Learning High-frequency Feature Enhancement and Alignment for Pan-sharpening

Yingying Wang
Institute of Artificial Intelligence,
Xiamen University, Xiamen, China
wangyingying7@stu.xmu.edu.cn

Yunlong Lin
School of Informatics, Xiamen
University, Xiamen, China
lyl047853860@gmail.com

Ge Meng
School of Informatics, Xiamen
University, Xiamen, China
mengg@stu.xmu.edu.cn

Zhenqi Fu
School of Informatics, Xiamen
University, Xiamen, China
fuzhenqi@stu.xmu.edu.cn

Yuhang Dong
School of Informatics, Xiamen
University, Xiamen, China
dongyh@stu.xmu.edu.cn

Linyu Fan
School of Informatics, Xiamen
University, Xiamen, China
flyannie@stu.xmu.edu.cn

Hedeng Yu
School of Informatics, Xiamen
University, Xiamen, China
yuhedeng@stu.xmu.edu.cn

Xinghao Ding*
Institute of Artificial Intelligence,
Xiamen University, Xiamen, China
School of Informatics, Xiamen
University, Xiamen, China
dxh@xmu.edu.cn

Yue Huang
Institute of Artificial Intelligence,
Xiamen University, Xiamen, China
School of Informatics, Xiamen
University, Xiamen, China
yhuang2010@xmu.edu.cn

ABSTRACT

Pan-sharpening aims to utilize the high-resolution panchromatic (PAN) image as a guidance to super-resolve the spatial resolution of the low-resolution multispectral (MS) image. The key challenge in pan-sharpening is how to effectively and precisely inject high-frequency edges and textures from the PAN image into the low-resolution MS image. To address this issue, we propose a High-frequency Feature Enhancement and Alignment Network (HFEAN) for effectively encouraging the high-frequency learning. To implement it, three core designs are customized: a Fourier convolution based efficient feature enhancement module (FEM), an implicit neural alignment module (INA), and a preliminary alignment module (Pre-align). To be specific, FEM employs the fast Fourier convolution with attention mechanism to achieve the mixed global-local receptive field on each scale of the high-frequency domain, thus yielding the informative latent codes. INA leverages implicit neural function to precisely align the latent codes from different scales in the continuous domain. In this way, the high frequency signals at different scales are represented as functions of continuous coordinates, enabling a precise feature alignment in a resolution-free manner. Pre-align is developed to further address the inherent misalignment between PAN and MS pairs. Extensive experiments over multiple satellite datasets validate the effectiveness of the proposed network and demonstrate its favorable performance against the existing state-of-the-art methods both visually and quantitatively. Code is available at: https://github.com/Gracewangyy/HFEAN.

CCS CONCEPTS

• **Computing methodologies → Hyperspectral imaging**.

KEYWORDS

pan-sharpening, high-frequency, alignment, enhancement, implicit neural function, fast Fourier convolution

ACM Reference Format:
Yingying Wang, Yunlong Lin, Ge Meng, Zhenqi Fu, Yuhang Dong, Linyu Fan, Hedeng Yu, Xinghao Ding, and Yue Huang. 2023. Learning High-frequency Feature Enhancement and Alignment for Pan-sharpening. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611937

1 INTRODUCTION

Satellite images are widely used in military, environmental surveillance, and mapping systems. However, due to technological and physical limitations, it is challenging for imaging devices to capture both high-spatial resolution and multi-spectral images at the same time. To address this issue, pan-sharpening has emerged as an important technology that combines the complementary information between the high-spatial resolution (PAN) image and the multispectral (MS) image to generate the desirable high-spectral and high-spatial resolution (HRMS) image.

In the past few decades, a variety of classic pan-sharpening techniques have been proposed, which can be broadly categorized into three groups: component substitution (CS) [6, 16, 22], multi-resolution analysis (MRA) [28, 35], and variational optimization (VO) [2, 20]. Although traditional methods have produced satisfactory outcomes, they still encounter challenges in accurately

[*] Xinghao Ding is the corresponding author.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611937

restoring spatial and spectral details in HRMS images. This limitation can be attributed to their dependence on manually crafted features and inadequate modeling of prior knowledge. With the advancement of deep learning techniques, convolutional neural networks (CNNs) have been increasingly applied in pan-sharpening due to their highly nonlinear mapping capability [4, 39, 43]. Masi et al. [29] built the Pan-sharpening Neural Network (PNN) model based on a three-layer CNN, which achieved significantly better results in terms of spatial and spectral feature fidelity than traditional methods and became a benchmark for deep learning based pan-sharpening. Since then, a significant amount of complicated and deeper models have been proposed to enhance the mapping capability of pan-sharpening, including PanNet [44], LHFNet [53], NormNet [54], Panformer [51], MutNet [58], SFIIN [55] and so on.

Pan-sharpening aims to improve the spatial resolution of the MS image by leveraging the PAN image as a guidance. The key challenge is how to effectively and precisely inject the high-frequency edges and textures from the PAN image into the low-resolution MS image. Efficient feature extraction, enhancement, and precise alignment are crucial for achieving successful pan-sharpening. To accurately represent the high-frequency information, which includes edges and intricate details, both global and local features must be captured. However, most current pan-sharpening methods only rely on convolutional encoders with limited receptive fields for high-frequency feature extraction, neglecting to fully investigate the long-range dependencies of high-frequency features. Additionally, precise alignment is vital for seamlessly integrating high-frequency information into the MS image. Unfortunately, the resolution difference between PAN and MS image pairs can inherently cause misalignment. Moreover, the various localizations of high-frequency information extracted through multi-scale Gaussian kernels will further intensify the misalignment issue in pan-sharpening. Despite these challenges, many deep learning-based pan-sharpening methods solely rely on ground truth as a constraint to implicitly address these misalignment issues, without explicitly exploring the alignment function.

In this paper, we propose a novel pan-sharpening framework that explicitly addresses the misalignment issues and effectively extracts and enhances high-frequency features. Our network consists of three core designs: an efficient high-frequency feature extraction and enhancement module (FEM), an implicit neural alignment module (INA), and a preliminary alignment module (Pre-align). Specifically, FEM employs the fast Fourier convolution with attention mechanism to obtain the mixed global-local receptive field on the high-frequency information extracted via each scale Gaussian kernel. The interaction between global and local features can enhance the feature representation, leading to more accurate extraction of high-frequency information. This effective feature extraction and enhancement process enables the generation of informative latent codes. INA leverages implicit neural functions to precisely align latent codes across different scales in the continuous domain. By utilizing implicit neural representations, feature values can be queried at any resolution, enabling efficient and accurate aggregation of high-frequency features across scales. By representing high-frequency signals at different scales as functions of continuous coordinates, INA facilitates precise alignment of features in a resolution-free manner. Moreover, we introduce the Pre-align module to address the inherent resolution mismatch between PAN and MS modalities. This module is a trainable upsampling operator, which is more effective than traditional fixed bilinear or bicubic interpolations in resolving the misalignment issues.

