Learning correspondence in Frequency Domain by a Latent-Space Similarity Loss for Multispectral Pansharpening

Yinghui Xing\textsuperscript{a,*}, Yan Zhang\textsuperscript{a}, Houjun He\textsuperscript{b}, Xiuwei Zhang\textsuperscript{a}, Yanning Zhang\textsuperscript{a}

\textsuperscript{a}School of Computer Science, Northwestern Polytechnical University, Xi’an, Shaanxi, China, 710072
The National Engineering Laboratory for Integrated Aerospace-Ground-Ocean Big Data Application Technology, Northwestern Polytechnical University, Xi’an, Shaanxi, China, 710072
\textsuperscript{b}The Information Center of Yellow River Conservancy Commission, Zhengzhou, Henan, China

Abstract

The process of fusing a high spatial resolution (HR) panchromatic (PAN) image and a low spatial resolution (LR) multispectral (MS) image to obtain an HRMS image is known as pansharpening. With the development of convolutional neural networks, the performance of pansharpening methods has been improved, however, the blurry effects and the spectral distortion still exist in their fusion results due to the insufficiency in details learning and the mismatch between the high-frequency (HF) and low-frequency (LF) components. Therefore, the improvements of spatial details at the premise of reducing spectral distortion is still a challenge. In this paper, we propose a frequency-aware network (FAN) together with a novel latent-space similarity loss to address above mentioned problems. FAN is composed of three modules, where the frequency feature extraction module aims to extract features in the frequency domain with the help of discrete wavelet transform (DWT) layers, and the inverse DWT (IDWT) layers are then utilized in the frequency feature fusion module to reconstruct the features. Finally, the fusion results are obtained through the reconstruction module. In order to learn the correspondence, we also propose a latent-

*Corresponding author.
Email address: xyh_7491@nwpu.edu.cn (Yinghui Xing)
space similarity loss to constrain the HF features derived from PAN and MS branches, so that HF features of PAN can reasonably be used to supplement that of MS. Experimental results on three datasets at both reduced- and full-resolution demonstrate the superiority of the proposed method compared with several state-of-the-art pansharpening models, especially for the fusion at full resolution.

**Keywords:** Pansharpening, Image fusion, Frequency feature extraction, Frequency correspondence, Remote sensing

---

**Fig. 1.** Visual comparisons among reference and fusion images of different pansharpening methods as well as their Fourier spectrum. It is clear to see that Fourier spectrum of GSA [6] and MTF-GLP-CBD [11] results have some unexpected frequency components, so they suffer from obvious spectral distortions, while that of GTP-PNet [25] and FusionNet [26] results miss some specific frequency components, resulting in blurry effects and the lose of important boundaries of objects. (One can zoom in for more details)

1. **Introduction**

   With the increasing demand of earth observation and monitoring, many optical satellites were launched, such as GaoFen-2, QuickBird and WorldView-2, whose sensors can produce bundled multispectral (MS) and panchromatic (PAN) images of the same scene. MS images have high spectral resolution but relatively low spatial resolution, while the single-band PAN images have
reverse characteristics. Due to physical limitations of the satellite sensors [4], neither MS nor PAN images have the high resolution both in spatial and spectral domains. However, such a high-quality image is urgently required in practical to effectively facilitate the visual interpretation and other applications, e.g., object detection [1], land-cover classification [2] and change detection [3], etc. Therefore, researchers resorted to pansharpening technology, a feasible solution to obtain high-resolution (HR) MS images via the fusion of low-resolution (LR) MS and corresponding HR PAN images.

Pansharpening has been developed for nearly 40 years, and it can be mainly divided into four categories [5], i.e., component substitution (CS)-based, multiresolution analysis (MRA)-based, variational optimization (VO)-based and deep learning (DL)-based methods.

CS-based methods, also known as spectral methods, are based on the transformation of MS images to project them into a space, where the spatial and the spectral components are assumed to be separated to each other, then the spatial components of MS images are substituted by histogram-matched PAN images. Finally, the inverse projection is applied to obtain fusion results [6, 7, 8, 9]. Generally, the fusion results of CS-based methods have abundant spatial details, but they are easily affected by spectral distortions.

MRA-based methods rely on the extraction and injection of spatial information, thus are also referred to spatial methods. They extract high-frequency components from PAN images by multiresolution analysis tools [10, 11, 12, 13], and inject them into the interpolated MS images to get HRMS images. In contrast to CS-based methods, these methods preserve spectral information well, but tend to suffer from some spatial degradations.

VO-based methods explore the relationships among PAN, LRMS and HRMS images, and design an energy function with various regularization terms like total-variation [14], sparse representation [15], and low-rank [16] to obtain the HRMS images. Most of them design spatial and spectral correlation terms based on the similarity of spatial information between HRMS and PAN, and the consistence of spectral information between HRMS and MS [17, 18]. In
general, VO-based methods can obtain high-quality fusion images, but they have high computational complexity.

Recently, DL-based methods show superior performance due to the powerful ability of automatic nonlinear representation learning. Among them, convolutional neural networks (CNNs) based pansharpening methods have attracted more attention. The first CNN-based pansharpening neural network (PNN) [19] adopted a simple three-layer convolutional structure [20], and obtained desirable fusion results. After that, many CNN-based pansharpening methods were proposed by incorporating some advanced feature extraction structures and fusion strategies. However, their results are easily affected by blurry effects because of the insufficient details learning. Actually, neural networks tend to fit low-frequency components with a high priority [21, 22], which results in the loss of some high-frequency edges and details. In addition, commonly used loss functions, such as mean squared error (MSE) and mean absolute error (MAE), are insensitive to small perturbations, which further leads to the loss of spatial details [29]. Although many works utilized residual learning strategy to enforce networks to learn high-frequency components [23, 24, 25, 26], they still suffered from spatial and spectral distortions. As can be seen in Fig. 1, where the spectra of different fusion results have discrepancy when compared with that of reference image. We think the missing or redundant frequency components attribute to insufficient details learning and the frequency inconsistency between low- and high-frequency components.

To deal with the problem mentioned above, we propose a frequency-aware network (FAN) together with a novel latent-space similarity (LSS) loss for pansharpening. FAN can preserve the high-frequency details by explicitly learning the frequency-aware features with the assistance of DWT/IDWT layers, and also improve the frequency consistency by considering the LSS loss to align the features in the frequency domain. The main contributions of this paper can be summarized as follows. First, we propose a frequency-aware network to emphasize the details learning in pansharpening task. In our work, the frequency features are extracted, aligned and fused automatically. FAN is the first DL-based
pansharpening method that extracts and updates the frequency components explicitly. Second, a novel latent-space similarity (LSS) loss is proposed to learn the correspondence between MS and PAN features, and to further restrict the high-frequency features of PAN images as equivalent as possible to the missing high-frequency parts of MS images. Third, DWT and IDWT layers are introduced to our network to enforce FAN work in the frequency domain, and also make it adaptively learn the frequency-aware features through the backward propagations.

