Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)

Hybrid Frequency Modulation Network for Image Restoration

Yuning Cui[1], **Mingyu Liu**[1], **Wenqi Ren**[2*] and **Alois Knoll**[1]
[1]Technical University of Munich
[2]Shenzhen Campus of Sun Yat-sen University
{yuning.cui, liumi, knoll}@in.tum.de, renwq3@mail.sysu.edu.cn

Abstract

Image restoration involves recovering a high-quality image from its corrupted counterpart. This paper presents an effective and efficient framework for image restoration, termed CSNet, based on "channel + spatial" hybrid frequency modulation. Different feature channels include different degradation patterns and degrees, however, most current networks ignore the importance of channel interactions. To alleviate this issue, we propose a frequency-based channel feature modulation module to facilitate channel interactions through the channel-dimension Fourier transform. Furthermore, based on our observations, we develop a multi-scale frequency-based spatial feature modulation module to refine the direct-current component of features using extremely lightweight learnable parameters. This module contains a densely connected coarse-to-fine learning paradigm for enhancing multi-scale representation learning. In addition, we introduce a frequency-inspired loss function to achieve omni-frequency learning. Extensive experiments on nine datasets demonstrate that the proposed network achieves state-of-the-art performance for three image restoration tasks, including image dehazing, image defocus deblurring, and image desnowing. The code and models are available at https://github.com/c-yn/CSNet.

1 Introduction

As a fundamental task in computer vision, image restoration involves recovering a high-quality image from its corrupted counterpart by removing degradations and restoring content details [Cui *et al.*, 2023b]. As this task plays an important role in many fields, such as transportation systems, unmanned platforms, and photography, it increasingly garners attention from the industrial community and academia. Due to its ill-posedness property, manifold traditional approaches have been proposed based on various hand-crated features to reduce the solution space. However, these attempts are inapplicable when the assumption is not satisfied.

*Corresponding author

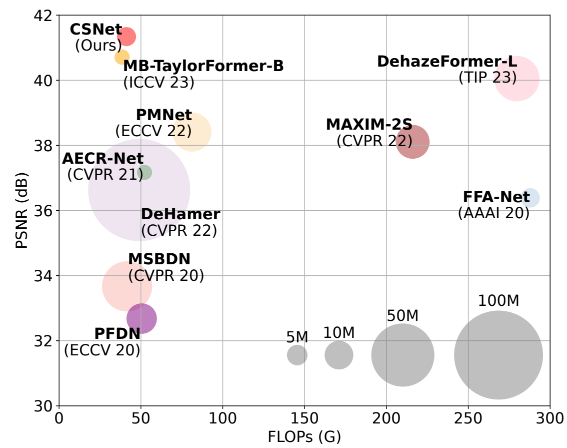

Figure 1: Comparisons between our CSNet and the state-of-the-art algorithms on the SOTS dataset [Li *et al.*, 2018] for image dehazing. The circle size indicates the number of parameters.

Recently, methods based on convolutional neural networks (CNNs) have significantly ameliorated the above issue and produced more promising results than conventional approaches. This is achieved by steering clear of human knowledge and learning generalizable priors from large-scale collected datasets. To improve the performance of these networks, many ingenious functional modules have been devised to obtain high-quality predicted images. For example, Qin *el al.* utilize various attention units, such as channel and pixel attention, for image dehazing, considering totally different weighted information contained in different channel features and uneven haze distribution among the spatial pixels [Qin *et al.*, 2020]. Son *et al.* leverage multiple atrous convolutions with different rates to deal with spatially-varying defocus blurs [Son *et al.*, 2021]. These advanced mechanisms have significantly boosted the performance of image restoration. Nevertheless, the inherent drawback of convolution, *i.e.,* local connectivity, prohibits its further applications.

Fortunately, inspired by the success of Transformer models in natural language processing and high-level vision tasks, such as object detection and segmentation, many efforts have been made to tailor Transformer for image restoration problems. For instance, Guo *et al.* first introduce the strengths of Transformer for haze removal [Guo *et al.*, 2022]. To improve the efficiency of Transformer models, some researche attempt to reduce the operation regions of self-attention [Ts *et al.*, 2022]. Zamir *et al.* creatively apply the self-attentio operator to the channel dimension rather than spatial r gions [Zamir *et al.*, 2022]. Thanks to the powerful abili of self-attention, these models can effectively capture lon range dependencies and have remarkably advanced state-c the-art performance for image restoration tasks. Nonetheles efficiency is a key factor for practical applications, and how reduce the complexity of self-attention remains a formidab challenge.

To pursue global perceptive fields while remaining high efficient, a few recent works resort to embedding frequen processing in deep networks. For example, Mao *et al.* i corporate the fast Fourier transform (FFT) into the residu block to enable both low- and high-frequency learning for ir age deblurring [Mao *et al.*, 2021]. Guo *et al.* propose a nov window-based frequency attention to solve the frequency re olution mismatch problem [Guo *et al.*, 2023]. According the Fourier theorem, these algorithms can effectively mod global signals while bridging the frequency gap between d graded and sharp image pairs. However, they only apply t Fourier transform to the spatial dimension, ignoring the ir portance of channel interactions [Zhang *et al.*, 2023].

To alleviate the above issues, we propose an efficient aı effective network based on hybrid dual-dimension frequen modulation. Concretely, to facilitate information exchan between channels that contain different degradation patterr we develop a novel frequency-based channel feature modul tion module by applying FFT to the channel dimension. E doing this, each pixel can perceive signals encompassing di ferent degradation patterns from the same location across a channels. Compared to the pure channel attention that on learns a single attention weight for each channel [Hu *et a* 2018], our approach operates at the pixel-wise granularit offering efficacy in managing spatially varying degradatior Additionally, our module can achieve channel-dimension i teractions in multiple spectral spaces by using FFT.

