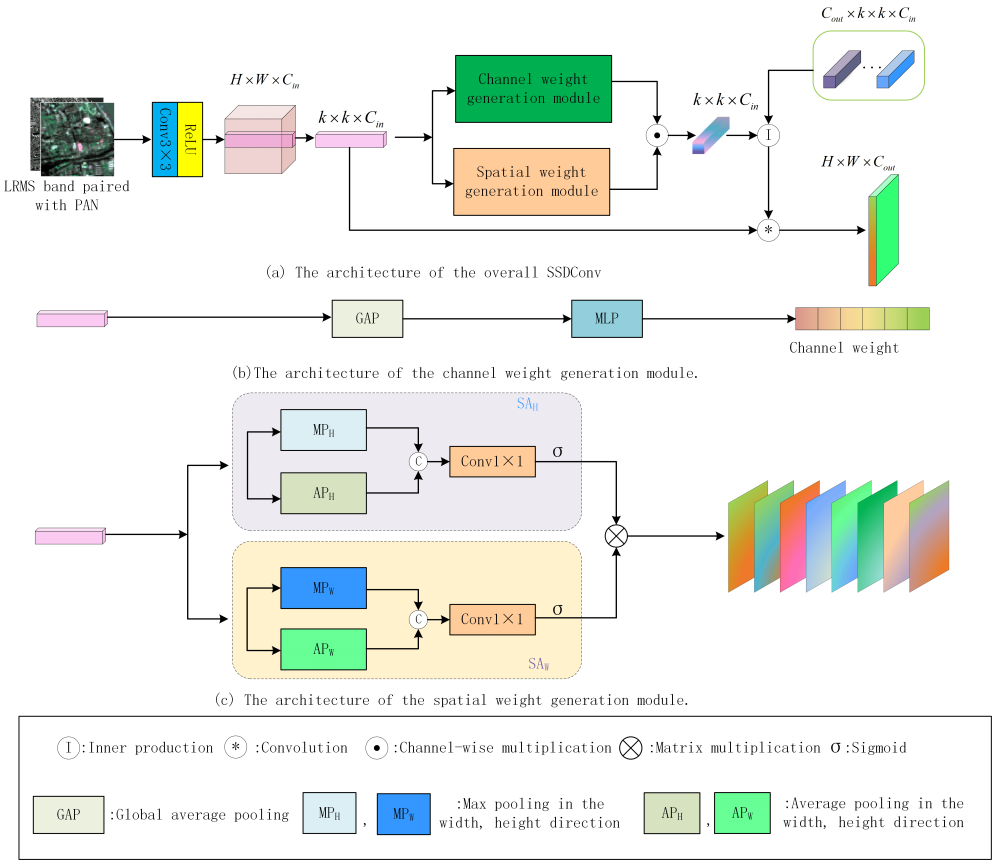


Figure 5. The architecture of the overall and partial SSDConv.

2.4. Space–Frequency Domain Feature Interaction Fusion Module

Most existing methodologies primarily focus on designing fusion strategies within the spatial domain, employing techniques such as direct addition, concatenation, or the application of adaptive weights. While these approaches have yielded satisfactory fusion results, there has been a lack of corresponding network designs that address frequency domain information. Furthermore, although certain methods, such as FAFNet [41], extract features in the frequency domain, they tend to overlook the information present in the spatial domain. To enhance the accuracy of the fusion results, this study developed SFFIM (Figure 6). The research used DFT to $\boldsymbol{L}_{out}$ and $\boldsymbol{P}_{out}$ individually to derive their respective amplitudes ($\boldsymbol{A}_m$, $\boldsymbol{A}_p$) and phases ($\boldsymbol{P}_m$, $\boldsymbol{P}_p$). Subsequently, this study focused on the ampli-

tudes and phases from the different branches separately, thereby integrating the amplitude information $\boldsymbol{A}$ and phase information $\boldsymbol{P}$. This is performed to ensure that the amplitude and phase information in the frequency domain of $\boldsymbol{L}_{out}$ and $\boldsymbol{P}_{out}$ are fully communicated. If it is spatial domain interaction between $\boldsymbol{L}_{out}$ and $\boldsymbol{P}_{out}$, then only conventional strategies such as addition or concatenation can be used. Therefore, we consider the interaction between $\boldsymbol{L}_{out}$ and $\boldsymbol{P}_{out}$ in the frequency domain as an effective way to integrate the information from these two branches. Finally, IDFT was performed on the combined A and $\boldsymbol{P}$ to obtain the frequency domain feature $\boldsymbol{F}_{mp}$ (Equation (7)):

$$\begin{aligned} \boldsymbol{A}_m, \boldsymbol{P}_m &= DFT(\boldsymbol{L}_{out}) \\ \boldsymbol{A}_p, \boldsymbol{P}_p &= DFT(\boldsymbol{P}_{out}) \\ \boldsymbol{A} &= Con\big(Concat[\boldsymbol{A}_m, \boldsymbol{A}_p]\big) \\ \boldsymbol{P} &= Con\big(Concat[\boldsymbol{P}_m, \boldsymbol{P}_p]\big) \\ \boldsymbol{F}_{mp} &= IDFT(\boldsymbol{A}, \boldsymbol{P}) \end{aligned} \tag{7}$$

where Con $(\cdot)$ represents the convolutional layer of two consecutive layers in Figure 6.

For $\boldsymbol{C}_{out}$, two successive convolutional layers are employed to extract the spatial domain feature $\boldsymbol{S}_{mp}$. If $\boldsymbol{C}_{out}$ is processed in the frequency domain by simply performing DFT and IDFT transformations, there is no information exchange involved. Following this, $\boldsymbol{S}_{mp}$ and $\boldsymbol{F}_{mp}$ are fused by separately applying Sigmoid activation functions to each of them to derive their respective feature map weights, $\boldsymbol{M}_S$ and $\boldsymbol{M}_F$. Subsequently, the weights $\boldsymbol{M}_S$ and $\boldsymbol{M}_F$ are cross-multiplied with $\boldsymbol{S}_{mp}$ and $\boldsymbol{F}_{mp}$, respectively, and the results are concatenated to produce the output feature $\boldsymbol{F}_{out}$. The detailed process is expressed in Equation (8):

$$\boldsymbol{F}_{out} = Concat[\boldsymbol{S}_{mp} \cdot \sigma(\boldsymbol{F}_{mp}), \boldsymbol{F}_{mp} \cdot \sigma(\boldsymbol{S}_{mp})] \tag{8}$$

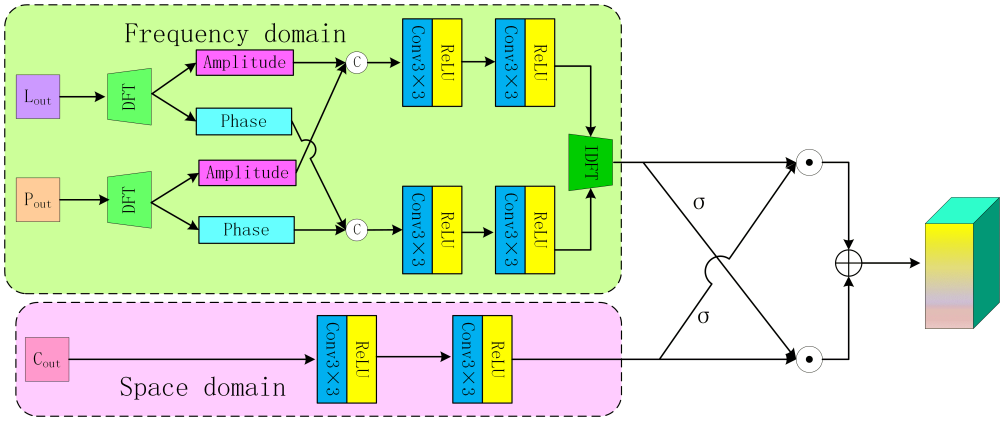


Figure 6. The architecture of the SFFIM.