In summary, the contributions of this work are as follows:

- We propose a novel pan-sharpening framework to explicitly address the misalignment issues in pan-sharpening. To the best of our knowledge, this is the first attempt to explore the precise alignment in pan-sharpening.
- We design a fast Fourier convolution based module with attention mechanism dedicated for pan-sharpening tasks. This module achieves the mixed global-local receptive field on each scale of the high-frequency domain, enhancing feature representation and resulting in more accurate extraction of high-frequency information.
- We introduce an implicit neural alignment module to precisely align multi-scale high-frequency features in the continuous domain via implicit neural representation.
- To further address the inherent misalignment between PAN and MS pairs, we develop a preliminary alignment module in a simple yet effective manner.

2 RELATED WORK

2.1 Traditional Pan-sharpening

Traditional pan-sharpening methods can be mainly classified into three categories: CS [6, 16, 22], MRA [28, 35], and VO [2, 20]. The CS approach was first proposed by Chavez [22] to improve the spatial resolution by projecting spatial information from MS images into a transform domain and replacing it with corresponding information from PAN images. Although this technique yielded satisfactory spatial feature fidelity, it still caused some loss of spectral feature information. To mitigate the issue of spectral distortion, MRA employed multi-resolution decomposition techniques like Laplacian pyramid [35] and decimated wavelet transform [28] to extract spatial information from the PAN images and incorporate it into the upsampled MS images. This method reduced spectral distortion compared to the CS approach, but may resulted in higher distortion of the spatial features. The VO-based method utilized prior knowledge to reasonably constrain the model and achieved final panchromatic sharpening results through an efficient algorithm with typical schemes including P+XS [2] and PHLP [20]. Although the VO method provided superior spatial and spectral feature fidelity, its computational speed was significantly reduced. In general, traditional pan-sharpening methods heavily relied on manually designed features [8, 15, 34] , which often led to degradation in results due to the lack of sufficient priors [14].

2.2 Deep learning based methods

Due to the powerful capabilities of deep neural networks (DNN) in nonlinear fitting and feature extraction, an increasing number of DNN-based methods have been applied in various areas, such as image restoration [9–11, 45–47] and pan-sharpening [17, 41, 52, 56, 57, 59]. Masi et al. [29] first introduced a simple three-layer CNN structure called PNN for multi-spectral pan-sharpening learning. This model achieved excellent fusion results and became a benchmark method for deep learning-based pan-sharpening tasks.

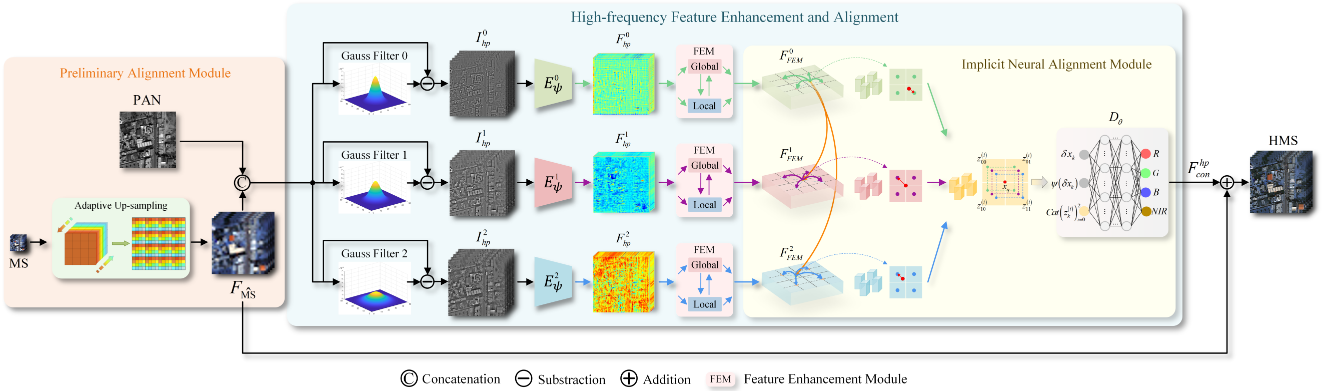

Figure 1: The overall framework of the proposed network, which consists of three main modules: 1) Preliminary Alignment Module (Pre-align). 2) Feature Enhancement Module (FEM). 3) Implicit Neural Alignment Module (INA).

Yuan et al. [48] put forward MSDCNN, which addressed the multi-scale problem by incorporating multi-scale modules into the network to boost its performance. Li et al. [24] introduced SIPSA-Net, an alignment-aware network that was the first method to tackle the challenge of significant misalignment in moving object regions in pan-sharpening. To facilitate joint feature learning across the PAN and MS modalities, Zhou et al. [51] proposed a custom transformer architecture and an information-lossless invertible neural module, enabling the modeling of long-range dependencies and effective feature fusion in pan-sharpening. Additionally, Zhou et al. [54] introduced a normalization-based feature selection and restitution mechanism to accurately select consistent features and propagate them, while effectively handling inconsistent ones between PAN and MS modalities. This mechanism can filter out inconsistent features while promoting the learning of consistent ones. To explore the potential solution of pan-sharpening in both spatial and frequency domains, Zhou et al. [55] introduced the Spatial-Frequency Information Integration Network (SFIIN) as a means of integrating global and local information from the two modalities of PAN and MS images, which significantly improved the performance.

2.3 Implicit Neural Representation

Implicit Neural Representation (INR) is an innovative research field that challenges the traditional approach of discretely representing signals. Instead of relying on discrete grids of pixels for images, or using voxels, point clouds and meshes for representing 3D shapes [26, 27, 42], INR employs continuous functions to map the input domain of the signal (such as pixel coordinates in an image) to a representation of color, occupancy, or density at that precise input location. Chan et al. [7] utilized a neural representation with periodic activation functions and volume rendering to depict the scene as a radial field that aligned with the viewpoint. Park et al. introduced DeepSDF [31] for learning a set of continuous signed distance functions to represent shapes. Mildenhall et al. put forward NeRF [30], a versatile approach for synthesizing new perspectives of complex scenes. Chen et al. [12] was motivated by the achievements of INR in 3D reconstruction and proposed to integrate local latent codes into INR as a means of restoring intricate details in natural and complex images for super-resolution tasks. Besides, INR was also considered as a promising approach for achieving accurate feature alignment in image segmentation [18, 33]. Motivated by this, we employ the implicit neural representation to precisely align multi-scale high-frequency features in the continuous domain.

3 METHODS

The overall framework is clearly presented in Figure 1, which consists of three modules: 1) Preliminary Alignment Module (Pre-align). 2) Feature Enhancement Module (FEM). 3) Implicit Neural Alignment Module (INA). The details are illustrated as below.

