This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Fourier Prior-Based Two-Stage Architecture for Image Restoration

Hemkant Nehete, Amit Monga, Partha Kaushik, Brajesh Kumar Kaushik
Indian Institute of Technology Roorkee, India
nehete_h@ece.iitr.ac.in, amit_m@ece.iitr.ac.in, p_kaushik@cs.iitr.ac.in, bkk23fec@iitr.ac.in

Abstract

This work presents a novel two stage architecture designed to enhance degraded images affected by environmental factors such as haze, blur, fog, and rain. Despite the dominance of deep Convolutional Neural Networks (CNNs) and Transformers in single image restoration tasks, existing methods neglect the intrinsic priors for physical properties of degradation. To enhance the generalization ability of image restoration models, we propose Fourier prior based on a key observation that substituting the Fourier amplitude of degraded images with that of clean images effectively mitigates degradation. Therefore, amplitude contains degradation information, while the phase retains background structures. Consequently, a two-stage model is proposed, that consists of Amplitude Refinement Unit (ARU) and the Phase Refinement Unit (PRU), that separately restore both amplitude and phase information, respectively. ARU and PRU leverage a CNN-Transformer-based architecture to extract local and global features, overcoming computational constraints posed by large image sizes in Transformers. Additionally, a multi-scale approach in ARU refines amplitude features at coarse and fine levels, improving restoration efficiency. Experimental results across multiple image restoration tasks, like image deraining, dehazing, and low-light enhancement, indicate that the proposed architecture improved the performance in terms of PSNR, SSIM, and computational efficiency compared to state-of-the-art Transformer approaches.

1. Introduction

Image restoration is a vital discipline within computer vision and digital image processing, dedicated to enhancing and recovering degraded images. In various domains such as photography, medical imaging, satellite imagery, and more, images frequently encounter distortions, noise, blurriness, or other imperfections that degrade their visual quality and informational content. Image restoration techniques aim to rectify these issues restoring them to improve their visual quality.

Traditionally, image restoration methods have relied on mathematical models and signal processing techniques to address various issues including blur, noise, and atmospheric distortions. These methods often utilize different image priors, such as the dark channel prior for image dehazing [13] and the non-local mean prior for image denoising [10], to guide the restoration process by formulating the restoration image as an optimisation problem to constraint the solution space. This optimisation typically involves minimizing a cost function that penalizes deviations from the prior assumptions about the image structure and degradation model. Although effective in many scenarios, traditional approaches often struggle when confronted with real-world scenarios characterized by complex degradations and variations.

Deep learning has transformed image restoration domain. Unlike traditional methods, deep learning can directly learn from data, overcoming limitations and adapting to new scenarios. These methods excel at extracting complex image features and generalizing the relationship between degraded and clean images. Convolutional Neural Networks (CNNs) have been the backbone of this revolution, using various architectures like encoder-decoder [8], U-Nets [24] and residual blocks [14] to tackle specific restoration tasks. However, CNNs struggle with capturing long-range dependencies within images due to their limited receptive field. Additionally, they lack flexibility in adapting to diverse content as they rely on fixed weights during inference. Attention mechanisms were introduced to address this, allowing models to focus on relevant parts of the image. However, they introduce computational inefficiencies and scalability issues. Transformers offer a promising alternative, excelling at capturing long-range dependencies and adapting to different content. Their core component, self-attention, enables efficient learning and parallel processing. While transformers have been used for image deraining and deblurring, their computational cost increases significantly with image size, posing a challenge for practical implementation.

CNNs and Transformer-based networks have demonstrated impressive capabilities in restoring clean images

