Article

Hybrid Attention Based Residual Network for Pansharpening

Qin Liu ¹, Letong Han ¹ , Rui Tan ¹, Hongfei Fan ¹ , Weiqi Li ¹, Hongming Zhu ^{1,*}, Bowen Du ^{1,2} and Sicong Liu ³

¹ School of Software Engineering, Tongji University, 4800 Caoan Road Jiading District, Shanghai 201804, China; qin.liu@tongji.edu.cn (Q.L.); 2031543@tongji.edu.cn (L.H.); 2031570@tongji.edu.cn (R.T.); fanhongfei@tongji.edu.cn (H.F.); 1853401@tongji.edu.cn (W.L.); B.Du@Warwick.ac.uk (B.D.)

² Department of Computer Science, University of Warwick, Gibbet Hill Road, Coventry CV47 4L, UK

³ School of Geodesy and Geomatics, Tongji University, 1239 Siping Road Yangpu District, Shanghai 200082, China; sicong.liu@tongji.edu.cn

* Correspondence: zhu_hongming@tongji.edu.cn

Abstract: Pansharpening aims at fusing the rich spectral information of multispectral (MS) images and the spatial details of panchromatic (PAN) images to generate a fused image with both high resolutions. In general, the existing pansharpening methods suffer from the problems of spectral distortion and lack of spatial detail information, which might prevent the accuracy computation for ground object identification. To alleviate these problems, we propose a Hybrid Attention mechanism-based Residual Neural Network (HARNN). In the proposed network, we develop an encoder attention module in the feature extraction part to better utilize the spectral and spatial features of MS and PAN images. Furthermore, the fusion attention module is designed to alleviate spectral distortion and improve contour details of the fused image. A series of ablation and contrast experiments are conducted on GF-1 and GF-2 datasets. The fusion results with less distorted pixels and more spatial details demonstrate that HARNN can implement the pansharpening task effectively, which outperforms the state-of-the-art algorithms.

Keywords: deep learning; HARNN; hybrid attention mechanism; image fusion; remote sensing

Citation: Liu, Q.; Han, L.; Tan, R.; Fan, H.; Li, W.; Zhu, H.; Du, B.; Liu, S. Hybrid Attention Based Residual Network for Pansharpening. *Remote Sens.* **2021**, *13*, 1962. <https://doi.org/10.3390/rs13101962>

Academic Editor: Giuseppe Scarpa

Received: 13 April 2021

Accepted: 13 May 2021

Published: 18 May 2021

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

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Remote sensing technology has played an important role in economic, political, military and other fields since the successful launch of the first human-made earth resources satellite. With the development of remote sensing technology, existing remote sensing satellites are able to obtain images with higher and higher spatial, temporal and spectral resolution [1]. However, due to the restrictions of technical conditions and hardware limitations [2], optical remote sensing satellites can only provide high-resolution PAN images and low-resolution MS images. PAN image has only one spectral channel, which means it cannot express RGB colors, and on the contrary, MS image carries high expression ability of color [3–6]. Therefore, the fusion of high spatial resolution of PAN and high spectral resolution of MS, called pansharpening, is proposed and proven to be an effective method.

The existing pansharpening methods can be roughly divided into traditional fusion algorithms [7–9] and deep learning based fusion algorithms [10,11]. As the focus of this paper, deep learning based methods have been developed to refine the spatial resolution via substituting components [12,13], or transforming features into another vector space [14]. Although these previous works have achieved fusion accuracy to some extent, the extraction of spectral and spatial features from MS and PAN images could be further promoted to improve the spatial resolution and alleviate spectral distortion. The spectral distortions are caused by the large numerical difference between pixel values of MS and PAN images, since the surface features have discrepant value in different spectral bands. As shown in Figure 1b, which is generated by PNN, if there are spectral distortion regions in the fused

image, the identification and analysis of ground objects will be affected, such as the identification of rivers, which mainly relies on color expression and spectral features of the fused image. As for the problem that the spatial resolution is not high enough [12], the accuracy of target segmentation will also be influenced by the indistinct edges of buildings and arable lands. In addition, the problem of high computational complexity leads to high hardware requirements and high time consumption [14].

Figure 1. Examples of spectral distortion in GF-1 and GF-2 MS images: (a) The original images without spectral distortion. (b) The distorted multi-spectral images generated by PNN.

To handle the problems of spectral distortion and low spatial resolution, we propose HARNN for pansharpening task. The proposed method is based on ResNet with a novel hybrid attention mechanism. The inputs of the network consists of two feature extraction branches to better extract spectral and spatial detail information from MS and PAN images. In order to extract multi-scale features from remote sensing images obtained by different satellites and reduce the complexity of training network, the extracted features are downsampled once [15,16] through convolutional operation and then rescaled to the original resolution after being fused. Moreover, to ease the problem of spectral distortion and improve the spatial resolution of the fused image, the encoder attention module and hybrid attention module are designed as parts of the network. Finally, we conduct extensive experiments on two remote sensing image datasets collected by Gaofen-1 (GF-1) and Gaofen-2 (GF-2) satellites, and the experimental results demonstrate that the proposed method could achieve promising results compared with other state-of-the-art methods.

Specifically, the main contributions of this paper include:

1. A feature extraction network is designed with two branches including an encoder attention module to extract the spectral correlation between MS image channels and the advanced texture features from PAN images;
2. A hybrid attention mechanism with the truncation normalization is proposed in the feature fusion network to alleviate the problem of spectral distortion, and to improve the spatial resolution simultaneously;
3. Extensive experiments are conducted to verify the effectiveness of the attention mechanism in the proposed method, which could provide a comparative baseline for related research work.

The rest of this paper is organized as follows. Section 2 reviews the related work in the pansharpening field. Section 3 describes the proposed network architecture, the utilized loss function, as well as the hybrid attention mechanism. Section 4 then introduces the experimental dataset, the evaluation metrics, and presents the experiment results of different pansharpening methods. Finally, the overall conclusion of this paper is summarized in Section 5.

2. Related Work

Existing pansharpening methods fall into two main categories: traditional fusion algorithms and deep learning (DL) based methods, among which the former is divided into component substitution algorithms (CS), multi-resolution analysis algorithms (MRA), and sparse representation (SR) based algorithms [3–6].