3.1 Preliminary Alignment Module (Pre-align)

Pan-sharpening aims to fuse the complementary information between the MS image ($MS \in R^{H/r \times W/r \times C}$) and the PAN image ($P \in R^{H \times W \times 1}$) to generate the desirable high spatial resolution MS image ($HMS \in R^{H \times W \times C}$). H and W represent the height and width of the image, C refers to spectral bands, and the ratio r is equal to 4. One of the main difficulties in pan-sharpening is the inherent misalignment between PAN and MS image pairs, which often leads to artifacts and blurring in output HRMS images. Existing approaches usually upscale the input MS image via bilinear or bicubic interpolation before fusing it with the corresponding PAN image. However, this kind of fixed interpolations are insufficient for effectively resolving the misalignment problem. To address this issue, we propose a preliminary alignment module to ensure the alignment between the PAN and MS pairs. Specifically, given the MS image ($MS \in R^{H/r \times W/r \times C}$) , our method firstly applies a 3×3 convolution to obtain the MS shallow feature $F_{MS} \in R^{H/r \times W/r \times C \cdot r^2}$

$$F_{MS} = Conv_{3\times3}(MS). \tag{1}$$

Then we employ a trainable periodic shuffling operator $\mathcal{PS}$ to convert the shallow feature F_{MS} into the upscaled feature $F_{\widehat{MS}} \in R^{H \times W \times C}$. This operator enables the generation of adaptively upsampled $F_{\widehat{MS}}$ according to the PAN image

$$F_{\widehat{MS}} = \mathcal{PS}\left(F_{MS}\right). \tag{2}$$

3.2 High-frequency Feature Enhancement and Alignment

Multi-scale high-frequency feature extraction. To extract the high-frequency information, we employ a low-pass Gaussian filter, which is subsequently subtracted from the input signal to yield the high-frequency component, the Gaussian filter is denoted as below

$$\mathcal{G}(x, y) = \frac{1}{2\pi\sigma^2} \mathrm{e}^{-\frac{\mathrm{x}^2+\mathrm{y}^2}{2\sigma^2}}, \tag{3}$$

where (x, y) is the 2D image-level coordinates.

To obtain accurate high-frequency information, relying on a single-scale filter is insufficient. A small kernel size Gaussian filter is suitable for capturing finer details and textures, while a larger kernel size Gaussian filter is more effective in preserving the overall structure and producing a smoother output. Therefore, we adopt multi-scale high-frequency feature extraction approach to capture the complementary high-frequency information. After the preliminary alignment, the adaptively upsampled MS feature $F_{\widehat{MS}}$ and the PAN image are concatenated as ($I_{cat} \in R^{H\times W\times(C+1)}$) and then subtract the results from three different scale Gaussian filters $\mathcal{G}(x, y)$ with kernel sizes of 5, 27 and 41, respectively, to derive the high-frequency component $I_{hp}^{(i)}$

$$I_{hp}^{(i)} = I_{cat}^{(i)} - \mathcal{G}\left(x_{I_{cat}^{(i)}}, y_{I_{cat}^{(i)}}\right)_{i\in\{0,1,2\}}, \tag{4}$$

where i is the scale index of the Gaussian filter.

Following, a gated convolution encoder Hornet [32] is utilized to facilitate efficient feature interaction and fusion. Suppose that the joint high frequency component of PAN and MS are denoted as $I_{hp}^{(i)} \in R^{H\times W\times(C+1)}$, the expression can be simplified as follows

$$F_{hp}^{(i)} = E_{\psi}^{(i)}\left(I_{hp}^{(i)}\right), \tag{5}$$

where $E_{\psi}^{(i)}$ denotes the encoder, $F_{hp}^{(i)} \in R^{H\times W\times C}$ denotes the output high-frequency feature on each scale.

Feature Enhancement Module (FEM). The spectral convolution theorem in Fourier theory reveals that updating a point in the spectral domain has a global impact on all input features involved in the Fourier transform. This insight can be leveraged to achieve the global receptive field. Inspired by FFC [13], we design a fast Fourier convolution module with attention mechanism dedicated for pan-sharpening tasks.

The architecture of the FEM module is presented in Figure 2. To enhance the localization of high-frequency features and improve the feature extraction, we integrate RCAB attention mechanism [50] at both the beginning and the end of the FEM module. Subsequently, we split the high-frequency features $F_{hp}^{(i)} \in R^{H\times W\times C}$ along the dimension of feature channels to generate the global and local features respectively, with split ratio of 0.5. The global part $F_g^{(i)} \in R^{H\times W\times\frac{C}{2}}$ is designed to capture long-range context, while the local part $F_l^{(i)} \in R^{H\times W\times\frac{C}{2}}$ is expected to learn the local details

$$\begin{aligned} F_{hp\_rcab}^{(i)} &= RCAB\left(F_{hp}^{(i)}\right), \\ F_g^{(i)}, F_l^{(i)} &= \mathrm{split}\left(F_{hp\_rcab}^{(i)}\right). \end{aligned} \tag{6}$$

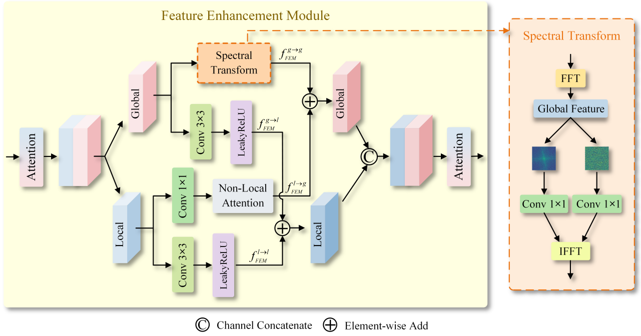

Figure 2: The architecture of the Feature Enhancement Module (FEM), with global-global, global-local, local-global and local-local four branches.

To effectively extract multimodal PAN and MS high-frequency features, we simultaneously leverage spatial and spectral information to achieve mixed receptive fields. The FEM module consists of four branches. The local-to-local branch captures small-scale information using a 3 × 3 convolution operation, the local-to-global branch leverages a non-local attention mechanism [5], the global-to-local branch employs a 3 × 3 convolution operation, and the global-to-global branch utilizes spectral transform. The procedures can be described as follows

$$\begin{aligned} f_{FEM}^{g\to g}\left(F_g^{(i)}\right) &= \mathcal{ST}\left(F_g^{(i)}\right), \\ f_{FEM}^{g\to l}\left(F_g^{(i)}\right) &= \mathrm{Conv}_{3\times3}^{g\to l}\left(F_g^{(i)}\right), \\ f_{FEM}^{l\to g}\left(F_l^{(i)}\right) &= \mathcal{NL}\left(\mathrm{Conv}_{1\times1}^{l\to g}\left(F_l^{(i)}\right)\right), \\ f_{FEM}^{l\to l}\left(F_l^{(i)}\right) &= \mathrm{Conv}_{3\times3}^{l\to l}\left(F_l^{(i)}\right), \end{aligned} \tag{7}$$

where $\mathcal{NL}$ and $\mathcal{ST}$ denote the non-local attention and the spectral transform respectively.

