# Fourier-enhanced Implicit Neural Fusion Network for Multispectral and Hyperspectral Image Fusion

**Yu-Jie Liang**, **Zihan Cao**, **Shangqi Deng** and **Liang-Jian Deng***
UESTC
yujieliang0219@gmail.com, iamzihan@std.uestc.edu.cn, dengsq5856@126.com,
liangjian.deng@uestc.edu.cn

## Abstract

Recently, implicit neural representations (INR) have made significant strides in various vision-related domains, providing a novel solution for Multispectral and Hyperspectral Image Fusion (MHIF) tasks. However, INR is prone to losing high-frequency information and is confined to the lack of global perceptual capabilities. To address these issues, this paper introduces a Fourier-enhanced Implicit Neural Fusion Network (FeINFN) specifically designed for MHIF task, targeting the following phenomena: *The Fourier amplitudes of the HR-HSI latent code and LR-HSI are remarkably similar; however, their phases exhibit different patterns.* In FeINFN, we innovatively propose a spatial and frequency implicit fusion function (Spa-Fre IFF), helping INR capture high-frequency information and expanding the receptive field. Besides, a new decoder employing a complex Gabor wavelet activation function, called Spatial-Frequency Interactive Decoder (SFID), is invented to enhance the interaction of INR features. Especially, we further theoretically prove that the Gabor wavelet activation possesses a time-frequency tightness property that favors learning the optimal bandwidths in the decoder. Experiments on two benchmark MHIF datasets verify the state-of-the-art (SOTA) performance of the proposed method, both visually and quantitatively. Also, ablation studies demonstrate the mentioned contributions. The code will be available on Anonymous GitHub after possible acceptance.

## 1 Introduction

Hyperspectral imaging captures scenes across contiguous spectral bands, offering intricate details compared to traditional single or limited-band images, and improving computer vision application accuracy, such as target recognition, classification, tracking, and segmentation [Fauvel *et al.*, 2012; Uzair *et al.*, 2013; Van Nguyen *et al.*, 2010; Tarabalka *et al.*, 2009]. However, practical optical sensors face challenges in balancing spatial resolution and spectral precision. Images with over 100 bands often exhibit lower spatial resolution, while those with fewer bands display higher spatial resolution. Efforts for MHIF are underway to fuse high spatial-resolution multispectral images (HR-MSI) with low spatial-resolution hyperspectral images (LR-HSI) to finally obtain high spatial-resolution hyperspectral images (HR-HSI). Actually, MHIF technology could fuse hyperspectral images with multispectral images, extracting information not detectable by HR-MSI to enhance richness and precision. Recent MHIF literature explores model-based approaches [Dian and Li, 2019; Dian *et al.*, 2019; Xu *et al.*, 2022] and deep learning methods [Huang *et al.*, 2022; Dong *et al.*, 2021; Cao *et al.*, 2024]. While model-based methods leverage image priors, challenges persist in obtaining high-fidelity, low-distortion HR-HSI due to the lack of large-scale training datasets. Among deep-learning approaches, CNN-based networks for HR-MSI and LR-HSI tend to be limited and lack interpretability for MHIF tasks and Transformer frameworks [Hu *et al.*, 2022a; Deng *et al.*, 2023a] address the small receptive field of CNN but bring greater computational overhead.

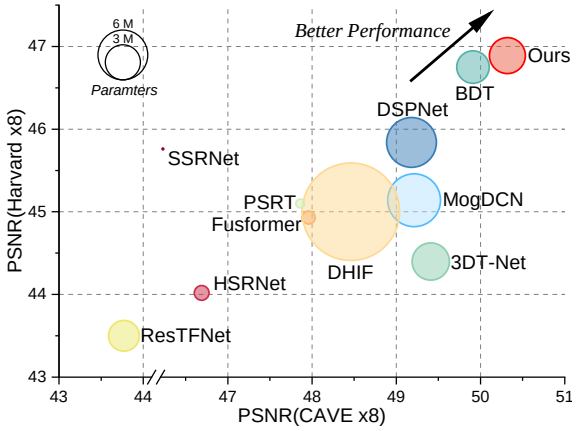


Figure 1: Comparison of our method with other methods on the CAVE($\times$ 8) and Harvard($\times$ 8) datasets. Closer to the top-right corner indicates better performance and the size of the circle indicates the number of parameters in the model.

In recent years, implicit representations of 3D scenes have garnered significant attention from researchers. For instance,

*Corresponding author.

Neural Radiance Field [Wang *et al.*, 2021] models 3D static scenes by mapping coordinates to signals through a neural network. Inspired by this, researchers have revisited image representation for 2D tasks. Recent studies [Chen *et al.*, 2021; Lee and Jin, 2022; Sitzmann *et al.*, 2020; Chen *et al.*, 2023] have achieved arbitrary-scale super-resolution (SR) by replacing commonly used upsampling layers with local implicit image functions. Though these methods demonstrate superior performance in 2D tasks, they still have some drawbacks. *Firstly*, INR calculates the RGB values of a queried coordinate based on the relative distances to the surrounding four pixels, treating it as a local operation in space that lacks consideration for global information. *Additionally*, the MLP-ReLU structure used in traditional INR inherent high-frequency information bias [Rahaman *et al.*, 2019] which is challenging to be eliminated during training.

To address these issues, we propose implicit fusion functions tailored for the MHIF task as a novel fusion paradigm. We first employ encoders to extract prior information from LR-HSI and HR-MSI, which is then fed into the implicit fusion functions in the form of latent codes. Unlike traditional INR, we transform latent codes into the Fourier domain and simultaneously perform spatial and frequency fusion in a unified network. This approach not only rectifies the high-frequency insensitivity induced by the MLP but also effectively extends the receptive field, encompassing a more comprehensive scope of global information. To integrate spatial and frequency domain representations efficiently, we design a decoder with time-frequency tightness, mapping features on both domains to pixel space. The contributions of this work are three folds:

- We define a novel fusion framework based on INR, which innovatively extracts information from the spatial and Fourier domains, effectively enhances the representation ability of high-frequency information, and expands the receptive field.
- We propose a new decoder employing a Gabor wavelet activation function to enhance the interaction of INR features. Furthermore, we theoretically prove that the complex Gabor wavelet activation possesses a time-frequency tightness property, which facilitates the decoder in learning the optimal bandwidths.
- The proposed network reaches state-of-the-art (SOTA) performance on the MHIF task across two widely used hyperspectral datasets at various fusion ratios. Fig. 1 provides a fair comparison with other SOTA methods.

## 2 Related works

### 2.1 Implicit Neural Representation (INR)

Unlike traditional discrete representations, neural implicit representation (INR) provides a more elegant and continuous parameterized approach. Initially applied in 3D modeling tasks, NeRF [Wang *et al.*, 2021] revolutionized 3D computer vision by representing intricate three-dimensional scenes with just 2D pose images. This line of work extends to the 2D imaging domain, where INR performs a weighted average on adjacent sub-codes to ensure output value continuity. LIIF [Chen *et al.*, 2021] recently introduces a local implicit image function for SR, leveraging MLP to sample pixel signals across the spatial domain. Several improvements focus on decoding networks; for example, UltraSR [Xu *et al.*, 2021] incorporates residual networks, merging spatial coordinates and depth encoding. DIINN [Nguyen and Beksi, 2023] utilizes a dual-interactive implicit neural network to decouple content and position features, improving decoding capabilities. JIIF [Tang *et al.*, 2021] proposes joint implicit image functions for multimodal learning, extracting priors from guided images. Regarding activation functions in the MLP, SIREN [Sitzmann *et al.*, 2020] recommends utilizing periodic activation functions for continuous INR to fit complex signals. On the other hand, WIRE [Saragadam *et al.*, 2023] further employs continuous complex Gabor wavelet activation functions to activate non-linearity, focusing more on spatial frequencies. However, there is limited research dedicated to designing INR architectures specifically for the MHIF task. The unique characteristics of hyperspectral images pose challenges for INR networks, in their insensitivity to high-frequency information.