Furthermore, we observe that replacing the direct-curre (DC) component of the degraded image with that of t ground truth results in a sharper image, as illustrated in Fi ure 2. This fact inspires us to propose a frequency-based sp tial feature modulation module that only refines the DC cor ponent in spectra. To achieve this, we utilize the global a erage pooling technique to obtain the DC part of the featur and then apply lightweight learnable parameters to refine which bypasses the use of FFT and IFFT, saving computatio overhead. We further inject this mechanism into a dense connected coarse-to-fine paradigm for multi-scale represe tation learning. Moreover, as the DC component is a kind of low-frequency signal, to facilitate omni-frequency learning, we enhance high-frequency information learning by introducing a frequency-based loss function.

By incorporating the above designs into a U-shaped CNN, the proposed CSNet achieves state-of-the-art performance on several image restoration tasks. For dehazing, our model outperforms the recent Transformer-based method, MB-TaylorFormer-B [Qiu *et al.*, 2023] by 0.63 dB PSNR on the widely-used SOTS [Li *et al.*, 2018] dataset with similar computation overhead, as illustrated in Figure 1. Also, our CSNet shows the strong capability of defocus deblurring by providing a gain of 0.07 dB PSNR over FocalNet [Cui *et al.*, 2023a] in the combined category of the DPDD [Abuolaim and Brown, 2020] dataset. Furthermore, CSNet achieves 38.13 dB PSNR on the CSD [Chen *et al.*, 2021] dataset, an improvement of 0.95 dB over FocalNet [Cui *et al.*, 2023a].

Figure 2: Replacing the direct-current component of the hazy image with that of the ground truth leads to a cleaner result. From left to right: hazy images, the obtained results, and ground truth.

Overall, we summarize the main contributions of this article as follows:

- We propose a frequency-based channel feature modulation module to enhance channel interactions in multiple spectral spaces using the Fourier transformer, enabling each pixel to perceive different degradation patterns from other channels.
- We introduce a multi-scale frequency-based spatial feature modulation module that refines the direct-current component at multiple scales to bring the degraded image closer to ground truth. We also present a frequency-based loss for omni-frequency representation learning.
- Employing hybrid dual-dimension frequency learning, the proposed network achieves state-of-the-art performance on nine datasets for three image restoration tasks.

2 Related Work

Image Restoration Networks. As a fundamental computer vision task, image restoration involves recovering missing details and removing degradations in corrupted images [Cui *et al.*, 2023d]. Traditional approaches mainly stand on various hand-crafted features and assumptions, inevitably facing the issue of inappropriateness in practical scenarios [He *et al.*, 2010]. With the rapid development of deep learning, multifarious CNN-based frameworks have been proposed for diverse image restorations, such as image dehazing [Qin *et al.*, 2020], desnowing [Liu *et al.*, 2018], and deblurring [Son *et al.*, 2021], showcasing more promising performance than the conventional predecessors [Cui *et al.*, 2023c]. Various advanced units and modules have been devised to further boost restoration performance [Cui *et al.*, 2024]. The recent success of Transformer models in high-level vision tasks facil-

