Frequency-Adaptive Pan-Sharpening with Mixture of Experts

Xuanhua He[1,2*], Keyu Yan[1,2*], Rui Li[1], Chengjun Xie[1], Jie Zhang[1†], Man Zhou[3†]

[1]Hefei Institutes of Physical Science, Chinese Academy of Sciences
[2]University of Science and Technology of China
[3]Nanyang Technological University
{hexuanhua,keyu}@mail.ustc.edu.cn, {lirui,cjxie,zhangjie}@iim.ac.cn,manzhountu@gmail.com

Abstract

Pan-sharpening involves reconstructing missing high-frequency information in multi-spectral images with low spatial resolution, using a higher-resolution panchromatic image as guidance. Although the inborn connection with frequency domain, existing pan-sharpening research has not almost investigated the potential solution upon frequency domain. To this end, we propose a novel Frequency Adaptive Mixture of Experts (FAME) learning framework for pan-sharpening, which consists of three key components: the Adaptive Frequency Separation Prediction Module, the Sub-Frequency Learning Expert Module, and the Expert Mixture Module. In detail, the first leverages the discrete cosine transform to perform frequency separation by predicting the frequency mask. On the basis of generated mask, the second with low-frequency MOE and high-frequency MOE takes account for enabling the effective low-frequency and high-frequency information reconstruction. Followed by, the final fusion module dynamically weights high-frequency and low-frequency MOE knowledge to adapt to remote sensing images with significant content variations. Quantitative and qualitative experiments over multiple datasets demonstrate that our method performs the best against other state-of-the-art ones and comprises a strong generalization ability for real-world scenes. Code will be made publicly at https://github.com/alexhe101/FAME-Net.

Introduction

The demand for high-resolution multispectral (HRMS) images is increasing in various industries such as agriculture, mapping services, and environmental protection. However, direct acquisition of HRMS images using satellite sensors is often not feasible due to technology and hardware limitations. Instead, a common approach is to use two distinct sensors on satellites to capture high-resolution panchromatic (PAN) and low-resolution multispectral (LRMS) images. These images are then fused through the pan-sharpening process to generate HRMS images suitable for specific applications.

Recent years have witnessed significant progress in maintaining both spectral and spatial details over pan-sharpening as a consequence of the rapid progress of deep learning technology. The PNN (Masi et al. 2016), which takes inspiration from the SRCNN (Dong et al. 2016) and employs a similar network architecture, is one of the first deep learning solutions in this field. Despite its simplicity, the PNN has achieved remarkable improvements in various performance metrics, showcasing the strong capabilities of deep learning. Since then, explosive pan-sharpening networks have been proposed, leveraging advanced network architectures to attain superior visual performance. However, existing pan-sharpening methods have overlooked the discrepancies between various frequency components of multi-spectral image and relied on a uniform approach across the entire image, limiting the potential for further spatial detail enhancement. As shown in the previous study (Fuoli, Van Gool, and Timofte 2021), there is a significant correlation between super-resolution and frequency information. Considering that pan-sharpening is essentially a super-resolution process, it is reasonable to investigate how the interaction between various frequency components in two modal images can be utilized to improve the performance of pan-sharpening models.

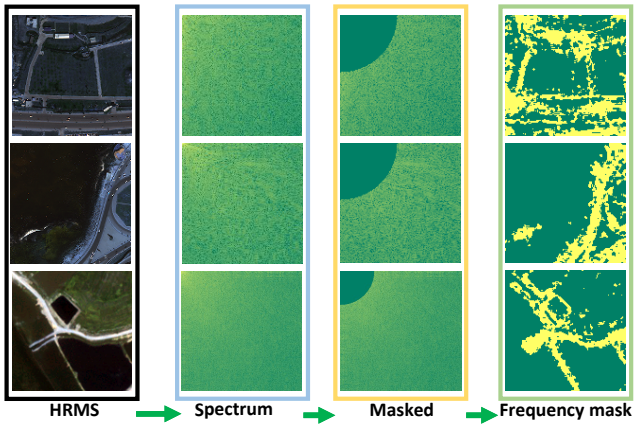

Figure 1: Generation process of frequency mask. Firstly, a discrete cosine transform is applied to the image. Then, the upper left part of the DCT spectrum is masked using manually selected thresholds. Finally, the frequency mask is generated through inverse transformation.

*Co-first authors contributed equally. † Corresponding author.

Our motivation. Our goal is to improve the performance of pan-sharpening methods by effectively recovering high-frequency information, benefiting for generating clear images with fine textures. Previous convolution network-based approaches have struggled to learn high-frequency details, as CNNs are inherently inclined towards low-frequency information (Magid et al. 2021). Recovering high-frequency information is of great importance in generating clear images. The discrete cosine transform (DCT) (Ahmed, Natarajan, and Rao 1974; Xie et al. 2021) provides a powerful tool for frequency domain analysis of images, as illustrated in Figure 1. Initially, we apply the DCT to the image to obtain the second column, where the low-frequency components are concentrated in the upper left corner of the DCT image. Subsequently, we obtain the frequency mask of the image by masking the upper left corner and employing the inverse discrete cosine transform. As shown in the fourth column of the Figure 1, the frequency mask decomposes the original image into high-frequency and low-frequency parts. This characteristic enables different modules of the network to focus on the high and low frequency parts of the image separately, explicitly encouraging the network to learn high-frequency information and generate pan-sharpened images with clear textures. Considering the significant variability in content among different remote sensing images, utilizing a dynamic network structure can enhance the model's generalization performance. The Mixture of Experts (MOE) (Jordan and Jacobs 1994) has demonstrated efficacy in various vision tasks by leveraging expert knowledge of different parts and employing a dynamic network structure. By utilizing frequency experts to facilitate the separate learning of high- and low-frequency information and adapting to different inputs through a dynamic network structure, we can significantly enhance the performance of the pan-sharpening model.