### 2.2 Latent Enhancement by Fourier Transform

Fourier transform is a commonly used time-frequency analysis technique in signal processing, which converts signals from the time domain to the frequency domain. The Fourier domain has global statistical properties, and in recent years, many researchers have focused on processing frequency domain information in Fourier space. Many works use the Fourier transform to enhance the representation ability of neural networks. For example, FDA [Zhao *et al.*, 2023] proposes exchanging amplitude and phase components in Fourier space between images to enhance and adjust frequency information. FFC [Chi *et al.*, 2020] introduces a novel convolution module that internally fuses cross-scale information to capture global features in Fourier space. Similarly, GFNet [Rao *et al.*, 2021] uses 2D discrete Fourier transform and inverse transform to extract features, implements learnable global filtering, and replaces the self-attention layer in Transformer. UHDFour [Li *et al.*, 2023] embeds Fourier transform into the image enhancement network to model global information. Together, these studies demonstrate the utility of frequency domain information in improving performance on visual tasks. We exploit the architecture of FeINFN to transform latent codes into the frequency domain, implicitly integrating representations of amplitude and phase components, and enhancing high-frequency injection.

### 2.3 Motivation

[Rahaman *et al.*, 2019] finds that most neural networks exhibit a phenomenon of spectral bias through Fourier analysis. This includes neural networks such as MLP, which tend to learn low-frequency information during the early stages of training and are insensitive to high-frequency information. Moreover, we found this issue occurs in the MHIF task according to an experimental analysis as shown in Fig. 2, where HR-HSI and LR-HSI were concatenated with HR-MSI and fed into a trained encoder to obtain latent codes. These codes

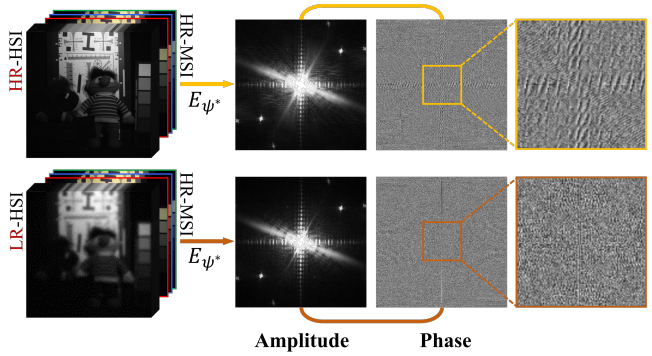


Figure 2: The amplitude of latent code from the encoder fed by HR-HSI and LR-HSI (combined with HR-MSI) share a similarity, but the phases differ from each other. $E_{\psi^*}$ is a trained encoder.

were transformed into the frequency domain to visualize the amplitude and phase. It can be observed that the amplitudes from HR-HSI and LR-HSI are very similar, while the phases differ significantly. The phase of HR-HSI should naturally contain more texture than LR-HSI, a hypothesis validated by the visualized phase maps. Based on this finding, we transformed the latent codes into the Fourier domain to separately process amplitude and phase, to enhance the global learning of high-frequency information in the images.

# 3 Methodology

In this section, we first present the preliminary of INR and then provide the proposed framework tailored for MHIF task. Subsequently, we elaborate on the implementation details of the composited modules of the proposed FeINFN.

## 3.1 Preliminary: Implicit Neural Representation

Neural Radiance Fields [Wang *et al.*, 2021] is represented by integral construction scenes. The value of a pixel in a certain viewing angle image is regarded as the integral of the characteristics of the sampling point from the proximal end to the far end of the ray. During actual training, the integral needs to be discretized. Extended to 2D image representation [Chen *et al.*, 2021], it is sampled pixel by pixel from the vicinity of the query target. Taking the low-resolution (LR) image $\mathbf{I} \in \mathbb{R}^{h\times w\times 3}$ upsampling to the high-resolution (HR) image $\hat{\mathbf{I}} \in \mathbb{R}^{H\times W\times 3}$ as an example, the process of generating the RGB values of the target coordinates $\mathbf{x}_q \in \mathbb{R}^2$ can be regarded as interpolation form, expressed as:

$$\hat{\mathbf{I}}(\mathbf{x}_q) = \sum_{i\in\mathcal{N}_q} w_{q,i}\mathbf{v}_{q,i}, \tag{1}$$

where $\mathbf{v}_{q,i} \in \mathbb{R}^{4\times4\times3}$ is the interpolation pixel of $i$ interpolated by $q$'s surrounding pixels $\mathcal{N}_q \in \mathbb{R}^4$ and $w_{q,i} \in \mathbb{R}$ signifies the interpolation weight. In the implicit representation of local image features, the weights $w_{q,i} = S_i/S$, where $S_i$ represents the area formed by $q$ and $i$ in the diagonal region and $S$ denotes the total area enclosed by the set $\mathcal{N}_q$.

The interpolation value $\mathbf{v}_{q,i}$ is effectively generated by a basis function:

$$\mathbf{v}_{q,i} = \phi_\theta(\mathbf{z}_i, \mathbf{x}_q - \mathbf{x}_i), \tag{2}$$

where $\phi_\theta$ is typically an MLP, $\mathbf{z}_i$ is the latent code generated by an encoder for the coordinates $\mathbf{x}_i$, and $\mathbf{x}_q - \mathbf{x}_i$ represents the relative coordinates. From the above equations, it can be inferred that the interpolation features can be represented by a set of local feature vectors in the LR domain. Typically, interpolation-based methods [Press, 2007; Keys, 1981] achieve upsampling by querying $\mathbf{x}_q - \mathbf{x}_i$ in the arbitrary SR task. See more details in [Chen *et al.*, 2021].

## 3.2 Overview of the FeINFN Framework

In this work, we propose the FeINFN, which adopts a novel framework for simultaneously performing neural implicit representation in both the spatial and frequency domains to execute the MHIF task. Fig. 3 provides an overview of the proposed framework, designed to fuse LR-HSI $\mathbf{I}^{LR} \in \mathbb{R}^{h\times w\times S}$ and HR-MSI $\mathbf{I}^{HR} \in \mathbb{R}^{H\times W\times s}$ to generate HR-HSI images $\widetilde{\mathbf{I}} \in \mathbb{R}^{H\times W\times S}$ based on a given upsampling scale $r$.