The remainder of the paper is outlined as follows. Section 2 introduces the related works and the motivations of proposed model. In Section 3 we present the architecture of proposed method as well as the designed loss function in detail. The experimental results and analysis are reported in Section 4. Finally, conclusions are drawn in Section 5.

2. Related Works and Motivations

In this section, we first review several CNN-based pansharpening methods relevant to ours and then provide the motivations of proposed method.

2.1. CNN-Based Pansharpening Methods for details learning

To solve the problem of insufficient details learning, many CNN-based methods resort to learning the high-frequency residuals through residual learning (RL) strategy. The adoption of RL strategy aims to make networks focus on the learning of high-frequency details and thus to address spatial degradation problem in fusion results. Pan-sharpening network (PanNet) [23] was the first method that adopted RL strategy to learn the high-frequency components, and its fusion results preserved spatial details well. Similarly, a target-adaptive CNN (TA-PNN) model [24] was proposed and the effectiveness of RL strategy was verified through experiments on four datasets. In [25], the authors introduced the gradient transformation prior (GTP) into RL architecture, called GTP-PNet, which is composed of two networks, i.e., gradient transformation network
(TNet) and pan-sharpening network (PNet). TNet tried to explore the nonlinear mapping between the gradients of PAN and HRMS images. Then PNet is optimized under the guidance of the trained TNet to learn the residuals. FusionNet \cite{26} took the differences between duplicated PAN and up-sampled MS images as input, and learned the spatial details through a structure-preservation network. These methods reconstructed spatial details by calculating the residuals between source images and target HRMS image.

Apart from residual learning (RL) strategy, some other works adopted the progressive fusion strategy to compensate the spatial details. A Laplacian pyramid pan-sharpening network (LPPN) \cite{27} was designed under the Laplacian pyramid framework and utilized the recursive structure to progressively fuse spatial information at different scales. A dynamic cross feature fusion network (DCFNet) \cite{28} contained a high-resolution branch served as the mainbranch and two parallel low-resolution branches to progressively supplement information to the mainbranch. However, few of them directly learn frequency-aware features, resulting in the loss of high-frequency information because of the inherent bias of CNNs that tend to fit the low-frequency information \cite{21, 22}. In our model, we introduce 2D DWT and IDWT layers to explicitly learn the frequency-aware features, so as to preserve more high-frequency details in the fusion results.

2.2. Motivations

Recently, many computer vision tasks explicitly and adaptively extracted features in frequency domain for efficient feature learning. Wavelet integrated CNNs (WaveCNets) \cite{33} designed DWT and IDWT layers and integrated them into commonly used CNNs for noise-robust image classification, but it dealt with high-level vision tasks, and did not take the details preservation into consideration. Style and wavelet based generative adversarial network (SWAGAN) \cite{29} predicted wavelet coefficients at different scales to make the generated images have more HF details. However, it utilized normal discrete wavelet transforms which lacked the ability of backward propagations when dealing with downstream tasks. Inspired by these tasks, we propose a frequency-aware network
(FAN) that can adaptively learn the frequency components through backward propagations of not only convolutional layers but also DWT and IDWT layers. Furthermore, a novel latent-space similarity loss is applied to high-frequency features of PAN and MS to improve the consistence between the high- and low-frequency components contained in fusion results.

**Fig. 2.** The architecture of proposed FAN, where FFE and F^3 are the acronyms of “Frequency Feature Extraction” and “Frequency Feature Fusion”, respectively. L (H) in the superscript represents low-frequency (high-frequency) related modules and M (P) in the subscript denotes the modules in MS (PAN) branch.

### 3. Proposed Method

#### 3.1. Notations

MS and PAN images are represented as $M \in \mathbb{R}^{S/r \times T/r \times B}$ and $P \in \mathbb{R}^{S \times T}$, where $S$ and $T$ is width and height of an image, respectively, and $B$ denotes number of image bands. $r$ is the ratio of spatial resolution between PAN and MS. The pansharpened results and the reference images are represented by $\hat{M} \in \mathbb{R}^{S \times T \times B}$ and $R \in \mathbb{R}^{S \times T \times B}$. In this paper, the up-sampled MS images obtained by bicubic interpolation is denoted by $\tilde{M} \in \mathbb{R}^{S \times T \times B}$. 


3.2. Network Overview

The overall architecture of proposed frequency-aware network (FAN) is shown in Fig. 2. It consists of three modules: frequency feature extraction (FFE) module, frequency feature fusion (F³) module and the reconstruction module. Following the paradigm of traditional MRA-based methods, FAN fuses the MS and PAN image also through three steps: feature extraction, feature fusion, and feature reconstruction, where FFE module extracts frequency features of MS and PAN with the assistance of DWT layers, and these features are fused in the frequency domain by F³ module. Finally, the fused frequency features are reconstructed by the reconstruction module with the corresponding IDWT layers. To make the frequency features consistent with each other, we also proposed a latent-space similarity (LSS) loss to learn the correspondence of them. Though follows the similar paradigm as traditional MRA-based methods, the proposed FAN has two advantages: 1) FAN is based on a nonlinear model, and it can well represent the misalignment in the spectral range between MS and PAN sensors. 2) FAN learns the correspondence in the wavelet domain, where the wavelet space allows us to have smooth filters with fewer parameters, and the learnability provides us more flexibility. Next, we will introduce the architecture and operations of each module in detail.

3.3. Frequency Feature Extraction module

Frequency Feature Extraction (FFE) module extracts the features in the frequency domain, therefore, we first utilize the discrete wavelet transform (DWT) layer to transform the images into frequency domain. Discrete Wavelet Transform (DWT) is an efficient multi-resolution analysis tool in signal processing [43], which has the ability of anti-aliasing [33], and can also preserve more structural information in image processing [44]. Motivated by the success of [29] [33] and the advantages of DWT, we resort to DWT layers to adaptively learn frequency components. Compared with normal DWT, DWT layers are compatible with convolution layers and their gradients can be back-propagated,
which not only make the network learn in the frequency domain, but also have the flexibility.

FFE module is composed of a PAN branch and an MS branch, in which DWT layers are coupled with convolutional blocks (CBs) to transform original images into frequency domain and then to learn frequency-aware features. MS and PAN branches share the same architecture but different parameters in this module to keep the consistency, preliminarily.