The goal of the **spectral transform** $\mathcal{ST}$ is to efficiently enlarge the receptive field of the convolution to the full resolution of the input feature maps. We first adopt 2D Fast Fourier Transform (FFT) to transform global spatial features into spectral domain

$$\mathcal{A}\left(F_g^{(i)}\right), \mathcal{P}\left(F_g^{(i)}\right) = \mathcal{F}\left(F_g^{(i)}\right), \tag{8}$$

where $\mathcal{A}(\cdot)$ and $\mathcal{P}(\cdot)$ indicate the amplitude and phase. Then, we use two groups of operations, $\mathcal{OA}(\cdot)$ and $\mathcal{OP}(\cdot)$, which consist of 1 × 1 convolution and ReLU activation function, respectively. These operations are applied to the corresponding amplitude and phase components, providing the enhanced global feature representation

$$\begin{aligned} \mathcal{A}(F_g^{(i)}) &= \mathcal{OA}\left(\mathcal{A}\left(F_g^{(i)}\right)\right), \\ \mathcal{P}(F_g^{(i)}) &= \mathcal{OP}\left(\mathcal{P}\left(F_g^{(i)}\right)\right). \end{aligned} \tag{9}$$

Next, we apply the inverse DFT to transform the amplitude and phase components of $\mathcal{A}(F_g)$ and $\mathcal{P}(F_g)$ back to the spatial domain

$$F_{g\_st}^{(i)} = F^{-1}(\mathcal{A}(F_g^{(i)}), \mathcal{P}(F_g^{(i)})). \tag{10}$$

By summing up the global-to-global and local-to-global branches, we obtain the global feature $Y_g^{(i)} \in R^{H\times W\times\frac{C}{2}}$. Similarly, we can also get the local feature $Y_l^{(i)} \in R^{H \times W \times \frac{C}{2}}$

$$\begin{aligned} Y_g^{(i)} &= f_{FEM}^{l \to g}\left(F_l^{(i)}\right) + f_{FEM}^{g \to g}\left(F_g^{(i)}\right), \\ &= f_{FEM}^{l \to g}\left(F_l^{(i)}\right) + F_{g-st}^{(i)}, \end{aligned} \tag{11}$$

$$Y_l^{(i)} = f_{FEM}^{l \to l}\left(F_l^{(i)}\right) + f_{FEM}^{g \to l}\left(F_g^{(i)}\right). \tag{12}$$

Afterwards, the global and local features are concatenated to produce the enhanced high-frequency feature $F_{FEM}^{(i)}$

$$F_{FEM}^{(i)} = \mathrm{Cat}\left(Y_g^{(i)}, Y_l^{(i)}\right). \tag{13}$$

The FEM module is able to obtain the mixed global-local receptive field on each scale of the high-frequency domain, enhancing feature representation and resulting in more accurate feature extraction. For better feature enhancement, two FEM modules are implemented sequentially on each scale to generate the final features.

Implicit Neural Alignment Module (INA). To precisely inject the high-frequency features from the PAN image into the low-resolution MS image, accurate localization is essential. One of the main challenges in aggregating multi-scale high-frequency features arises from their different localization due to various Gaussian kernel sizes. A smaller kernel size is preferable for capturing finer details with relatively precise localization, while a larger kernel size prioritizes the overall structure and produces a smoother output with relatively less precise localization.

Implicit neural function defines a decoding function $D_\theta^{(i)}$ (typically an MLP) over a discrete feature map to obtain the continuous feature map. We can utilize the implicit neural function to precisely align multi-scale high-frequency features at a unified resolution. For a certain scale, given the discrete feature map $F_{FEM}^{(i)}$, the high-frequency feature vectors can be viewed as latent codes $z^{(i)}$ evenly distributed on $F_{FEM}^{(i)}$, each of the latent code is assigned with a 2D coordinate. The high-frequency feature value at arbitrary query coordinate $x_q^{(i)}$ on the continuous feature map $F_{con}^{(i)}$ is defined by

$$F_{con}^{(i)}\left(x_q^{(i)}\right) = D_\theta^{(i)}\left(z^{(i)}, x_q^{(i)} - x^{(i)}\right), \tag{14}$$

where $z^{(i)}$ is the nearest latent code from $x_q^{(i)}$, $x^{(i)}$ is the coordinate of latent code $z^{(i)}$, i is the multi-scale index, $D_\theta^{(i)}(\cdot)$ is the decoding function. In practice, $D_\theta^{(i)}$ is jointly trained with the feature encoder $E_\psi^{(i)}$ and the FEM module, so that the features are learned to precisely represent continuous fields of information.

Neural networks tend to prioritize learning low-frequency signals during training, while displaying inferior performance in representing high-frequency signals. The learning power of neural networks gets limited when directly operated on xy coordinates. To overcome this limitation, we propose to encode the coordinates with a position encoding function before feeding them into the network. Formally, the position encoding function is as below

$$\begin{aligned} \psi(x) = (&\sin(\omega_1 x), \cos(\omega_1 x), \ldots, \\ &\sin(\omega_L x), \cos(\omega_L x)). \end{aligned} \tag{15}$$

During the training process, the frequency values ω_l are initialized as $\omega_l = 2e^l$ for $l \in \{1, \ldots, L\}$ and can be further fine-tuned. The 2D coordinates will be expanded into an encoding with 2L dimensions via position encoding. The implicit feature function is

$$F_{con}^{(i)}\left(x_q^{(i)}\right) = D_\theta^{(i)}\left(z^{(i)}, \psi\left(x_q^{(i)} - x^{(i)}\right), x_q^{(i)} - x^{(i)}\right). \tag{16}$$

Then, we employ the neighboring four latent codes to predict the signal on certain coordinate in the continuous domain

$$\begin{aligned} F_{con}^{(i)}\left(x_q^{(i)}\right) = \sum_{k \in \{00,01,10,11\}} \frac{w_k^{(i)}}{w^{(i)}} D_\theta^{(i)}\left(z_k^{(i)}, \psi\left(x_q^{(i)} - x_k^{(i)}\right),\right. \\ \left. x_q^{(i)} - x_k^{(i)}\right), \end{aligned} \tag{17}$$

where $z_k^{(i)}$ $(k \in \{00, 01, 10, 11\})$ is the nearest latent code of $x_q^{(i)}$ in top-left, top-right, bottom-left, bottom-right sub-spaces, $x_k^{(i)}$ is the coordinate of $z_k^{(i)}$, $\psi(\cdot)$ is the position encoding function, $w_k^{(i)}$ is the weight of sub-space k (the area value of the box between $x_k^{(i)}$ and $x_q^{(i)}$), $w^{(i)}$ is the aggregation weight (neighboring total area).

To accurately align multi-scale high-frequency features, one promising approach is to use implicit neural function to convert discrete feature maps from different scales into a continuous one. This method enables feature retrieval at arbitrary coordinates, facilitating precise alignment. Intuitively, each latent code represents a field of features, D_θ can decode features from different scales and align them in the continuous domain. The concatenated latent codes, relative coordinates and position encoding results are fed into an MLP to generate the continuous high-frequency features

$$\begin{aligned} F_{con}^{hp}\left(x_q\right) = \sum_{k \in \{00,01,10,11\}} \frac{w_k}{w} D_\theta\left(Cat(z_k^{(i)})_{i=0}^2, \psi(\delta x_k), \delta x_k\right), \\ \delta x_k = x_q - x_k, \end{aligned} \tag{18}$$

where $Cat(z_k^{(i)})_{i=0}^2$ is the concatenated latent codes from different scales, i denotes the scale index. By concatenating multi-scale high-frequency latent codes $z_k^{(i)}$, a unified coordinate is defined for query. δx_k represents the relative coordinate between x_q and x_k in the unified coordinates, $\psi(\delta x_k)$ is the position encoding of δx_k. The weight of the sub-space k is defined by w_k, and the aggregation weight is w. The aligned multi-scale high-frequency features in the continuous domain is denoted as F_{con}^{hp}.