Taking into account the above-discussed insights, we present an innovative Frequency Adaptive Mixture of Experts network, named FAMEnet. By blending the MOE technique with frequency domain information, it is able to guide the network to learn image features at different frequencies, particularly high-frequency information. Furthermore, by utilizing dynamic network structures, our proposed FAMEnet can adapt to remote sensing images with significant content variance, thereby enhancing its generalization ability. The FAMEnet comprises three key modules: Frequency Mask predictor, Sub-frequency learning experts module, and Experts Mixture module. The Mask predictor is responsible for generating frequency masks that segregate the image into high-frequency and low-frequency parts, thus enabling the effective processing of the image content. The Frequency experts consist of two MOE components, namely low-frequency MOE and high-frequency MOE, which are exclusively utilized for processing low-frequency and high-frequency information of the image. With the aid of the expert network, it can distinctly focus on the high- and low-frequency components of the image to achieve targeted processing. The final experts mixture part is responsible for dynamically fusing high- and low-frequency features, as well as PAN and MS features, to adapt to remote sensing images with significant content variations. The final output is obtained by dynamically adding multiple different frequency experts. By encouraging the network to process high- and low-frequency information separately and dynamically fuse features, the generated images have clearer textures and better generalization.

Our contribution can be summarized as follows: 1) In this work, we devised a method that combines MOE (Mixture of Experts) with frequency domain information. In this way, we enable the network to learn and adapt to the high-frequency information present in remote sensing images in a dynamic manner. 2) The proposed method comprises of a frequency separation mask predictor, MOE-based frequency adapatively learning module, and experts mixture module. This design allows the pan-sharpening network to effectively capture high-frequency information, leading to high-quality pan-sharpening results. 3) Our proposed Mixture of Experts framework surpasses existing methods and achieves state-of-the-art results in pan-sharpening. The output is characterized by clear textures, accurate spectra, and strong generalization ability, as evidenced by qualitative and quantitative experiments conducted on multiple datasets.

Related Work

Pan-sharpening

A plethora of research has emerged in the community of pan-sharpening. Existing methods can be classified into traditional and deep learning-based approaches. Traditional methods include component substitution-based (Haydn et al. 1982; Gillespie, Kahle, and Walker 1987; Laben and Brower 2000; Liao et al. 2017), multi-resolution analysis-based (Mallat 1989; Nunez et al. 1999; Vivone et al. 2014; Schowengerdt 1980), and model-based methods (Fasbender, Radoux, and Bogaert 2008; Palsson, Sveinsson, and Ulfarsson 2013). However, these methods are limited by insufficient feature representation, and it is difficult to achieve satisfactory results. The success of convolutional neural networks has sparked interest in the field of pan-sharpening. PNN (Masi et al. 2016) was the first to introduce CNNs, which achieved significant improvements compared to traditional methods. PANNET (Yang et al. 2017) further improved the performance by introducing residual design. Since then, more complex designs and deeper networks have been used to enhance the performance of pan-sharpening task, such as MSDCNN (Yuan et al. 2018) for capturing multi-scale information and SRPPNN (Cai and Huang 2021) with a very deep super-resolution architecture. Recently, GPPNN (Xu et al. 2021) and MMNet (Yan et al. 2022b) were designed to enhance interpretability through deep unrolling methods. ARFNet (Yan et al. 2022a) further explored the convergence of the unrolling process. MutNet (Zhou et al. 2022c) introduced information theory to minimize mutual information redundancy. Inspired by the wide-spread application of transformer, INN-former (Zhou et al. 2022a) combines CNN and Transformer to promote the combination of local and global information. SFINet (Zhou et al. 2022b) utilizes the Fourier transform to implicitly learn high-frequency features, yet it lacks explicit incentives for

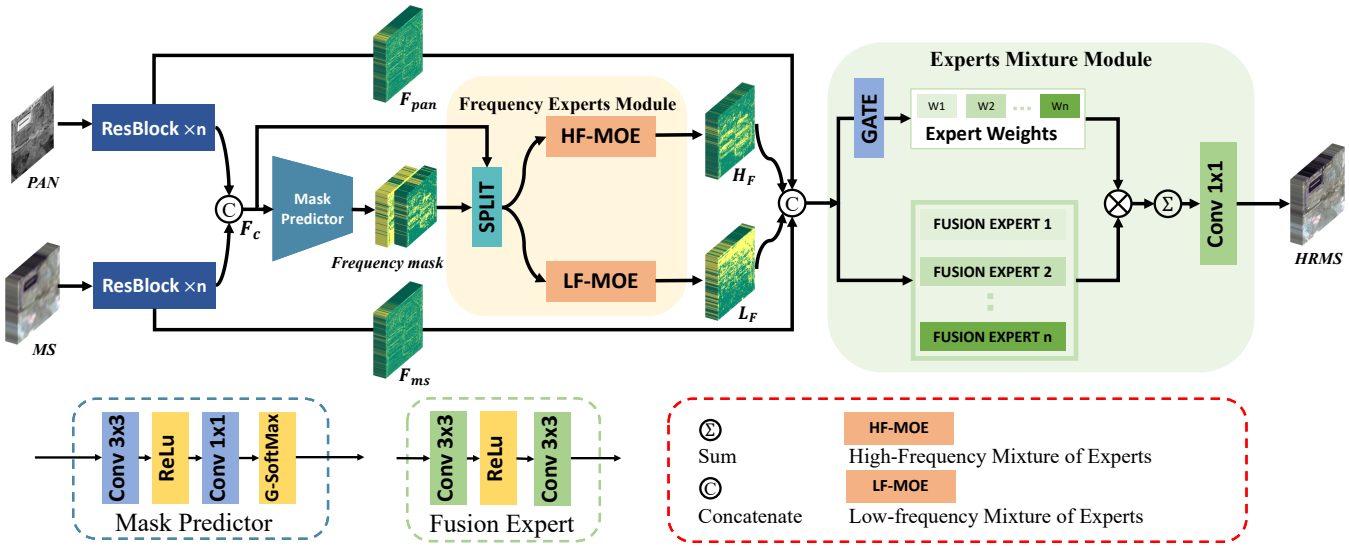

Figure 2: The overall structure of FAMEnet, which is composed of three main components: Mask predictor, Frequency Experts Module, and Experts Mixture Module.

the network to effectively harness this information, leading to suboptimal outcomes. These methods are limited in their ability to leverage high-frequency information, resulting in less clear generated textures.