Due to the fact that MS branch has the same structure as PAN branch, we take PAN branch as an example to introduce the FFE module. As can be seen from Fig. 2, PAN image $\mathbf{P} \in \mathbb{R}^{S \times T}$ is fed into PAN branch, and it is decomposed by the first DWT layer to obtain the LF components $G_p^{L(1)}$ and the HF components $G_p^{H(1)}$. Then, two parallel convolutional blocks $\text{CB}^L_p$ and $\text{CB}^H_p$ whose architecture can be seen in Fig. 2 are used to extract LF-aware features $F_p^{L(1)}$ and HF-aware features $F_p^{H(1)}$,

\[
\begin{align*}
G_p^{L(1)}, G_p^{H(1)} &= \text{DWT}^1_p(\mathbf{P}) \\
F_p^{L(1)} &= \text{CB}^L_p(G_p^{L(1)}) \\
F_p^{H(1)} &= \text{CB}^H_p(G_p^{H(1)}),
\end{align*}
\]

(1)

where $\text{DWT}^1_p$ denotes the first DWT layer of the PAN branch.

Similar to traditional multiscale wavelet decomposition, another DWT layer is applied to the LF component to further conduct the finer decomposition. Two convolutional blocks $\text{CB}^2_p^L$ and $\text{CB}^2_p^H$ are adopted in the same way to obtain the second level of LF- and HF-aware features,

\[
\begin{align*}
F_p^{x(2)} &= \text{CB}^2_p^x(\text{DWT}^2_p(F_p^{L(1)})),
\end{align*}
\]

(2)

where $x \in \{L, H\}$.

For MS branch, it takes interpolated MS image $\tilde{\mathbf{M}} \in \mathbb{R}^{S \times T \times B}$ as input, and produces LF- and HF-aware features via a cascade of DWT and convolutional layers.
3.4. Frequency Feature Fusion (F\textsuperscript{3}) module

After FFE module, we have four PAN-related features at two scales, i.e., \( F_{P}^{L(1)} , F_{P}^{H(1)} , F_{P}^{L(2)} , F_{P}^{H(2)} \), and four corresponding MS-related features \( F_{M}^{L(1)} , F_{M}^{H(1)} , F_{M}^{L(2)} , F_{M}^{H(2)} \). Then we should consider the way of combining them to obtain the fused features. In traditional MRA-based methods, LF components are obtained directly from MS image, and HF components are acquired from both PAN and MS images, as shown in Eq. (3),

\[
\hat{M}_{b} = \tilde{M}_{b} + g_{b} (P - P_{L})
\]

\[
= \tilde{M}_{b}^{L} + \tilde{M}_{b}^{H} + g_{b} (P - P_{L})
\]

\[
= \tilde{M}_{b}^{L} + \tilde{M}_{b}^{H} + g_{b} P_{H},
\]

where \( b \) indexes the \( b \)th bands of multispectral images and \( g_{b} \) denotes injection coefficient with respect to \( b \)th bands.

In order to avoid introducing unexpected frequency components, in this paper, we propose a frequency feature fusion (F\textsuperscript{3}) module to fuse these features. Note that F\textsuperscript{3} module plays the role of inverse process of FFE module, therefore, it has four convolutional blocks and two corresponding IDWT layers. Similar to Eq. (3), the HF components \( F_{H(i)}^{H} \) come from both the MS and PAN features. In our F\textsuperscript{3} module, the HF features of MS and PAN at the same scale are first concatenated along the channel dimension, then the concatenated features are transformed by a \( 1 \times 1 \) convolutional layer and a HF-aware convolutional block \( CB_{2H}^{H} \), that is,

\[
F_{(i)}^{H} = CB_{2H}^{H} (\text{Conv} (\text{Concat} (F_{P}^{H(i)}, F_{M}^{H(i)}))), i = 1, 2,
\]

where \( \text{Concat}(\cdot, \cdot) \) denotes the operation that concatenates features along the channel dimension.

The LF components of IDWT layers are obtained from the MS branch, where the LF features should firstly pass through a convolutional block \( CB_{2L}^{L} \) \((i = 1, 2)\). The LF components of the first IDWT layer is,

\[
F_{(2)}^{L} = CB_{2L}^{L} (F_{M}^{L(2)}),
\]
and that of the second layer is

\[ F^L_{(i)} = \text{CB}^L_2(F^2). \]  

(6)

The inverse wavelet transform through the IDWT layers can be formulated as:

\[ F^i = \text{IDWT}^i(F^L_{(i)}, F^H_{(i)}), i = 1, 2, \]  

(7)

where \( F^1 \) is the final fused features.

3.5. Reconstruction module

The reconstruction module takes the fused features as inputs and transforms them from the feature space to the original image space. It is composed of a convolutional block \( \text{CB}^1 \), which converts \( F^1 \) back to the image space to obtain the high-quality fused result \( \hat{M} \),

\[ \hat{M} = \text{CB}^1(F^1). \]  

(8)

It should be noted that the filter size of each convolutional layer in our model is \( 3 \times 3 \) whose padding size is 1, except for the \( 1 \times 1 \) convolutional layer with zero-padding, and the output features all have 48 channels. The activation functions used in FAN except for the last layer are Leaky Rectified Linear Unit (LReLU) with \( \alpha = 0.2 \), and that of the last layer is tanh.

3.6. Latent-Space Similarity Loss

As mentioned before, HF components come from both the MS and PAN branches, while LF components are only obtained from MS one. Therefore, there is no guarantee that the HF and LF components can completely be consistent with each other, leading to spectral distortion. To address this problem, we design a latent-space similarity (LSS) loss aiming to constrain the correspondence between HF-aware features of PAN and MS so as to align the HF-aware features of PAN to that of MS.

In order to calculate the LSS loss, the HF-aware features of MS and PAN branches are firstly transformed by a multi-layer perceptron (MLP) as Fig. 3.
Fig. 3. Structure of multi-layer perceptron (MLP) adopted in the latent-space similarity (LSS) loss, where FC, BN and LReLU represent the fully-connected, batch-normalization and leaky rectified linear unit layers, respectively.

shows. Then we compute the cross-correlation matrix of them to represent the correlations,

\[ C = CC(MLP(R(F_1)), MLP(R(F_2))), \]

where \( R(\cdot) \) is the operator that reshapes a fourth-order tensor feature with the shape of \( \mathbb{R}^{D_1 \times D_2 \times D_3 \times D_4} \) to a matrix with the shape of \( \mathbb{R}^{D_1 \times D_2 D_3 D_4} \). \( MLP(\cdot) \) represents the operations of MLP structure, and \( CC(\cdot, \cdot) \) denotes the calculation of cross-correlation matrix \( C \).

There are two reasons why the LF-aware features are transformed by the MLP structure. Firstly, we use MLP structure to reduce the dimension of LF-aware features to facilitate the computation of cross-correlation matrix. The second reason is that MLP structure can further aggregate the non-local information of features, resulting in more discriminative feature representations.