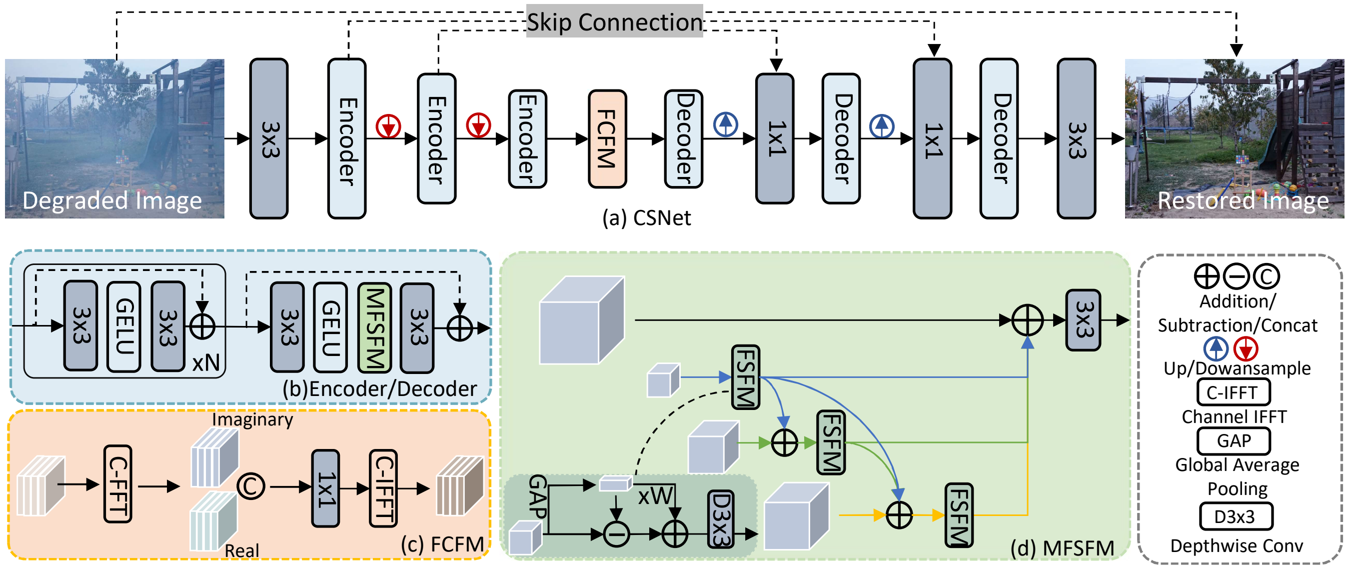

Figure 3: The architectural details of the proposed CSNet. (a) CSNet consists of three encoder blocks and three decoder blocks. (b) The Encoder/Decoder block comprises N regular residual blocks and a modified one that includes our multi-scale frequency-based spatial feature modulation module (MFSFM). (c) The frequency-based channel feature modulation module (FCFM) performs FFT across the channel dimension. (d) MFSFM refines the direct-current component of spatial features in a coarse-to-fine manner with dense connection.

itates the paradigm shift from CNN to Transformer models in image restoration [Qiu *et al.*, 2023; Guo *et al.*, 2022; Song *et al.*, 2022; Valanarasu *et al.*, 2022]. These models have significantly advanced state-of-the-art performance by effectively providing long-distance information interactions.

Attention Mechanisms. Attention mechanisms have been commonly adopted in image restoration tasks to attend to informative regions. For example, Liu *et al.* adopt the channel-wise attention mechanism to flexibly fuse features from different scales for image dehazing [Liu *et al.*, 2019]. Zamir *et al.* leverage the supervised attention module to control information flow between different stages [Zamir *et al.*, 2021]. Cui *et al.* introduce a strip attention module to harvest multi-scale contextual information [Cui *et al.*, 2023e]. On the other hand, the recent self-attention module has put the Transformer models in the spotlight [Song *et al.*, 2022; Guo *et al.*, 2022]. Despite a few remedies, the high complexity of self-attention is still an intractable problem. Different from these attention-based methods, we develop lightweight attention modules from the perspective of frequency and adhere to the "channel + spatial" paradigm. Our proposed modules can model global information in spatial and channel dimensions and leverage frequency discrepancies between degraded and clean image pairs.

Spectral Networks. Recently, spectral networks have produced promising results for image restoration by revitalizing frequency processing, which is widely used in traditional algorithms [Mao *et al.*, 2021; Yu *et al.*, 2022; Guo *et al.*, 2023]. The common practice adopted by these methods is first to transfer the spatial features into the spectral domain using the Fourier or Wavelet transforms. The resulting spectra are then refined by a few convolutional layers. The inverse transforms are finally utilized to convent the modulated representation into the spatial domain. The above process is mostly applied in the spatial dimension. In this paper, we present a hybrid dual-dimension frequency learning strategy to enhance channel interactions and refine spatial global information.

3 Method

In this section, we first introduce the overall pipeline of our network. Then, we present the details of the core components: frequency-based channel feature modulation (FCFM), multi-scale frequency-based spatial feature modulation (MFSFM), and frequency-based loss function (FLF).

3.1 Overall Pipeline

As illustrated in Figure 3, our CSNet adopts the widely-used encoder-decoder architecture and consists of three scales for effective multi-scale representation learning. Figure 3 (a) shows that each decoder/encoder contains N normal residual blocks and a modified one that accommodates our MFSFM in the middle of two 3×3 convolutions. Our FCFM is employed in the bottleneck position of CSNet.

Given any input degraded image of shape $3 \times H \times W$, where 3 denotes the number of channels and $H \times W$ specifies the spatial locations in each channel, CSNet first leverages a single 3×3 convolutional layer to produce embedding features of size $C \times H \times W$, which are then fed into three encoders to obtain the in-depth features. During this process, the channels are expanded, and the resolutions are downsampled using the strided convolutions with the kernel size of 3 and stride of 2. After being refined by our FCFM, the resulting features pass through three decoder blocks to recover sharp features. In this process, the resolution is gradually restored to the original size through transposed convolutions (*kernel size*=4, *stride*=2), while the channel capability is reduced. The residual sharp image is produced via a 3×3 convolution, and the final output of CSNet is generated by additionally adding the original degraded input image. Next, we delineate the core components.

3.2 FCFM

Our frequency-based channel feature modulation module (FCFM) is illustrated in Figure 3 (c). It applies FFT among the channel dimensions to enhance channel interactions. As a result, each pixel can integrate information from the same location across all channels in multiple spectral spaces. Furthermore, our module operates at the pixel-wise granularity, which is conducive to managing spatially varying blurs. Specifically, given any input features $X \in \mathbb{R}^{C \times H \times W}$, FFT is used across channels to obtain the real and imaginary components, which are then concatenated among the channel dimension. The concatenated spectra are modulated through a 1×1 convolution layer. The output of FCFM is yielded via the channel-dimension IFFT. The above process can be formally expressed as:

$$\hat{X} = \mathrm{C-IFFT}(Conv_{1\times1}([\mathcal{R}(X), \mathcal{I}(X)])), \tag{1}$$

where $\mathcal{R}$ and $\mathcal{I}$ are the real and imaginary components, respectively; $[\cdot, \cdot]$ denotes the concatenation operation; $Conv_{1\times1}$ represents a convolutional layer with the kernel size of 1×1; and C-IFFT applies IFFT among the channel dimension. Although our lightweight FCFM has simple operation, it significantly improves performance over the baseline model, which will be shown in the ablation studies.