Mixture of Experts

MOE (Jordan and Jacobs 1994; Gross, Ranzato, and Szlam 2017) is a widely-used technique that follows the divide-and-conquer strategy to decompose tasks into multiple parts and utilizes task-specific experts to handle them, with the final output obtained by weighting the experts. MOE's gate network can dynamically adjust the network structure according to the inputs, making it highly generalizable and widely applicable in various domains, including natural language processing (Shazeer et al. 2017), image classification (Zhang et al. 2019), and Re-ID (Dai et al. 2021) and image fusion (Cao et al. 2023). In contrast to these approaches, we propose dividing images into frequency specialists to encourage the network to capture high-frequency information. This represents the first attempt to employ MOE in the pan-sharpening community.

Method

Our proposed methodology involves utilizing DCT to generate frequency masks, and employing the FAMEnet network architecture. This section presents a brief overview of DCT, followed by a detailed description of the network structure used in this paper.

Discrete Cosine Transform

The DCT is a valuable tool for analyzing frequency domains, offering several advantages over the commonly used Fourier transform. With its simpler form and superior energy compression characteristics, the DCT enables the concentration of most of the energy in a few small coefficients. As a result, the DCT is highly suitable for both image compression and image frequency division. Given an image $x \in \mathbb{R}^{H \times W}$, its cosine transformation process can be defined as:

$$\mathrm{D(u,v)} = \sum_{h=0}^{\mathrm{H}-1} \sum_{w=0}^{\mathrm{W}-1} x_{h,w} \cos\left(\frac{\pi u}{H}\left(h+\frac{1}{2}\right)\right) \cos\left(\frac{\pi v}{W}\left(w+\frac{1}{2}\right)\right) \tag{1}$$

As illustrated in Figure 1, the image is transformed to the frequency domain via cosine transformation. This transformation results in the majority of energy being concentrated in the upper left corner of the frequency domain, which represents the low-frequency component, with the remaining high-frequency portion elsewhere. To generate the corresponding frequency mask, we remove the low-frequency component using a manually selected radius, and then perform an inverse transformation. By utilizing frequency masks, the network can effectively concentrate on the high-frequency information and learn fine-grained details. However, since these masks are generated based on manually selected threshold values, they are not robust to different image content and are sensitive to noise. To address this issue, we propose to learn frequency masks from the network.

Network Framework

The overall architecture of the network is depicted in Figure 2. The input comprises upsampled LRMS and PAN images, from which we extract the features using ResBlock to obtain $\mathbf{F}_{\mathrm{ms}}$ and $\mathbf{F}_{\mathrm{pan}}$. These features are concatenated to yield F_c, which is passed through the mask predictor for frequency mask prediction. Subsequently, the frequency mask $\mathbf{M} \in \mathbb{R}^{\mathrm{H} \times \mathrm{W} \times 2}$ and $\mathbf{F}_{\mathrm{c}}$ are fed into the Frequency experts module, where $\mathbf{F}_{\mathrm{c}}$ is separated based on the frequency mask using LF-MOE and HF-MOE to process low and high frequency features, respectively. Finally, the output of the Frequency experts module is combined with $\mathbf{F}_{\mathrm{ms}}$ and $\mathbf{F}_{\mathrm{pan}}$ and

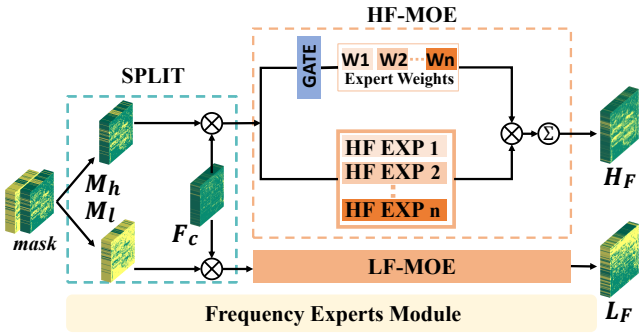

Figure 3: The architecture of the Frequency Experts Module. The frequency mask splits F_c into high-frequency and low-frequency parts, which are processed separately by HF-MOE and LF-MOE.

works with the frequency experts mixture to generate HRMS images.

Key Components

Mask Predictor. Our network comprises a lightweight mask predictor module, as illustrated in Figure 2. The mask predictor learns high-frequency and low-frequency components adaptively based on the image content for generating frequency masks. We utilized Gumbel-Softmax (Jang, Gu, and Poole 2017) to ensure differentiability in mask prediction. The frequency mask M for the input $\mathbf{F}_\mathrm{c}$ can be generated using the following process:

$$\mathbf{P} = \mathbf{C}_1 \circ \mathbf{ReLu} \circ \mathbf{C}_3(\mathbf{F}_\mathrm{c}) \tag{2}$$

$$\mathbf{M} = \mathbf{GumbelSoft}(\mathbf{P}) \tag{3}$$

Here, $\mathbf{C}_1$ and $\mathbf{C}_3$ are convolution blocks with kernel sizes of 1×1 and 3×3, respectively. The Gumbel-Softmax function $\mathbf{GumbelSoft}(.)$ is applied to generate masks for the high-frequency and low-frequency components, which are represented by the 2 channels of $\mathbf{M} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times 2}$, and $\mathbf{P} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times 2}$ is intermediate feature for generating M. The use of the Gumbel Softmax technique ensures the differentiability of the mask generation process, as opposed to the non-differentiable $\mathbf{argmax}$ operation.