Once we obtain the high-frequency feature F_{con}^{hp} in the continuous domain, we can add it with the adaptively upsampled feature $F_{\widehat{MS}}$ to produce the final HRMS result

$$HMS = F_{con}^{hp} + F_{\widehat{MS}}. \tag{19}$$

3.3 Loss Function

In order to achieve the satisfying pan-sharpening results, we propose a joint loss to supervise the network training. Let HMS and GT denote the entire network output and the ground truth, respectively. We first adopt the $\mathcal{L}_1$ loss

$$\mathcal{L}_1 = \|HMS - GT\|_1. \tag{20}$$

To further improve the quality of the fusion results, we employ the perceptual loss to enhance the semantic similarity between the HMS and GT. The perceptual loss [21] evaluates the distance

Table 1: Quantitative comparison. Best results are highlighted by red. ↑ indicates that the larger the value, the better the performance, and ↓ indicates that the smaller the value, the better the performance.

Method	WorldView-II				GaoFen2				Worldview-III			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
Brovey [16]	35.8646	0.9216	0.0403	1.8238	37.7974	0.9026	0.0218	1.3720	22.5060	0.5466	0.1159	8.2331
GFPCA [25]	34.5581	0.9038	0.0488	2.1411	37.9443	0.9204	0.0314	1.5604	22.3344	0.4826	0.1294	8.3964
GS [23]	35.6376	0.9176	0.0423	1.8774	37.2260	0.9034	0.0309	1.6736	22.5608	0.5470	0.1217	8.2433
IHS [6]	35.2962	0.9027	0.0461	2.0278	38.1754	0.9100	0.0243	1.5336	22.5579	0.5354	0.1266	8.3616
PNN [29]	40.7550	0.9624	0.0259	1.0646	43.1208	0.9704	0.0172	0.8528	29.9418	0.9121	0.0824	3.3206
PanNet [44]	40.8176	0.9626	0.0257	1.0557	43.0659	0.9685	0.0178	0.8577	29.6840	0.9072	0.0851	3.4263
MSDCNN [48]	41.3355	0.9664	0.0242	0.9940	45.6847	0.9827	0.0135	0.6389	30.3038	0.9184	0.0782	3.1884
SRPPNN [3]	41.4538	0.9679	0.0233	0.9899	47.1998	0.9877	0.0106	0.5586	30.4346	0.9202	0.0770	3.1553
GPPNN [40]	41.1622	0.9684	0.0244	1.0315	44.2145	0.9815	0.0137	0.7361	30.1785	0.9175	0.0776	3.2593
SFIIN [55]	41.7244	0.9725	0.0220	0.9506	47.4712	0.9901	0.0102	0.5462	30.5971	0.9236	0.0741	3.0798
Ours	42.2275	0.9732	0.0212	0.9012	47.6184	0.9902	0.0100	0.5274	30.6768	0.9246	0.0741	3.0402

Table 2: Evaluation on the real-world full-resolution scenes from GaoFen2 dataset. The best values are highlighted by red. The up or down arrow indicates higher or lower metric corresponding to better images.

Metrics	Brovey	GFPCA	GS	IHS	PNN	PanNet	MSDCNN	SRPPNN	GPPNN	SFIIN	Ours
D_λ↓	0.1378	0.0914	0.0696	0.0770	0.0746	0.0737	0.0734	0.0767	0.0782	0.0685	0.0673
D_S↓	0.2605	0.1635	0.2456	0.2985	0.1164	0.1224	0.1151	0.1162	0.1253	0.1121	0.1110
QNR↑	0.6390	0.7615	0.7025	0.6485	0.8191	0.8143	0.8251	0.8173	0.8073	0.8463	0.8472

between features extracted from both the predicted and target images using a pre-trained VGG19 network. The perceptual loss is more flexible than other loss functions in that it does not require an exact reconstruction, thus allowing for variations in the generated image. The perceptual loss is computed by

$$\mathcal{L}_{\text{percep}} = \sum_i \frac{1}{C_i H_i W_i} \left\| \phi_i\left(HMS\right) - \phi_i\left(GT\right) \right\|_2^2, \tag{21}$$

where $\phi_i(\cdot)$ represents a VGG19 network, C_i, H_i, and W_i denote the number of channels, height, and width of the feature maps respectively at the i^{th} feature layer.

Finally, the overall loss function is formulated as follows

$$\mathcal{L} = \mathcal{L}_1 + \lambda \mathcal{L}_{\text{percep}}, \tag{22}$$

where λ is a weight factor and is empirically set to 0.1.

4 EXPERIMENTS

4.1 Baseline methods

We demonstrate the effectiveness of our method by comparing its performance with several traditional and deep learning-based pansharpening approaches. Specifically, we select four traditional pansharpening methods, including Brovey [16], GFPCA [25], GS [23], IHS [6], and six deep learning-based methods, PNN [29], PANNET [44], MSDCNN [48], SRPPNN [3], GPPNN [40], and SFIIN [55].

4.2 Datasets and Implementation details

Datasets Due to the lack of sufficient ground-truth images in pansharpening, we employ the Wald protocol tool [37] to generate the training set. Given an origin high-resolution MS image $HMS \in R^{H\times W\times C}$ and its corresponding PAN image $P_{org} \in R^{rH\times rW\times 1}$, both of them are downsampled with ratio r to obtain image pairs $MS \in R^{\frac{H}{r}\times\frac{W}{r}\times C}$ and $P \in R^{H\times W\times 1}$, where r is equal to 4. Within the training set, MS and P are treated as inputs, while HMS serves as the ground truth. The PAN images are cropped into patches with the size of 128 × 128, while the MS patches are cropped as 32 × 32. For this study, three satellite image datasets, namely Worldview-II (WV2), Worldview-III (WV3), and GaoFen2 (GF2) are examined to evaluate the performance of our proposed approach.

Implementation details. In our experiments, we utilize a single NVIDIA GeForce GTX 3090 GPU running on a personal computer, and construct our networks in Python by Pytorch framework. During the training phase, the Adam optimizer is used to update the network parameters for 1000 epochs with a batch size of 4. The learning rate is initialized with 8×10^{-4}. A StepLR learning rate adjustment strategy is employed to reduce the learning rate by half after every 200 iterations.