The diagonal elements in \( C \) denote the correlations of two representations from corresponding branches, therefore, the LSS loss is formulated as,

\[ \mathcal{L}_{LSS} = \frac{1}{N} \sum_i^N (1 - c_{ii})^2 + \lambda \frac{1}{N(N-I)} \sum_i^N \sum_{j \neq i} c_{ij}^2, \]  

where \( c_{ii} \) and \( c_{ij} \) denote the diagonal and off-diagonal elements of cross-correlation matrix \( C \in \mathbb{R}^{N \times N} \). \( \lambda \) is the balance parameter and is set to \( 5 \times 10^{-3} \). The diagonal elements of \( C \) are constrained as close as 1, which means two representations from the same group of PAN and MS should as close as possible. At the same time, off-diagonal ones of \( C \) are constrained as close as 0, which aims to decorrelate two representations from different groups to produce instance-specific representations.
3.7. Overall Loss

Because our FAN model has two DWT layers, we can compute two LSS losses $L_{\text{LSS}}^1$ and $L_{\text{LSS}}^2$:

$$L_{\text{LSS}}^i = L_{\text{LSS}}(F_H^{(i)}(P), F_H^{(i)}(M)), i = 1, 2. \quad (11)$$

The LSS losses are combined with the mean absolute error (MAE) loss to constitute the overall loss function,

$$\mathcal{L} = \mathcal{L}_{\text{MAE}} + \beta (L_{\text{LSS}}^1 + L_{\text{LSS}}^2), \quad (12)$$

where $\mathcal{L}_{\text{MAE}} = (1/B) \sum_{b=1}^{B} ||R_b - \hat{M}_b||_1$ is the MAE loss. $\beta$ is the balance parameter.

4. Experimental Results and Analysis

In order to verify the effectiveness of the proposed method, we compare our model with nine state-of-the-art methods, including Gram-Schmidt adaptive (GSA) \[6\], band-dependent spatial-detail with physical constraints (BDSD-PC) \[9\], Modulation Transfer Function-generalized Laplacian pyramid with context-based decision (MTF-GLP-CBD) \[11\], additive wavelet luminance proportional with haze-correction (AWLP-H) \[13\], PCA/Wavelet model-based fusion (PWMBF) \[35\], PanNet \[23\], FusionNet \[26\], GTP-PNet \[25\] and LPPN \[27\]. Experiments are conducted on three satellite datasets at both reduced and full resolution. In the reduced-resolution experiments, the widely used evaluation indexes, spectral angle mapper (SAM) \[36\], erreur relative globale adimensionnelle de synthèse (ERGAS) \[37\], Q2\[n\] \[38\] and the spatial correlation coefficient (SCC) \[39\] are adopted, while in the full-resolution experiments, the spectral distortion index ($D_\lambda$), the spatial distortion index ($D_s$) and the quality with no reference (QNR) \[40\] index are used to evaluate the quality of fusion results. The optimal values for SAM, ERGAS, $D_\lambda$ and $D_s$ are 0, while for SCC, QNR, and Q2\[n\] are 1.
### Table 1: Details of Used Datasets.

| Sensor                  | Spatial resolution | Training data | Validation data | Test data  |
|-------------------------|--------------------|---------------|----------------|------------|
|                         | Size               | Number        | Size           | Number     | Size          | Number     |
| WV-4                   | 1.2m MS            | 64 × 64 × 4 MS | 7938           | 64 × 64 × 4 MS | 772 256 × 256 × 4 MS | 271 |
|                         | 0.3m MS            | 256 × 256 PAN  | 6943           | 256 × 256 PAN | 743 1024 × 1024 PAN | 156 |
| QB                     | 2.4m MS            | 64 × 64 × 4 MS | 6943           | 64 × 64 × 4 MS | 743 256 × 256 × 4 MS | 156 |
|                         | 0.6m MS            | 256 × 256 PAN  | 6943           | 256 × 256 PAN | 743 1024 × 1024 PAN | 156 |
| WV-2                   | 2.0m MS            | 64 × 64 × 8 MS | 9641           | 64 × 64 × 8 MS | 945 256 × 256 × 8 MS | 136 |
|                         | 0.5m MS            | 256 × 256 PAN  | 9641           | 256 × 256 PAN | 945 1024 × 1024 PAN | 136 |

#### 4.1. Datasets

Experiments are conducted on three datasets acquired from WorldView-4 (WV-4), QuickBird (QB) and WorldView-2 (WV-2) satellites. All datasets are composed of paired multi-bands MS and single-band PAN images. The MS images of WV-4 and QB datasets have four spectral bands (i.e., blue, green, red, and near infrared) while the MS images of WV-2 dataset have not only the blue, green, red and near infrared bands, but also the coastal, yellow, red edge, and near-infrared 2 bands. The details of the datasets are listed in Table 1.

In the reduced-resolution experiment, the original MS and PAN images are firstly filtered by a 5 × 5 Gaussian smooth kernel and then downsampled by a factor of 4. The original MS images are treated as references, and these downsampled pairs of MS and PAN images are inputs of all the comparison methods [41]. In the full-resolution experiment, original MS and PAN pairs are directly taken as inputs.

#### 4.2. Training Details

The training samples are obtained under the Wald’s protocol [41], where the original MS and PAN patches are filtered by a 5 × 5 Gaussian kernel and downsampled by a factor of 4, then the original MS patches are treated as references. The sizes of training MS, PAN and reference patches are 16 × 16 × B, 64 × 64, and 64 × 64 × B (B = 4 for WV-4 and QB datasets, and B = 8 for WV-2 dataset), respectively.

Our proposed model is achieved under the PyTorch framework [42] and the
parameters are updated by the Adam optimizer\(^1\). The learning rate is initialized to \(1 \times 10^{-4}\). The batch size is empirically set to 32 and the training process terminates after 2000 epochs. Additionally, we train four CNN-based methods, PanNet \(^{23}\), FusionNet \(^{26}\), GTP-PNet \(^{25}\) and LPPN \(^{27}\), on our datasets according to the default settings provided by the authors to guarantee the fair comparisons. All experiments are supported by an Nvidia GTX 1080Ti GPU.

4.3. Experiments at Reduced Resolution

In this section, experiments at reduced resolution are conducted on the WV-4, QB and WV-2 datasets. Both the visual inspections and the quantitative assessments are presented to fully verify the superiority of our approach. We magnify representative regions to show the details of the fusion results. Furthermore, we also compute the absolute error maps (AEMs) between the fusion results and the reference image, which provide more insights about the performance of the compared methods. For quantitative assessments, we report the mean values and the standard deviation values of the indexes in Tables \(^2\) and \(^3\), where the best results are marked in bold and the second-best results are underlined.