Initially, the LR-HSI is fed into encoder $E_\chi$ to extract spectral features $\mathbf{Z}_{spe} \in \mathbb{R}^{h\times w\times C}$. Simultaneously, the concatenated bicubic interpolation LR-HSI $\mathbf{I}^{LR}_{up} \in \mathbb{R}^{H\times W\times S}$ and $\mathbf{I}^{HR}$, are fed into encoder $E_\psi$ to extract spatial features $\mathbf{Z}_{spa} \in \mathbb{R}^{H\times W\times C}$. Additionally, the pixel's central position is represented as the coordinate point. The coordinate map is normalized into a two-dimensional grid $[-1,1]\times[-1,1]$, obtaining a HR normalized 2D coordinate map $\mathbf{X}^{HR} \in \mathbb{R}^{H\times W\times 2}$. The extracted $\mathbf{Z}_{spe}$ and $\mathbf{Z}_{spa}$, along with the 2D coordinates of $\mathbf{I}^{HR}$, are forwarded to Spatial-Frequency Implicit Fusion Function (Spa-Fre IFF), outputting spatial domain features $\mathcal{E}_s \in \mathbb{R}^{H\times W\times S}$ and frequency domain features $\mathcal{E}_f \in \mathbb{R}^{H\times W\times S}$. The $\mathcal{E}_s$ and $\mathcal{E}_f$ serve as inputs to a pixel space mapping decoder which generates the residual image $\mathbf{I}^{HR}_r \in \mathbb{R}^{H\times W\times S}$. Finally, the residual image $\mathbf{I}^{HR}_r$ is combined with the bicubicly upsampled image $\mathbf{I}^{LR}_{up}$ via element-wise addition, yielding the ultimate fusion image $\widetilde{\mathbf{I}}$.

## 3.3 INR Encoder Networks

Analogous to local implicit representation functions [Chen *et al.*, 2021; Lee and Jin, 2022; Sitzmann *et al.*, 2020; Chen *et al.*, 2023], the initial step involves extracting latent code representations. For the MHIF task, we address the challenges of both upsampling and fusion simultaneously, employing implicit neural representations as the solution. The INR encoders try to extract spatial and spectral latent codes $\mathbf{Z}_{spa} \in \mathbb{R}^{H\times W\times C}, \mathbf{Z}_{spe} \in \mathbb{R}^{h\times w\times C}$: one is extracted from $\mathbf{I}^{LR}$, serving as the carrier for spectral information; the other is encoded from the concatenation of $\mathbf{I}^{LR}_{up}$ and $\mathbf{I}^{HR}$, aiding in spatial information during the fusion process. This process can be denoted as:

$$\begin{cases} \mathbf{Z}_{spe} = E_\chi(\mathbf{I}^{LR}), \\ \mathbf{Z}_{spa} = E_\psi\left(\mathrm{Cat}(\mathbf{I}^{LR}_{up}, \mathbf{I}^{HR})\right), \end{cases} \tag{3}$$

where $E_\chi$ is the spectral encoder parameterized by $\chi$, $E_\psi$ is the spatial encoder parameterized by $\psi$, and $\mathrm{Cat}(\mathbf{I}^{LR}_{up}, \mathbf{I}^{HR})$ denotes the concatenation along the channel dimension. In practice, we utilize EDSR [Lim *et al.*, 2017] as INR encoder networks.

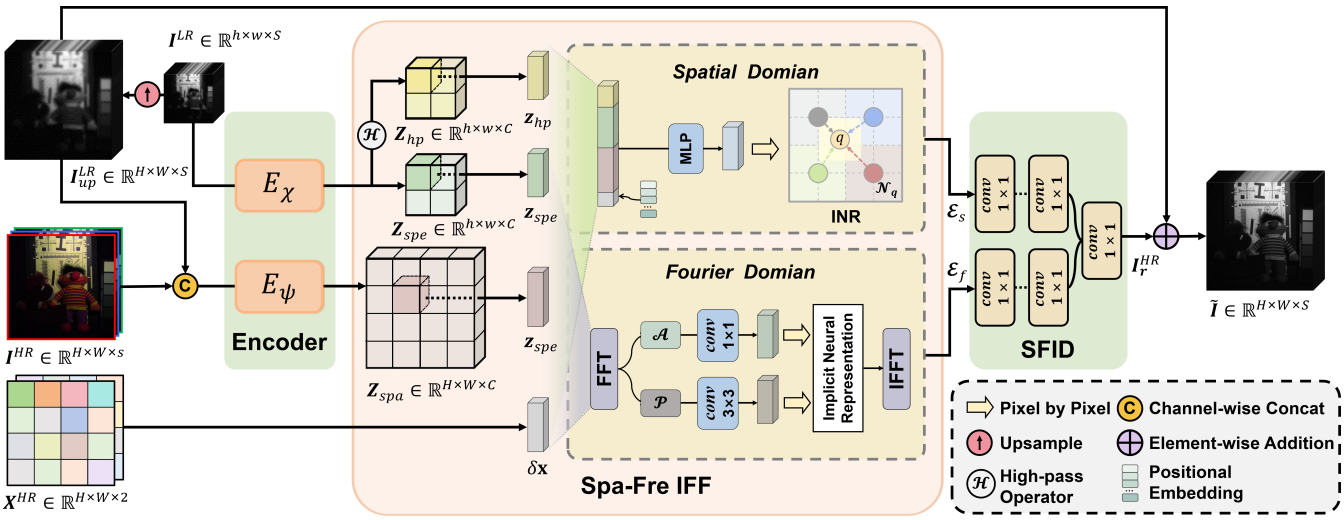


Figure 3: The flowchart of the FeINFN framework which is composed of a spectral encoder $E_\chi$, a spatial encoder $E_\psi$, MHIF task-designed spatial and Fourier domains implicit fusion functions, and a pixel space mapping decoder. Please note that $\mathbf{I}^{LR}$ is the LR-HSI, $\mathbf{I}^{HR}$ is the HR-MSI, $\mathbf{I}_{up}^{LR}$ is the bicubic interpolation LR-HSI, and $\mathbf{X}^{HR}$ is the HR normalized 2D coordinate map. $\mathbf{z}_{spe}$, $\mathbf{z}_{spa}$, $\mathbf{z}_{hp}$, $\delta\mathbf{x}$ correspond to individual pixel units, $\mathcal{A}$ and $\mathcal{P}$ represents amplitude and phase, respectively.

### 3.4 Spatial-Frequency Implicit Fusion Function

To address the mentioned issues 2.3, we propose Spatial-Frequency Implicit Fusion Function, dubbed Spa-Fre IFF which is a dual-branch fusion function and utilized for computing the fusion feature of $\mathbf{Z}_{spe}$ and $\mathbf{Z}_{spa}$ in the spatial and frequency domains, respectively. Given a queried HR coordinate $\mathbf{x}_q \in \mathbf{X}^{HR}$ of a pixel unit $q$, Spa-Fre IFF estimates spatial feature vector $\boldsymbol{\varepsilon}_s \in \mathbb{R}^{1\times1\times S}$ ($\boldsymbol{\varepsilon}_s \in \mathcal{E}_s$) and frequency feature vector $\boldsymbol{\varepsilon}_f \in \mathbb{R}^{1\times1\times S}$ ($\boldsymbol{\varepsilon}_f \in \mathcal{E}_f$)as follows:

$$\boldsymbol{\varepsilon}_s, \boldsymbol{\varepsilon}_f = \text{Spa-Fre IFF}(\mathbf{z}_{spe}, \mathbf{z}_{spa}, \delta\mathbf{x}), \tag{4}$$

where $\mathbf{z}_{spe} \in \mathbb{R}^{1\times1\times C}$ represents the spectral latent code vector corresponding to $\mathbf{x}_q$, and $\mathbf{z}_{spa} \in \mathbb{R}^{4\times4\times C}$ is the spatial latent code vector. $\delta\mathbf{x}$ denotes the set of local relative coordinates, expressed by the following formula:

$$\delta\mathbf{x} = \{\mathbf{x}_q - \mathbf{x}_{q,i}\}_{i\in\mathcal{N}_q}, \tag{5}$$

where $\mathbf{x}_{q,i}$ refers to the coordinates most proximate to the query coordinate $\mathbf{x}_q$, representing the four corner pixels closest to $q$ in the HR space.