3.3 MFSFM

In addition to FCFM, which provides channel-dimension modulation, we further propose a multi-scale frequency-based spatial feature modulation module (MFSFM) to refine spatial features from the frequency perspective. From Figure 2, we can see that replacing the direct-current (DC) component of the degraded image with that of the ground truth leads to a sharper result. This fact inspires us to refine the DC component for spatial feature modulation, which can also model global information. Furthermore, we inject this mechanism, termed FSFM, into a densely connected coarse-to-fine learning paradigm to achieve multi-scale spatial feature refinement. In the following, we first introduce FSFM and then present its multi-scale version.

Similarly, with the input features $X \in \mathbb{R}^{C \times H \times W}$, we utilize the global average pooling to directly obtain the DC component instead of using FFT, resulting in lower computation overhead. The resulting DC part is recalibrated by the lightweight attention parameters optimized by backpropagation. Next, the improved DC part is fused with the remaining features, followed by a depthwise convolution for refinement. The above process can be expressed as:

$$\begin{aligned} \hat{X} &= \mathcal{F}_{FSFM}(X) \\ &= DConv_{3\times3}(X - \mathrm{GAP}(X) + W \odot \mathrm{GAP}(X)), \end{aligned} \tag{2}$$

where GAP is the global average pooling operation; $W \in \mathbb{R}^{C \times 1 \times 1}$ is the channel-wise attention parameters; $DConv_{3\times3}$ is a 3×3 depthwise convolution; and $\odot$ denotes the element-wise multiplication, where the size difference between W and features are bridged via the broadcast mechanism of the programming framework.

To enhance multi-scale representation learning, we inject FSFM into a densely connected coarse-to-fine learning paradigm, illustrated in Figure 3 (d). Specifically, the input features are downsampled to multiple feature spaces and then refined by the above FSFM. The resulting features of a branch are delivered to all subsequent branches for feature fusion and coarse-to-fine restoration. The final output of MFSFM is generated by applying a 3×3 convolution to the added features from all branches. Similarly, given any input tensor X, the process of MFSFM can be formally expressed as:

$$\hat{X} = Conv_{3\times3}(\sum_i \hat{X}_i \uparrow_{2^{4-i}} + X), \tag{3}$$

$$\hat{X}_i = \mathcal{F}_{FSFM}(X \downarrow_{2^{4-i}} + \sum_{j=1}^{i-1} \hat{X}_j), \tag{4}$$

where $i \in \{1, 2, 3\}$ indexes the branch; $\hat{X}_j$ represents the upsampled result of a preceding branch; $\uparrow_{2^{4-i}}$ denotes the upsampling operator with the rate of 2^{4-i}; and $\hat{X}_0$=$\mathbf{0}$.

3.4 FLF

As the DC component can be considered as a kind of low-frequency signal, we further propose a frequency-based loss function (FLF) as a complementary part of FSFM to refine omni-frequency signals. Denoting the predicted image of CSNet and the ground truth as $\hat{I}$ and G, respectively, FLF is given by:

$$\mathcal{L}_{FLF} = \|\hat{I} - \mathrm{GAP}(\hat{I}) - (G - \mathrm{GAP}(G))\|_1, \tag{5}$$

where GAP is the global average pooling technique. By doing this, we bring the high-frequency component of $\hat{I}$ closer to that of G. In our case, the high frequency is obtained by removing the DC part from the images.

4 Experiments

In this section, we first introduce the implementation details and evaluation metrics. Then, we compare our results with state-of-the-art algorithms on nine different datasets for three representative image restoration tasks: image dehazing, image defocus deblurring and image desnowing. Ablation experiments are performed in the final part. In the tables, the top-performing scores are highlighted in purple. In the figures, PSNR is computed for comparisons.

4.1 Implementation Details

Unless mentioned otherwise, we adopt the following hyper-parameters to train our CSNet. Specifically, the model is trained using the Adam optimizer on 256×256 patches with a batch size of 8. The initial learning rate is 2^{-4}, which is gradually reduced to $1e^{-6}$ with cosine annealing. We adopt the horizontal flips for data augmentation. According to the complexity of different tasks, we set N (Figure 3 (b)) to 3 for

11.12 dB	33.89 dB	29.37dB	32.85 dB	40.74 dB	+∞
Input	FFA-Net	MAXIM-2S	DehazeFormer-L	Ours	GT

Figure 4: Qualitative comparisons on the SOTS [Li *et al.*, 2018] dataset for image dehazing.

Methods	Venue	SOTS		Dense-Haze		NH-HAZE		O-HAZE	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DehazeNet [Cai *et al.*, 2016]	TIP'16	19.82	0.821	13.84	0.43	16.62	0.52	17.57	0.77
AOD-Net [Li *et al.*, 2017]	ICCV'17	20.51	0.816	13.14	0.41	15.40	0.57	15.03	0.54
GridDehazeNet [Liu *et al.*, 2019]	ICCV'19	32.16	0.984	-	-	13.80	0.54	-	-
MSBDN [Dong *et al.*, 2020]	CVPR'20	33.67	0.985	15.37	0.49	19.23	0.71	24.36	0.75
FFA-Net [Qin *et al.*, 2020]	AAAI'20	36.39	0.989	14.39	0.45	19.87	0.69	22.12	0.77
AECR-Net [Wu *et al.*, 2021]	CVPR'21	37.17	0.990	15.80	0.47	19.88	0.72	-	-
PFDN [Dong and Pan, 2020]	ECCV'20	32.68	0.976	-	-	-	-	-	-
DeHamer [Guo *et al.*, 2022]	CVPR'22	36.63	0.988	16.62	0.56	20.66	0.68	-	-
MAXIM-2S [Tu *et al.*, 2022]	CVPR'22	38.11	0.991	-	-	-	-	-	-
FSDGN [Yu *et al.*, 2022]	ECCV'22	38.63	0.990	16.91	0.58	19.99	0.73	-	-
PMNet[Ye *et al.*, 2022]	ECCV'22	38.41	0.990	16.79	0.51	20.42	0.73	24.64	0.83
DehazeFormer-L [Song *et al.*, 2022]	TIP'23	40.05	0.996	-	-	-	-	-	-
SANet [Cui *et al.*, 2023e]	IJCAI'23	40.40	0.996	-	-	-	-	-	-
MB-TaylorFormer-B [Qiu *et al.*, 2023]	ICCV'23	40.71	0.992	16.66	0.56	-	-	25.05	0.79
FocalNet [Cui *et al.*, 2023a]	ICCV'23	40.82	0.996	17.07	0.63	20.43	0.79	25.50	0.94
CSNet	**Ours**	41.34	0.996	17.33	0.65	20.43	0.80	25.60	0.94