2.5. Loss Function Design

Most pansharpening networks primarily utilize ℓ_1 loss function during the training process to evaluate the discrepancy between the fusion output and the reference image. This methodology neglects the supervision of the outputs from the intermediate layers of the network [42]. The unavoidable presence of redundant information across various branches can adversely impact the quality of the fusion results. To mitigate the redundancy inherent in different modalities, this study proposed the incorporation of mutual information constraints, which aim to minimize the mutual information between different branches ($MI(\cdot, \cdot)$ in Equation (9) [28]. This approach is intended to reduce redundancy and enhance the feature representation capabilities of the various branches. The mutual information loss function is presented in Equation (9).

$$\boldsymbol{L}_{MI} = MI(\boldsymbol{L}_{out}, \boldsymbol{P}_{out}) + MI(\boldsymbol{L}_{out}, \boldsymbol{C}_{out}) + MI(\boldsymbol{C}_{out}, \boldsymbol{P}_{out}) \tag{9}$$

Additionally, it was considered that the frequency domain characteristics of the output should closely match those of the ground truth (GT). To address this, a frequency domain loss function was designed to impose constraints on the frequency domain features (Equation (10)):

$$\boldsymbol{L}_F = \|A(\boldsymbol{HRMS}) - A(\boldsymbol{GT})\|_1 + \|P(\boldsymbol{HRMS}) - P(\boldsymbol{GT})\|_1 \tag{10}$$

where $A(\boldsymbol{HRMS})$ and $P(\boldsymbol{HRMS})$ represent the amplitude and phase of the fusion result, respectively, while $A(\boldsymbol{GT})$ and $P(\boldsymbol{GT})$ represent the amplitude and phase of the reference image, respectively.

In summary, the developed loss function is expressed as follows:

$$\boldsymbol{L} = \|\boldsymbol{HRMS} - \boldsymbol{GT}\|_1 + \alpha \boldsymbol{L}_F + \beta \boldsymbol{L}_{MI} \tag{11}$$

where $\alpha = \beta = 0.1$, as indicated by [43,44].

3. Experimental Results

3.1. Datasets

To assess the efficacy of the suggested methodology, the research carried out validations using simulated as well as real datasets acquired from the IKONOS, WV3, and WV4 satellites. The LRMS images from IKONOS and WV4 contain four spectral bands, while the LRMS images from WV3 include eight spectral bands. The PAN images from all three satellite datasets are composed of single-band imagery. Specifically, the spatial resolutions for the LRMS and PAN images in the IKONOS dataset are 4 m and 1 m, respectively. In the case of the WV4 and WV3 datasets, the spatial resolutions for both LRMS and PAN images are 1.24 m and 0.31 m, respectively. Additional details and acquisition approaches for these three datasets are described in [45]. The assessment encompasses experiments with both simulated as well as actual data to objectively validate the approach's effectiveness. While there exists no directly relevant simulation dataset for the simulated data assessment, training pansharpening networks relying on CNN is carried out on simulated datasets. As a result, we apply the Wald protocol, which involves applying MTF [46] filtering to the original LRMS and PAN images, followed by 4× down-sampling. By following this process, we obtain 1200, 1500, and 960 pairs of simulated images from the IKONOS, WV4, and WV3 datasets, respectively. The original LRMS image serves as the reference image. For every simulated dataset, 80% is used for training, 10% for validation, and 10% for testing. The dimensions of the employed LRMS and PAN images are 64 × 64 × 4 (64 × 64 × 8) and 256 × 256, respectively. For the real data assessment, the original LRMS as well as PAN images (with dimensions of 256 × 256 × 4 (256 × 256 × 8) and 1024 × 1024, respectively) are directly used as network inputs for testing (the original image pairs are only employed as the testing dataset), with every satellite's testing dataset having the same amount of image pairs as the simulated data assessment.

3.2. Evaluation Indicators and Comparison Methods

On the simulated dataset, we use ERGAS (erreur relative globale adimensionnelle de synthese) [47], UIQI (universal image quality index) [48], SAM (spectral angle mapping) [49], SCC (structural correlation coefficient) [50], and Q4 (Q8) [47]. as evaluation metrics. It is worth noting that the higher the values of UIQI, SCC, and Q4(Q8), the higher the quality, while the lower the values of ERGAS, RMSE, RASE, and SAM, the better the performance. On the real dataset, we use the quality without reference (QNR), spectral distortion index (D_λ), and spatial distortion index (D_s) to evaluate the full-resolution per-

formance. The larger the QNR, the better the quality, while the smaller the D_λ and D_s [51], the smaller the distortion. This article compares our proposed method with some classic methods (GS [8], HPF [12], PanNet [52], MSDCNN [20], TFNet [21], GPPNN [53], SSAF [22], SSE [24], CIKA [30], Band aware [31], and PAPS [32]) to verify the effectiveness of our proposed method. Among them, GS and HPF are traditional methods, while the rest are DL methods.

3.3. Experimental Details

All traditional methods were obtained by running MATLAB R2023a on a computer with Intel (R) Core (TM) i5-13500H processors and 16GB of memory. All DL-based network training and testing are implemented using the PyTorch 1.13.1 framework and accelerated using a GPU with a GeForce RTX 4090 graphics card. For the proposed method, the Adam optimizer is used to train the model with 1000 training rounds and a learning rate set to 1×10^{-4}. The model decays to the original 0.1 for every 200 training rounds, Set the batch size to 4.

3.4. Experimental Results on Simulated Datasets

Figures 7 and 8 depict the fusion outcomes of different approaches on the IKONOS simulated dataset as well as their residual plots with the reference image, respectively. Overall, the fusion results of the HPF approach seem quite blurry, and there is a certain degree of loss of spectral information in the fusion results of the GPPNN method. Upon analyzing the magnified region, the color of the road in the GS approach's fusion outcome exhibits some deviation from the surrounding area, and the spectral data in the SSE approach's fusion outcome demonstrate certain discrepancies compared with the reference image. The fusion result of the PanNet method shows that the circular texture along the highway edge is very blurry. The fusion outcomes of the SSAF and MSDCNN approaches show some fuzzy characteristics at the highway edges. Among the remaining approaches, the fusion outcomes of Band aware, CIKA, and our proposed approach bear relatively close resemblance to the reference image, and the errors of these three approaches are also comparable in Figure 8. However, the quantitative analysis in Table 1 reveals that all indicators of our proposed approach surpassed those of other approaches. The best and suboptimal values are highlighted in bold and underlined, respectively(as noted in subsequent tables).