Table 2: Average Quantitative Results on 271 Pairs of Test Data from WV-4 Dataset.

| Methods   | ERGAS  | Q4    | SAM   | SCC   | QNR  | \(D_s\) | \(D_f\) |
|-----------|--------|-------|-------|-------|------|--------|--------|
| Reduced resolution |       |       |       |       |      |        |        |
| Ideal value | 0     | 1     | 0     | 1     |      |        |        |
| GSA       | 3.0470±0.3345 | 0.8760±0.0014 | 3.0679±0.4112 | 0.8915±0.0014 | 0.8031±0.0013 | 0.0691±0.0002 | 0.0730±0.0007 |
| HDS-PF    | 3.2301±0.1084 | 0.8646±0.0017 | 3.3227±1.3760 | 0.8601±0.0016 | 0.894±0.0016 | 0.0148±0.0002 | 0.0739±0.0009 |
| MTF-GLP-CHD | 3.1908±0.3141 | 0.8605±0.0015 | 3.1438±1.4091 | 0.8851±0.0031 | 0.8875±0.0013 | 0.0446±0.0002 | 0.0713±0.0007 |
| AWLP-II   | 2.8441±0.2822 | 0.861±0.0008  | 2.919±1.2919  | 0.9179±0.0004 | 0.913±0.0006  | 0.044±0.0003  | 0.047±0.0002  |
| PWMBF     | 3.2870±0.2400 | 0.8460±0.0014 | 3.2756±1.5659 | 0.8827±0.0008 | 0.864±0.0012  | 0.0367±0.0004 | 0.0782±0.0003 |
| PanNet    | 1.8679±0.1350 | 0.8368±0.0047 | 2.5000±1.5736 | 0.9201±0.0032 | 0.908±0.0013  | 0.0363±0.0004 | 0.0768±0.0009 |
| FusionNet | 1.6754±0.4311 | 0.862±0.0074  | 2.3900±1.5702 | 0.954±0.0086  | 0.942±0.0015  | 0.0363±0.0007 | 0.0730±0.0003 |
| GTP-PNet  | 2.3218±0.2699 | 0.828±0.0018  | 3.014±1.6521  | 0.872±0.0006  | 0.934±0.0008  | 0.021±0.0003  | 0.045±0.0005  |
| LPPN      | 1.593±0.6560  | 0.835±0.0049  | 2.092±1.4801  | 0.956±0.0069  | 0.943±0.0009  | 0.022±0.0002  | 0.0108±0.0004 |
| FAN       | 1.792±0.1547  | 0.933±0.0012  | 1.805±0.4730  | 0.904±0.0013  | 0.964±0.0002  | 0.013±0.0001  | 0.022±0.0001  |

Fusion results and the corresponding AEMs of WV-4 dataset are shown in Figs. \(^1\) and Figs. \(^5\) respectively. It can be observed that the fusion results

\(^1\) Code will be available soon.
Fig. 4. Visual comparison of different methods on WorldView-4 (WV-4) dataset at reduced resolution.

of GSA [6], BDSD-PC [9], MTF-GLP-CBD [11], AWLP-H [13], PWMBF [35], PanNet [23], and GTP-PNet [25] suffer from varying degrees of spectral distortion. Although the results of FusionNet [26] and LPPN [27] have better spectral fidelity in terms of the color of soil, roof and vegetations, their boundaries between different objects tend to be blurry. Compared with them, the proposed FAN preserves the spatial and spectral information better, which can be seen especially from the zoomed-in regions in Fig. 4. The AEMs shown in Fig. 5 also verify the effectiveness of the proposed method in terms of the preservation of spatial and spectral information. Table 2 lists the quantitative assessments on 271 pairs of test data from the WV-4 dataset. From Table 2, it is clear that our proposed method obtains the best result regarding to the Q4, SAM, SCC indexes and ranks second in terms of ERGAS, which are in consistent with
Fig. 5. Absolute error maps (AEMs) between reference image and fusion images/reference on WV-4 dataset at reduced resolution.

visual inspections.

Table 3: Average Quantitative Results on 156 Pairs of Test Data from QB Dataset.

| Method          | ERGAS   | Q4     | SAM    | SCC     | QMR    | \( D_s \) | \( D_r \) |
|-----------------|---------|--------|--------|---------|--------|----------|----------|
| Reduced resolution |        |        |        |         |        |          |          |
| Full resolution  |        |        |        |         |        |          |          |
| Ideal value     | 0.0000 | 1.0000 | 0.0000 | 1.0000  | 0.0000 | 0.0000   | 0.0000   |
| GSA             | 2.7983 ± 0.057 | 0.6400 ± 0.0079 | 4.3239 ± 0.1018 | 0.9204 ± 0.0006 | 8.4320 ± 0.0032 | 0.6833 ± 0.0004 | 0.1789 ± 0.0000 |
| BDSD-PC         | 3.6900 ± 0.8889 | 0.8773 ± 0.0014 | 4.0325 ± 0.3408 | 0.9295 ± 0.0064 | 8.6941 ± 0.0016 | 0.1384 ± 0.0005 | 0.0739 ± 0.0000 |
| MTF-GLP-CBD     | 3.6705 ± 0.7075 | 0.8737 ± 0.0021 | 4.3265 ± 0.2050 | 0.9219 ± 0.0006 | 8.6650 ± 0.0017 | 0.0939 ± 0.0003 | 0.0801 ± 0.0000 |
| AWLP-H          | 3.4514 ± 1.3352 | 0.9053 ± 0.0008 | 3.5204 ± 0.7985 | 0.9374 ± 0.00010 | 8.9776 ± 0.0009 | 0.0521 ± 0.0003 | 0.0637 ± 0.0004 |
| PWMBF           | 3.6714 ± 1.7377 | 0.8571 ± 0.0026 | 4.3109 ± 1.2973 | 0.9229 ± 0.0064 | 8.2874 ± 0.0011 | 0.0701 ± 0.0005 | 0.0955 ± 0.0006 |
| PanNet          | 1.8231 ± 0.1917 | 0.9102 ± 0.0019 | 2.8717 ± 0.4195 | 0.9533 ± 0.0002 | 0.8911 ± 0.0010 | 0.0468 ± 0.0002 | 0.0774 ± 0.0004 |
| FusionNet       | 1.5123 ± 0.1034 | 0.9395 ± 0.0009 | 2.9084 ± 0.2044 | 0.9750 ± 0.0008 | 0.9107 ± 0.0006 | 0.0209 ± 0.0002 | 0.0851 ± 0.0003 |
| GTP-PNet        | 2.8529 ± 0.5879 | 0.8289 ± 0.0015 | 3.2133 ± 0.6455 | 0.9009 ± 0.0005 | 0.9231 ± 0.0015 | 0.0318 ± 0.0002 | 0.0469 ± 0.0008 |
| LPPN            | 1.7729 ± 0.2179 | 0.9223 ± 0.0010 | 2.9830 ± 0.306 | 0.9779 ± 0.0004 | 0.9416 ± 0.0006 | 0.0281 ± 0.0002 | 0.0332 ± 0.0001 |
| FAN             | 1.8988 ± 0.2361 | 0.9071 ± 0.0004 | 2.6237 ± 0.3136 | 0.9743 ± 0.0001 | 0.9754 ± 0.0001 | 0.0223 ± 0.0000 | 0.0124 ± 0.0000 |

Figs. 6-7 provide the fusion results and AEMs on the QB dataset. The pansharpened results of GSA [6], BDSD-PC [9], MTF-GLP-CBD [11], AWLP-H [13], PWMBF [35], PanNet [23], GTP-PNet [25], FusionNet [26] and LPPN [27] are affected by different levels of spectral distortion and spatial degradation, especially for the color and the shape of buildings in the zoomed-in regions and small objects on the playground. The fusion product of FAN is the closest to the reference. Additionally, absolute error error maps (AEMs) between fusion results and the reference image are demonstrated in Fig. 7. From Fig. 7 we can find that
FAN and FusionNet [26] have less information loss. The same conclusion can also be drawn by the quantitative assessments shown in Table 3 that FusionNet [26] and the proposed FAN obtain the best and the second-best indexes. All the indexes listed in Table 3 are calculated on the 156 pairs of test data from the QB dataset, from which we can conclude that FAN is competitive to FusionNet [26] and superior to other methods on QB dataset at the reduced resolution.