2.1. Traditional Algorithms

One of the most popular pansharpening methods is CS-based algorithms. The basic idea of CS is to extract the spectral and spatial information of MS images by applying a pre-determined transformation [3,17]. Then the spatial component is replaced with the high resolution part generated from PAN image, and the final result is constructed by the inverse operation. Some representative CS-based methods include Principal Component Analysis (PCA) [7], Gram-Schmidt (GS) algorithm [18], Intensity-Hue-Saturation (IHS) algorithm [8], and improved IHS algorithms such as Adaptive IHS (AIHS) [19] and Non-linear IHS (NIHS) [20]. These color-transformation-based methods are popular because of the fast transformation process and high spatial resolution of the fused images [21]. However, due to the direct substitution of components of CS methods, though retaining the spatial details of PAN, the difference in scales of pixel values of PAN and MS images leads to spectral distortion and color deviation [8].

In 2004, Benz U C applied MRA in remote sensing data analysis for the first time [22]. Since then, many MRA-based pansharpening methods such as Wavelet Transform (WT) [23], Discrete Wavelet Transform (DWT) [24], and Laplacian pyramid method [25] have been proposed to solve the problems above. Different from CS methods, these algorithms extract the high frequency information such as spatial details of PAN and inject it into MS images through multi-resolution transformations to reconstruct a fused image of high resolution. Since MRA methods only leverage the high-frequency details of the PAN image, the consistency can be maintained in terms of color characteristics. For example, the DWT method [24] decomposes the origin PAN and MS images into high and low frequency components, and performs inverse transformation after fusing these components at different resolutions. Despite the high spectral resolution, these MRA-based methods have the disadvantage of ignoring edge information, which results in the low spatial resolution of the fused image.

Apart from CS and MRA based approaches, the SR based algorithms such as [26] are designed to improve the spectral and spatial resolution of the fusion result at the same time. These SR-based algorithms use high and low resolution dictionaries to sparsely represent MS and PAN images. The maximum fusion rule is adopted to partially replace the coefficients with the sparse representation coefficients of the PAN image. Then the spatial details of sparse representation coefficients are injected into the MS image. Finally, through image reconstruction, the fused image is obtained [27]. Compared to mentioned methods, the SR-based methods alleviate spectral distortion and increase the spatial detail information, but suffer from the excessive sharpening [28].

2.2. Deep Learning Based Algorithms

In recent years, DL-based methods have been introduced into image processing fields such as image fusion [29,30], object segmentation [31,32], and video recognition [33,34], and they have been proven to be effective. On the issue of pansharpening, the DL-based

methods are mainly divided into two categories: single-branch neural network and dual-branch neural network [35].

As for the single-branch architecture, the PAN image is considered as another spectral band of MS image and concatenated into it. Then, the composite image is delivered into CNN modules as one input, and transformed to a higher resolution version. For example, Zhong et al. [36] proposed to combine the GS algorithm and the SRCNN model [37,38] in the super-resolution domain and perform GS transform on high resolution MS and PAN images. However, the fusion results of GS-SRCNN still suffer from spectral distortion and lack of spatial details. Creatively, Masi et al. [10] proposed the PNN model based on convolutional neural network (CNN) [39] for the first time, which is composed of three convolution layers using kernel size (9, 5, 5). To improve the fusion accuracy, this same lab introduced nonlinear radiation index into PNN [40], yet the fused images still contained spectral distorted pixels and unclear edges, which implies the limitation of single-branch networks. Furthermore, to improve the full-resolution performance, Vitale et al. [41] proposed to introduce perpetual loss (PL) into network training process of pansharpening. By introducing an additional loss term, the training phase is optimized and the visual perception ability of CNN is promoted.

With regard to the dual-branch networks, the MS and PAN images are generally processed by two feature extraction networks separately. Then, the extracted features are fused through a fusion network and finally the high resolution image is generated. Compared to single-branch architecture, the dual-branch networks are able to better extract spectral and spatial features from MS and PAN images, respectively, without influencing the spectral correlation of MS image.

Gaetano et al. proposed a two-branch deep fusion network called RSIFNN [42] in 2018. They considered that there was redundant information between MS and PAN images which lead to a residual mask, and treat the entire network as a residual unit. The predicted mask was then superimposed on the original MS image to obtain the fusion result, while it also had the problem of spectral distortion and blurry contours. Furthermore, another two-branch fusion network named PSGAN [11] was proposed, in which the concept of Generative Adversarial Networks (GAN) [43] was introduced into pansharpening task and the fusion accuracy was improved because of the introduction of discriminators. In addition, the concept of residual module [44] was introduced into two branch pansharpening network by Liu et al. [12] to make better use of the feature extraction and fusion ability of deep neural network, but its fusion result still has some spectral abnormal pixels which effect the fusion quality. Besides, the attention module was adopted in the recent proposed networks to improve the ability of feature extraction and proved to be effective in pansharpening task [45]. As a consequence, the residual module was adopted in the proposed network, and a hybrid attention module was designed to further enhance the spatial and spectral resolution of the fused image, which will be discussed in the following section.

3. Proposed Methods

For the sake of brevity and clarity, the notations below in Table 1 will be used in subsequent sections to describe the proposed network in detail:

Table 1. The notations in the following sections.

Notation	Explanation
M	Original low-resolution MS image
P	Original high-resolution PAN image
$\tilde{M} \downarrow$	Downsampled blurred low-resolution MS image
$\tilde{P} \downarrow$	Downsampled high-resolution PAN image
$\hat{M}$	Fused high-resolution MS image
F	Feature maps extracted from feature extraction network
$\tilde{F}$	fused feature maps

3.1. Network Architecture

We proposed a hybrid attention mechanism based dual-branch residual convolutional neural network called HARNN, which consists of feature extraction network, feature fusion network and image reconstruction network. The feature extraction part in HARNN is split into MS and PAN feature extraction branches.