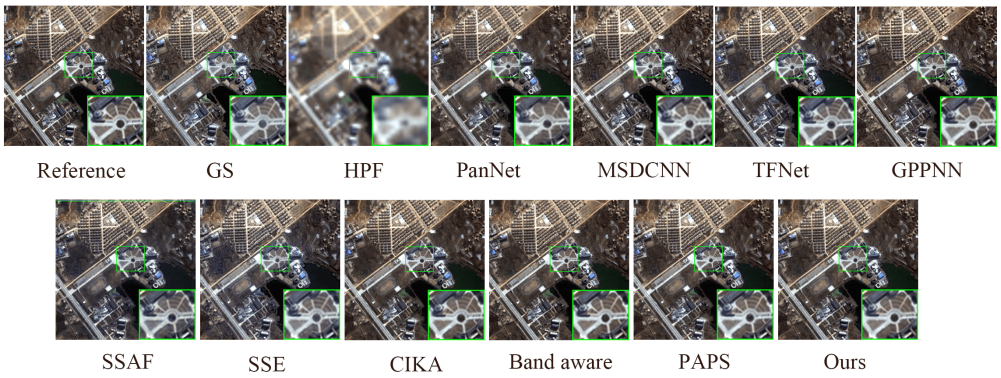


Figure 7. The fusion results of various methods on the IKONOS simulation dataset.

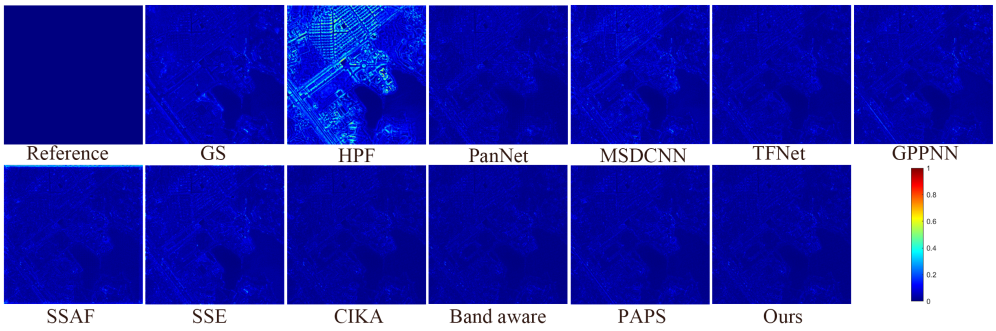


Figure 8. The absolute error maps between the fusion results of all methods and the reference image on the IKONOS simulation dataset.

Table 1. Quantitative evaluations using different methods on the IKONOS dataset.

	ERGAS	UIQI	SAM	SCC	Q4	QNR	D_λ	D_s
GS	2.7475	0.9328	3.8596	0.8661	0.6477	0.7926	0.0794	0.1411
HPF	2.9985	0.8793	2.9933	0.7100	0.5092	0.7942	0.0979	0.1252
PanNet	1.5637	0.9743	2.1368	0.9341	0.7941	0.8925	0.0461	0.0657
MSDCNN	2.0111	0.9565	2.6961	0.9002	0.7182	0.8997	0.0600	0.0438
TFNet	1.5739	0.9752	2.0298	0.9424	0.7829	0.8774	0.0727	0.0568
GPPNN	1.5834	0.9742	2.0745	0.9423	0.7948	0.8456	0.0841	0.0838
SSAF	2.1664	0.9393	2.7219	0.9125	0.7064	0.8141	0.0658	0.1331
SSE	1.8113	0.9656	2.4252	0.9252	0.7507	0.8645	0.0702	0.0746
CIKA	1.3815	0.9799	1.7580	0.9501	0.8315	0.9094	0.0403	0.0542
Band aware	1.3523	0.9805	1.7095	0.9521	0.8361	0.9072	**0.0332**	0.0650
PAPS	1.4452	0.9780	1.8512	0.9475	0.8152	0.8728	0.0700	0.0642
Ours	**1.2795**	**0.9825**	**1.6102**	**0.9567**	**0.8492**	**0.9181**	0.0377	**0.0478**

Figures 9 and 10 illustrate the fusion results for different techniques employed on the simulated WV4 dataset, in addition to their residual plots in comparison to the reference image. The HPF approach's fusion outcomes are severely impacted by spatial data loss, resulting in a blurry appearance of all buildings. The GS and MSDCNN approaches' fusion outcomes depict slightly brighter roof colors on the building in the upper-right corner, which ought to have been a dark gray shade. A small portion of the building's roof has been magnified as the area of interest and placed in the bottom-right corner for closer inspection. The enlarged area of the fusion result of the PanNet method is relatively blurry. The SSE and GPPNN approach's fusion outcome retains only a faint orange color on the roof, while the TFNet approach's outcome also fails to capture the corresponding roof color data accurately. The PAPS approach's fusion outcome exhibits significant color discrepancies in all roofs when compared with the reference image. The CIKA approach's fusion outcome features smoothed roof edges, while the other approaches' fusion outcomes bear a close resemblance to the reference image. Figure 10 clearly demonstrates that our proposed approach incurs the least amount of data loss in comparison to other techniques. Moreover, our approach achieved the highest indicator values across different simulated dataset indicators, as shown in Table 2.

Figures 11 and 12 present the fusion results of different approaches applied to the simulated WV3 dataset, in addition to their residual plots compared with the reference image. Figure 11 shows the HPF approach's fusion outcome appears significantly blurry overall. The industrial park's roof color, which should have been blue, is notably distorted in the fusion outcomes of the MSDCNN, GPPNN, TFNet, and SSE approaches, deviating considerably from the reference image. The GS approach's fusion outcomes also exhibit some degree of spectral distortion in the roof color. Upon analyzing the magnified area, it is

evident that the PanNet and SSAF's fusion outcome have a slightly lighter color, while the other approaches' fusion outcomes closely match the reference image within the enlarged region. Figure 12 clearly demonstrates that our proposed approach produces the smallest residual compared with other techniques. The quantitative analysis shown in Table 3 further confirms the outstanding performance of our introduced approach.

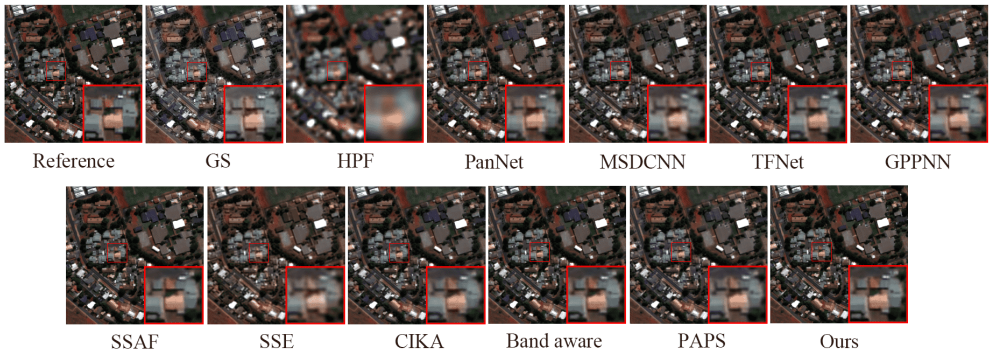


Figure 9. The fusion results of various methods on the WV4 simulation dataset.