To verify the effectiveness of proposed method for eight-band MS image pansharpening, we conduct experiments on the WV-2 dataset. The experimental results are presented in Fig. 8. The results of GSA [6], BDSD-PC [9], MTF-GLP-CBD [11] and PWMBF [35] methods have sufficient spatial details, but they are affected by serious spectral distortion, which can be observed from the zoomed-in regions. Although AWLP-H [13], PanNet [23], GTP-PNet [25] and LPPN [27] suffer from less spectral distortion, they are affected by the spatial
Fig. 7. Absolute error maps (AEMs) between reference image and fusion images/reference on QB dataset at reduced resolution.

Table 4: Average Quantitative Results on 136 Pairs of Test Data from WV-2 Dataset.

| Methods     | ERGAS       | Q4       | SAM       | SCC | QNR   | $D_s$ | $D_a$ |
|-------------|-------------|----------|-----------|-----|-------|-------|-------|
| Reduced resolution |           |          |           |     |       |       |       |
| Ideal value | 0          | 1        | 1         | 1   | 1     | 1     | 1     |
| GSA         | 4.8996±0.6014 | 0.8017±0.0445 | 6.0521±1.2636 | 0.8592±0.0067 | 0.8307±0.0035 | 0.9031±0.0021 | 0.1056±0.0036 |
| BDSD-PC     | 4.9243±0.6125 | 0.8022±0.0150 | 6.4459±1.8612 | 0.8762±0.0002 | 0.9054±0.0018 | 0.0961±0.0027 | 0.0066±0.0058 |
| MTF-GLP-CBD | 4.9737±0.6721 | 0.8021±0.0149 | 6.6062±1.3094 | 0.8606±0.0007 | 0.8732±0.0006 | 0.0536±0.0025 | 0.0089±0.0044 |
| AWLP-H      | 4.7903±0.6615 | 0.8094±0.0148 | 5.9112±1.1113 | 0.8900±0.0001 | 0.8336±0.0157 | 0.0956±0.0062 | 0.0941±0.0061 |
| PWMBF       | 5.1137±0.7221 | 0.7273±0.0140 | 7.2528±3.0394 | 0.8556±0.0004 | 0.8110±0.0068 | 0.0820±0.0020 | 0.1196±0.0048 |
| PanNet      | 4.5570±0.5361 | 0.8220±0.0160 | 5.6769±1.7453 | 0.9029±0.0003 | 0.9067±0.0133 | 0.0475±0.0038 | 0.0328±0.0054 |
| FusionNet   | 4.5827±0.5451 | 0.8280±0.0165 | 5.6969±1.7453 | 0.9029±0.0003 | 0.9067±0.0133 | 0.0475±0.0038 | 0.0328±0.0054 |
| GTP-PNet    | 4.5570±0.5361 | 0.8220±0.0160 | 5.6769±1.7453 | 0.9029±0.0003 | 0.9067±0.0133 | 0.0475±0.0038 | 0.0328±0.0054 |
| LPPN        | 4.4662±0.5606 | 0.8365±0.0185 | 5.7387±1.7201 | 0.8951±0.0001 | 0.8833±0.0139 | 0.0475±0.0030 | 0.0474±0.0047 |
| FAN         | 3.9358±0.4900 | 0.8516±0.0173 | 4.6238±2.1948 | 0.9240±0.0001 | 0.9067±0.0141 | 0.0325±0.0015 | 0.0310±0.0011 |

degradation. From Fig. 9, we can see that the error maps of FusionNet and FAN are comparable, but the advantages of FAN can be observed from the zoomed-in regions of Fig. 8(m) that the color of road, roof, grasses and the shape of objects (such as the line between the road, the white part of roof) are more reliable. In addition, the quality assessments on 136 pairs of test data from WV-2 dataset are shown in Table 4, from which we can also conclude that the proposed FAN outperforms the compared state-of-the-art methods with respect to four adopted evaluation metrics.
Fig. 8. Visual comparison of different methods on WorldView-2 (WV-2) dataset at reduced resolution.

4.4. Experiments at Full Resolution

Due to the fact that the model is trained on simulated training samples obtained through the Wald’s protocol [41], there is a resolution gap between the training samples and the test samples at full resolution, which makes the fusion of full-resolution images more challenging. In this section, we conduct experiments on full-resolution WV-4, QB and WV-2 datasets to verify the effectiveness of proposed model. Similarly, both the visual inspections and the quantitative assessments are used in the comparison. Figs. 10-12 show the visual comparisons of different methods on WV-4, QB, and WV-2 datasets. Because there are no reference images, original PAN and the up-sampled MS images are shown in these figures to assist the visual inspections.

Fig. 10 shows the fusion results on WV-4 dataset. Similarly, PAN and the up-sampled MS are shown firstly, followed by the results of compared methods.
It can be found from the zoomed-in regions that BDSD-PC [9], PWMBF [35] and FusionNet [26] yield obvious spectral distortion, while the results of GSA [6], MTF-GLP-CBD [11], AWLP-H [13], PanNet [23], and GTP-PNet [25] are affected by the blurry effects. Concerning both the fidelity of spectral and spatial information, LPPN [27] and FAN achieve superior performance, but the fusion product of LPPN [27] contains more artifacts. Table 2 reports the average quantitative results on 271 pairs of test data from the WV-4 dataset. From Table 2, we can observe that FAN obtains the best fusion results compared with other state-of-the-art methods.

Fig. 11 displays a group of pansharpened images produced by different methods on the QB dataset. As shown in Fig. 11, the GSA [6], BDSD-PC [9], MTF-GLP-CBD [11], AWLP-H [13], PWMBF [35] and FusionNet [26] preserve spatial details well, but they yield serious spectral distortion. The result of GTP-PNet [25] is blurry, such as the building areas. On the contrary, the fusion results of PanNet [23] and LPPN [27] are affected by severe aliasing artifacts. Our proposed FAN produces high-quality pansharpened result with abundant details, clear boundaries and high spectral fidelity. The average quantitative
results on 156 pairs of test data are reported in Table 3, which indicates that FAN obtains the best values in all three metrics.