Figure 2 illustrates the semantic framework of DL-based pansharpening network, which is composed of three parts that highlighted by red, green and blue boxes. In the feature extraction part (red box), the original MS image is blurred using Gaussian Blur function and downsampled by 4 times, then rescaled to the initial resolution through bicubic interpolation, resulting in the blurred image $\tilde{M} \downarrow$. In addition, the original PAN image is also downsampled to the same resolution of MS represented by $P \downarrow$ according to Wald Protocol. After these image preprocessing steps, $\tilde{M} \downarrow$ and $P \downarrow$ are sent into feature extraction branches, and the corresponding spectral and spatial features can be denoted as:

$$F_l = \phi(w_l \times F_{l-1} + b_l) \quad (1)$$

where w_l and b_l represent the weight and bias vectors of the l th layer of network, $\times$ stands for the convolution denotation, ϕ is the activation function, and F_l denotes the feature map after convolution and activation operation.

Figure 2. Network architecture of HARNN.

Compared to single-branch feature extraction network and super-resolution-based pansharpening network, this dual-branch network has improved performance for making better use of the spatial information contained in MS and PAN and eliminating the redundant information between different bands of MS image. When extracting features from $\tilde{M} \downarrow$, the result feature maps represent the spatial characteristics of each channel of MS image on a two-dimensional plane, as well as the spectral correlation between these channels. Correspondingly, the outline and texture information are better extracted while mining spatial features from $P \downarrow$, which means that the extracted features contain spatial and spectral information from both of the images.

Subsequently, as the network architecture of HARNN shown in Figure 2, the feature maps extracted from two branches are fused via feature fusion network, in which the feature maps are downsampled by 2 times by pooling layer to get features with scale invariance and concatenated in the channel dimension before sent into network. After being processed by several residual blocks and attention modules, the fused feature maps can be represented as:

$$\tilde{F} = \phi(\Theta(F_M \oplus F_P)) \quad (2)$$

where F_M and F_P denotes the feature maps extracted from respective feature extraction branch, $\oplus$ represents channel-wise concatenation, Θ and ϕ are the convolution and activation operation, respectively. By using this concept of Fusion instead of Detail Injection,

the characteristics of CNN are efficiently utilized and the high-level abstract features are better extracted via deep network.

Unfortunately, according to equation(2) mentioned above, it is inevitable that the gradient of deep network will disappear or explode during the process of back propagation, which can be denoted as:

$$\Delta\omega = \frac{\partial Loss}{\partial \omega_i} = \frac{\partial Loss}{\partial f_n} \times \frac{\partial f_n}{\partial f_{n-1}} \times \dots \times \frac{\partial f_i}{\partial \omega_i} \quad (3)$$

where f_n and $Loss$ represents the activation function and the error calculated in the n th layer, and $\Delta\omega$ is the derivative when passing $Loss$ back into the i th layer which could be quite large or close to zero. Hence the architecture of ResNet is adopted in HARNN to solve this problem. In residual blocks, the mapping between inputs and residuals are learned by network via skip connection, so $Loss$ can be propagated directly to the lower layers without intermediate calculation. In the proposed model, the pre-activated residual block [44] without batch normalization is introduced to prevent damaging the contrast information of original images.

Furthermore, in order to enhance spatial resolution and alleviate spectral distortion of the pansharpening result, the combined loss function and hybrid attention module is also adopted in the proposed network and will be discussed in detail in subsequent subsections.

3.2. Loss Function

On the basis of a reasonable network architecture, the selection of loss function also affects the pansharpening results. In some pansharpening papers [10,46], MSE is selected as the single loss function for its faster converge rate. However, it is sensitive to outliers because it calculates the sum of squares of the errors, which causes that the loss value cannot reflect the overall error of the fused image. On the contrary, MAE is more robust to outliers, but with slower convergence rate. Inspired by [47], we make a compromise and adapt the weighed combination of MSE and MAE as loss function in order to achieve the balance of the converge rate and robustness. The combined loss function can be defined as:

$$\begin{cases} Combined_loss(y_true, y_pred) = MAE + \frac{1}{10} MSE \\ MAE = \frac{1}{n} \sum_{i=1}^n |y_true_i, y_pred_i| \\ MSE = \frac{1}{n} \sum_{i=1}^n (y_true_i, y_pred_i)^2 \end{cases} \quad (4)$$

In the contrast experiment, we found that the network has the best performance when MAE and MSE has the proportion of 10:1 on our dataset. Accordingly, we chose this proportion to boost the fusion accuracy.

3.3. Hybrid Attention Mechanism

The hybrid attention mechanism will be introduced in detail in this section, and the schematic diagram of attention module is represented in Figure 3. The core concept of hybrid attention mechanism is inspired by the idea of depthwise separable convolution proposed by [48,49], in which the standard convolution is divided into depthwise convolution and pointwise convolution, and the number of parameters and computation will be decreased significantly without compromising fusion accuracy.

In the proposed network, the concept of depthwise convolution is applied into the selection of image spatial features and it constitutes the spatial part of the hybrid attention module. In addition, the proposed attention mechanism consists of encoder attention module and fusion attention module. The encoder attention module is designed to extract implicit information in the feature extraction network, and the fusion attention module is designed to select more informative features. Suppose the feature maps $\tilde{F}$ fused through

feature fusion network with N channels (N filters in the last convolution layer), then they are divided into N groups, which implies each group consists of only one feature map. Inside each group, the *DepthwiseConv2D* layer is applied to extract spatial features such as outlines and texture information in two dimensions. After being calculated by spatial attention block, feature maps are converted to weighted ones with strengthened texture features, which will be able to contribute more to the fusion result. The process of calculating spatial attention features is expressed as:

$$\hat{F}_{-s_i} = \tilde{\Theta}_3(\phi_2(\tilde{\Theta}_2(\phi_1(\tilde{\Theta}_1(\tilde{F}_i)))))) \quad (5)$$

where $\tilde{F}_i$ denotes the i th group of feature maps, $\tilde{\Theta}_i \phi_i$ represents depthwise convolution and activation function of the i th layer, and $\hat{F}_{-s_i}$ stands for the i th weighted feature map after spatial attention module.

Figure 3. Hybrid attention module of HARNN.

As another part of the hybrid attention mechanism, channel attention module also plays an important role in optimizing fusion results by screening more informative feature maps. Unlike spatial attention, the processing objects of channel attention module are the complete set of the feature maps by adding all convolution results of each feature map. This process is essentially equivalent to performing Fourier Transform to feature maps by convolution kernels, if we consider the transformation from features to features in another dimension is similar to the transformation from time domain to frequency domain. After applying channel attention to $\tilde{F}$, the eigencomponents of different convolution kernels will be represented, which implies the devotion of feature maps to improve fusion accuracy and can be denoted as $\hat{F}_{-c_i}$.