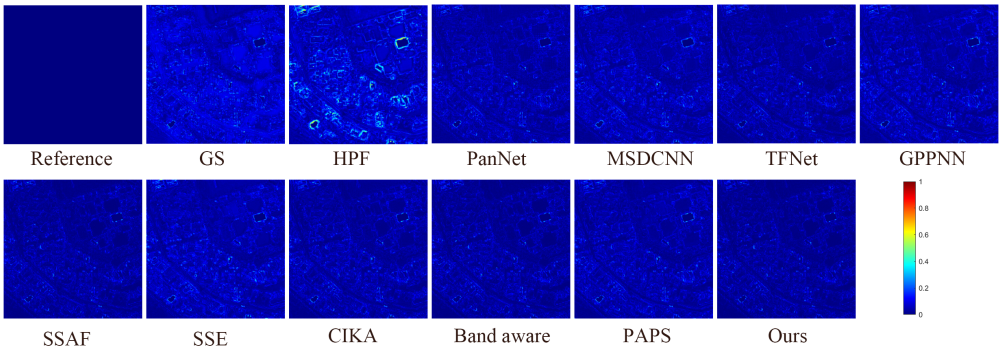


Figure 10. The absolute error maps between the fusion results of all methods and the reference image on the WV4 simulation dataset.

Table 2. Quantitative evaluations using different methods on the WV4 dataset.

	ERGAS	UIQI	SAM	SCC	Q4	QNR	D_λ	D_s
GS	2.7630	0.9667	2.8919	0.8823	0.6508	0.8581	0.0539	0.0946
HPF	3.1100	0.9594	2.3687	0.7802	0.5569	0.8872	0.0557	0.0626
PanNet	1.9192	0.9832	1.9704	0.9052	0.7448	0.9152	0.0357	0.0454
MSDCNN	2.3100	0.9801	2.5487	0.8904	0.6761	0.8906	0.0621	0.0511
TFNet	1.8147	0.9871	1.8626	0.9250	0.7381	0.8914	0.0610	0.0514
GPPNN	2.1848	0.9816	2.3348	0.9042	0.7250	0.8819	0.0606	0.0630
SSAF	1.5764	0.9879	1.6118	0.9364	0.8025	0.9279	0.0341	0.0396
SSE	2.3583	0.9780	2.7840	0.8831	0.6932	0.8426	0.0911	0.0753
CIKA	1.9091	0.9848	1.9392	0.9067	0.7566	0.9205	0.0379	0.0434
Band aware	1.5609	0.9898	1.6113	0.9374	0.7983	0.8953	0.0434	0.0646
PAPS	2.0560	0.9836	2.1098	0.9124	0.7388	0.9002	0.0616	0.0415
Ours	**1.3773**	**0.9924**	**1.4797**	**0.9522**	**0.8246**	**0.9336**	**0.0310**	**0.0375**

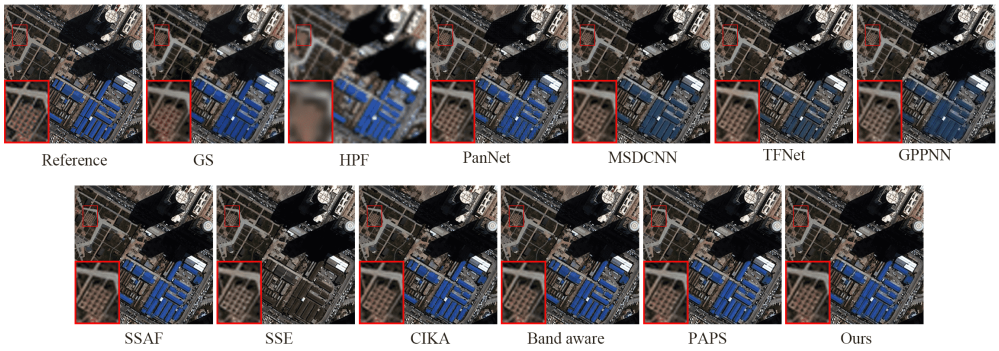


Figure 11. The fusion results of various methods on the WV3 simulation dataset.

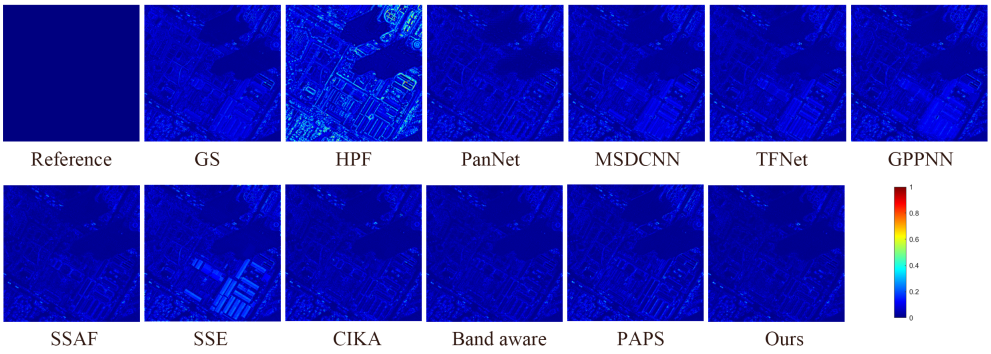


Figure 12. The absolute error maps between the fusion results of all methods and the reference image on the WV3 simulation dataset.

Table 3. Quantitative evaluations using different methods on the WV3 dataset.

	ERGAS	UIQI	SAM	SCC	Q8	QNR	D_λ	D_s
GS	6.0708	0.9088	6.9071	0.8641	0.6323	0.8277	0.0722	0.1231
HPF	8.0068	0.8368	6.8444	0.5927	0.4292	0.8134	0.0958	0.1176
PanNet	5.0879	0.9512	6.2120	0.8696	0.7055	0.8499	0.0668	0.1023
MSDCNN	4.7723	0.9560	5.5764	0.8929	0.7297	0.8276	0.0978	0.0968
TFNet	4.7073	0.9558	5.4749	0.8998	0.7089	0.8275	0.0884	0.1030
GPPNN	4.8313	0.9538	5.7267	0.8909	0.6999	0.7908	0.1092	0.1346
SSAF	3.0722	0.9814	4.1068	0.9598	0.7949	0.8383	0.0862	0.0988
SSE	4.9640	0.9522	5.9466	0.8893	0.7124	0.8127	0.0993	0.1135
CIKA	3.4862	0.9759	4.6022	0.9423	0.7862	0.8591	0.0744	0.0831
Band aware	3.0315	0.9826	4.1554	0.9612	0.8000	0.7761	0.1005	0.1430
PAPS	2.9046	0.9835	3.9959	0.9652	0.8028	0.8731	0.0588	0.0834
Ours	**2.6121**	**0.9865**	**3.6363**	**0.9730**	**0.8130**	**0.8886**	**0.0539**	**0.0657**

3.5. Experimental Results on Real Datasets

Figure 13 showcases the outcomes of different fusion techniques employed on the real IKONOS dataset. Due to the absence of reference images, spatial data are sourced from PAN images, while spectral data are derived from MS images. To facilitate the examination of small objects in every sub-figure, a heart-shaped building complex has been magnified as the region of interest and positioned in the top–left corner. The fusion outcomes generated by the GPPNN, SSAF, SSE, and MSDCNN approaches display a yellowish tint on the building rooftops, whereas the TFNet approach's outcome exhibits an orange hue, markedly diverging from the red color observed in the MS images. The PanNet and PAPS approaches yield lighter color in their fusion outcome, and the HPF approach's outcomes show some blurriness. The other approaches have largely succeeded in preserving both spectral and

spatial data. Our proposed approach demonstrates competitiveness, as evidenced by the quantitative assessment shown in Table 1.