**Spatial Implicit Fusion Function:** The Spatial Implicit Fusion Function aims to leverage the powerful representation capabilities of INR to achieve implicit fusion in the spatial domain, as shown in Fig. 3 (see branch "Spatial Domain"). Specifically, we employ high-pass operators $\mathcal{H}$ to filter the spectral latent codes, as a complement to the high-frequency information on the spectrum:

$$\mathbf{z}_{hp} = \mathcal{H}(\mathbf{z}_{spe}), \tag{6}$$

where $\mathbf{z}_{hp} \in \mathbb{R}^{1\times1\times C}$ represents the high-frequency latent code of $\mathbf{I}^{LR}$. Also, we suggest frequency encoding for relative positional coordinates as follows:

$$\begin{aligned}\gamma(\delta\mathbf{x}) =&[\sin(2^0\delta\mathbf{x}), \cos(2^0\delta\mathbf{x}), \cdots, \\ &\sin(2^{L-1}\delta\mathbf{x}), \cos(2^{L-1}\delta\mathbf{x})],\end{aligned} \tag{7}$$

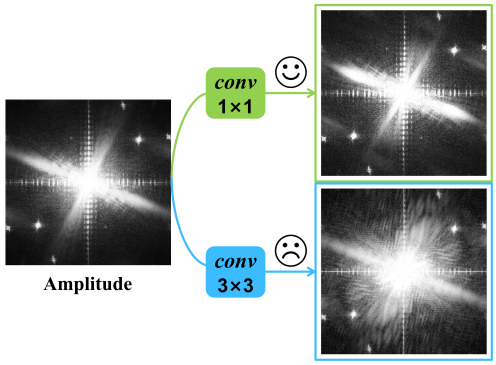


Figure 4: $3\times3$ convolution would suffer from the issue of spectrum leakage, which can be alleviated by $1\times1$ convolution.

where $L$ is a hyperparameter, in practice, we set $L$ to 10. Additionally, leveraging the graph attention mechanism [Tang *et al.*, 2021], we parameterize the solution for interpolation weights $\mathbf{w}_{q,i} \in \mathbb{R}^{1\times S}$, and the implicit fusion function simultaneously outputs fusion interpolation values $\mathbf{v}_{q,i} \in \mathbb{R}^{4\times4\times S}$ and interpolation weights $\mathbf{w}_{q,i}$. The implicit fusion function is specifically expressed as:

$$\mathbf{w}_{q,i}, \mathbf{v}_{q,i} = \phi_\theta(\mathbf{z}_{spe}, \mathbf{z}_{spa}, \mathbf{z}_{hp}, \gamma(\delta\mathbf{x})), \tag{8}$$

where $\phi_\theta$ is an MLP parameterized by $\theta$. The spatial implicit fusion interpolation, as shown in Eq. (1), yields the fused spatial feature $\boldsymbol{\varepsilon}_s \in \mathbb{R}^{1\times1\times S}$ and can be described as follows:

$$\boldsymbol{\varepsilon}_s = \sum_{i\in\mathcal{N}_q} \overline{\mathbf{w}}_{q,i} * \mathbf{v}_{q,i}. \tag{9}$$

**Frequency Implicit Fusion Function:** From Fig. 2, we observed characteristics in the frequency features between LR-HSI and HR-HSI. Hence, we design a frequency implicit fusion function to express global features continuously in the Fourier domain. Notably, directly applying static kernel convolution in the frequency domain would only enhance a

specific frequency range, which is inappropriate for the fusion task. However, by learning feature content to generate weights, INR can be seen as a dynamic interpolation method in continuous space, adaptively enhancing information in the frequency domain without overly altering the frequency distribution. Therefore, introducing INR into the Fourier domain is reasonable. Since amplitude and phase exhibit different forms, as shown in Fig. 2, we handle them separately.

With the considerations mentioned above, as illustrated in Fig. 3 (see branch "Fourier Domain"), we initially employ FFT to transform latent codes $\mathbf{z}_{spe}$ and $\mathbf{z}_{spa}$ from the spatial domain to the frequency domain, obtaining $\mathbf{f}_{spe} \in \mathbb{R}^{1\times1\times C}$ and $\mathbf{f}_{spa} \in \mathbb{R}^{4\times4\times C}$. After the transformation, we further obtain amplitude components $\mathcal{A}(\mathbf{f}_{spe})$ and $\mathcal{A}(\mathbf{f}_{spa})$, as well as phase components $\mathcal{P}(\mathbf{f}_{spe})$ and $\mathcal{P}(\mathbf{f}_{spa})$.

*For the amplitude*, as shown in Fig. 4, the amplitude distribution of LR-HSI and HR-HSI are very similar, and the non-point-wise convolution (*e.g.* Conv $3 \times 3$) causes an issue of spectrum leakage, confusing channel information. In contrast, point-wise convolution does not span multiple locations in the frequency domain and has no overlap allowing it to capture information across channels effectively. Thus the fusion function for amplitude components is more suitable when applying point-wise convolution:

$$\mathbf{w}^{\mathcal{A}}_{q,i}, \mathbf{v}^{\mathcal{A}}_{q,i} = \phi^{\mathcal{A}}_{\alpha}(\mathcal{A}(\mathbf{f}_{spe}), \mathcal{A}(\mathbf{f}_{spa}), \delta\mathbf{x}), \tag{10}$$

where $\mathbf{w}^{\mathcal{A}}_{q,i} \in \mathbb{R}^{1\times S}$ and $\mathbf{v}^{\mathcal{A}}_{q,i} \in \mathbb{R}^{4\times4\times S}$ are the weights and interpolated values for the corresponding amplitude component, and $\phi^{\mathcal{A}}_{\alpha}$ is a simple network composed of two layers of point convolutions parameterized by $\alpha$. Similar to operations in the spatial domain, implicit fusion interpolation is performed after obtaining interpolated values $\mathbf{v}^{\mathcal{A}}_{q,i}$ and the normalized weights $\overline{\mathbf{w}}^{\mathcal{A}}_{q,i}$:

$$\mathcal{A}'_f = \sum_{i\in\mathcal{N}_q} \overline{\mathbf{w}}^{\mathcal{A}}_{q,i} * \mathbf{v}^{\mathcal{A}}_{q,i}, \tag{11}$$

where $\mathcal{A}'_f \in \mathbb{R}^{1\times1\times S}$ is the integrated amplitude component.