Specifically, Gumbel Softmax can be expressed as follows:

$$\mathbf{Z}_\mathrm{i} = \frac{\exp((\mathbf{P}_\mathrm{i} + \mathbf{g}_\mathrm{i})/\tau)}{\sum_{\mathrm{c}=1}^{\mathrm{C}} \exp((\mathbf{P}_{\mathrm{i,c}} + \mathbf{g}_{\mathrm{i,c}})/\tau)} \tag{4}$$

$$\mathbf{M}_\mathrm{i} = \zeta \circ \arg\max_{\mathrm{c}} \mathbf{Z}_{\mathrm{i,c}} \tag{5}$$

Here, $\mathbf{C}$ corresponds to the number of channels in $\mathbf{P}$, which is 2. $\mathbf{g}_i$ is the noise sampled from the Gumbel distribution, and τ is the temperature coefficient. Additionally, the function $\zeta(.)$ is utilized to represent the onehot encoding. Two channels are generated in $\mathbf{M}$, one for the high-frequency mask and the other for the low-frequency mask. During the backward process, we utilize the gradient of $\mathbf{Z}$ as the gradient of $\mathbf{M}$ approximately, since $\mathbf{M}$ is non-differentiable.

Frequency Experts Module. The targeted processing of high and low frequency components of the image can enhance the network's ability to capture frequency domain information, as illustrated in Figure 3. Our frequency experts module, which comprises split, LF-MOE (low frequency - mixture of experts), and HF-MOE (high frequency - mixture of experts), performs this task. The split operation separates the input into high-frequency and low-frequency components based on the frequency mask, which are then processed by HF-MOE and LF-MOE, respectively, to extract high and low-frequency features. The HF expert in HF-MOE consists of HIN (Half-Instance Normalization) (Chen et al. 2021) blocks, while the LF expert in LF-MOE uses 3x3 convolutions. To handle the complexity of high-frequency feature extraction, more sophisticated modules are used in the HF expert. Adaptive adjustments are made to the dynamic weights of HF-MOE and LF-MOE to choose the most suitable experts for processing the high and low frequency information that varies significantly across remote sensing images. To define the split process, we begin with the input $\mathbf{M}$ and $\mathbf{F}_\mathrm{c}$. The process is as follows:

$$\mathbf{M}_\mathrm{h}, \mathbf{M}_\mathrm{l} = \mathbf{M}(:,:,0), \mathbf{M}(:,:,1) \tag{6}$$

$$\mathbf{F}_\mathrm{h}, \mathbf{F}_\mathrm{l} = \mathbf{M}_\mathrm{h} \odot \mathbf{F}_\mathrm{c}, \mathbf{M}_\mathrm{l} \odot \mathbf{F}_\mathrm{c} \tag{7}$$

The mask $\mathbf{M}$ comprises two channels that correspond to high-frequency and low-frequency masks, respectively. By multiplying $\mathbf{F}_\mathrm{c}$ with these masks, we obtain the high-frequency and low-frequency components of $\mathbf{F}_\mathrm{c}$, denoted by $\mathbf{F}_\mathrm{h}$ and $\mathbf{F}_\mathrm{l}$. We then feed these two sets of features into HF-MOE and LF-MOE, respectively, to obtain high-frequency and low-frequency features $\mathbf{H}_\mathrm{F}$ and $\mathbf{L}_\mathrm{F}$. HF-MOE and LF-MOE share the similar structure. To be specific, the high-frequency feature extraction process for HF-MOE is defined as follows:

$$\mathbf{W}_\mathrm{h} = \mathbf{Gate}(\mathbf{F}_\mathrm{h}) \tag{8}$$

$$\mathbf{H}_\mathrm{F} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbf{W}_\mathrm{h}^\mathrm{i} \cdot \phi_\mathbf{i}(\mathrm{F}_\mathrm{h}) \tag{9}$$

In this context, the function $\mathbf{Gate}(.)$ produces the gate weights $\mathbf{W}_\mathrm{h} \in \mathbb{R}^\mathrm{N}$, which are then utilized as the weighting coefficients for a linear combination of the different HF-EXPERT outputs. Here, $\phi_\mathbf{i}(.)$ represents the i-th HF-EXPERT block. Depending on the input, the $\mathbf{Gate}(.)$ generates different weights, allowing the network structure to be dynamically adjusted. The details of $\mathbf{Gate}(.)$ will be discussed in the Experts Mixture part.

Experts Mixture Module. We have designed the Experts Mixture module, as shown in Figure 4, which adopts the MOE architecture and leverages multiple frequency learning experts to fuse input features and adapt to the diverse content of remote sensing images. Gate generates different weights for feature fusion, selecting the most suitable experts based on the input feature. To prevent homogenization among experts, the gate generates sparse weights. Given the input feature $\mathbf{F}_\mathrm{f}$, which is obtained by concatenating $\mathbf{F}_\mathrm{ms}$, $\mathbf{F}_\mathrm{pan}$, $\mathbf{L}_\mathrm{F}$, and $\mathbf{H}_\mathrm{F}$, the weight generation process is defined

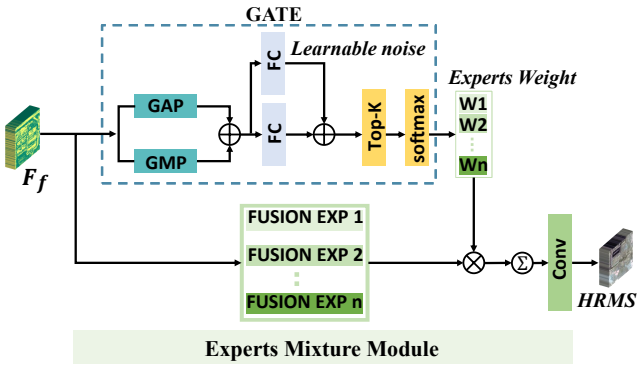

Figure 4: The architecture of Experts Mixture module, which includes the gating mechanism responsible for generating sparse weights based on input features, and the selection of appropriate fusion expert outputs based on the weights.