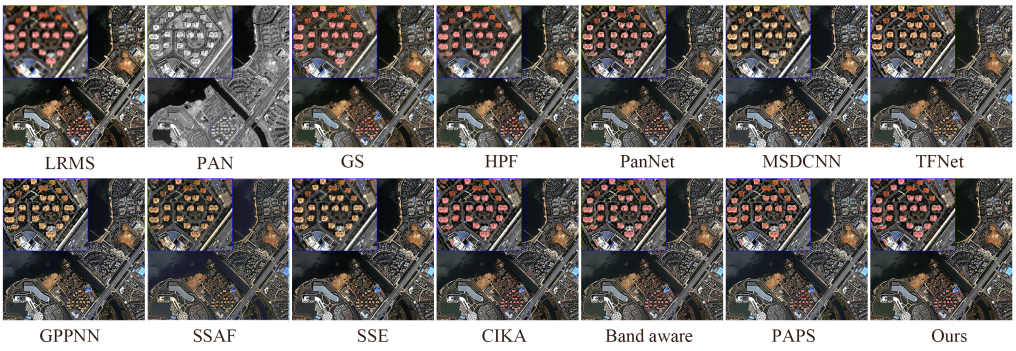


Figure 13. The fusion results of various methods on the IKONOS real dataset.

Figure 14 shows the fusion results of various methods on the real dataset of WV4. The vegetation color in the lower right corner of the river in the fusion result of the MSDCNN and GPPNN methods is slightly black, and the color of the corresponding area in the fusion result of the SSE method is also significantly different from that in the MS image. We choose a section of river and its surrounding banks as the area of interest, enlarge it, and place it in the upper left corner. The river color in the fusion result of the GS method is significantly lighter, while the river color in the fusion result of the SSAF method lacks brightness. The fusion result of the TFNet method shows that the soil color on the riverbank deviates from the soil color at the corresponding position in the MS image. The fusion results of Band aware and CIKA methods show a fuzzy phenomenon within the river. The fusion results of other methods almost retain the original spectral information and spatial phenomena, but the river lines in the fusion results of our proposed method are more natural. Table 2 also demonstrates that the proposed method outperforms other methods.

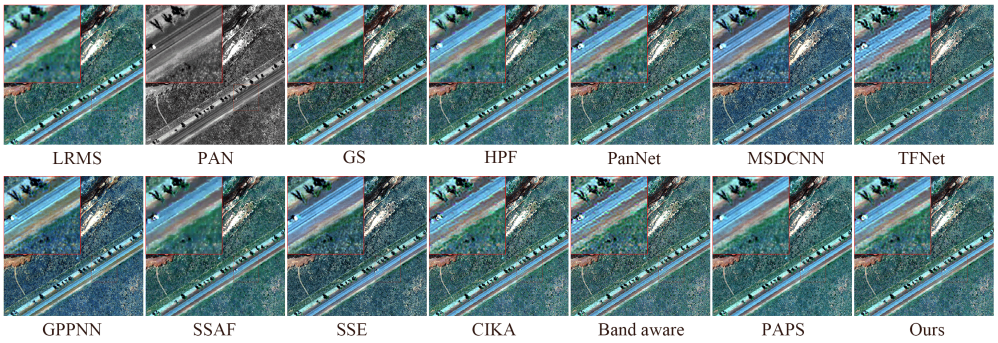


Figure 14. The fusion results of various methods on the WV4 real dataset.

Figure 15 presents the results of different fusion approaches applied to the real WV3 dataset. Akin to Figures 13 and 14, the real dataset does not include reference images. The PAN images capture spatial details, while the MS images contain spectral data. Upon closer inspection, it becomes apparent that the SSAF fusion output exhibits an overall green tint in the ocean's hue, deviating from the true color. Zooming in on specific regions reveals further disparities. The bevery color in the MSDCNN and SSE fusion outcomes fails to align with the spectral data accordingly within the MS image. Strikingly, the GPPNN and SSE approach's output completely omits the color data of the red vehicles, while the TFNet approach's fusion outcome preserves only a faint trace. Similarly, the GS fusion output displays inconsistencies in the bevery color's spectral data when compared with the MS image. Although other approaches successfully retain both spectral and spatial details

to varying degrees, our proposed approach stands out by generating a more vivid and visually appealing fused image. The superiority of the introduced approach is further corroborated by the overall advantages presented in Table 3.

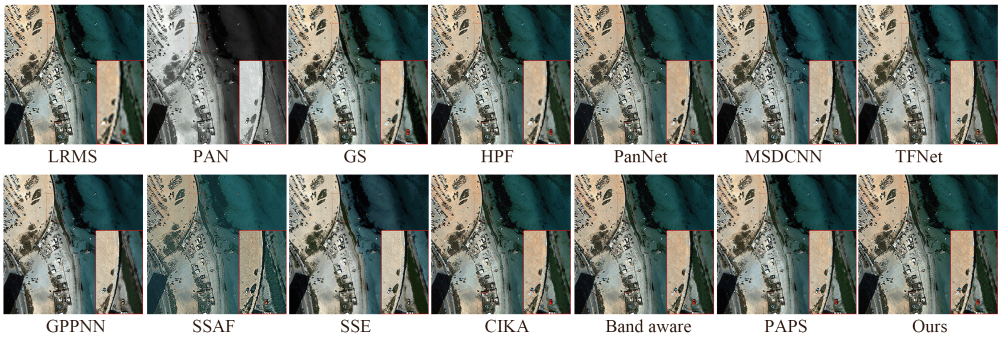


Figure 15. The fusion results of various methods on the WV3 real dataset.

3.6. Analysis of Ablation Experiment Results

In the ablation experiment, we will conduct network module ablation experiments and loss function partial ablation experiments separately and analyze the impact of each part on the overall performance of the model through experimental results.

We conducted ablation experiments on the IKONOS simulation dataset to analyze the contributions of each module in the proposed method. In this section, comparative experiments were conducted on removing different modules and their different combinations, with 'w/o' indicating that this module is not used. Figure 16 shows the fusion results after removing some modules, as well as the error maps between each fusion result and the reference image. There is not much difference between the several images, both overall and in the enlarged area. However, Table 4 shows the various indicators after removing each module. It can be seen that the network after removing the spatial frequency interaction module has the worst performance in all indicators, indicating that this module has the greatest impact on our proposed network model.

In addition to verifying the impact of network modules on the overall network, we also verified the influence of loss functions in each part on the fusion results. Similarly, this part was experimented on the IKONOS simulation dataset. 'W/o' means not using this loss function part. Figure 17 shows the mutual information loss function, spectral loss function, and fusion results after removing both from formula 10, as well as the error plots between each fusion result and the reference image. The quantitative evaluation of the fusion results is shown in Table 5. In Figure 17, overall, the fusion results of each version are not significantly different. From the enlarged area, it can be seen that there is a certain deviation between the green color of the fusion result of removing two loss functions and the reference image, and the soil part is almost not restored. The soil spectral information of the fusion result after removing L_{MI} also has a certain degree of loss, and the soil edge of the fusion result after removing L_F is not as clear as using the complete loss function. The various indicators in Table 5 further demonstrate the effectiveness of our designed loss function.