*For the phase*, which encapsulates information such as texture details, LR-HSI and HR-HSI often have different phase information. It is known that point convolutions fail to capture sufficient spatial representations. Therefore, we use a $3 \times 3$ convolution to learn phase information. Additionally, small changes in the frequency domain may result in significant variations in the spatial domain. We still consider using the form of INR interpolation for phase learning. The handling of the phase components $\mathcal{P}(\mathbf{f}_{spe})$ and $\mathcal{P}(\mathbf{f}_{spa})$ are formally similar to Eqs. (10) and (11):

$$\mathbf{w}^{\mathcal{P}}_{q,i}, \mathbf{v}^{\mathcal{P}}_{q,i} = \phi^{\mathcal{P}}_{\beta}(\mathcal{P}(\mathbf{f}_{spe}), \mathcal{P}(\mathbf{f}_{spa}), \delta(\mathbf{x})), \tag{12}$$

$$\mathcal{P}'_f = \sum_{i\in\mathcal{N}_q} \overline{\mathbf{w}}^{\mathcal{P}}_{q,i} * \mathbf{v}^{\mathcal{P}}_{q,i}. \tag{13}$$

The simple network $\phi^{\mathcal{P}}_{\beta}$ consists of two layers of $3 \times 3$ convolutions parameterized by $\beta$. $\mathcal{P}'_f \in \mathbb{R}^{1\times1\times S}$ represents the integrated phase component.

Finally, IFFT is applied to map the frequency features $\mathcal{A}'_f$ and $\mathcal{P}'_f$ back to the image space, obtaining the frequency domain feature $\varepsilon_f \in \mathcal{E}_f$. Since in frequency space, one frequency point may correspond to multiple pixels at different positions in the spatial domain, the receptive field of INR in the frequency domain is enlarged in the spatial domain.

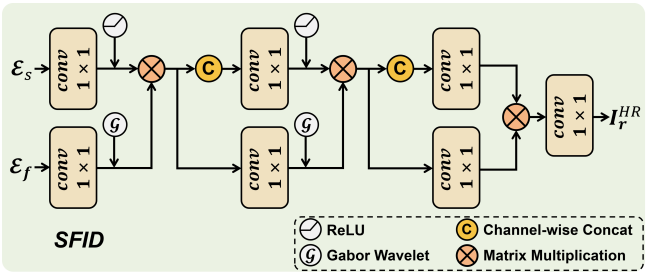


Figure 5: Detailed composition of the proposed SFID.

### 3.5 Spatial-Frequency Interactive Decoder

After obtaining the spatial feature map and frequency domain feature map, it is essential to consider how to integrate them seamlessly. Firstly, our decoder needs to have dual input and interactive capabilities. Secondly, it is necessary to focus on representing images in the spatial-frequency domain. With this in mind, we introduce the complex Gabor wavelet activation function with good time-frequency tightness and propose the Spatial-Frequency Interactive Decoder (SFID). Specifically, SFID consists of three layers, taking spatial and frequency domain features as inputs. The outputs $\mathbf{I}^{HR}_r$ and $\mathbf{I}^{HR}_{up}$ contribute to the final fused image $\mathbf{I}$. The decoding process is illustrated in Fig. 5. The complex Gabor wavelet function is defined as:

$$\mathcal{G}(\mathbf{x}) = e^{j\omega_0\mathbf{x}} e^{-|\upsilon_0\mathbf{x}|^2}, \tag{14}$$

where $\omega_0$ is the center frequency in the frequency domain, $\upsilon_0$ is a constant that is considered as the standard deviation of the Gaussian function, and $\mathbf{x}$ is a vector in the time (or spatial) domain. In what follows, we provide a theorem below that this Gabor wavelet activation has time-frequency tightness [Blu and Lebrun, ], which is helpful for the decoder's information interaction.

**Theorem 1.** *The complex Gabor wavelet activation in Eq.* (14) *has the time-frequency tightness property. Moreover, from the perspective of signal spectrum analysis, this activation helps the decoder learn the optimal bandwidths.*

*Proof:* The detailed proof can be found in the Supplementary.

## 4 Experiments

**Datasets:** To evaluate the efficacy of our model, we conducted experiments using the CAVE and Harvard datasets. The CAVE dataset comprises 32 Hyperspectral Images (HSIs) with 31 spectral bands spanning from 400 nm to 700 nm at 10 nm intervals. We randomly selected 20 images for training and used the remaining 11 for testing. The Harvard dataset consists of 77 HSIs depicting indoor and outdoor scenes, covering the spectral range from 420 nm to 720 nm. We standardized the data by cropping the upper left sections of 20 Harvard images, with 10 for training and the rest for testing. Details can be found in the Supplementary.

**Implementation details:** We implement the proposed methods FeINFN with Pytorch [Paszke *et al.*, 2019] on a workstation with an Intel I9 CPU and two 3090 GPUs. The optimizer

| Methods | CAVE $\times 4$ | | | | | Harvard $\times 4$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR(↑) | SAM(↓) | ERGAS(↓) | SSIM(↑) | #params | PSNR(↑) | SAM(↓) | ERGAS(↓) | SSIM(↑) | #params |
| Bicubic | 34.33±3.88 | 4.45±1.62 | 7.21±4.90 | 0.944±0.029 | – | 38.71±4.33 | 2.53±0.67 | 4.45±1.81 | 0.948±0.027 | – |
| CSTF-FUS [Li *et al.*, 2018] | 34.46±4.28 | 14.37±5.30 | 8.29±5.29 | 0.866±0.075 | – | 39.15±3.45 | 6.93±2.69 | 4.66±1.81 | 0.914±0.049 | – |
| LTTR [Dian *et al.*, 2019] | 35.85±3.49 | 6.99±2.55 | 5.99±2.92 | 0.956±0.029 | – | 40.88±3.94 | 4.01±1.27 | 4.03±2.18 | 0.957±0.035 | – |
| LTMR [Dian and Li, 2019] | 36.54±3.30 | 6.71±2.19 | 5.39±2.53 | 0.963±0.021 | – | 42.06±3.56 | 3.51±0.99 | 3.59±2.03 | 0.970±0.020 | – |
| IR-TenSR [Xu *et al.*, 2022] | 35.61±3.45 | 12.30±4.68 | 5.90±3.05 | 0.945±0.027 | – | 40.47±3.04 | 4.36±1.52 | 5.57±1.57 | 0.963±0.014 | – |
| ResTFNet [Liu *et al.*, 2020] | 45.58±5.47 | 2.82±0.70 | 2.36±2.59 | 0.993±0.006 | 2.387M | 45.94±4.35 | 2.61±0.69 | 2.56±1.32 | 0.985±0.008 | 2.387M |
| SSRNet [Zhang *et al.*, 2020] | 48.62±3.92 | 2.54±0.84 | 1.63±1.21 | 0.995±0.002 | **0.027M** | 48.00±3.36 | 2.31±0.60 | 2.30±1.42 | 0.987±0.007 | **0.027M** |
| HSRNet [Hu *et al.*, 2022b] | 50.38±3.38 | 2.23±0.66 | 1.20±0.75 | 0.996±0.001 | 0.633M | 48.29±3.03 | 2.26±0.56 | 1.87±0.81 | 0.988±0.006 | 0.633M |
| MogDCN [Dong *et al.*, 2021] | 51.63±4.10 | 2.03±0.62 | 1.11±0.82 | 0.997±0.002 | 6.840M | 47.89±4.09 | 2.11±0.52 | 1.89±0.82 | 0.988±0.007 | 6.840M |
| Fusformer [Hu *et al.*, 2022a] | 49.98±8.10 | 2.20±0.85 | 2.50±5.21 | 0.994±0.011 | 0.504M | 47.87±5.13 | 2.84±2.07 | 2.04±0.99 | 0.986±0.010 | 0.467M |
| DHIF [Huang *et al.*, 2022] | 51.07±4.17 | 2.01±0.63 | 1.22±0.97 | 0.997±0.002 | 22.462M | 47.68±3.85 | 2.32±0.53 | 1.95±0.92 | 0.988±0.007 | 22.462M |
| PSRT [Deng *et al.*, 2023a] | 50.47±6.19 | 2.19±0.64 | 2.06±3.71 | 0.996±0.003 | 0.247M | 47.96±3.21 | 2.18±0.55 | 1.89±0.86 | 0.988±0.006 | 0.247M |
| 3DT-Net [Ma *et al.*, 2023] | 51.38±4.18 | 2.16±0.70 | 1.14±1.00 | 0.996±0.003 | 3.464M | 47.78±4.42 | **2.04±0.51** | 1.98±0.86 | **0.989±0.006** | 3.464M |
| DSPNet [Sun *et al.*, 2023] | 51.18±3.92 | 2.15±0.64 | 1.13±0.82 | 0.997±0.002 | 6.064M | 48.29±3.16 | 2.30±0.55 | 1.93±0.93 | 0.988±0.006 | 6.064M |
| BDT [Deng *et al.*, 2023b] | 52.30±3.98 | 1.93±0.55 | 1.02±0.77 | 0.997±0.001 | 2.668 M | 48.83±3.45 | 2.07±0.49 | 1.83±0.81 | 0.989±0.007 | 2.668 M |
| FeINFN(Ours) | **52.47±4.10** | **1.91±0.59** | **0.98±0.74** | **0.998±0.002** | 3.165 M | **49.06±3.15** | 2.10±0.53 | **1.78±0.75** | 0.989±0.007 | 3.165 M |