as follows:

$$\mathbf{F}_{\mathrm{e}} = \mathbf{GAP}(\mathbf{F}_{\mathrm{f}}) + \mathbf{GMP}(\mathbf{F}_{\mathrm{f}}) \tag{10}$$

$$\epsilon = \mathbf{SoftPlus}(\mathbf{A}_1 \times \mathbf{F}_{\mathrm{e}}) \tag{11}$$

$$\mathbf{V} = \mathbf{A}_2 \times \mathbf{F}_{\mathrm{e}} + \epsilon \tag{12}$$

$$\mathbf{W}_{\mathrm{f}} = \mathbf{Softmax} \circ \mathrm{Topk}(\mathbf{V}) \tag{13}$$

Here, $\mathbf{GAP}(.)$ and $\mathbf{GMP}(.)$ correspond to average pooling and maximum pooling operations, respectively, while $\mathbf{A}_1$ and $\mathbf{A}_2$ are learnable matrices, specifically the fully connected layers shown in the Figure 4. First, we process the features using $\mathbf{GAP}(.)$ and $\mathbf{GMP}(.)$, and then we sum them to obtain $\mathbf{F}_{\mathrm{e}}$. Next, $\mathbf{F}_{\mathrm{e}}$ is passed through a fully connected layer, and learnable noise ϵ is added to produce $\mathbf{V} \in \mathbb{R}^{\mathrm{N}}$. By applying the Top-K operation, we select the k positions with the highest weight value from among the n weights, assign negative infinity to the remaining positions, and finally obtain the expert weight $\mathbf{W}_{\mathrm{f}}$ using $\mathbf{softmax}(.)$, with unselected expert weights reset to 0. The use of learnable noise ensures that the probability of each expert being selected can be made almost equal.

To obtain the output HRMS image, we first multiply the output of Expert weight and frequency learning experts, and then adjust the channel through convolution. The process can be defined as follows:

$$\mathbf{HRMS} = \mathbf{C}_1 \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbf{W}_{\mathrm{f}}^{\mathrm{i}} \cdot \psi_{\mathbf{i}}(\mathbf{F}_{\mathrm{f}}) \tag{14}$$

Here, $\psi_{\mathbf{i}}$ represents the i-th fusion expert, and $\mathbf{C}_1$ is the convolution block with a 1×1 kernel size.

Loss Function

Our loss function comprises three parts: reconstruction loss, mask loss, and load loss. Reconstruction loss is used to minimize the difference between the model output and the ground truth. Mask loss is used to learn the appropriate frequency mask, and load loss balances the load of different experts in MOE and prevents some experts from being ignored during training. Let the model output be denoted as Y, the ground truth as G, and the reconstruction loss as the L1 loss between the two, given by:

$$\mathcal{L}_{rec} = ||\mathbf{Y} - \mathbf{G}||_1 \tag{15}$$

For the mask learning process, we first follow the procedure shown in Figure 1 and use manually selected rough thresholds to generate the frequency mask label of the training data, which we refer to as M_{gt}. We minimize the L1 distance between the mask output from the mask predictor and M_{gt}. We adopt an annealing strategy to adjust the weight of the mask loss, ensuring that the network no longer relies on the mask label to generate more accurate masks after learning mask information. The mask loss is defined as follows:

$$\mathcal{L}_{mask} = ||\mathbf{M} - \mathbf{M}_{\mathrm{gt}}||_1 \tag{16}$$

To balance the load of experts, we use the square of the coefficient of variation (SCV) as the load loss. Given the weight W, SCV can be computed as:

$$\mathbf{SCV}(\mathrm{W}) = (\sigma(\mathrm{W})/\bar{\mathrm{W}})^2 \tag{17}$$

Here, $\sigma(W)$ and $\bar{W}$ denote the standard deviation and mean of the elements in the weight vector, respectively.
To balance the workload of the experts in our network, we use a load loss that is the sum of the SCV values of their respective weight vectors. This is given by:

$$\mathcal{L}_{\mathrm{load}} = \mathbf{SCV}(\mathbf{W}_{\mathrm{h}}) + \mathbf{SCV}(\mathbf{W}_{\mathrm{l}}) + \mathbf{SCV}(\mathbf{W}_{\mathrm{f}}) \tag{18}$$

where $\mathbf{W}_{\mathrm{h}}$, $\mathbf{W}_{\mathrm{l}}$, and $\mathbf{W}_{\mathrm{f}}$ represent the weight vectors of the HF-MOE, LF-MOE, and Experts Mixture module, respectively.

The total loss function is given by:

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \alpha\mathcal{L}_{\mathrm{mask}} + \beta\mathcal{L}_{\mathrm{load}} \tag{19}$$

where the initial value of α is 0.001, which is attenuated using an annealing strategy. After 70% of the training epochs, α decreases to 0 while β is set to 0.1.

Experiments

Datasets and Benchmark

Our experiments were conducted on three typical datasets, namely WorldView-II (WV2), Gaofen2 (GF2), and WorldView-III (WV3), which include various natural and urban scenarios. Since ground truth is unavailable, we follow the Wald protocol (Wald, Ranchin, and Mangolini 1997) to generate training samples. We compare our proposed method with cutting-edge deep learning methods, including PANNET, MSDCNN, SRPPNN, GPPNN, MutNet, INN-former and SFINet and some classic methods such as GFPCA (Liao et al. 2017), GS (Laben and Brower 2000), IHS (Haydn et al. 1982), Brovey (Gillespie, Kahle, and Walker 1987) and SFIM (Liu. 2000).