In order to get the mask of hybrid attention module, the sigmoid function $\sigma(\cdot)$ is applied to the final feature maps which concatenating $\hat{F}_{-s_i}$ and $\hat{F}_{-c_i}$. After multiplying mask and $\tilde{F}$, the weight of the feature maps $\hat{F}$ are reassigned both in the spatial and the channel dimension, which can be represented as:

$$\hat{F} = \tilde{F} \otimes \sigma(\hat{F}_{-s_i} \oplus \hat{F}_{-c_i}) \quad (6)$$

To sum up, the hybrid attention mechanism extracts the informative parts from the fused feature maps, including spatial features that are conducive to enhance the texture expression ability of fusion results and a certain feature map that contributes more compared to other ones. By introducing the hybrid attention mechanism into the proposed model, the problems of spectral distortion and lack of texture details could be alleviated. The ablation study of this mechanism which verifies its effectiveness will be presented in the following section.

4. Results and Discussion

In this section, we design and conduct three groups of experiments to verify the following assumptions:

1. The encoder attention and fusion attention modules could improve pansharpening accuracy.
2. The residual hybrid attention mechanism is able to alleviate the problem of spectral distortion.
3. The proposed network outperforms other state-of-the-art pansharpening methods in spatial resolution.

The details of experiments are illustrated as follows.

4.1. Datasets

To evaluate the former mentioned objects, the following experiments were implemented on the dataset of MS and PAN images obtained by Gaofen-2 satellite (GF-2) and Gaofen-1 satellite (GF-1), of which the MS images consisted of four bands (Red, Green, Blue and Near Infrared Band) and had the image size of 6000*6000 and 4500*4500, respectively. In order to enhance the diversity of data and improve the fusion accuracy of the model to various ground objects, we selected images covering different landforms including urban buildings, roads, fields, mountains and rivers, etc. The Ground Sample Distance (GSD), i.e., the real distance between each two adjacent pixels in the image, is used to describe the spatial resolution of remote sensing images. The GSD of MS and PAN images of GF-1 satellite are 8 m and 2 m, respectively. Correspondingly, the GF-2 remote sensing images have the GSD of 3.24 m and 0.81 m, where the resolution is twice larger than that of GF-1 images. The detailed information of corresponding spectral wave length is listed in Table 2, and these two datasets cover the landscape of mountains, settlements and vegetation.

Table 2. The detailed description of Gf-1 and GF-2 datasets.

Satellite	Gaofen-1	Gaofen-2
Spectral Bands	4	4
GSD-MS	2 m	0.81 m
GSD-PAN	8 m	3.24 m
Image size	4500*4500	6000*6000
Image num	1	5
Tile num	5625	48,956
Landscape	Mountains, Settlements	Settlements, Vegetation, Water

To improve the training efficiency of HARNN and obtain multiscale features, we performed the downsampling operation on the original images according to Wald Protocol [50] (the original MS images were used as reference images), and cut them into 64×64 small tiles with the overlap ratio of 0.125. After preprocessing the images, these 58,581 samples were divided on a scale of 8:2, of which the random selected 80% part was used for model training and the rest part used for validation.

The experiments implemented in this section were carried out on a remote server installing Ubuntu 18.04.4 LTS. To improve the computational efficiency, the multi-GPU training strategy was adopted via four NVIDIA RTX2080Ti GPU, which was realized by data parallel method allocating batches of training data to different GPUs. The total batch size was set to 32, the initial learning rate was set to 0.0001 and the total number of iterations was 30 k.

4.2. Comparison Methods and Evaluation Indices

In this section, we select 6 widely used pansharpening methods including traditional CS-based method PCA [7], MRA-based method Wavelet [9] and four recently proposed state-of-the-art DL-based methods, i.e., PNN [10], SRCNN [37], RSIFNN [42] and TFCNN [15]. To comprehensively verify the effectiveness of two-branch feature extraction network, we choose two single-branch networks PNN and SRCNN, and two dual-branch networks RSIFNN and TFCNN for comparison.

In order to evaluate and analyze the performance of fusion algorithms in all aspects, we selected nine widely used evaluation indices in the pansharpening task, which can be divided into referenced and no-referenced indices based on whether to calculate using reference images or not. Referenced indices include relative dimensionless global error in synthesis (ERGAS), universal image quality index (UIQI), the Q metric, correlation coefficient (CC), spectral angle mapper (SAM), structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR). Among them, *ERGAS*, *UIQI* and *Q* are metrics that describe the overall quality of the fusion image, *CC* and *SAM* are the spectral quality metrics, and *SSIM* denotes the structural similarity. As for *PSNR*, it represents the ratio of valid information to noise and is calculated by dividing the difference between the max and min pixel value and MSE of two images.

According to Wald Protocol [50], the referenced indicators can be calculated with the reference MS image, but cannot be used in the real data experiments. Correspondingly, since the no-referenced indices are calculated using only images before and after fusion, the spectral distortion D_λ index and the spatial distortion D_s index are used for both downsampled data and real data experiments.

4.3. Comparison of the Efficiency of Different Attention Strategies

To verify the effectiveness of the encoder and hybrid attention strategies of the proposed method, we compared the qualitative and the quantitative performance of models adopting different attention strategies on the downsampled GF-2 dataset of Henan province, China. The Plain network refers to the network that has the same structure of HARNN, but the encoder attention modules and hybrid attention modules are not adopted in it. The Plain network is utilized as the baseline of this series of comparison experiments. The second network introduces encoder attention modules into its feature fusion network without hybrid attention module, so it is called the Encoder network. Correspondingly, the third Fusion network only adopts hybrid attention without encoder attention modules, and the Proposed network uses these two modules simultaneously.