Table 1: Quantitative comparisons on the synthetic and real-world datasets for image dehazing.

Methods	PSNR↑	SSIM↑
NDIM [Zhang *et al.*, 2014]	14.31	0.526
GS [Li *et al.*, 2015]	17.32	0.629
MRPF [Zhang *et al.*, 2017]	16.95	0.667
MRP [Zhang *et al.*, 2017]	19.93	0.777
OSFD [Zhang *et al.*, 2020]	21.32	0.804
HCD [Wang *et al.*, 2024]	23.43	0.953
FocalNet [Cui *et al.*, 2023a]	25.35	0.969
CSNet (Ours)	26.13	0.971

Table 2: Image dehazing comparisons on the NHR [Zhang *et al.*, 2020] dataset for nighttime scenes.

Methods	PSNR↑	SSIM↑
GS [Li *et al.*, 2015]	21.02	0.639
MRP [Zhang *et al.*, 2017]	20.92	0.646
Ancuti *et al* [Ancuti *et al.*, 2016]	20.59	0.623
CycleGAN [Zhu *et al.*, 2017]	21.75	0.696
Yan *et al* [Yan *et al.*, 2020]	27.00	0.850
Jin *et al* [Jin *et al.*, 2023]	30.38	0.904
CSNet (Ours)	31.55	0.914

Table 3: Nighttime image dehazing comparisons on the GTA5 [Yan *et al.*, 2020] dataset.

dehazing and desnowing, and 15 for deblurring. All experiments are carried out on an NVIDIA Tesla A100 GPU.

We measure the widely-used peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [Wang *et al.*, 2004] for all datasets, and additionally, adopt the mean absolute error (MAE) and learned perceptual image patch similarity (LPIPS) [Zhang *et al.*, 2018] for the DPDD [Abuolaim and Brown, 2020] dataset.

4.2 Results

Image Dehazing. We evaluate our model on both daytime and nighttime dehazing datasets. For daytime scenes, the numerical results on a synthetic (SOTS [Li *et al.*, 2018]) and three real-world datasets, *i.e.,* Dense-Haze [Ancuti *et al.*, 2019], NH-HAZE [Ancuti *et al.*, 2020], and O-HAZE [Ancuti *et al.*, 2018], are presented in Table 1. As seen, our model produces top-performing results on most metrics across synthetic and real-world datasets. In particular, CSNet outperforms the recent strong Transformer-based algorithm, MB-TaylorFormer-B [Qiu *et al.*, 2023], by 0.63 dB PSNR on the SOTS dataset with similar computation overhead, as illustrated in Figure 1. Compared to another recent algorithm, FocalNet [Cui *et al.*, 2023a], our method is more effective in removing real-world haze degradations by providing performance gains of 0.26 dB and 0.10 dB PSNR on Dense-Haze and O-HAZE, respectively. Figure 4 shows the visual results on SOTS [Li *et al.*, 2018]. The image produced by our model is much closer to the ground truth.

In addition, we provide nighttime dehazing results on two datasets, NHR [Zhang *et al.*, 2020] and GTA5 [Yan *et al.*, 2020], in Table 2 and Table 3, respectively. Our model is superior to FocalNet [Cui *et al.*, 2023a] with a performance

Figure 5: Qualitative comparisons on the NHR [Zhang *et al.*, 2020] dataset for image dehazing.

Figure 6: Qualitative comparisons on the GTA5 [Yan *et al.*, 2020] dataset for image dehazing.

Figure 7: Qualitative comparisons on the DPDD [Abuolaim and Brown, 2020] dataset for image defocus deblurring.