Table 1: The average and standard deviation calculated for all the compared approaches on 11 CAVE examples and 10 Harvard examples simulating a scaling factor of 4. The best results are in bold, second-best in underline. "M" refers to millions.

is chosen as AdamW [Kingma and Ba, 2014] and we use a Cosine anneal learning rate scheduler. The base channel number of the encoder is 128, that of the proposed implicit fusion function is 32 and in the decoder, the channel number is 31.

**Benchmark:** To evaluate FeINFN's performance, we compare it with MHIF methods on the CAVE and Harvard datasets. The bicubic-interpolated result of the upsampled LR-HSI in Tab. 1 serves as our baseline. Various model-based techniques, including the CSTF-FUS [Li *et al.*, 2018], LTTR [Dian *et al.*, 2019], LTMR [Dian and Li, 2019], and IR-TenSR [Xu *et al.*, 2022] approaches, are considered. Additionally, we compare our approach with various deep learning methods, such as SSRNet [Zhang *et al.*, 2020], ResTFNet [Liu *et al.*, 2020], HSRNet [Hu *et al.*, 2022b], MoGDCN [Dong *et al.*, 2021], Fusformer [Hu *et al.*, 2022a], and DHIF [Huang *et al.*, 2022], PSRT [Deng *et al.*, 2023a], 3DT-Net [Ma *et al.*, 2023], DSPNet [Sun *et al.*, 2023], BDT [Deng *et al.*, 2023b]. We compare our method with other methods using different image quality metrics to validate the image fusion capability of our model, including SAM [Yuhas *et al.*, 1992], ERGAS [Wald, 2002], PSNR [Horé and Ziou, 2010], and SSIM [Wang *et al.*, 2004].

**Results on CAVE Dataset:** In this section, we evaluate the effectiveness of FeINFN on the CAVE dataset and compare it with five traditional methods and some state-of-the-art deep learning-based approaches. As shown in Tab. 1 on the left, our method achieves optimal performance in the tasks of $\times 4$ in all metrics. In the $\times 4$ experiment, compared to currently leading methods such as DSPNet [Sun *et al.*, 2023], 3DT-Net [Ma *et al.*, 2023], and BDT [Deng *et al.*, 2023b], our approach demonstrates improvements in PSNR by 1.29dB/1.09dB/0.17dB, respectively. The $\times 8$ experiment is detailed in the Supplementary. To illustrate the advantages of our method, we provide visual comparisons in Fig. 6, including close-ups and error maps to highlight specific details. Our fusion results closely match the ground truth, achieving the best quality. In comparing error maps, the darker colors indicate closer proximity to the original image. In contrast to other excellent methods, the error maps of FeINFN distinctly exhibit superior restoration effects on details.

| Methods | PSNR(↑) | SAM(↓) | ERGAS(↓) | SSIM(↑) |
|---|---|---|---|---|
| Bilinear | 52.23±4.40 | 1.92±0.60 | 1.03±0.86 | 0.997±0.0021 |
| Bicubic | 52.22±4.31 | 1.95±0.61 | 1.02±0.82 | 0.997±0.0021 |
| Pixel Shuffle | 52.26±4.37 | **1.90±0.59** | 1.02±0.85 | 0.997±0.0022 |
| Our | **52.47±4.10** | 1.91±0.59 | **0.98±0.74** | **0.998±0.0015** |

Table 2: Quantitative comparisons with other upsampling methods on the CAVE ($\times 4$) dataset.

**Results on Harvard Dataset:** In Tab. 1, the right columns present the comparison results of our FeINFN with other methods on the Harvard dataset at scale factors 4. Our method performs exceptionally well, with only SAM being slightly surpassed by 3DT-Net [Ma *et al.*, 2023] and BDT [Deng *et al.*, 2023b]. FeINFN exhibits significant gains in PSNR/ERGAS/SSIM metrics compared to the current state-of-the-art [Deng *et al.*, 2023b], with improvements of 0.14dB/0.16/0.001, respectively. The results with a scale factor of 8 can be found in the Supplementary. As depicted in Fig. 1, our model outperforms others, highlighting the crucial role of FeINFN's continuous representation capability in high-scale factor scenarios. To better visualize the performance gap, Fig. 6 illustrates the fused images and error maps, confirming that our FeINFN maintains high fidelity in recovering the texture details of the images.

## 4.1 Ablation Studies

**Upsampling methods:** Implicit image representation can be seen as an advanced interpolation algorithm, offering additional spatial information and parameterized weight generation. In this section, we compare INR with other upsampling methods. We replace INR with pixel-shuffle [Shi *et al.*, 2016] and traditional CNN interpolation methods, presenting a comparative analysis. As seen in Tab. 2, our approach outperforms other methods in MHIF tasks.