Implement Details

We trained our model using the Python framework on an RTX 3060 GPU, with four experts (n=4), k=2 for each MOE, and a total of 1000 epochs. We used the Adam optimizer with a learning rate of 5e-4 and linearly decreased the

Method	WorldView-II				GaoFen2				Worldview-III			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
PANNET	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	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	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	41.1622	0.9684	0.0244	1.0315	44.2145	0.9815	0.0137	0.7361	30.1785	0.9175	0.0776	3.2593
MutNet	41.6773	0.9705	0.0224	0.9519	47.3042	0.9892	0.0102	0.5481	30.4907	0.9223	0.0749	3.1125
INNformer	41.6903	0.9704	0.0227	0.9514	47.3528	0.9893	0.0102	0.5479	30.5365	0.9225	0.0747	3.0997
SFINet	41.7244	**0.9725**	0.0220	0.9506	47.4712	**0.9901**	0.0102	0.5479	30.5901	0.9236	0.0741	3.0798
Ours	**42.0262**	0.9723	**0.0215**	**0.9172**	**47.6721**	0.9898	**0.0098**	**0.5242**	**30.9903**	**0.9287**	**0.0697**	**2.9531**

Table 1: Quantitative comparison on three datasets. Best results highlighted in bold. ↑ indicates that the larger the value, the better the performance, and ↓ indicates that the smaller the value, the better the performance.

Metric	Brovey	GS	IHS	GFPCA	PNN	PANNET	MSDCNN	SRPPNN	GPPNN	MutNet	INNformer	SFINet	Ours
$D_\lambda \downarrow$	0.1378	0.0696	0.0770	0.0914	0.0746	0.0737	0.0734	0.0767	0.0782	0.0694	0.0697	0.0681	**0.0674**
$D_S \downarrow$	0.2605	0.2456	0.2985	0.1635	0.1164	0.1224	0.1151	0.1162	0.1253	**0.1118**	0.1128	0.1119	0.1121
QNR ↑	0.6390	0.0725	0.6485	0.7615	0.8191	0.8143	0.8215	0.8173	0.8073	0.8259	0.8253	0.8276	**0.8291**

Table 2: Evaluation of the proposed method on real-world full-resolution scenes from the GaoFen2 dataset.

weight of the loss (α) during training. For training samples, we cropped LRMS patches of size 32x32 and PAN patches of size 128x128 from the images, with a batch size of 4. We have employed a comprehensive set of evaluation metrics to assess the performance of our approach, including well-established measures such as PSNR, SSIM, SAM (J. R. H. Yuhas and Boardman 1992), and ERGAS, as well as non-reference metrics such as D_s, D_λ and QNR to evaluate the generalization performance of our model.

Comparison with State-of-the-Art Methods

Evaluation on Reduced-Resolution Scene. The evaluation results on the three datasets are presented in Table 1, clearly demonstrating the superior performance of our proposed method over the SOTA methods in all metrics. Our model exhibits significant improvements in PSNR across all three datasets, with 0.301dB improvement on WV2 and 0.4dB improvement on WV3 compared to the INNformer, respectively. These results validate the consistency of our method with the ground truth, and other metrics further confirm the effectiveness of our approach. Qualitative experiments in Figure 5 showcase representative samples from the WV3 datasets, where the residual plot of our method has the least brightness, indicating its closeness to the ground truth. Our method provides clear edges and accurate spectra, further highlighting its superiority over other methods.

Evaluation on Full-Resolution Scene. To assess the generalization ability of our method, we evaluated it on the full Gaofen2 dataset using no-reference metrics. This dataset consists of images from the reserved part of the Gaofen2 dataset, which were not downsampled. The experimental results, as shown in Table 2, demonstrate that our method outperformed other approaches on all three metrics, indicating the exceptional adaptability of the MOE architecture to remote sensing images.

Ablation Experiments

The core components of our network comprise of the Mask Predictor, Frequency Experts module, and Experts Mixture module. The former two are responsible for improving the network's frequency domain perception, while the latter enables dynamic feature fusion. We conducted two sets of ablation experiments on three datasets, the results of which are presented in Table 3.

Mask Predictor. The mask predictor serves as the core component for frequency domain perception. In the first set of experiments, we conducted ablation by removing the mask predictor and the split operation, and directly feed the F_c into LF-MOE and HF-MOE. Due to the loss of the frequency mask, both of them were unable to process high and low-frequency information in a targeted manner. The experimental results, shown in the first row of Table 3, demonstrate a significant decrease in various indicators, proving that the targeted processing of high-frequency and low-frequency information can promote the network's learning of high-frequency information to improve detail perception.

Experts Mixture Module. In the second set of experiments, we replaced the Experts Mixture module with a resblock having similar parameter number for feature fusion, thereby eliminating the dynamic feature fusion capability of the network. The experimental results in the second row of the Table 3 clearly demonstrate that the deletion of the Experts Mixture module led to a decrease in various evaluation indicators on the three datasets.