The experimental results on the full-resolution data from WV-2 are shown in Fig. 12. From the enlarged red and green regions in Figs. 12(c)-(l), we can observe that the fused results of BDSD-PC [9], AWLP-H [13] and PWMBF [35] have different levels of spatial degradation, especially for the texture of objects, such as roads, trees and buildings, while the result of LPPN [27] suffers from serious checkerboard artifacts. The fusion results of GSA [9], MTF-GLP-CBD [11], PanNet [23], FusionNet [26], and GTP-PNet [25] have different degrees of spectral distortion. Compared with them, the proposed FAN has high fidelity in both the spectral and spatial domain. The quantitative evaluations are listed in Table 4 where the average indices of fusion results on 136 pairs of test data from the WV-2 dataset are demonstrated, from which we can conclude that FAN performs better on the eight-band WV-2 data compared with other methods.

![Fig. 10. Visual comparison of different methods on WorldView-4 (WV-4) dataset at full resolution.](image)

4.5. Ablation Studies

To verify the effectiveness of the DWT and IDWT layers as well as the proposed latent-space similarity (LSS) loss, we conduct several ablation studies on WV-4 dataset at both reduced- and full-resolution. Taking the proposed FAN in Fig. 2 as Baseline, the experimental results are listed in Table 5.
4.5.1. Effectiveness of DWT/IDWT layers

We replace the DWT and IDWT layers with normal Haar DWT and IDWT to evaluate the effects of DWT/IDWT layers. This setting is denoted as Variant 1. By comparing quantitative evaluations of Variant 1 with that of Baseline in Table 5, we find that the introduction of DWT and IDWT layers has positive impacts on all evaluation indexes.

4.5.2. Effectiveness of LSS loss

To investigate the effects of LSS loss, we neglect the LSS loss from the loss function. This setting is denoted as Variant 2. From the comparison between Variant 2 and Baseline, we can find that the evaluation indices are all degraded.
especially ERGAS and SAM, which concludes that the LSS loss is of great importance and has the ability of restricting the HF features of PAN as equivalent as possible to the missing HF parts of MS images.

It should be noted that the configurations of above variants are the same as Baseline, except the architectures. From above analyses, we can conclude that the DWT and the IDWT layers, together with the LSS loss are all essential for FAN, and each of them can bring specific improvements. By working together, the optimal fusion results both at reduced and full resolution are obtained.

### 4.5.3. Impacts of different $\beta$ values

In this subsection, we conduct a series of experiments on WV-4 dataset to investigate the impacts of $\beta$ on fusion results. Concretely, we change the values of $\beta$ from 10 to 50 and present the variation of average four indexes with respect to different combinations. The results are shown in Fig. 13, from which, we can find that the optimal result obtained when $\beta=20$. Therefore, we set $\beta=20$ when train FAN on WV-4 dataset. Similarly, the configuration of $\beta=10$ is adopted in the experiments on QB dataset, and $\beta=20$ is used for WV-2 dataset.

### 4.6. Computational Time

In this section, we compare the average computational time in seconds for different fusion methods at two scales, i.e. PAN sizes at both reduced resolution ($256 \times 256$) and full resolution ($1024 \times 1024$). The traditional methods are implemented by MATLAB R2018b on the computer with Intel(R) Core(TM) i5-1135G7 processor. For CNN-based methods, they are tested on the computer with an Nvidia GTX 1080Ti GPU and an Intel(R) Xeon(R) W-2123 CPU. Table
lists the average running time in seconds of different methods, from which, we can see that proposed FAN model is efficient when compared with most of state-of-the-art methods. This is partially due to the relatively simple architecture of FAN.

Table 6: Comparison Of Average Computational Time in Seconds on Different Size.

| PAN Size | GSA | BSD-PC | MTF-GLP-CBD | AWLP-H | PWMBF | PanNet | FusionNet | GTP-PNet | LPPN | FAN |
|----------|-----|--------|-------------|--------|-------|--------|-----------|----------|------|-----|
| 256 × 256 | 0.0545 | 0.0663 | 0.0988 | 0.0536 | 0.3400 | **0.0027** | 0.1686 | 0.2556 | 0.8215 | **0.0241** |
| 1024 × 1024 | 0.7708 | 0.4418 | 1.3236 | 0.3747 | 3.5229 | **0.0042** | 2.4387 | 3.9525 | 13.2468 | **0.0718** |

5. Conclusion

In this paper, we propose a frequency-aware network (FAN) along with a novel latent-space similarity (LSS) loss to learn the correspondence in frequency domain for multispectral pansharpening. In order to enable FAN to learn sufficient high-frequency information, DWT/IDWT layers are introduced, and FAN works directly in frequency domain to explicitly and adaptively extract and process high-frequency features. To reduce the spectral distortion, LSS loss is further designed to align high-frequency features of PAN and MS. After the constraint of LSS loss, the high-frequency features of PAN can appropriately
complement to that of MS. Extensive experiments on three datasets at both reduced and full resolution demonstrate the advantages of FAN, especially for the full-resolution experiments, which further proves that FAN has strong generalization ability in the real scenes. FAN is an attempt of learning frequency features by CNNs, and we believe that it is a trend for many low-level computer vision tasks.

References

[1] M. Vakalopoulou, K. Karantzalos, N. Komodakis, N. Paragios, Building detection in very high resolution multispectral data with deep learning features, in: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2015, pp. 1873–1876. doi:10.1109/IGARSS.2015.7326158

[2] M. Bouziani, K. Goita, D.-C. He, Rule-based classification of a very high resolution image in an urban environment using multispectral segmentation guided by cartographic data, IEEE Transactions on Geoscience and Remote Sensing 48 (8) (2010) 3198–3211. doi:10.1109/TGRS.2010.2044508

[3] M. Gong, T. Zhan, P. Zhang, Q. Miao, Superpixel-based difference representation learning for change detection in multispectral remote sensing images, IEEE Transactions on Geoscience and Remote Sensing 55 (5) (2017) 2658–2673. doi:10.1109/TGRS.2017.2650198

[4] H. Yin, Pscsc-net: A deep coupled convolutional sparse coding network for pansharpening, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–16. doi:10.1109/TGRS.2021.3088313

[5] X. Meng, Y. Xiong, F. Shao, H. Shen, W. Sun, G. Yang, Q. Yuan, R. Fu, H. Zhang, A large-scale benchmark data set for evaluating pansharpening performance: Overview and implementation, IEEE Geoscience and Remote Sensing Magazine 9 (1) (2021) 18–52. doi:10.1109/MGRS.2020.2976696
[6] B. Aiazzi, S. Baronti, M. Selva, Improving component substitution pancharpening through multivariate regression of ms + pan data, IEEE Transactions on Geoscience and Remote Sensing 45 (10) (2007) 3230–3239. doi:10.1109/TGRS.2007.901007