Methods	Indoor Scenes				Outdoor Scenes				Combined			
	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓
EBDB [Karaali and Jung, 2017]	25.77	0.772	0.040	0.297	21.25	0.599	0.058	0.373	23.45	0.683	0.049	0.336
DMENet [Lee *et al.*, 2019]	25.50	0.788	0.038	0.298	21.43	0.644	0.063	0.397	23.41	0.714	0.051	0.349
JNB [Shi *et al.*, 2015]	26.73	0.828	0.031	0.273	21.10	0.608	0.064	0.355	23.84	0.715	0.048	0.315
DPDNet [Abuolaim and Brown, 2020]	26.54	0.816	0.031	0.239	22.25	0.682	0.056	0.313	24.34	0.747	0.044	0.277
KPAC [Son *et al.*, 2021]	27.97	0.852	0.026	0.182	22.62	0.701	0.053	0.269	25.22	0.774	0.040	0.227
IFAN [Lee *et al.*, 2021]	28.11	0.861	0.026	0.179	22.76	0.720	0.052	0.254	25.37	0.789	0.039	0.217
DRBNet [Ruan *et al.*, 2022]	-				-				25.73	0.791	-	0.183
Restormer [Zamir *et al.*, 2022]	28.87	0.882	0.025	0.145	23.24	0.743	0.050	0.209	25.98	0.811	0.038	0.178
NRKNet [Quan *et al.*, 2023]	-				-				26.11	0.810	-	0.210
FocalNet [Cui *et al.*, 2023a]	29.10	0.876	0.024	0.173	23.41	0.743	0.049	0.246	26.18	0.808	0.037	0.210
CSNet (Ours)	29.20	0.881	0.023	0.147	23.45	0.752	0.049	0.206	26.25	0.815	0.037	0.178

Table 4: Image defocus deblurring comparisons on the DPDD [Abuolaim and Brown, 2020] dataset.

gain of 0.78 dB in terms of PSNR on NHR. Also, our CSNet outperforms the recent algorithm [Jin *et al.*, 2023] by 1.17 dB PSNR, although it is specially designed for nighttime scenes. Visual comparisons in Figure 5 and Figure 6 illustrate that our model is robust in nighttime scenarios.

Image Defocus Deblurring. For this task, we conduct experiments on the widely used DPDD [Abuolaim and Brown, 2020] dataset. The quantitative results are shown in Table 4. The proposed network outperforms state-of-the-art algorithms in most scenes. Specifically, our model significantly outperforms the strong Transformer-based Restormer [Zamir *et al.*, 2022] by 0.27 dB PSNR in the combined category. In comparison to the recently proposed methods, NRKNet [Quan *et al.*, 2023] and FocalNet [Cui *et al.*, 2023a], our method continues to achieve superior scores, surpassing them by 0.14 dB and 0.07 dB PSNR, respectively. The visual results are shown in Figure 7. We can see that our method recovers more details from the hard blurs than competitors.

Image Desnowing. We evaluate our model on two widely adopted datasets for image desnowing. The numerical scores on the CSD [Chen *et al.*, 2021] and Snow100K [Liu *et al.*, 2018] are presented in Table 5. CSNet generates a 0.18 dB

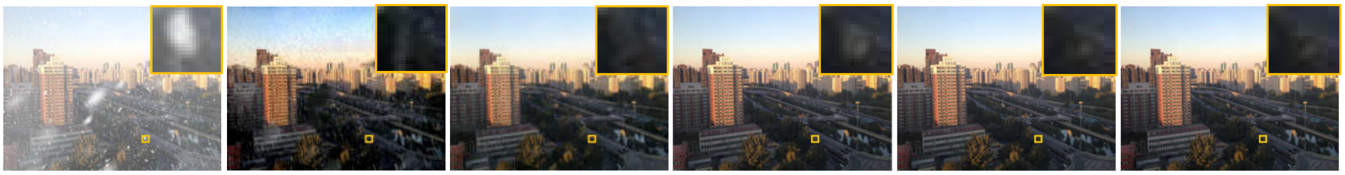

Figure 8: Qualitative comparisons on the CSD [Chen *et al.*, 2021] dataset for image desnowing.

Methods	CSD		Snow100K	
	PSNR	SSIM	PSNR	SSIM
DesnowNet [Liu *et al.*, 2018]	20.13	0.81	30.50	0.94
All in One [Li *et al.*, 2020]	26.31	0.87	26.07	0.88
JSTASR [Chen *et al.*, 2020]	27.96	0.88	23.12	0.86
HDCW-Net [Chen *et al.*, 2021]	29.06	0.91	31.54	0.95
SMGARN [Cheng *et al.*, 2022]	31.93	0.95	31.92	0.93
TransWeather [Valanarasu *et al.*, 2022]	31.76	0.93	31.82	0.93
IRNeXt [Cui *et al.*, 2023c]	37.29	0.99	33.61	0.95
FocalNet [Cui *et al.*, 2023a]	37.18	0.99	33.53	0.95
CSNet (Ours)	38.13	0.99	33.71	0.95

Table 5: Image desnowing comparisons on the CSD [Chen *et al.*, 2021] and Snow100K [Liu *et al.*, 2018] datasets.

Method	a	b	c	d	e	f	Fourmer
Baseline	✓	✓	✓	✓	✓	✓	
FLF		✓	✓	✓	✓	✓	
FCFM			✓	✓	✓	✓	
FSFM†				✓	✓	✓	
MFSFM†					✓	✓	
MFSFM						✓	
PSNR	31.32	31.66	33.42	34.87	37.80	37.87	37.32
SSIM	0.984	0.984	0.987	0.990	0.993	0.993	0.990
GFLOPs	15.44	15.44	15.71	15.78	19.38	19.38	20.6

Table 6: Ablation studies for the proposed components. † denotes models that do not utilize dense connections by eliminating the delivery of features from the first branch to the third branch. It is worth mentioning that our tiny version outperforms the recent Fourmer [Zhou *et al.*, 2023] with lower FLOPs.

PSNR gain over the FocalNet [Cui *et al.*, 2023a] algorithm. On the CSD dataset containing more complex scenes, the advantage is further expanded, suggesting the superiority of our model for snow removal. The visual comparisons on the CSD dataset are illustrated in Figure 8. The result of FocalNet still remains snow blurs. In contrast, our result is much closer to the ground truth and obtains a higher PSNR value.