Evaluation Metrics. To evaluate the performance, we adopt the following image quality assessment (IQA) metrics: peak signal-to-noise ratio (PSNR) [19], structural similarity index (SSIM) [38], spectral angle mapper (SAM) [49], relative dimensionless global

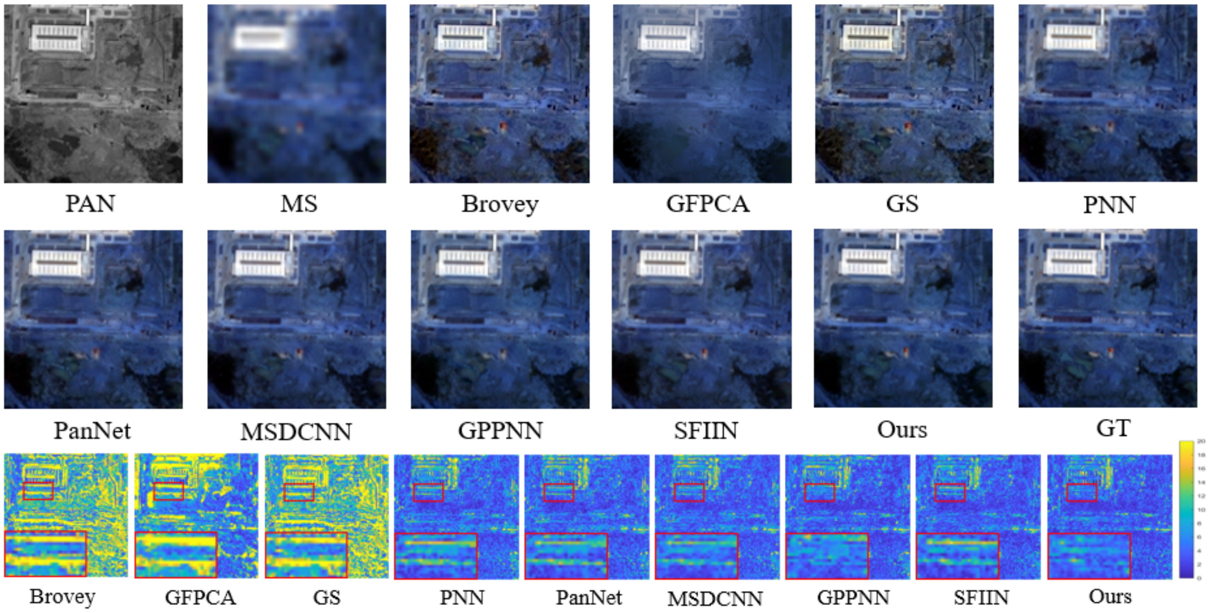

Figure 3: Visual comparison of all methods on WorldView-II. The last row visualizes the MSE residues between the pan-sharpening results and the ground truth.

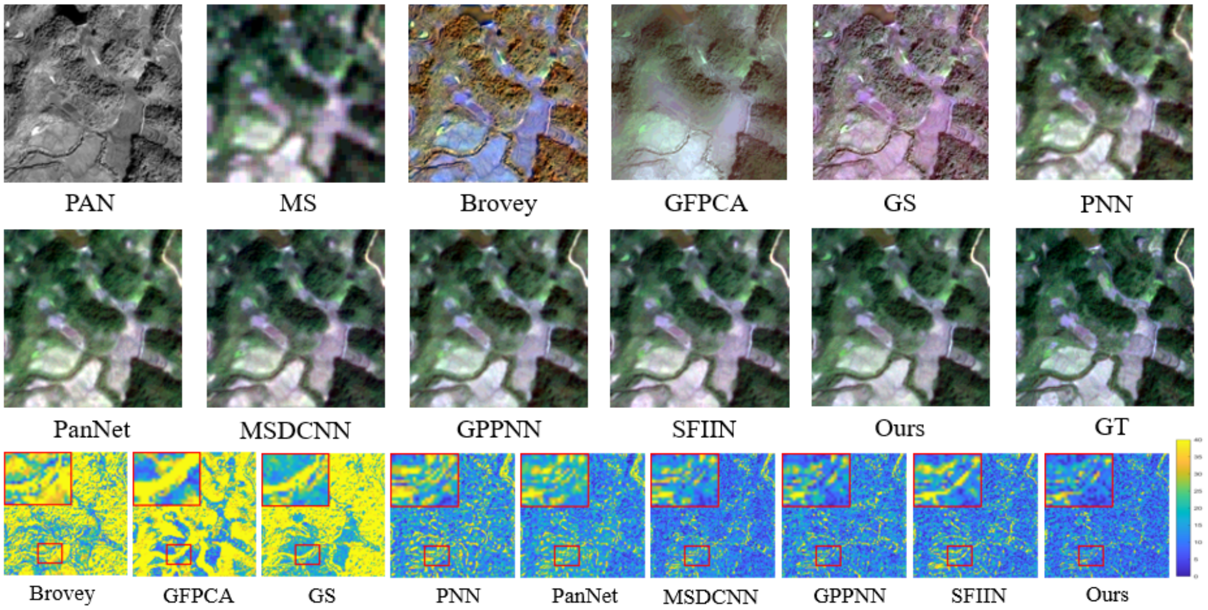

Figure 4: Visual comparison of all methods on GaoFen2. The last row visualizes the MSE residues between the pan-sharpening results and the ground truth.

error in synthesis (ERGAS) [36], spectral distortion index D_λ, spatial distortion index D_S and the quality with no reference (QNR) [1].

4.3 Comparison with state-of-the-art methods

Evaluation on reduced-resolution scene. The overview of the assessment metrics over three datasets is presented in Table 1, where the best results are highlighted in red for better visibility. From the table, it is evident that our method outperforms previous algorithms on three satellite datasets across all assessment metrics, with 0.50 dB, 0.15 dB and 0.08 dB improvements in PSNR compared to the second-best results obtained by other methods. Additionally, we also provide the visual results shown in Figures 3 and 4 for

Table 3: Evaluating generalization performance of the pre-trained model from WorldView-III dataset to the new satellite WorldView-II.

Method	WorldView-II		
	D_λ ↓	D_S ↓	QNR ↑
PNN [29]	0.1186	0.1228	0.7741
PanNet [44]	0.1090	0.1227	0.7826
MSDCNN [48]	0.1329	0.1228	0.7621
SRPPNN [3]	0.1371	0.1273	0.7543
GPPNN [40]	0.1193	0.1195	0.7764
SFIIN [55]	0.1209	0.1196	0.7751
Ours	0.0892	0.1052	0.8160

WorldView-II and GaoFen2 datasets, respectively. In the last row, we present the Mean Squared Error (MSE) residues between the pan-sharpened results and the ground truth. Our proposed strategy yields more accurate results than existing pan-sharpening techniques in terms of MSE residues, resulting in superior performance.

Evaluation on full-resolution scene. To assess the performance and generalization ability of our network on full-resolution scenes, we apply a pre-trained model trained on GaoFen2 data to unseen full-resolution GaoFen2 satellite datasets. For evaluation purposes, we utilize an additional real-world dataset of 200 samples captured over the newly selected GaoFen2 satellite. The results are summarized in Table 2. It is apparent that our proposed method outperforms other traditional and deep learning-based methods across all performance metrics.

Generalization to the new satellite. Training the network in the high-frequency domain, rather than the commonly used image domain, can enhance the network's ability to generalize well to other satellites without the need for retraining [44]. To demonstrate this, we employ a pre-trained model from WorldView-III data and directly tested it on WorldView-II, as shown in Table 3. The results clearly indicate that our model exhibits significantly better generalization ability to the new satellite WorldView-II.