**Spatial domain and Fourier domain:** To assess the dual-domain model's efficacy, we performed model reduction, preserving spatial and Fourier domains independently. As shown in Tab. 3, FeINFN excels by using both spatial and Fourier

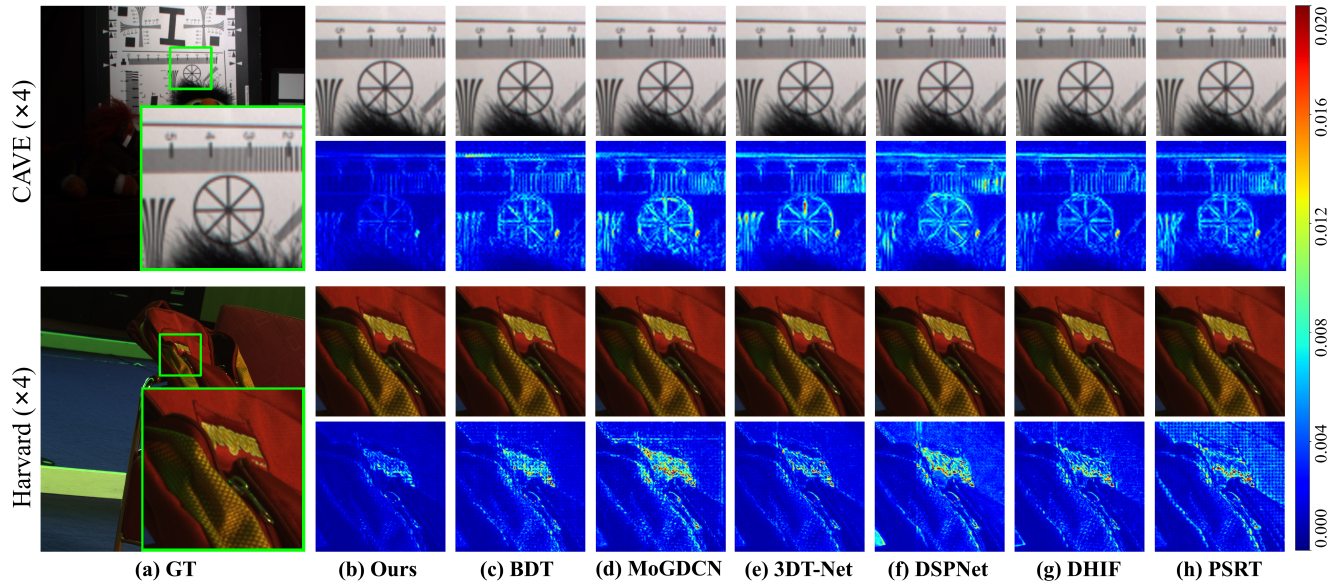


Figure 6: The upper and lower parts respectively showcase the results of "*Chart and Stuffed Toy*" from the CAVE dataset and "*Backpack*" from the Harvard dataset using pseudo-color representation. Green rectangles depict some close-up shots. The second and fourth rows show the residuals between the ground truth (GT) and the fusion products.

| $\mathcal{S}$ | $\mathcal{F}$ | PSNR(↑) | SAM(↓) | ERGAS(↓) | SSIM(↑) |
|---|---|---|---|---|---|
| ✓ | ✗ | 52.11±4.22 | 1.95±0.59 | 1.04±0.82 | 0.998±0.0017 |
| ✗ | ✓ | 47.86±3.42 | 3.49±1.30 | 1.67±1.13 | 0.995±0.0020 |
| ✓ | ✓ | **52.47±4.10** | **1.91±0.59** | **0.98±0.74** | **0.998±0.0015** |

Table 3: Quantitative comparisons with reduced models on the CAVE (×4) dataset. $\mathcal{S}$ & $\mathcal{F}$ mean the domain difference.

| Nonlinear | PSNR(↑) | SAM(↓) | ERGAS(↓) | SSIM(↑) |
|---|---|---|---|---|
| ReLU | 52.03±3.84 | 2.00±0.59 | 1.02±0.74 | **0.998±0.0013** |
| GELU | 51.96±3.88 | 2.01±0.60 | 1.03±0.75 | 0.998±0.0014 |
| Leaky ReLU | 51.98±3.92 | 2.01±0.60 | 1.03±0.76 | 0.998±0.0014 |
| Our | **52.47±4.10** | **1.91±0.59** | **0.98±0.74** | 0.998±0.0015 |

Table 4: Quantitative comparisons with different activation functions in SFID on the CAVE (×4) dataset.

domains concurrently, underscoring the positive impact of Fourier domain integration on overall network performance.

Spectral deviation occurs during training, where the network tends to prioritize low-frequency information, capturing high-frequency details only in later stages. To validate our resolution of this issue, we remove the "Fourier Domain" from Spa-Fre IFF, or retain it, and the corresponding training data is illustrated in Fig. 7. Our FeINFN, which incorporates Fourier domain fusion, leads to faster PSNR convergence and overall higher efficiency. The visual comparison of high-frequency details in "*chart and stuffed toy*" from the cave dataset at 80k iterations further supports the significant improvement achieved with our results.

**Decoder with Different Nonlinear:** In this section, we evaluate the impact of different activation functions in SFID, aiming to match SFIFF. Our dual-input decoder incorporates a complex Gabor wavelet activation function to facilitate the fusion of spatial and frequency domain features.

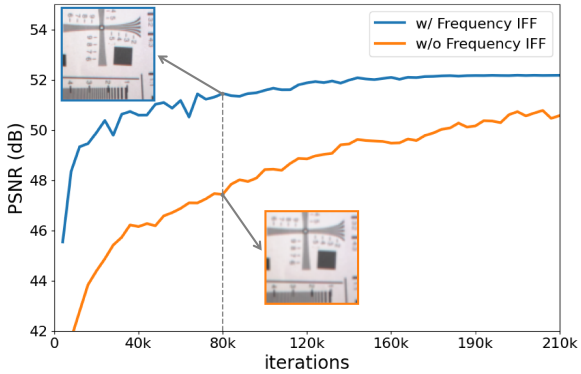


Figure 7: Changes in PSNR on the CAVE dataset of our FeINFN over iterations with and without the "Fourier Domain". The Frequency IFF can help the network learn the high-frequency details and converge faster.

Through experiments, we replaced the Gabor wavelet activation with other activations, presenting the results in Tab. 4. The findings distinctly demonstrate the enhanced fusion quality achieved with the complex Gabor wavelet activation. This emphasizes the critical role of wavelet activation in promoting robust and reliable learning in SFID.

## 5 Conclusion

Inspired by the distinct behaviors of LR-HSI and HR-HSI in the Fourier domain, we introduce a novel Fourier-enhanced Implicit Neural Fusion Network (FeINFN) based on INR. Through Fourier transformation, latent features are converted into the frequency domain, allowing the modeling of frequency components to enrich high-frequency information in images. Additionally, we propose a spatial-frequency decoding module, achieving a unified representation of both spatial

and frequency domains using a time-frequency-tight activation function. Thanks to the unique design of our network, it outperforms state-of-the-art methods in MHIF with appealing efficiency. We desire that our work will inspire future research on frequency fusion-based MHIF methods.

## References


[Blu and Lebrun, ] Thierry Blu and Jérôme Lebrun. *Linear Time-Frequency Analysis II: Wavelet-Type Representations*, chapter 4, pages 93–130. John Wiley and Sons, Ltd.

[Cao *et al.*, 2024] zihan Cao, Shiqi Cao, Liang-Jian Deng, Xiao Wu, Junming Hou, and Gemine Vivone. Diffusion model with disentangled modulations for sharpening multispectral and hyperspectral images. *Inf. Fusion.*, 104:102–158, 2024.

[Chen *et al.*, 2021] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8628–8638, 2021.