4.3 Ablation Studies

We experiment to demonstrate the efficacy of the proposed components by training and testing on RESIDE [Li *et al.*, 2018] and SOTS [Li *et al.*, 2018], respectively. The model is trained for 300 epochs with $N = 0$. Other configurations remain the same as that of our final dehazing model. We obtain the baseline model by removing all proposed components in this tiny CSNet.

Methods	PSNR
Squeeze-and-Excitation Block [Hu *et al.*, 2018]	32.28
Simplified Channel Attention [Chen *et al.*, 2022]	32.35
FCFM (Ours)	33.42

Table 7: Comparisons with Alternative to FSFM.

Effects of Individual Components. The ablation results for the proposed components are shown in Table 6. The baseline model attains 31.32 dB PSNR on the SOTS [Li *et al.*, 2018] dataset. Our FLF achieves a gain of 0.34 dB PSNR over the baseline model. The channel-dimension processing module, FCFM, significantly improves the performance to 33.42 dB PSNR. Without employing the dense connection, FSFM and MFSFM continue to generate performance improvements, demonstrating the effectiveness of our spatial feature modulation module and multi-scale learning paradigm, respectively. Our complete model obtains the best performance, providing a performance boost of 6.55 dB PSNR over the baseline model. It is worth mentioning that our complete model outperforms the recent Fourmer [Zhou *et al.*, 2023] algorithm with lower complexity.

Comparisons with alternatives to FCFM. We compare our FCFM with popular channel attention mechanisms, such as the squeeze-and-excitation block [Hu *et al.*, 2018] and simplified channel attention [Chen *et al.*, 2022]. The results in Table 7 demonstrate that our method shows superiority to these alternatives by facilitating channel interactions in the spectral domain.

5 Conclusion

In this paper, we present an effective and efficient network, named CSNet, for image restoration based on hybrid frequency modulation, *i.e.,* "channel + spatial" dual-dimension representation learning. Specifically, we propose a frequency-based channel feature modulation model to enhance interactions between all channels based on the Fourier transform. Moreover, inspired by our observation, a multi-scale frequency-based spatial feature modulation module is developed to refine the direct-current component, which can model the global information and bridge the frequency discrepancies between degraded and sharp image pairs. To achieve omni-frequency learning, a frequency-based loss function is further introduced to train the network. Extensive experiments on nine different benchmark datasets demonstrate that the proposed network achieves state-of-the-art performance for three image restoration tasks.

Acknowledgements

This work was supported partly by the National Natural Science Foundation of China (Grant No.62322216, 62172409), 2023 CCF-Tencent Rhino-Bird Young Faculty Open Research Fund, and partly by the project "VIDETEC-2" (Grant No.19F2232A) and the Federal Ministry for Digital and Transport of Germany (BMDV).

References

[Abuolaim and Brown, 2020] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *ECCV*, 2020.

[Ancuti *et al.*, 2016] Cosmin Ancuti, Codruta O Ancuti, Christophe De Vleeschouwer, and Alan C Bovik. Night-time dehazing by fusion. In *TIP*, 2016.

[Ancuti *et al.*, 2018] Codruta O Ancuti, Cosmin Ancuti, Radu Timofte, and Christophe De Vleeschouwer. O-haze: a dehazing benchmark with real hazy and haze-free outdoor images. In *CVPRW*, 2018.

[Ancuti *et al.*, 2019] Codruta O Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In *TIP*, 2019.

[Ancuti *et al.*, 2020] Codruta O. Ancuti, Cosmin Ancuti, and Radu Timofte. Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *CVPRW*, 2020.

[Cai *et al.*, 2016] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *TIP*, 2016.

[Chen *et al.*, 2020] Wei-Ting Chen, Hao-Yu Fang, Jian-Jiun Ding, Cheng-Che Tsai, and Sy-Yen Kuo. Jstasr: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal. In *ECCV*, 2020.

[Chen *et al.*, 2021] Wei-Ting Chen, Hao-Yu Fang, Cheng-Lin Hsieh, Cheng-Che Tsai, I Chen, Jian-Jiun Ding, Sy-Yen Kuo, et al. All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In *ICCV*, 2021.

[Chen *et al.*, 2022] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022.

[Cheng *et al.*, 2022] Bodong Cheng, Juncheng Li, Ying Chen, Shuyi Zhang, and Tieyong Zeng. Snow mask guided adaptive residual network for image snow removal. *arXiv preprint arXiv:2207.04754*, 2022.

[Cui *et al.*, 2023a] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Focal network for image restoration. In *ICCV*, 2023.

[Cui *et al.*, 2023b] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Image restoration via frequency selection. *TPAMI*, 2023.

[Cui *et al.*, 2023c] Yuning Cui, Wenqi Ren, Sining Yang, Xiaochun Cao, and Alois Knoll. Irnext: Rethinking convolutional network design for image restoration. In *ICML*, 2023.

[Cui *et al.*, 2023d] Yuning Cui, Yi Tao, Zhenshan Bing, Wenqi Ren, Xinwei Gao, Xiaochun Cao, Kai Huang, and Alois Knoll. Selective frequency network for image restoration. In *ICLR*, 2023.

[Cui *et al.*, 2023e] Yuning Cui, Yi Tao, Luoxi Jing, and Alois Knoll. Strip attention for image restoration. In *IJCAI*, 2023.

[Cui *et al.*, 2024] Yuning Cui, Wenqi Ren, and Alois Knoll. Omni-kernel network for image restoration. In *AAAI*, 2024.