4.4 Ablation experiments

We also conduct ablation studies using the WorldView-III satellite dataset to analyze the impact of the developed modules. Specifically, we investigated the effectiveness of three key components: INA, FEM and Pre-align. We evaluated all experimental data via PSNR [19], SSIM [38], SAM [49] and ERGAS [36] metrics.

Three key components. To assess the impact of three key components, we follow a stepwise approach. Initially, we introduce the INA module to achieve precise alignment of multi-scale high-frequency features. Next, we focus on the FEM module with or without attention mechanism, examining its effectiveness. Finally, we add the Pre-align module to address the inherent misalignment between the PAN and MS modalities to further improve the performance. Table 4 illustrates the relative importance of each module.

FEM module with and without attention mechanism. The attention mechanism can enhance the localization of high-frequency features and improve the feature extraction of the FEM module. We further evaluate the impact of the attention mechanism in Table 5. As seen from the table, the attention mechanism can significantly boost the performance of the FEM module.

Table 4: Ablation study on three key components.

Config	INA	FEM	Pre-align	PSNR↑	SSIM↑	SAM↓	ERGAS↓
(I)	✗	✗	✗	30.0877	0.9160	0.0816	3.2635
(II)	✓	✗	✗	30.2862	0.9192	0.0783	3.1864
(III)	✓	✓	✗	30.4111	0.9206	0.0773	3.1360
Ours	✓	✓	✓	30.6768	0.9246	0.0741	3.0402

Table 5: Ablation study on the FEM module with and without attention mechanism.

Config	FEM module	PSNR↑	SSIM↑	SAM↓	ERGAS↓
(I)	w/o attention	30.4528	0.9207	0.0768	3.1288
Ours	w/ attention	30.6768	0.9246	0.0741	3.0402

Table 6: Ablation study on single-scale and multi-scale feature extraction approaches.

Config	Feature extraction	PSNR↑	SSIM↑	SAM↓	ERGAS↓
(I)	Kernel size = 5	30.4486	0.9203	0.0770	3.1303
(II)	Kernel size = 27	30.4812	0.9210	0.0766	3.1250
(III)	Kernel size = 41	30.5243	0.9215	0.0756	3.1032
Ours	Multi-scale	30.6768	0.9246	0.0741	3.0402

Single-scale and multi-scale feature extraction. To effectively extract complementary high-frequency features, we adopt multi-scale feature extraction approach. Table 6 compares the high-frequency feature extraction using single-scale and multi-scale feature extraction strategy, with single-scale Gaussian kernel size in 5, 27 and 41, respectively. It is evident that the multi-scale method can significantly improve the performance.

5 CONCLUSION

In this paper, we propose a novel pan-sharpening method to explicitly address the misalignment issues and precisely inject the high-frequency features into MS images. Our core idea is to utilize the mixed global-local receptive field to enhance the representation of high-frequency features and achieve accurate feature extraction. These enhanced features are then precisely aligned in the continuous domain via implicit neural representation. Extensive experiments over different satellite datasets demonstrate the effectiveness of our proposed method.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants (No.82172033, No.U19B2031, No.61971369, No.52105126).

REFERENCES

[1] Luciano Alparone, Bruno Aiazzi, Stefano Baronti, Andrea Garzelli, Filippo Nencini, and Massimo Selva. 2008. Multispectral and panchromatic data fusion assessment without reference. *Photogrammetric Engineering & Remote Sensing* 74, 2 (2008), 193–200.

[2] Coloma Ballester, Vicent Caselles, Laura Igual, Joan Verdera, and Bernard Rougé. 2006. A variational model for P+ XS image fusion. *International Journal of Computer Vision* 69, 1 (2006), 43.

[3] Jiajun Cai and Bo Huang. 2020. Super-Resolution-Guided Progressive Pansharpening Based on a Deep Convolutional Neural Network. *IEEE Transactions on Geoscience and Remote Sensing* 59, 6 (Aug 2020), 5206–5220. https://doi.org/10.1109/tgrs.2020.3015878

[4] Xiangyong Cao, Xueyang Fu, Danfeng Hong, Zongben Xu, and Deyu Meng. 2021. PanCSC-Net: A model-driven deep unfolding method for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–13.

[5] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. 2019. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops.* 0–0.

[6] Wjoseph Carper, Thomasm Lillesand, Ralphw Kiefer, et al. 1990. The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data. *Photogrammetric Engineering and remote sensing* 56, 4 (1990), 459–467.

[7] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 5799–5809.

[8] Chen Chen, Yeqing Li, Wei Liu, and Junzhou Huang. 2015. SIRF: Simultaneous satellite image registration and fusion in a unified framework. *IEEE Transactions on Image Processing* 24, 11 (2015), 4213–4224.

[9] Sixiang Chen, Tian Ye, Yun Liu, Erkang Chen, Jun Shi, and Jingchun Zhou. 2022. Snowformer: Scale-aware transformer via context interaction for single image desnowing. *arXiv preprint arXiv:2208.09703* (2022).

[10] Sixiang Chen, Tian Ye, Yun Liu, Taodong Liao, Yi Ye, and Erkang Chen. 2022. MSP-Former: Multi-Scale Projection Transformer for Single Image Desnowing. *arXiv preprint arXiv:2207.05621* (2022).

[11] Sixiang Chen, Tian Ye, Jun Shi, Yun Liu, JingXia Jiang, Erkang Chen, and Peng Chen. 2023. DEHRFormer: Real-Time Transformer for Depth Estimation and Haze Removal from Varicolored Haze Scenes. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 1–5.

[12] Yinbo Chen, Sifei Liu, and Xiaolong Wang. 2021. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 8628–8638.

[13] Lu Chi, Borui Jiang, and Yadong Mu. 2020. Fast fourier convolution. *Advances in Neural Information Processing Systems* 33 (2020), 4479–4488.

[14] Liang-Jian Deng, Gemine Vivone, Mercedes E Paoletti, Giuseppe Scarpa, Jiang He, Yongjun Zhang, Jocelyn Chanussot, and Antonio Plaza. 2022. Machine Learning in Pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine* 10, 3 (2022), 279–315.

[15] Xueyang Fu, Zihuang Lin, Yue Huang, and Xinghao Ding. 2019. A variational pan-sharpening with local gradient constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10265–10274.

[16] Alan R Gillespie, Anne B Kahle, and Richard E Walker. 1987. Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques. *Remote Sensing of Environment* 22, 3 (1987), 343–365.

[17] Xuanhua He, Keyu Yan, Jie Zhang, Rui Li, Chengjun Xie, Man Zhou, and Danfeng Hong. 2023. Multi-Scale Dual-Domain Guidance Network for Pan-sharpening. *IEEE Transactions on Geoscience and Remote Sensing* (2023).

[18] Hanzhe Hu, Yinbo Chen, Jiarui Xu, Shubhankar Borse, Hong Cai, Fatih Porikli, and Xiaolong Wang. 2022. Learning implicit feature alignment function for semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX.* Springer, 487–505.