[Chen *et al.*, 2023] Hao-Wei Chen, Yu-Syuan Xu, Min-Fong Hong, Yi-Min Tsai, Hsien-Kai Kuo, and Chun-Yi Lee. Cascaded local implicit transformer for arbitrary-scale super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18257–18267, 2023.

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

[Deng *et al.*, 2023a] Shang-Qi Deng, Liang-Jian Deng, Xiao Wu, Ran Ran, Danfeng Hong, and Gemine Vivone. PSRT: Pyramid shuffle-and-reshuffle transformer for multispectral and hyperspectral image fusion. *IEEE Trans. Geosci. Remote Sens.*, 61:1–15, 2023.

[Deng *et al.*, 2023b] Shangqi Deng, Liang-Jian Deng, Xiao Wu, Ran Ran, and Rui Wen. Bidirectional dilation transformer for multispectral and hyperspectral image fusion. In *the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.

[Dian and Li, 2019] Renwei Dian and Shutao Li. Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization. *IEEE Trans. Image Process.*, 28(10):5135–5146, 2019.

[Dian *et al.*, 2019] Renwei Dian, Shutao Li, and Leyuan Fang. Learning a low tensor-train rank representation for hyperspectral image super-resolution. *IEEE Trans. Neural Netw. Learn. Syst.*, 30(9):2672–2683, 2019.

[Dong *et al.*, 2021] Weisheng Dong, Chen Zhou, Fangfang Wu, Jinjian Wu, Guangming Shi, and Xin Li. Model-guided deep hyperspectral image super-resolution. *IEEE Trans. Image Process.*, 30:5754–5768, 2021.

[Fauvel *et al.*, 2012] Mathieu Fauvel, Yuliya Tarabalka, Jon Atli Benediktsson, Jocelyn Chanussot, and James C Tilton. Advances in spectral-spatial classification of hyperspectral images. *IEEE*, 101(3):652–675, 2012.

[Horé and Ziou, 2010] Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *International Conference on Pattern Recognition (ICIP)*, pages 2366–2369, 2010.

[Hu *et al.*, 2022a] Jinfan Hu, Tingzhu Huang, Liangjian Deng, Hongxia Dou, Danfeng Hong, and Gemine Vivone. Fusformer: A transformer-based fusion network for hyperspectral image super-resolution. *IEEE Geosci. Remote Sens. Lett.*, 19:1–5, 2022.

[Hu *et al.*, 2022b] Jinfan Hu, Tingzhu Huang, Liangjian Deng, Taixiang Jiang, Gemine Vivone, and Jocelyn Chanussot. Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[Huang *et al.*, 2022] Tao Huang, Weisheng Dong, Jinjian Wu, Leida Li, Xin Li, and Guangming Shi. Deep hyperspectral image fusion network with iterative spatio-spectral regularization. *IEEE Trans. Comput Imaging.*, 8:201–214, 2022.

[Keys, 1981] Robert Keys. Cubic convolution interpolation for digital image processing. *IEEE Trans. Acoust. Speech Signal Process.*, 29(6):1153–1160, 1981.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Lee and Jin, 2022] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1929–1938, 2022.

[Li *et al.*, 2018] Shutao Li, Renwei Dian, Leyuan Fang, and José M Bioucas-Dias. Fusing hyperspectral and multispectral images via coupled sparse tensor factorization. *IEEE Trans. Image Process.*, 27(8):4118–4130, 2018.

[Li *et al.*, 2023] Chongyi Li, Chun-Le Guo, Man Zhou, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Embedding fourier for ultra-high-definition low-light image enhancement. In *International Conference on Learning Representations (ICLR)*, 2023.

[Lim *et al.*, 2017] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE/CVPR Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 136–144, 2017.

[Liu *et al.*, 2020] Xiangyu Liu, Qingjie Liu, and Yunhong Wang. Remote sensing image fusion based on two-stream fusion network. *Inf. Fusion.*, 55:1–15, 2020.

[Ma *et al.*, 2023] Qing Ma, Junjun Jiang, Xianming Liu, and Jiayi Ma. Learning a 3d-cnn and transformer prior for hyperspectral image super-resolution. *Inf. Fusion.*, page 101907, 2023.

[Nguyen and Beksi, 2023] Quan H Nguyen and William J Beksi. Single image super-resolution via a dual interactive implicit neural network. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4936–4945, 2023.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (Neurips)*, pages 8024–8035. Curran Associates, Inc., 2019.

[Press, 2007] William H Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[Rahaman *et al.*, 2019] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning (ICML)*, pages 5301–5310. PMLR, 2019.

[Rao *et al.*, 2021] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in Neural Information Processing Systems (Neurips)*, 34:980–993, 2021.

[Saragadam *et al.*, 2023] Vishwanath Saragadam, Daniel LeJeune, Jasper Tan, Guha Balakrishnan, Ashok Veeraraghavan, and Richard G Baraniuk. Wire: Wavelet implicit neural representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18507–18516, 2023.

[Shi *et al.*, 2016] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016.

[Sitzmann *et al.*, 2020] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems (Neurips)*, 33:7462–7473, 2020.

[Sun *et al.*, 2023] Yucheng Sun, Han Xu, Yong Ma, Minghui Wu, Xiaoguang Mei, Jun Huang, and Jiayi Ma. Dual spatial-spectral pyramid network with transformer for hyperspectral image fusion. *IEEE Trans. Geosci. Remote Sensing*, 2023.

[Tang *et al.*, 2021] Jiaxiang Tang, Xiaokang Chen, and Gang Zeng. Joint implicit image function for guided depth super-resolution. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, pages 4390–4399, 2021.

[Tarabalka *et al.*, 2009] Yuliy Tarabalka, Jocelyn Chanussot, and Jón Atli Benediktsson. Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers. In *IEEE Transactions on Systems, Man, and Cybernetics (TSMC)*, volume Part B (Cybernetics) 40, pages 1267–1279, 2009.

[Uzair *et al.*, 2013] Muhammad Uzair, Arif Mahmood, and Ajmal S Mian. Hyperspectral face recognition using 3d-dct and partial least squares. In *The British Machine Vision Conference (BMVC)*, volume 1, page 10, 2013.

[Van Nguyen *et al.*, 2010] Hien Van Nguyen, Amit Banerjee, and Rama Chellappa. Tracking via object reflectance using a hyperspectral video camera. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops(CVPR)*, pages 44–51. IEEE, 2010.

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

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

[Wang *et al.*, 2021] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021.

[Xu *et al.*, 2021] Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution. *arXiv preprint arXiv:2103.12716*, 2021.

[Xu *et al.*, 2022] Ting Xu, Tingzhu Huang, Liangjian Deng, and Naoto Yokoya. An iterative regularization method based on tensor subspace representation for hyperspectral image super-resolution. *IEEE Trans. Geosci. Remote Sens.*, 60:1–16, 2022.

[Yuhas *et al.*, 1992] Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *JPL AGW-3 Vol. 1: AVIRIS Workshop.*, 1992.

[Zhang *et al.*, 2020] Xueting Zhang, Wei Huang, Qi Wang, and Xuelong Li. SSR-NET: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion. *IEEE Trans. Geosci. Remote Sens.*, 59(7):5953–5965, 2020.

[Zhao *et al.*, 2023] Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. *arXiv preprint arXiv:2311.16845*, 2023.