[Dong and Pan, 2020] Jiangxin Dong and Jinshan Pan. Physics-based feature dehazing networks. In *ECCV*, 2020.

[Dong *et al.*, 2020] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *CVPR*, 2020.

[Guo *et al.*, 2022] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *CVPR*, 2022.

[Guo *et al.*, 2023] Shi Guo, Hongwei Yong, Xindong Zhang, Jianqi Ma, and Lei Zhang. Spatial-frequency attention for image denoising. *arXiv preprint arXiv:2302.13598*, 2023.

[He *et al.*, 2010] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *TPAMI*, 2010.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[Jin *et al.*, 2023] Yeying Jin, Beibei Lin, Wending Yan, Yuan Yuan, Wei Ye, and Robby T Tan. Enhancing visibility in nighttime haze images using guided apsf and gradient adaptive convolution. In *ACM MM*, 2023.

[Karaali and Jung, 2017] Ali Karaali and Claudio Rosito Jung. Edge-based defocus blur estimation with adaptive scale selection. *TIP*, 2017.

[Lee *et al.*, 2019] Junyong Lee, Sungkil Lee, Sunghyun Cho, and Seungyong Lee. Deep defocus map estimation using domain adaptation. In *CVPR*, 2019.

[Lee *et al.*, 2021] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *CVPR*, 2021.

[Li *et al.*, 2015] Yu Li, Robby T Tan, and Michael S Brown. Nighttime haze removal with glow and multiple light colors. In *ICCV*, 2015.

[Li *et al.*, 2017] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *ICCV*, 2017.

[Li *et al.*, 2018] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *TIP*, 2018.

[Li *et al.*, 2020] Ruoteng Li, Robby T. Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *CVPR*, 2020.

[Liu *et al.*, 2018] Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *TIP*, 2018.

[Liu *et al.*, 2019] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *ICCV*, 2019.

[Mao *et al.*, 2021] Xintian Mao, Yiming Liu, Wei Shen, Qingli Li, and Yan Wang. Deep residual fourier transformation for single image deblurring. *arXiv preprint arXiv:2111.11745*, 2021.

[Qin *et al.*, 2020] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *AAAI*, 2020.

[Qiu *et al.*, 2023] Yuwei Qiu, Kaihao Zhang, Chenxi Wang, Wenhan Luo, Hongdong Li, and Zhi Jin. Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In *ICCV*, 2023.

[Quan *et al.*, 2023] Yuhui Quan, Zicong Wu, and Hui Ji. Neumann network with recursive kernels for single image defocus deblurring. In *CVPR*, 2023.

[Ruan *et al.*, 2022] Lingyan Ruan, Bin Chen, Jizhou Li, and Miuling Lam. Learning to deblur using light field generated and real defocus images. In *CVPR*, 2022.

[Shi *et al.*, 2015] Jianping Shi, Li Xu, and Jiaya Jia. Just noticeable defocus blur detection and estimation. In *CVPR*, 2015.

[Son *et al.*, 2021] Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *ICCV*, 2021.

[Song *et al.*, 2022] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *arXiv preprint arXiv:2204.03883*, 2022.

[Tsai *et al.*, 2022] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *ECCV*, 2022.

[Tu *et al.*, 2022] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *CVPR*, 2022.

[Valanarasu *et al.*, 2022] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *CVPR*, 2022.

[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. *TIP*, 2004.

[Wang *et al.*, 2024] Tao Wang, Guangpin Tao, Wanglong Lu, Kaihao Zhang, Wenhan Luo, Xiaoqin Zhang, and Tong Lu. Restoring vision in hazy weather with hierarchical contrastive learning. *PR*, 2024.

[Wu *et al.*, 2021] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *CVPR*, 2021.

[Yan *et al.*, 2020] Wending Yan, Robby T Tan, and Dengxin Dai. Nighttime defogging using high-low frequency decomposition and grayscale-color networks. In *ECCV*, 2020.

[Ye *et al.*, 2022] Tian Ye, Yunchen Zhang, Mingchao Jiang, Liang Chen, Yun Liu, Sixiang Chen, and Erkang Chen. Perceiving and modeling density for image dehazing. In *ECCV*, 2022.

[Yu *et al.*, 2022] Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao. Frequency and spatial dual guidance for image dehazing. In *ECCV*, 2022.

[Zamir *et al.*, 2021] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021.

[Zamir *et al.*, 2022] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022.

[Zhang *et al.*, 2014] Jing Zhang, Yang Cao, and Zengfu Wang. Nighttime haze removal based on a new imaging model. In *ICIP*, 2014.

[Zhang *et al.*, 2017] Jing Zhang, Yang Cao, Shuai Fang, Yu Kang, and Chang Wen Chen. Fast haze removal for nighttime image using maximum reflectance prior. In *CVPR*, 2017.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[Zhang *et al.*, 2020] Jing Zhang, Yang Cao, Zheng-Jun Zha, and Dacheng Tao. Nighttime dehazing with a synthetic benchmark. In *ACM MM*, 2020.

[Zhang *et al.*, 2023] Jiale Zhang, Yulun Zhang, Jinjin Gu, Jiahua Dong, Linghe Kong, and Xiaokang Yang. Xformer: Hybrid x-shaped transformer for image denoising. *arXiv preprint arXiv:2303.06440*, 2023.

[Zhou *et al.*, 2023] Man Zhou, Jie Huang, Chun-Le Guo, and Chongyi Li. Fourmer: an efficient global modeling paradigm for image restoration. In *ICML*, 2023.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.