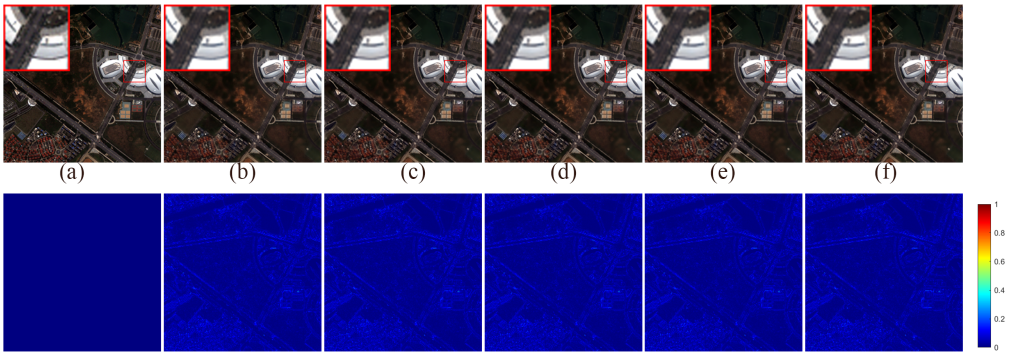


Figure 16. The fusion results after the removal of certain modules. (**a**) reference image, (**b**) w/o spectral feature extraction module, (**c**) w/o spatial feature extraction module, (**d**) w/o common feature extraction module, (**e**) w/o SFFIM, (**f**) ours.

Table 4. Quantitative evaluations of module ablation experiments on the simulated IKONOS dataset.

	ERGAS	UIQI	SAM	SCC	Q4
w/o Spectral feature extraction module	1.3125	0.9816	1.6760	0.9541	0.8430
w/o Spatial feature extraction module	1.3238	0.9813	1.6808	0.9535	0.8419
w/o Common feature extraction module	1.3076	0.9817	1.6599	0.9546	0.8445
w/o SFFIM	1.3293	0.9812	1.6944	0.9532	0.8397
ours	**1.2795**	**0.9825**	**1.6102**	**0.9567**	**0.8492**

Table 5. Quantitative evaluations of oss Abalation experiments on the simulated IKONOS dataset.

	ERGAS	UIQI	SAM	SCC	Q4
w/o L_{MI}	1.3744	0.9800	1.7581	0.9505	0.8331
w/o L_F	1.3614	0.9801	1.7309	0.9511	0.8350
only ℓ_1	1.4064	0.9794	1.8261	0.9492	0.8282
ours	**1.2795**	**0.9825**	**1.6102**	**0.9567**	**0.8492**

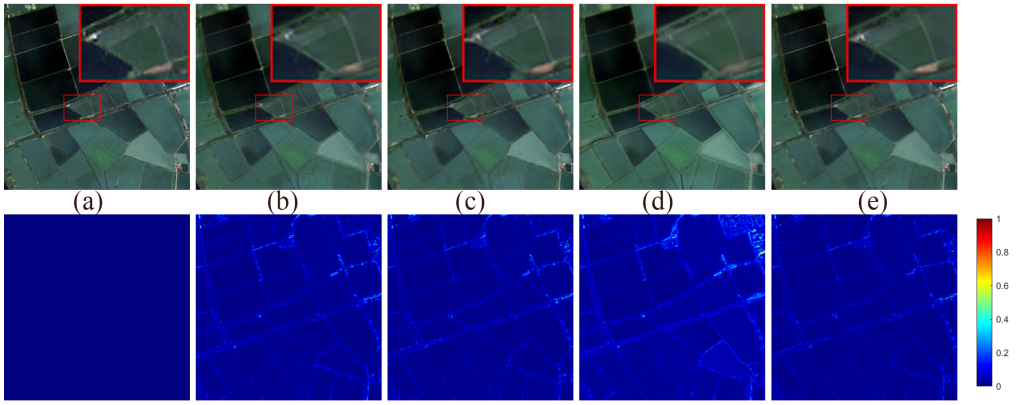


Figure 17. The fusion results after the removal of certain modules. (**a**) reference image, (**b**) w/o L_{MI}, (**c**) w/o L_F, (**d**) only ℓ_1, (**e**) ours.

3.7. Analysis of Experimental Results on Network Structure

In the network architecture experiment, we compared the performance differences between different network architectures and the fusion results of our proposed network architecture, which were tested on the IKONOS simulation dataset. Among them, single branch refers to the branch that retains the input of LRMS and PAN for band-by-band splicing, as well as the corresponding part of this branch in the fusion module after retention. Dual branch refers to the branch that retains the input of LRMS and the input of PAN,

as well as the corresponding part of this branch in the fusion module after retention. Direct splicing refers to not performing band-by-band splicing, and the rest is consistent with our proposed network model. Figure 18 shows the fusion results using different network structures and their errors with the reference image, while Table 6 presents the corresponding indicators. From the enlarged area, it can be seen that the colors of the two pools in the fusion result using a single-branch structure are significantly different from the reference image, and there are artifacts at the edge of the left pool. This is due to the differences in spatial and spectral information of the source image, which were not taken into account in the single-branch structure. The edge of the water pool in the fusion result using a dual branch structure is slightly smoothed, and the network structure using direct stitching is similar to our proposed model (possibly due to fully preserving the source image information), but slightly inferior in image clarity. The various indicators in Table 6 further validate the superiority of our designed network structure.

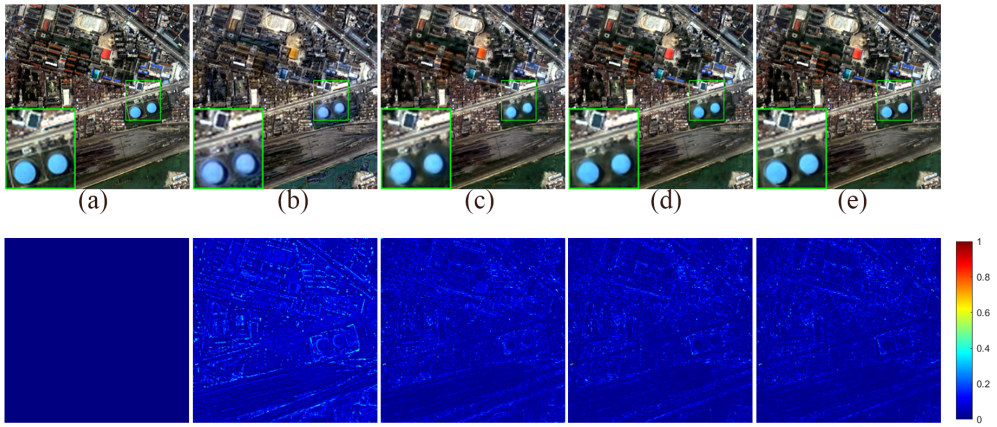


Figure 18. The fusion results after the removal of certain modules. (**a**) reference image, (**b**) single branch, (**c**) dual branch, (**d**) direct concatenation, (**e**) ours.

Table 6. Quantitative evaluations of loss ablation experiments on the simulated IKONOS dataset.