Visualization of Feature Maps

To further illustrate the capabilities of our model, we have visualized the feature maps generated by our network, as presented in the Figure 6. The columns from left to right show the PAN image feature maps $\mathbf{F}_{\mathrm{pan}}$, MS image feature maps $\mathbf{F}_{\mathrm{ms}}$, Mask predictor output $\mathbf{M}$, HF-MOE output $\mathbf{H}_{\mathrm{F}}$, LF-MOE output $\mathbf{L}_{\mathrm{F}}$, and Experts Mixture module output be-

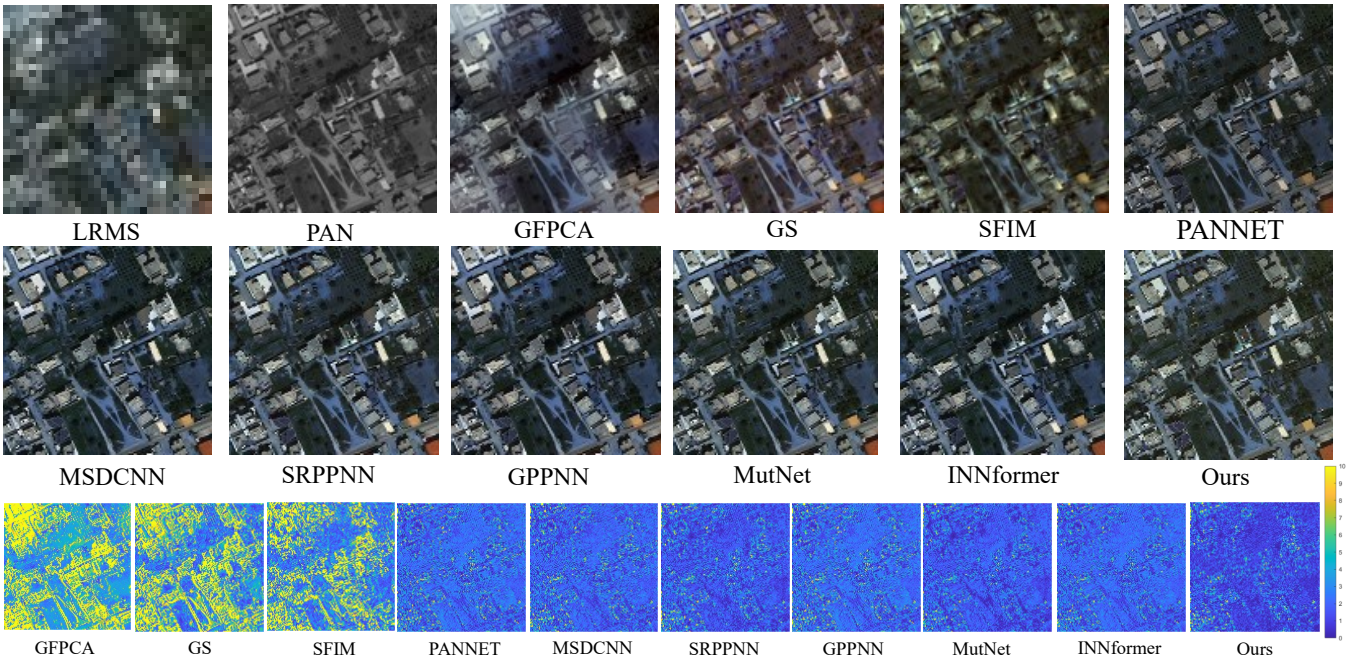

Figure 5: The result of our approach was compared against nine other methods on WorldView-III dataset.

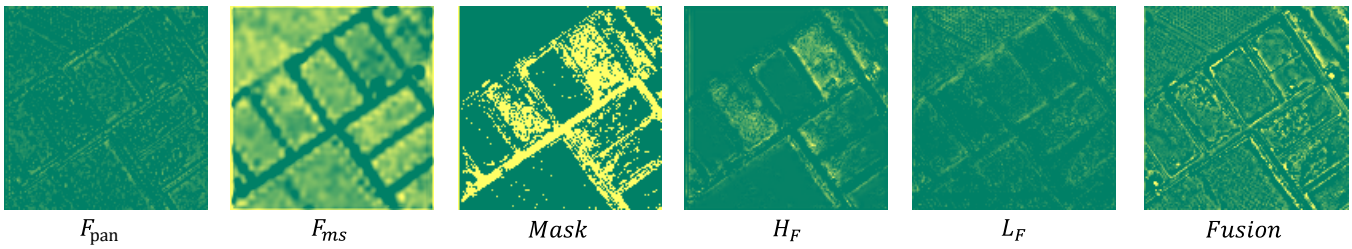

Figure 6: The network's feature map.

Config	Mask	Experts Mixture	WorldView-II				GaoFen2				WorldView-III			
			PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
(I)	✗	✓	41.8998	0.9720	0.0220	0.9322	47.5914	0.9896	0.0100	0.5312	30.8387	0.9261	0.0720	2.9969
(II)	✓	✗	41.8274	0.9714	0.0220	0.9358	47.3596	0.9888	0.0102	0.5429	30.7930	0.9261	0.0717	3.0134
Ours	✓	✓	**42.0261**	**0.9723**	**0.0215**	**0.9172**	**47.6721**	**0.9898**	**0.0098**	**0.5242**	**30.9903**	**0.9287**	**0.0697**	**2.9531**

Table 3: The results of the ablation experiments conducted on the three datasets.

fore channel adjustment, respectively. By observing the feature maps, it is evident that the mask predicted by the network can accurately distinguish between the high and low frequency components of remote sensing images, which is much more precise than the masks generated by the artificial threshold selection method. Moreover, it is clear from the $\mathbf{H}_{\mathrm{F}}$ and $\mathbf{L}_{\mathrm{F}}$ feature maps that the HF-MOE and LF-MOE components specifically learn the high and low frequency information of the image, respectively. Finally, the Experts Mixture module effectively integrates all the information. These feature maps demonstrate the targeted processing of information at different frequencies by our network.

Conclusion

This work introduces a new approach that employs a frequency mask for managing high and low-frequency data and a dynamic structure to adjust to the various content of remote sensing images. The proposed method utilizes a Frequency Adaptive Mixture of Experts (MOE) network, which is designed to target both high and low-frequency data while employing a dynamic network structure. Notably, our research is the first to apply the MOE structure in pan-sharpening. Extensive experiments indicate that our model outperforms state-of-the-art methods and exhibits robust generalization capabilities.

Acknowledgements

This work was supported by the Natural Science Foundation of Anhui Province (No.2208085MC57), and HFIPS Director's Fund, Grant No.2023YZGH04.