Figure 4 demonstrates the fusion results of the mentioned networks on downsampled GF-2 image of Henan province, China. Figure 4a,b are the blurred MS image and the downsampled PAN image to be fused. Figure 4c is the original MS image, and Figure 4d–g are the fusion results of networks adopting different attention strategies introduced above. Compared with the low-resolution and reference images, it is obvious that the proposed network which has both the encoder attention and hybrid attention modules has the most informative spatial features and the most accurate color expression, which represents the high spectral resolution. Further more, as shown in the lower left corner of the tiles in Figure 4, the bright yellow buildings are blurry in the results of Plain, Encoder and Fusion networks, and the edge of the playground is as clear as the Proposed network, which implies the effectiveness of the encoder and hybrid attention modules in improving spatial resolution without introducing spectral distortion.

To further verify the accuracy of the attention mechanism, the quantitative analyze results conducted on the downsampled GF-2 image of Henan province, China are listed in Table 3, which is the average result of five groups of experiments. In Table 3, the best performance for each index such as ERGAS, CC and SSIM is labeled in bold for convenience. As shown in the table, the proposed network adopting both encoder and hybrid attention modules outperform other control methods in almost all of the metrics, which is consistent with the qualitative results in Figure 4.

In conclusion, the proposed hybrid attention mechanism is proved to be effective in improving spatial resolution and in keeping high spectral resolution. The comparative figures and metrics values demonstrates the effectiveness of the proposed network in the pansharpening task. In the following experiments, we compare the fusion results of the proposed network with some state-of-the-art pansharpening methods to further verify the efficiency of it.

Figure 4. Comparison results of different attention strategies (downsampled GF-2 image of Henan Province, China). (a) LRMS image. (b) HRPAN image. (c) Reference MS image. (d–g) Pansharpening result of plain network, encoder attention, fusion attention and the proposed method.

Table 3. The quantitative results of different attention strategies (downsampled GF-2 image of Henan Province, China).

Evaluation Indices	Plain	Encoder	Fusion	Proposed
ERGAS↓	3.2986	2.961	2.9039	1.1603
Q↑	0.5027	0.4747	0.4793	0.8584
UIQI↑	0.9878	0.9915	0.9918	0.9984
CC↑	0.917	0.926	0.9233	0.9927
SAM↓	0.1065	0.1039	0.1043	0.0407
SSIM↑	0.7765	0.7766	0.7782	0.9638
PSNR↑	27.3061	27.8382	28.074	39.5128
D_{λ} ↓	0.0317	0.022	0.0287	0.0025
D_s ↑	0.9322	0.9239	0.9342	0.7318

4.4. Comparison of Spectral Distortion

In order to validate the ability of HARNN to alleviate spectral distortion, a series of experiments are designed as follows. These experiments are conducted on the simulated GF-2 dataset of Guangzhou Province, China, which is downsampled by four times according to Wald Protocol [50] and the original MS images are utilized as reference of the network. Figure 5 presents the qualitative results of this set of experiments where the lower right corner displays the zoomed details of the rooftop, and these false-color images are generated using red, green and blue channels of the fused MS images. Figure 5a–c shows the downsampled MS image, the downsampled PAN image and the original MS image, respectively, Figure 5d–j represents the fusion results of all of the pansharpening methods mentioned before including CS-based method PCA, MAR based method Wavelet and several DL-based pansharpening models proposed in recent years.

Figure 5. Comparison results obtained by different methods (downsampled GF-2 image of Guangzhou Province, China). (a) LRMS image. (b) HRPAN image. (c) Reference MS image. (d–j) Pansharpening result of RSIFNN, Wavelet, PCA, SRCNN, TFCNN, PNN and the proposed method.

Spatially, the fusion results of Wavelet and SRCNN suffer from the lack of spatial information and contour details, and the improvement of spatial resolution is not effective compared to the downsampled MS images. PCA and other DL-based methods successfully extract and fuse the spatial features into the final result, which is shown in the clear outline of the architecture. Spectrally, it is obvious that PCA has a color deviation, where the result is lighter in color than other images and has obvious synthetic traces. In the result of RSIFNN, SRCNN, TFCNN and PNN, there are several spectral distorted pixels shown on the bright rooftop, and these pixels look darker than the reference ones whose pixel value are close to 255. In contrast, the proposed method shows better effectiveness than former methods, both spatially and spectrally, for its fusion result having clear texture details in the circular area of the rooftop and having no obvious distorted pixels.

To further verify the effectiveness of alleviating spectral distortion of the proposed method, we calculate the number of distorted pixels in these tiles by contrasting the pixel values of the reference image and count the pixels whose differences of four channels are greater than 50. Table 4 lists the quantitative statistical results of the distorted pixels of the seven zoomed tiles of Figure 5d–j, with the 60*60 size of these enlarged tiles to calculate the percentage. As is shown in Table 4, PCA has the most distorted pixels whose percentage is more than 40%, and this confirm that PCA has weak spectral fusion ability though its spatial resolution is relatively high. SRCNN, PNN and Wavelet all have more than 10% distorted pixels, and compared to RSIFNN and TFCNN, the proposed method has fewer abnormal pixels which is less than 1%. Therefore, the percentage calculated according to image pixel values further verify that the proposed method has the ability of alleviating spectral distortion and preserving the spatial details of the image.

Table 4. The quantitative statistics of distorted pixels of DL-based pansharpening methods (down-sampled GF-2 image).

Methods	Distorted Pixels	Percentage
Wavelet	389	10.81%
PCA	1467	40.75%
RSIFNN	205	5.69%
SRCNN	583	16.19%
TFCNN	87	2.42%
PNN	428	11.89%
proposed	34	0.94%

Figure 6 shows other qualitative results of this set of experiments where the lower left corner displays the enlarged details of rooftops of factories, and is conducted on the GF-2 dataset of Beijing, China. Spatially, the results of Wavelet, PCA and SRCNN all have the problem of blurred details, and the edges of the roads next to factories are not clear. Due to the difference between the pixel values of MS and PAN images, the fusion result of PCA could be invalid like Figure 6f and has low spatial and spectral resolution. In contrast, HARNN shows better effectiveness than former methods both spatially and spectrally, for its fusion result having clear texture details in the circular area of the rooftop and no obvious distorted pixels. Spectrally, almost all of these DL-based methods suffer from spectral distortion, which is reflected in the abnormal green pixels of the fused images, and the proposed method performs the best among them. Through qualitative analysis, it can be found that PNN has the most severe distortion, and the fused image of the proposed method contains the least abnormal pixels.