	ERGAS	UIQI	SAM	SCC	Q4
Single branch	1.3618	0.9804	1.7364	0.9513	0.8346
Dual branch	1.3360	0.9810	1.7038	0.9527	0.8389
Direct Concatenation	1.2918	0.9822	1.6268	0.9558	0.8465
ours	**1.2795**	**0.9825**	**1.6102**	**0.9567**	**0.8492**

3.8. Complexity Analyses

To further compare the network performance, we also compared the number of parameters, running time, and floating point of operations (FLOPs) of our proposed method with other deep learning-based methods on the test set (all compared on the WV3 dataset), as shown in Table 7. Although the number of parameters in our proposed model is not the minimum, it remains within 0.5 M. In terms of runtime, the MSDCNN model has the fastest running speed, but the performance of our proposed model is far superior to MSDCNN. The slow running speed and large FLOPs value of our proposed model may be due to the presence of many matrix multiplication operations in the network.

Table 7. Model sizes and running times of DL-based pansharpening methods.

	Param (M)	Time (s)	Flops (G)
PanNet	0.08	1.0579	5.4211
MSDCNN	0.18	3.0727	14.9595
TFNet	2.36	3.5929	30.9760
GPPNN	0.24	3.5016	11.1065
SSAF	1.08	5.2443	70.7027
SSE	0.47	4.0478	30.7840
CIKA	0.07	3.7104	4.7322
Band aware	0.59	4.5207	54.4818
PAPS	0.48	6.0989	32.9523
Ours	0.29	3.9182	36.4856

4. Conclusions

This study introduced a novel three-branch pansharpening network that leverages interactions between the spatial and frequency domains. Each branch is designed to extract specific types of information: spectral features from LRMS images, spatial features from PAN images, and common features shared by both. These extracted features are exchanged across spatial and frequency domains before undergoing feature reconstruction. The proposed design includes a spectral feature extraction module capable of capturing nonlinear relationships between the spectral bands in LRMS images and a spatial feature extraction module that effectively captures texture details and embeds positional information, thereby enhancing the completeness of the spatial features. Additionally, the research proposed a dynamic convolution mechanism that adapts to spectral and spatial features, improving network flexibility. The proposed SSFM facilitates robust interaction between spatial and frequency domain information. Furthermore, this study developed a loss function aimed at reducing redundant information between branches while ensuring the frequency domain features closely align with the reference image. Comprehensive experiments, including comparative, ablation, and network structure evaluations across three datasets (IKONOS, WV3, and WV4), consistently demonstrate the superior performance of the proposed method.

It is important to acknowledge that, like many CNN-based approaches, the proposed method was trained on simulated datasets and evaluated on both simulated and real datasets. However, there is still room for improvement in terms of spatial details in our proposed method, and we will attempt to use more cutting-edge edge operator regions to extract detailed textures from the source image.In addition, its performance on real datasets with scale variations remains modest. Future work will focus on jointly training with both simulated and real datasets to improve performance and achieve a balanced effectiveness across diverse datasets.

Author Contributions: Conceptualization, X.W.; data curation, X.W.; investigation, X.W. and L.L.; methodology, X.W. and L.L.; resources, H.M.; software, X.W.; visualization, X.W.; writing—original draft, X.W.; writing—review and editing, H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Program of China under grant 2020YFA0711403.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author

Conflicts of Interest: The authors declare that this study was conducted without any commercial or financial relationships that could be perceived as potential conflicts of interest.

References

1. Ding, L.; Zhang, J.; Guo, H.; Zhang, K.; Liu, B.; Bruzzone, L. Joint Spatio-Temporal Modeling for Semantic Change Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
2. Lee, R.; Steele, S. Military Use of Satellite Communications, Remote Sensing, and Global Positioning Systems in the War on Terror. *J. Air Law Commer.* **2014**, *79*, 69–112.
3. Yang, J.; Gong, P.; Fu, R.; Zhang, M.; Chen, J.; Liang, S.; Xu, B.; Shi, J.; Dickinson, R. The role of satellite remote sensing in climate change studies. *Nat. Clim. Chang.* **2013**, *3*, 875–883. [CrossRef]
4. Kumar, U.; Cristina, M.; Nemani, R.; Basu, S. Multi-sensor multi-resolution image fusion for improved vegetation and urban area classification. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *40*, 51–58. [CrossRef]
5. Xie, B.; Zhang, H.K.; Huang, B. Revealing Implicit Assumptions of the Component Substitution Pansharpening Methods. *Remote Sens.* **2017**, *9*, 443. [CrossRef]
6. Yang, Y.; Wan, W.; Huang, S.; Lin, P.; Que, Y. A Novel Pan-Sharpening Framework Based on Matting Model and Multiscale Transform. *Remote Sens.* **2017**, *9*, 391. [CrossRef]
7. Vivone, G.; Dalla Mura, M.; Garzelli, A.; Restaino, R.; Scarpa, G.; Ulfarsson, M.O.; Alparone, L.; Chanussot, J. A New Benchmark Based on Recent Advances in Multispectral Pansharpening: Revisiting Pansharpening With Classical and Emerging Pansharpening Methods. *IEEE Geosci. Remote Sens. Mag.* **2021**, *9*, 53–81. [CrossRef]
8. Laben, C.A.; Brower, B.V. Process for Enhancing the Spatial Resolution of Multispectral Imagery Using Pan-Sharpening. U.S. Patent No. 6011875, 4 January 2000.
9. Tu, T.M.; Huang, P.; Hung, C.L.; Chang, C.P. A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery. *IEEE Geosci. Remote Sens. Lett.* **2004**, *1*, 309–312. [CrossRef]
10. Ghadjati, M.; Moussaoui, A.; Boukharouba, A. A novel iterative PCA–based pansharpening method. *Remote Sens. Lett.* **2019**, *10*, 264–273. [CrossRef]
11. Restaino, R.; Vivone, G.; Addesso, P.; Chanussot, J. A Pansharpening Approach Based on Multiple Linear Regression Estimation of Injection Coefficients. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 102–106. [CrossRef]
12. Vivone, G.; Restaino, R.; Chanussot, J. Full Scale Regression-Based Injection Coefficients for Panchromatic Sharpening. *IEEE Trans. Image Process.* **2018**, *27*, 3418–3431. [CrossRef] [PubMed]
13. Liu, J.G. Smoothing Filter-based Intensity Modulation: A spectral preserve image fusion technique for improving spatial details. *Int. J. Remote Sens.* **2000**, *21*, 3461–3472. [CrossRef]
14. Feng, X.; He, L.; Cheng, Q.; Long, X.; Yuan, Y. Hyperspectral and Multispectral Remote Sensing Image Fusion Based on Endmember Spatial Information. *Remote Sens.* **2020**, *12*, 1009. [CrossRef]
15. Yang, Y.; Wu, L.; Huang, S.; Sun, J.; Wan, W.; Wu, J. Compensation Details-Based Injection Model for Remote Sensing Image Fusion. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 734–738. [CrossRef]
16. Ayas, S.; Tunc Gormus, E.; Ekinci, M. An Efficient Pan Sharpening via Texture Based Dictionary Learning and Sparse Representation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2448–2460. [CrossRef]
17. Deng, L.J.; Feng, M.; Tai, X.C. The fusion of panchromatic and multispectral remote sensing images via tensor-based sparse modeling and hyper-Laplacian prior. *Inf. Fusion* **2019**, *52*, 76–89. [CrossRef]
18. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [CrossRef] [PubMed]
19. Masi, G.; Cozzolino, D.; Verdoliva, L.; Scarpa, G. Pansharpening by Convolutional Neural Networks. *Remote Sens.* **2016**, *8*, 594. [CrossRef]
20. Wei, Y.; Yuan, Q.; Meng, X.; Shen, H.; Zhang, L.; Ng, M. Multi-scale-and-depth convolutional neural network for remote sensed imagery pan-sharpening. In Proceedings of the International Geoscience and Remote Sensing Symposium, Fort Worth, TX, USA, 23–28 July 2017; pp. 3413–3416.
21. Liu, X.; Liu, Q.; Wang, Y. Remote sensing image fusion based on two-stream fusion network. *Inf. Fusion* **2020**, *55*, 1–15. [CrossRef]
22. Yang, Y.; Li, M.; Huang, S.; Lu, H.; Tu, W.; Wan, W. Multi-scale Spatial-Spectral Attention Guided Fusion Network for Pansharpening. In Proceedings of the 31st ACM International Conference on Multime, Ottawa, ON, Canada, 29 October–3 November 2023.
23. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11211, pp. 3–19.
24. Zhang, K.; Wang, A.; Zhang, F.; Diao, W.; Sun, J.; Bruzzone, L. Spatial and Spectral Extraction Network with Adaptive Feature Fusion for Pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 14–28. [CrossRef]