References

Ahmed, N.; Natarajan, T.; and Rao, K. R. 1974. Discrete cosine transform. *IEEE transactions on Computers*, 100(1): 90–93.

Cai, J.; and Huang, B. 2021. Super-Resolution-Guided Progressive Pansharpening Based on a Deep Convolutional Neural Network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6): 5206–5220.

Cao, B.; Sun, Y.; Zhu, P.; and Hu, Q. 2023. Multi-Modal Gated Mixture of Local-to-Global Experts for Dynamic Image Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23555–23564.

Chen, L.; Lu, X.; Zhang, J.; Chu, X.; and Chen, C. 2021. HINet: Half Instance Normalization Network for Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 182–192.

Dai, Y.; Li, X.; Liu, J.; Tong, Z.; and Duan, L.-Y. 2021. Generalizable person re-identification with relevance-aware mixture of experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16145–16154.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2016. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2): 295–307.

Fasbender, D.; Radoux, J.; and Bogaert, P. 2008. Bayesian data fusion for adaptable image pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 46(6): 1847–1857.

Fuoli, D.; Van Gool, L.; and Timofte, R. 2021. Fourier space losses for efficient perceptual image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2360–2369.

Gillespie, A. R.; Kahle, A. B.; and Walker, R. E. 1987. Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques - ScienceDirect. *Remote Sensing of Environment*, 22(3): 343–365.

Gross, S.; Ranzato, M.; and Szlam, A. 2017. Hard mixtures of experts for large scale weakly supervised vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6865–6873.

Haydn, R.; Dalke, G. W.; Henkel, J.; and Bare, J. E. 1982. Application of the IHS color transform to the processing of multisensor data and image enhancement. *National Academy of Sciences of the United States of America*, 79(13): 571–577.

J. R. H. Yuhas, A. F. G.; and Boardman, J. M. 1992. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. *Proc. Summaries Annu. JPL Airborne Geosci. Workshop*, 147–149.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*.

Jordan, M. I.; and Jacobs, R. A. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2): 181–214.

Laben, C.; and Brower, B. 2000. Process for Enhancing the Spatial Resolution of Multispectral Imagery Using Pan-Sharpening. *US Patent 6011875A*.

Liao, W.; Xin, H.; Coillie, F. V.; Thoonen, G.; and Philips, W. 2017. Two-stage fusion of thermal hyperspectral and visible RGB image by PCA and guided filter. In *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*.

Liu., J. G. 2000. Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing*, 21(18): 3461–3472.

Magid, S. A.; Zhang, Y.; Wei, D.; Jang, W.-D.; Lin, Z.; Fu, Y.; and Pfister, H. 2021. Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4288–4297.

Mallat, S. 1989. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7): 674–693.

Masi, G.; Cozzolino, D.; Verdoliva, L.; and Scarpa, G. 2016. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7): 594.

Nunez, J.; Otazu, X.; Fors, O.; Prades, A.; Pala, V.; and Arbiol, R. 1999. Multiresolution-based image fusion with additive wavelet decomposition. *IEEE Transactions on Geoscience and Remote sensing*, 37(3): 1204–1211.

Palsson, F.; Sveinsson, J. R.; and Ulfarsson, M. O. 2013. A new pansharpening algorithm based on total variation. *IEEE Geoscience and Remote Sensing Letters*, 11(1): 318–322.

Schowengerdt, R. A. 1980. Reconstruction of multispatial, multispectral image data using spatial frequency content. *Photogrammetric Engineering and Remote Sensing*, 46(10): 1325–1334.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations*.

Vivone, G.; Alparone, L.; Chanussot, J.; Dalla Mura, M.; Garzelli, A.; Licciardi, G. A.; Restaino, R.; and Wald, L. 2014. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5): 2565–2586.

Wald, L.; Ranchin, T.; and Mangolini, M. 1997. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric Engineering and Remote Sensing*, 63: 691–699.

Xie, W.; Song, D.; Xu, C.; Xu, C.; Zhang, H.; and Wang, Y. 2021. Learning frequency-aware dynamic network for

efficient super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4308–4317.

Xu, S.; Zhang, J.; Zhao, Z.; Sun, K.; Liu, J.; and Zhang, C. 2021. Deep Gradient Projection Networks for Pan-sharpening. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1366–1375.

Yan, K.; Zhou, M.; Huang, J.; Zhao, F.; Xie, C.; Li, C.; and Hong, D. 2022a. Panchromatic and Multispectral Image Fusion via Alternating Reverse Filtering Network. *Advances in Neural Information Processing Systems*, 35: 21988–22002.

Yan, K.; Zhou, M.; Zhang, L.; and Xie, C. 2022b. Memory-Augmented Model-Driven Network for Pansharpening. In *European Conference on Computer Vision*, 306–322. Springer.

Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; and Paisley, J. 2017. PanNet: A deep network architecture for pan-sharpening. In *IEEE International Conference on Computer Vision*, 5449–5457.

Yuan, Q.; Wei, Y.; Meng, X.; Shen, H.; and Zhang, L. 2018. A Multiscale and Multidepth Convolutional Neural Network for Remote Sensing Imagery Pan-Sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3): 978–989.

Zhang, L.; Huang, S.; Liu, W.; and Tao, D. 2019. Learning a mixture of granularity-specific experts for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8331–8340.

Zhou, M.; Huang, J.; Fang, Y.; Fu, X.; and Liu, A. 2022a. Pan-sharpening with customized transformer and invertible neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3553–3561.

Zhou, M.; Huang, J.; Yan, K.; Yu, H.; Fu, X.; Liu, A.; Wei, X.; and Zhao, F. 2022b. Spatial-frequency domain information integration for pan-sharpening. In *European Conference on Computer Vision*, 274–291. Springer.

Zhou, M.; Yan, K.; Huang, J.; Yang, Z.; Fu, X.; and Zhao, F. 2022c. Mutual information-driven pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1798–1808.