Figure 6. Comparison results obtained by different methods (downsampled GF-1 image). (a) LRMS image. (b) HRPAN image. (c) Reference MS image. (d–j) Pansharpening result of RSIFNN, Wavelet, PCA, SRCNN, TFCNN, PNN and the proposed method.

Table 5 lists the number of the distorted pixels of the zoomed tiles of Figure 6d–j, where the size of these enlarged tiles are 80*80. As shown in Table 5, PCA has 100% distorted pixels and PNN has 68.88%, which can also be observed in the figures. TFCNN contains

less than 1% of the abnormal pixels, and HRANN has only 0.02% of them, which verifies the high spectral fusing ability of the proposed method.

Table 5. The quantitative statistics of distorted pixels of DL-based pansharpening methods (down-sampled GF-1 image).

Methods	Distorted Pixels	Percentage
Wavelet	606	9.47%
PCA	6400	100%
RSIFNN	138	2.16%
SRCNN	689	10.76%
TFCNN	33	0.52%
PNN	4408	68.88%
proposed	1	0.02%

To summarize, the residual hybrid attention mechanism is verified to be able to alleviate spectral distortion on both GF-1 and GF-2 dataset. The distorted pixels are less in the fusion results of HARNN, which could be observed in Figures 5 and 6 and the statistical results in the tables above.

4.5. Comparison of Spatial Resolution

In order to further verify the effectiveness of improving spatial resolution of HARNN, we conduct this series of experiments on both downsampled and real dataset, and there is no reference image in the real dataset so we can only analyze the fusing result via qualitative analysis. Figure 7 shows the comparison result on a downsampled dataset of Qinghai Province, China of GF-2, where the detail tiles are enlarged to 50*50 and represented in the lower left corner. PCA has clear contour and edge details of shadow area, but suffers from slight color deviation, while RSIFNN, and TFCNN perform spectrally and are much clearer compared with Wavelet, SRCNN and PNN in terms of the expression of detailed information. Compared with all of these methods, the proposed method has the best comprehensive performance in spectral and spatial characteristics. For example, it has clear contour of the building shadow and the window on the house is also well sharpened.

Table 6 lists the quantitative evaluation results of this set of experiments with the average metric value of 25 different sample tiles of downsampled GF-2 images, where the best result is marked in bold. As is shown in Table 6, the proposed model HARNN outperforms other methods in all referenced indices like ERGAS, Q, UIQI, SSIM and so on, TFCNN has the best result in D_λ and RSIFNN has the highest value of D_s . Compared with the recently proposed method TFCNN, the ERGAS value of the proposed method is improved by 2.59%, and is 50.56% better than PNN, which verifies the effectiveness of the proposed method in improving the overall quality of the fused image. In addition, the proposed method HARNN has the SAM value of 0.0407, which is 4.67% better than TFCNN, and improved by 30.71% compared with RSIFNN. As for the non-referenced metrics, the proposed method does not perform the best, but still has competitive performance. By comparing the evaluation value of all of the pansharpening methods, it can be observed that the proposed method is superior to other methods in global image quality and spectral and spatial similarity, which is important in pansharpening task.

Figure 7. Comparison results obtained by different methods (downsampled GF-2 image of Qinghai Province, China). (a) LRMS image. (b) HRPAN image. (c) Reference MS image. (d–j) Pansharpening result of RSIFNN, Wavelet, PCA, SRCNN, TFCNN, PNN and the proposed method.

Table 6. The quantitative statistics of distorted pixels of DL-based pansharpening methods (downsampled GF-2 image).

Indices	Wavelet	PCA	RSIFNN	SRCNN	TFCNN	PNN	Proposed
ERGAS↓	2.3994	15.566	1.4773	2.1493	1.1924	1.7465	1.1603
Q↑	0.5999	0.3551	0.8049	0.6804	0.8503	0.7792	0.8584
UIQI↑	0.9949	0.8035	0.9977	0.9959	0.9984	0.9972	0.9984
CC↑	0.9674	0.6077	0.9877	0.9736	0.992	0.9803	0.9927
SAM↓	0.0873	0.2535	0.0532	0.0779	0.0426	0.0633	0.0407
SSIM↑	0.8558	0.5449	0.9422	0.8906	0.9604	0.9318	0.9638
PSNR↑	33.0496	19.8027	37.4499	33.9475	39.3568	35.2184	39.5128
D_λ ↓	0.0038	0.1518	0.0019	0.0024	0.0014	0.0034	0.0025
D_s ↑	0.7304	0.8025	0.7321	0.7299	0.7319	0.731	0.7318

To ensure the integrity and comprehensiveness of the experiments, the computation and time complexity of DL-based method are measured and listed in Table 7. Despite the weak pansharpening effectiveness, SRCNN and PNN have prominent time consumption of 308 us/step and 384 us/step because of their fewer network parameters. Correspondingly, the proposed HARNN method consumes 13 ms to process one step, which is inferior to other comparison algorithms and needs to be improved in the future.

Table 7. The time consumption and the number of training parameters of DL-based methods.

Algorithms	Time Consumption	Parameter Nums
RSIFNN	2 ms/step	4,132,324
SRCNN	308 us/step	26,084
TFCNN	1 ms/step	300,740
PNN	384 us/step	80,420
HARNN	13 ms/step	17,249,796

Besides, the training, validation loss and accuracy curves of model training process on GF-1 dataset are also recorded and presented in Figure 8. As is shown in Figure 8a, the combined-loss (as discussed in Section 3.2) curves of the DL-based methods are presented in different colors, where the proposed HARNN method is described in blue lines. Obviously, loss value of these five DL-based algorithms converge to a stable small value in less than 5 epochs. In general, the less parameters the model contains, the faster the loss converges. However, the proposed HARNN has better performance of training and validation loss than TFCNN, though having much more parameters. Similarly, the same trend reflects in Figure 8b, which demonstrates the accuracy curves of these methods. In 5 epochs, all of these algorithms are able to reach the accuracy close to 1, where the TFCNN is trained a little bit slower than others. To sum up, the proposed HARNN only presents a modest performance in time consumption and loss convergency, and needs to be further optimized.

Figure 8. Loss and accuracy curve of comparison DL-based methods on GF-1 dataset: (a) Loss Curve, (b) Accuracy Curve.