25. Xu, S.; Zhang, J.; Sun, K.; Zhao, Z.; Huang, L.; Liu, J.; Zhang, C. Deep Convolutional Sparse Coding Network For Pansharpening with Guidance Of Side Information. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6. [CrossRef]
26. Cao, X.; Fu, X.; Hong, D.; Xu, Z.; Meng, D. PanCSC-Net: A Model-Driven Deep Unfolding Method for Pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5404713. [CrossRef]
27. Yin, H. PSCSC-Net: A Deep Coupled Convolutional Sparse Coding Network for Pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5402016. [CrossRef]
28. Li, J.; Fan, W.; Lian, T.; Liu, F. Cross-Attention-Based Common and Unique Feature Extraction for Pansharpening. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 5002905. [CrossRef]
29. Li, Z.; Zhang, K.; Zhang, F.; Wan, W.; Sun, J. Pan-sharpening Based on Transformer with Redundancy Reduction. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5513205.
30. Zhang, P.; Mei, Y.; Gao, P.; Zhao, B. Cross-Interaction Kernel Attention Network for Pansharpening. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 5001505. [CrossRef]
31. Zhou, M.; Yan, K.; Fu, X.; Liu, A.; Xie, C. PAN-Guided Band-Aware Multi-Spectral Feature Enhancement for Pan-Sharpening. *IEEE Trans. Comput. Imaging* **2023**, *9*, 238–249. [CrossRef]
32. Jia, Y.; Hu, Q.; Dian, R.; Ma, J.; Guo, X. PAPS: Progressive Attention-Based Pan-sharpening. *IEEE/CAA J. Autom. Sin.* **2024**, *11*, 391–404. [CrossRef]
33. Zhang, F.; Yang, G.; Sun, J.; Wan, W.; Zhang, K. Triple disentangled network with dual attention for remote sensing image fusion. *Expert Syst. Appl.* **2024**, *245*, 123093. [CrossRef]
34. Hou, J.; Cao, Q.; Ran, R.; Liu, C.; Li, J.; Deng, L. Bidomain Modeling Paradigm for Pansharpening. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023.
35. Diao, W.; Zhang, F.; Wang, H.; Sun, J.; Zhang, K. Pansharpening via Triplet Attention Network With Information Interaction. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 3576–3588. [CrossRef]
36. Zhuo, Y.W.; Zhang, T.J.; Hu, J.F.; Dou, H.X.; Huang, T.Z.; Deng, L.J. A Deep-Shallow Fusion Network with Multidetail Extractor and Spectral Attention for Hyperspectral Pansharpening. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 7539–7555. [CrossRef]
37. Tang, L.; Xiang, X.; Zhang, H.; Gong, M.; Ma, J. DIVFusion: Darkness-free infrared and visible image fusion. *Inf. Fusion* **2023**, *91*, 477–493. [CrossRef]
38. Zhang, K.; Yang, G.; Zhang, F.; Wan, W.; Zhou, M.; Sun, J.; Zhang, H. Learning Deep Multiscale Local Dissimilarity Prior for Pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5406015. [CrossRef]
39. Qiu, Z.; Shen, H.; Yue, L.; Zheng, G. Cross-sensor remote sensing imagery super-resolution via an edge-guided attention-based network. *ISPRS J. Photogramm. Remote Sens.* **2023**, *199*, 226–241. [CrossRef]
40. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
41. Xing, Y.; Zhang, Y.; He, H.; Zhang, X.; Zhang, Y. Pansharpening via Frequency-Aware Fusion Network with Explicit Similarity Constraints. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5403614. [CrossRef]
42. Wang, H.; Gong, M.; Mei, X.; Zhang, H.; Ma, J. Deep Unfolded Network with Intrinsic Supervision for Pan-Sharpening. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 5419–5426.
43. Zhou, M.; Yan, K.; Huang, J.; Yang, Z.; Fu, X.; Zhao, F. Mutual Information-driven Pan-sharpening. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1788–1798.
44. Zhou, M.; Huang, J.; Yan, K.; Hong, D.; Jia, X.; Chanussot, J.; Li, C. A General Spatial-Frequency Learning Framework for Multimodal Image Fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, 1–18. [CrossRef]
45. Meng, X.; Xiong, Y.; Shao, F.; Shen, H.; Sun, W.; Yang, G.; Yuan, Q.; Fu, R.; Zhang, H. A Large-Scale Benchmark Data Set for Evaluating Pansharpening Performance: Overview and Implementation. *IEEE Geosci. Remote Sens. Mag.* **2021**, *9*, 18–52. [CrossRef]
46. 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]
47. Du, Q.; Younan, N.H.; King, R.L.; Shah, V.P. On the Performance Evaluation of Pan-Sharpening Techniques. *IEEE Geosci. Remote Sens. Lett.* **2007**, *4*, 518–522. [CrossRef]
48. Wang, Z.; Bovik, A. A universal image quality index. *IEEE Signal Process. Lett.* **2002**, *9*, 81–84. [CrossRef]
49. Yuhas, R.H.; Goetz, A.F.H.; Boardman, J.W. Discrimination among semi-arid landscape endmembers using the Spectral Angle Mapper (SAM) algorithm. In Proceedings of the Summaries of the Third Annual JPL Airborne Geoscience Workshop, Pasadena, CA, USA, 1–5 June 1992.

50. Zhou, J.T.; Civco, D.L.; Silander, J.A. A wavelet transform method to merge Landsat TM and SPOT panchromatic data. *Int. J. Remote Sens.* **1998**, *19*, 743–757. [CrossRef]
51. Alparone, L.; Aiazzi, B.; Baronti, S.; Garzelli, A.; Nencini, F.; Selva, M. Multispectral and panchromatic data fusion assessment without reference. *Photogramm. Eng. Remote Sens.* **2008**, *74*, 193–200. [CrossRef]
52. Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; Paisley, J. PanNet: A Deep Network Architecture for Pan-Sharpening. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1753–1761.
53. Xu, S.; Zhang, J.; Zhao, Z.; Sun, K.; Liu, J.; Zhang, C. Deep Gradient Projection Networks for Pan-sharpening. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 1366–1375.

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