[7] J. Choi, K. Yu, Y. Kim, A new adaptive component-substitution-based satellite image fusion by using partial replacement, IEEE Transactions on Geoscience and Remote Sensing 49 (1) (2011) 295–309. doi:10.1109/TGRS.2010.2051674

[8] A. Garzelli, F. Nencini, L. Capobianco, Optimal mse pan sharpening of very high resolution multispectral images, IEEE Transactions on Geoscience and Remote Sensing 46 (1) (2008) 228–236. doi:10.1109/TGRS.2007.907604

[9] G. Vivone, Robust band-dependent spatial-detail approaches for panchromatic sharpening, IEEE Transactions on Geoscience and Remote Sensing 57 (9) (2019) 6421–6433. doi:10.1109/TGRS.2019.2906073

[10] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis, IEEE Transactions on Geoscience and Remote Sensing 40 (10) (2002) 2300–2312. doi:10.1109/TGRS.2002.803623

[11] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, L. M. Bruce, Comparison of pansharpening algorithms: Outcome of the 2006 grs-s data-fusion contest, IEEE Transactions on Geoscience and Remote Sensing 45 (10) (2007) 3012–3021. doi:10.1109/TGRS.2007.904923

[12] X. Otazu, M. Gonzalez-Audicana, O. Fors, J. Nunez, Introduction of sensor spectral response into image fusion methods. application to wavelet-based methods, IEEE Transactions on Geoscience and Remote Sensing 43 (10) (2005) 2376–2385. doi:10.1109/TGRS.2005.856106
[13] G. Vivone, L. Alparone, A. Garzelli, S. Lolli, Fast reproducible pansharpening based on instrument and acquisition modeling: Awlp revisited, Remote Sensing 11 (19) (2019) 2315.

[14] F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, A new pansharpening algorithm based on total variation, IEEE Geoscience and Remote Sensing Letters 11 (1) (2014) 318–322. doi:10.1109/LGRS.2013.2257669

[15] M. R. Vicinanza, R. Restaino, G. Vivone, M. Dalla Mura, J. Chanussot, A pansharpening method based on the sparse representation of injected details, IEEE Geoscience and Remote Sensing Letters 12 (1) (2015) 180–184. doi:10.1109/LGRS.2014.2331291

[16] P. Liu, L. Xiao, T. Li, A variational pan-sharpening method based on spatial fractional-order geometry and spectral–spatial low-rank priors, IEEE Transactions on Geoscience and Remote Sensing 56 (3) (2018) 1788–1802. doi:10.1109/TGRS.2017.2768386

[17] C. Ballester, V. Caselles, L. Igual, J. Verdera, B. Rougé, A variational model for p+ xs image fusion, International Journal of Computer Vision 69 (1) (2006) 43–58.

[18] X. Fu, Z. Lin, Y. Huang, X. Ding, A variational pan-sharpening with local gradient constraints, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10257–10266. doi:10.1109/CVPR.2019.01051

[19] G. Masi, D. Cozzolino, L. Verdoliva, G. Scarpa, Pansharpening by convolutional neural networks, Remote Sensing 8 (7) (2016) 594.

[20] C. Dong, C. C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (2) (2016) 295–307. doi:10.1109/TPAMI.2015.2439281
[21] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, A. Courville, On the spectral bias of neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 5301–5310.

[22] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, R. Ng, Fourier features let networks learn high frequency functions in low dimensional domains, Advances in Neural Information Processing Systems 33 (2020) 7537–7547.

[23] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, J. Paisley, Pannet: A deep network architecture for pan-sharpening, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1753–1761. doi:10.1109/ICCV.2017.193

[24] G. Scarpa, S. Vitale, D. Cozzolino, Target-adaptive cnn-based pansharpening, IEEE Transactions on Geoscience and Remote Sensing 56 (9) (2018) 5443–5457. doi:10.1109/TGRS.2018.2817393

[25] H. Zhang, J. Ma, Gtp-pnet: A residual learning network based on gradient transformation prior for pansharpening, ISPRS Journal of Photogrammetry and Remote Sensing 172 (2021) 223–239.

[26] L.-J. Deng, G. Vivone, C. Jin, J. Chanussot, Detail injection-based deep convolutional neural networks for pansharpening, IEEE Transactions on Geoscience and Remote Sensing 59 (8) (2021) 6995–7010. doi:10.1109/TGRS.2020.3031366

[27] C. Jin, L.-J. Deng, T.-Z. Huang, G. Vivone, Laplacian pyramid networks: A new approach for multispectral pansharpening Information Fusion 78 (2022) 158–170.

[28] X. Wu, T.-Z. Huang, L.-J. Deng, T.-J. Zhang, Dynamic cross feature fusion for remote sensing pansharpening, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14687–14696.
[29] R. Gal, D. C. Hochberg, A. Bermano, D. Cohen-Or, Swagan: A style-based wavelet-driven generative model, ACM Transactions on Graphics (TOG) 40 (4) (2021) 1–11.

[30] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[31] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[32] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8110–8119.

[33] Q. Li, L. Shen, S. Guo, Z. Lai, Wavelet integrated cnns for noise-robust image classification, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7243–7252. doi:10.1109/CVPR42600.2020.00727.

[34] Y. Liang, P. Zhang, Y. Mei, T. Wang, Pmacnet: Parallel multiscale attention constraint network for pan-sharpening, IEEE Geoscience and Remote Sensing Letters 19 (2022) 1–5.

[35] F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, J. A. Benediktsson, Model-based fusion of multi- and hyperspectral images using pca and wavelets, IEEE Transactions on Geoscience and Remote Sensing 53 (5) (2015) 2652–2663. doi:10.1109/TGRS.2014.2363477.

[36] R. H. Yuhas, A. F. Goetz, J. W. Boardman, Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm, in: JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop, 1992.
[37] L. Wald, Data fusion: definitions and architectures: fusion of images of different spatial resolutions, Presses des MINES, 2002.

[38] A. Garzelli, F. Nencini, Hypercomplex quality assessment of multi/hyperspectral images, IEEE Geoscience and Remote Sensing Letters 6 (4) (2009) 662–665.

[39] J. Zhou, D. L. Civco, J. Silander, A wavelet transform method to merge landsat tm and spot panchromatic data, International journal of remote sensing 19 (4) (1998) 743–757.

[40] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, M. Selva, Multispectral and panchromatic data fusion assessment without reference, Photogrammetric Engineering & Remote Sensing 74 (2) (2008) 193–200.

[41] L. Wald, T. Ranchin, M. Mangolini, Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images, Photogrammetric engineering and remote sensing 63 (6) (1997) 691–699.

[42] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, 2017.

[43] S. Mallat, Wavelets for a vision, Proceedings of the IEEE 84 (4) (1996) 604–614.

[44] S. Han, A. Srivastava, C. Hurwitz, P. Sattigeri, D. D. Cox, not-so-biggan: Generating high-fidelity images on a small compute budget (2020).