In addition to the downsampled experiments, we also design a set of experiment on real data, which predict high-resolution fused image without reference. Figure 9 shows the fusion result on real GF-2 dataset of Beijing, China, and the detail tiles are enlarged to 50*50 and represented in the lower left corner in the red boxes. It is clearly observed that the fusion results of Wavelet, SRCNN, RSIFNN and PNN are not improved in spatial resolution and are still as blurry as the original MS image. The PCA fused image has as much spatial detail information as the original PAN images, but still has the problem of spectral deviation, which is reflected in the obvious difference between the color of the image and the original image. Compared with TFCNN, the proposed method is clearer in the building contours, and has less blur color blocks in the pink part of the building.

In conclusion, as shown in Figures 7–9, we verify that the proposed network HARNN performs well in improving spatial resolution of the fused images both on the downsampled data and the real data. The qualitative results present the high spatial resolution of fused images, and the quantitative evaluation results in the above tables lead to better performance of HARNN compared with other traditional and state-of-the-art methods.

Figure 9. Comparison results obtained by different methods (real GF-2 image of Beijing, China). (a) LRMS image. (b) HRPAN image. (c–i) Pansharpening result of RSIFNN, Wavelet, PCA, SRCNN, TFCNN, PNN and the proposed method.

5. Conclusions

In this paper, we propose a hybrid attention mechanism based network (HARNN) for the pansharpening task, which is proved to have the ability of alleviating the problem of spectral distortion and sharpening the edge contour of the fused image.

The main backbone of the proposed network is based on ResNet, and given the MS and PAN images as the network input, we design a dual-branch feature extraction network to extract spectral and spatial features from two inputs, respectively. To further improve the efficiency of the fusion network, the proposed network leverage the hybrid attention mechanism which enables to select more valuable spectral and spatial features from the extracted ones, and manages to solve the problems mentioned above.

In order to evaluate the performance of our proposed method, we conduct extensive experiments on the downsampled and real dataset of GF-1 and GF-2 satellites, and the experimental results demonstrate that the proposed method could achieve a competitive fusion result which further proves the effectiveness of the designed network. Besides, the time consumption and loss convergency experiments illustrate the shortcomings of HARNN, where the computational complexity should be reduced.

In the future work, we will focus on exploring the extraction and utilization of multi-scale features based on current deep convolutional network, work harder on reducing complexity of the network, and conduct more classification experiments based on the fused image to verify the applicability of this method.

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

Funding: This research was funded by the National Key R&D Program of China (No. 2018YFB0505001), the National Natural Science Foundation of China (No. 61702374).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Sweeting, M. Modern Small Satellites-Changing the Economics of Space. *Proc. IEEE* **2018**, *106*, 343–361. [\[CrossRef\]](#)
2. Zhang, Y. Understanding image fusion. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 657–661.
3. Javan, F.D.; Samadzadegan, F.; Mehravar, S.; Toosi, A.; Khatami, R.; Stein, A. A review of image fusion techniques for pan-sharpening of high-resolution satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *171*, 101–117. [\[CrossRef\]](#)
4. Meng, X.; Shen, H.; Li, H.; Zhang, L.; Fu, R. Review of the Pansharpening Methods for Remote Sensing Images Based on the Idea of Meta-analysis: Practical Discussion and Challenges. *Inf. Fusion* **2018**, *46*, 102–103. [\[CrossRef\]](#)
5. Kahraman, S.; Erturk, A. Review and Performance Comparison of Pansharpening Algorithms for RASAT Images. *J. Electr. Electron. Eng.* **2018**, *18*, 109–120. [\[CrossRef\]](#)
6. Vivone, G.; Alparone, L.; Chanussot, J.; MD Mura; Garzelli, A.; Licciardi, G. A.; Restaino, R.; Wald, L. A Critical Comparison Among Pansharpening Algorithms. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 2565–2586. [\[CrossRef\]](#)
7. Shahdoosti, H.R. Combining the spectral PCA and spatial PCA fusion methods by an optimal filter. *Inf. Fusion* **2016**, *27*, 150–160. [\[CrossRef\]](#)
8. Xu, Q.; Ding, L.; Zhang, Y.; Li, B. High-Fidelity Component Substitution Pansharpening by the Fitting of Substitution Data. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 7380–7392.
9. Dare, Z.P. Wavelet based image fusion techniques—An introduction, review and comparison. *IEEE Trans. Geosci. Remote Sens.* **2007**, *62*, 249–263.
10. Giuseppe, M.; Davide, C.; Luisa, V.; Giuseppe, S. Pansharpening by Convolutional Neural Networks. *Remote Sens.* **2016**, *8*, 594.
11. Liu, X.; Wang, Y.; Liu, Q. PSGAN: A Generative Adversarial Network for Remote Sensing Image Pan-Sharpener. *IEEE Trans. Geosci. Remote Sens.* **2018**. [\[CrossRef\]](#)
12. Zhang, Y.; Liu, C.; Sun, M.; Ou, Y. Pan-Sharpener Using an Efficient Bidirectional Pyramid Network. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5549–5563. [\[CrossRef\]](#)
13. Hallabia, H.; Hamam, H.; Hamida, A.B. An Optimal Use of SCE-UA Method Cooperated with Superpixel Segmentation for Pansharpening. *IEEE Geosci. Remote Sens. Lett.* **2020**. [\[CrossRef\]](#)
14. He, L.; Xi, D.; Li, J.; Zhu, J. A Spectral-Aware Convolutional Neural Network for Pansharpening. *Appl. Sci.* **2020**, *10*, 5809. [\[CrossRef\]](#)
15. Liu, X.; Wang, Y.; Liu, Q. *Remote Sensing Image Fusion Based on Two-Stream Fusion Network*; Springer: Cham, Switzerland, 2018.
16. Wei, Y.; Yuan, Q.; Shen, H.; Zhang, L. Boosting the Accuracy of Multispectral Image Pansharpening by Learning a Deep Residual Network. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1795–1799. [\[CrossRef\]](#)
17. Dadrasjavan, F.; Samadzadegan, F.; Fathollahi, F. Spectral and Spatial Quality assessment of IHS and Wavelet Based Pansharpening Techniques for High Resolution Satellite Imagery. *Adv. Image Video Process.* **2018**, *6*, 1.
18. Laben, C.A.; Brower, B.V. Process for Enhancing the Spatial Resolution of Multispectral Imagery Using Pan-Sharpener. U.S. Patent US6011875A, 4 January 2000.
19. Leung, Y.; Liu, J.; Zhang, J. An Improved Adaptive Intensity–Hue–Saturation Method for the Fusion of Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2013**, *11*, 985–989. [\[CrossRef\]](#)
20. Ghahremani, M.; Ghassemian, H. Nonlinear IHS: A Promising Method for Pan-Sharpener. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1606–1610. [\[CrossRef\]](#)
21. Vivone, G.; Alparone, L.; Id, A.; Garzelli, L.; Lolli, S. Fast Reproducible Pansharpening Based on Instrument and Acquisition Modeling: AWLP Revisited. *Remote Sens.* **2019**, *11*, 2315. [\[CrossRef\]](#)
22. Benz, U.C.; Hofmann, P.; Willhauck, G.; Lingenfelder, I.; Heynen, M. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *ISPRS J. Photogramm. Remote Sens.* **2004**, *58*, 239–258. [\[CrossRef\]](#)
23. Kim, Y.; Lee, C.; Han, D.; Kim, Y. Improved Additive-Wavelet Image Fusion. *IEEE Geosci. Remote Sens. Lett.* **2011**, *8*, 263–267. [\[CrossRef\]](#)
24. Pradhan, P.S.; King, R.L.; Younan, N.H.; Holcomb, D.W. Estimation of the Number of Decomposition Levels for a Wavelet-Based Multiresolution Multisensor Image Fusion. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 3674–3686. [\[CrossRef\]](#)