[19] Quan Huynh-Thu and Mohammed Ghanbari. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics letters* 44, 13 (2008), 800–801.

[20] Yiyong Jiang, Xinghao Ding, Delu Zeng, Yue Huang, and John Paisley. 2015. Pan-sharpening with a hyper-Laplacian penalty. In *Proceedings of the IEEE International Conference on Computer Vision.* 540–548.

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14.* Springer, 694–711.

[22] P Kwarteng and A Chavez. 1989. Extracting spectral contrast in Landsat Thematic Mapper image data using selective principal component analysis. *Photogramm. Eng. Remote Sens* 55, 1 (1989), 339–348.

[23] Craig A Laben and Bernard V Brower. 2000. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening. US Patent 6,011,875.

[24] Jaehyup Lee, Soomin Seo, and Munchurl Kim. 2021. Sipsa-net: Shift-invariant pan sharpening with moving object alignment for satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10166–10174.

[25] Wenzhi Liao, Xin Huang, Frieke Van Coillie, Guy Thoonen, Aleksandra Pižurica, Paul Scheunders, and Wilfried Philips. 2015. Two-stage fusion of thermal hyperspectral and visible RGB image by PCA and guided filter. In *2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS).* Ieee, 1–4.

[26] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. 2019. Point-voxel cnn for efficient 3d deep learning. *Advances in Neural Information Processing Systems* 32 (2019).

[27] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. 2019. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE transactions on pattern analysis and machine intelligence* 42, 10 (2019), 2624–2641.

[28] Stephane G Mallat. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence* 11, 7 (1989), 674–693.

[29] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. 2016. Pansharpening by convolutional neural networks. *Remote Sensing* 8, 7 (2016), 594.

[30] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

[31] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 165–174.

[32] Yongming Rao, Wenliang Zhao, Yansong Tang, Jie Zhou, Ser Nam Lim, and Jiwen Lu. 2022. Hornet: Efficient high-order spatial interactions with recursive gated convolutions. *Advances in Neural Information Processing Systems* 35 (2022), 10353–10366.

[33] Tiancheng Shen, Yuechen Zhang, Lu Qi, Jason Kuen, Xingyu Xie, Jianlong Wu, Zhe Lin, and Jiaya Jia. 2022. High Quality Segmentation for Ultra High-resolution Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1310–1319.

[34] Xin Tian, Yuerong Chen, Changcai Yang, and Jiayi Ma. 2021. Variational pansharpening by exploiting cartoon-texture similarities. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–16.

[35] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A Licciardi, Rocco Restaino, and Lucien Wald. 2014. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing* 53, 5 (2014), 2565–2586.

[36] Lucien Wald. 2002. *Data fusion: definitions and architectures: fusion of images of different spatial resolutions.* Presses des MINES.

[37] Lucien Wald, Thierry Ranchin, and Marc Mangolini. 1997. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images.

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[39] Han Xu, Jiayi Ma, Zhenfeng Shao, Hao Zhang, Junjun Jiang, and Xiaojie Guo. 2020. SDPNet: A deep network for pan-sharpening with enhanced information representation. *IEEE Transactions on Geoscience and Remote Sensing* 59, 5 (2020), 4120–4134.

[40] Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. 2021. Deep Gradient Projection Networks for Pan-sharpening. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* https://doi.org/10.1109/cvpr46437.2021.00142

[41] Keyu Yan, Man Zhou, Liu Liu, Chengjun Xie, and Danfeng Hong. 2022. When pansharpening meets graph convolution network and knowledge distillation. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–15.

[42] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. 2019. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision.* 4541–4550.

[43] Gang Yang, Man Zhou, Keyu Yan, Aiping Liu, Xueyang Fu, and Fan Wang. 2022. Memory-augmented deep conditional unfolding network for pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1788–1797.

[44] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. 2017. PanNet: A deep network architecture for pan-sharpening. In *Proceedings of the IEEE international conference on computer vision.* 5449–5457.

[45] Tian Ye, Sixiang Chen, Yun Liu, Yi Ye, Jinbin Bai, and Erkang Chen. 2022. Towards real-time high-definition image snow removal: Efficient pyramid network with asymmetrical encoder-decoder architecture. In *Proceedings of the Asian Conference on Computer Vision.* 366–381.

[46] Tian Ye, Sixiang Chen, Yun Liu, Yi Ye, Erkang Chen, and Yuche Li. 2022. Underwater light field retention: Neural rendering for underwater imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 488–497.

[47] Tian Ye, Yunchen Zhang, Mingchao Jiang, Liang Chen, Yun Liu, Sixiang Chen, and Erkang Chen. 2022. Perceiving and modeling density for image dehazing. In *European Conference on Computer Vision*. Springer, 130–145.

[48] Qiangqiang Yuan, Yancong Wei, Zisheng Zhang, Huanfeng Shen, and Liangpei Zhang. 2017. A Multi-Scale and Multi-Depth Convolutional Neural Network for Remote Sensing Imagery Pan-Sharpening.

[49] Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. 1992. 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*.

[50] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*. 286–301.

[51] Man Zhou, Jie Huang, Yanchi Fang, Xueyang Fu, and Aiping Liu. 2022. Pan-sharpening with customized transformer and invertible neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3553–3561.

[52] Man Zhou, Jie Huang, Xueyang Fu, Feng Zhao, and Danfeng Hong. 2022. Effective pan-sharpening by multiscale invertible neural network and heterogeneous task distilling. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–14.

[53] Man Zhou, Jie Huang, Chongyi Li, Hu Yu, Keyu Yan, Naishan Zheng, and Feng Zhao. 2022. Adaptively Learning Low-high Frequency Information Integration for Pan-sharpening. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3375–3384.

[54] Man Zhou, Jie Huang, Keyu Yan, Gang Yang, Aiping Liu, Chongyi Li, and Feng Zhao. 2022. Normalization-based Feature Selection and Restitution for Pan-sharpening. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3365–3374.

[55] Man Zhou, Jie Huang, Keyu Yan, Hu Yu, Xueyang Fu, Aiping Liu, Xian Wei, and Feng Zhao. 2022. Spatial-frequency domain information integration for pan-sharpening. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*. Springer, 274–291.

[56] Man Zhou, Jie Huang, Feng Zhao, and Danfeng Hong. 2022. Modality-Aware Feature Integration for Pan-Sharpening. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2022), 1–12.

[57] Man Zhou, Keyu Yan, Xueyang Fu, Aiping Liu, and Chengjun Xie. 2023. PAN-Guided Band-Aware Multi-Spectral Feature Enhancement for Pan-Sharpening. *IEEE Transactions on Computational Imaging* 9 (2023), 238–249.

[58] Man Zhou, Keyu Yan, Jie Huang, Zihe Yang, Xueyang Fu, and Feng Zhao. 2022. Mutual Information-Driven Pan-Sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1798–1808.

[59] Man Zhou, Keyu Yan, Jinshan Pan, Wenqi Ren, Qi Xie, and Xiangyong Cao. 2023. Memory-augmented deep unfolding network for guided image super-resolution. *International Journal of Computer Vision* 131, 1 (2023), 215–242.