25. Burt, P.J.; Adelson, E.H. The Laplacian pyramid as a compact image code. *IEEE Trans. Commun.* **1983**, *31*, 532–540. [\[CrossRef\]](#)
26. Li, S.; Yin, H.; Fang, L. Remote Sensing Image Fusion via Sparse Representations Over Learned Dictionaries. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 4779–4789. [\[CrossRef\]](#)
27. Vicinanza, M.R.; Restaino, R.; Vivone, G.; Mura, M.D.; Chanussot, J. A pansharpening method based on the sparse representation of injected details. *IEEE Geosci. Remote Sens. Lett.* **2014**, *12*, 180–184. [\[CrossRef\]](#)
28. Wei, Q.; Bioucas-Dias, J.; Dobigeon, N.; Tourneret, J.Y. Hyperspectral and Multispectral Image Fusion based on a Sparse Representation. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3658–3668. [\[CrossRef\]](#)
29. Amin-Naji, M.; Aghagolzadeh, A.; Ezoji, M. Ensemble of CNN for Multi-Focus Image Fusion. *Inf. Fusion* **2019**, *51*, 201–214. [\[CrossRef\]](#)
30. Prabhakar, K.R.; Srikar, V.S.; Babu, R.V. DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
31. Xu, K.; Wen, L.; Li, G.; Bo, L.; Huang, Q. Spatiotemporal CNN for Video Object Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
32. Ghosh, A.; Mishra, N.S.; Ghosh, S. Fuzzy clustering algorithms for unsupervised change detection in remote sensing images. *Inf. Sci.* **2011**, *181*, 699–715. [\[CrossRef\]](#)
33. Ullah, W.; Ullah, A.; Hussain, T.; Khan, Z.A.; Baik, S.W. An Efficient Anomaly Recognition Framework Using an Attention Residual LSTM in Surveillance Videos. *Sensors* **2021**, *21*, 2811. [\[CrossRef\]](#)
34. Ullah, A.; Muhammad, K.; Ding, W.; Palade, V.; Haq, I.U.; Baik, S.W. Efficient Activity Recognition using Lightweight CNN and DS-GRU Network for Surveillance Applications. *Appl. Soft Comput.* **2021**, *103*, 107102. [\[CrossRef\]](#)
35. Scarpa, G.; Vitale, S.; Cozzolino, D. Target-Adaptive CNN-Based Pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5443–5457. [\[CrossRef\]](#)
36. Zhong, J.; Yang, B.; Huang, G.; Fei, Z.; Chen, Z. Remote Sensing Image Fusion with Convolutional Neural Network. *Sens. Imaging* **2016**, *17*, 10. [\[CrossRef\]](#)
37. Chao, D.; Chen, C.L.; He, K.; Tang, X. *Learning a Deep Convolutional Network for Image Super-Resolution*; Springer International Publishing: New York, NY, USA, 2014.
38. 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. [\[CrossRef\]](#)
39. Lecun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code. *Neural Comput.* **1989**, *1*, 541–551. [\[CrossRef\]](#)
40. Masi, G.; Cozzolino, D.; Verdoliva, L.; Scarpa, G. CNN-based pansharpening of multi-resolution remote-sensing images. In Proceedings of the Urban Remote Sensing Event, Dubai, United Arab Emirates, 6–8 March 2017.
41. Vitale, S. A CNN-Based Pansharpening Method with Perceptual Loss. In Proceedings of the 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July 28–2 August 2019.
42. Shao, Z.; Cai, J. Remote Sensing Image Fusion with Deep Convolutional Neural Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 1656–1669. [\[CrossRef\]](#)
43. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Adv. Neural Inf. Process. Syst.* **2014**, *3*, 2672–2680. [\[CrossRef\]](#)
44. He, K.; Zhang, X.; Ren, S.; Jian, S. *Identity Mappings in Deep Residual Networks*; Springer: Cham, Switzerland, 2016.
45. Rui, X.; Cao, Y.; Kang, Y.; Song, W.; Ba, R. Maskpan: Mask Prior Guided Network for Pansharpening. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020.
46. Yuan, Q.; Wei, Y.; Meng, X.; Shen, H.; Zhang, L. A Multiscale and Multidepth Convolutional Neural Network for Remote Sensing Imagery Pan-Sharpener. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 978–989. [\[CrossRef\]](#)
47. Chen, H.; Teng, M.; Shi, B.; Wang, Y.; Huang, T. Learning to Deblur and Generate High Frame Rate Video with an Event Camera. *arXiv* **2020**, arXiv:2003.00847.
48. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
49. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
50. 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.