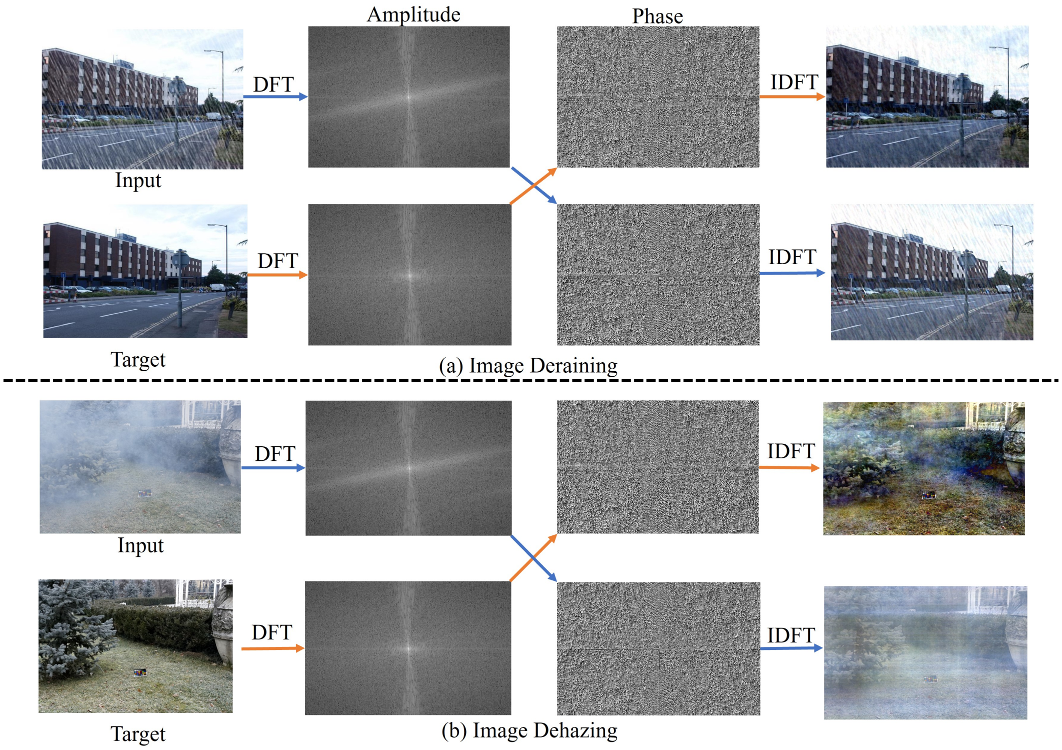


Figure 1. Interchanging the amplitude and phase components in the Fourier domain of degraded and ground truth images to separate the content information and degradation in form of (a) rain and (b) haze

from degraded images. However, these methods suffer from two main drawbacks. Firstly, they often neglect the intrinsic prior knowledge of the physical properties of degradation, leading to potential overfitting issues. Secondly, the computational constraints imposed by attention and Transformer architectures are very high. To address these challenges, this work proposes the FAPRNet architecture, that incorporates a Fourier prior. The motivation behind this lies in an observation made during the Fourier transformation, where the Fourier amplitude and phase spectrum of paired degraded and ground truth images were exchanged, as depicted in Figure 1. Degradation is greatly suppressed in images reconstructed with the phase of degraded images and the amplitude of ground truth, indicating that most of the degradation is preserved in the amplitude spectrum of degraded images. Second, intricate details present in clean images are effectively retained when substituting the phase spectrum with that of the degraded images, suggesting that the phase of degraded images maintains similar background structures as the ground truth. It is inherent that handling the amplitude spectrum of degraded images separately holds potential for efficient degradation removal. Additionally, the phase spectrum of degraded image offers the opportunity to enhance the structural details of the background. Consequently, the Fourier prior is achieved through the independent learning of the transformation for both the amplitude and phase spectrum in two different stages as shown in Figure 2.

The main contributions of this work are summarised as:

- This work proposes a two-stage image restoration architecture Fourier Amplitude and Phase Refinement Network (FAPRNet) comprising two essential components: the Amplitude Refinement Unit (ARU) and the Phase Refinement Unit (PRU). These blocks are designed to facilitate the transformation of the Fourier amplitude and phase spectrum.
- The architecture of ARU and PRU implements a hybrid CNN-Transformer-based image restoration technique that leverages the advantage of extracting local features utilizing CNN and global features utilizing the Transformer module thereby reducing the computational complexity. Furthermore, ARU consists of multiscale residual amplitude refinement transformer blocks to restore amplitude spectrum and feature fusion blocks to fuse multi-scale features from CNN and Transformer networks.
- We propose a novel loss function to incorporate intermediate loss during each stage of network i.e amplitude and phase refinement. These intermediate stage losses are finally added together with spatial loss and SSIM loss of

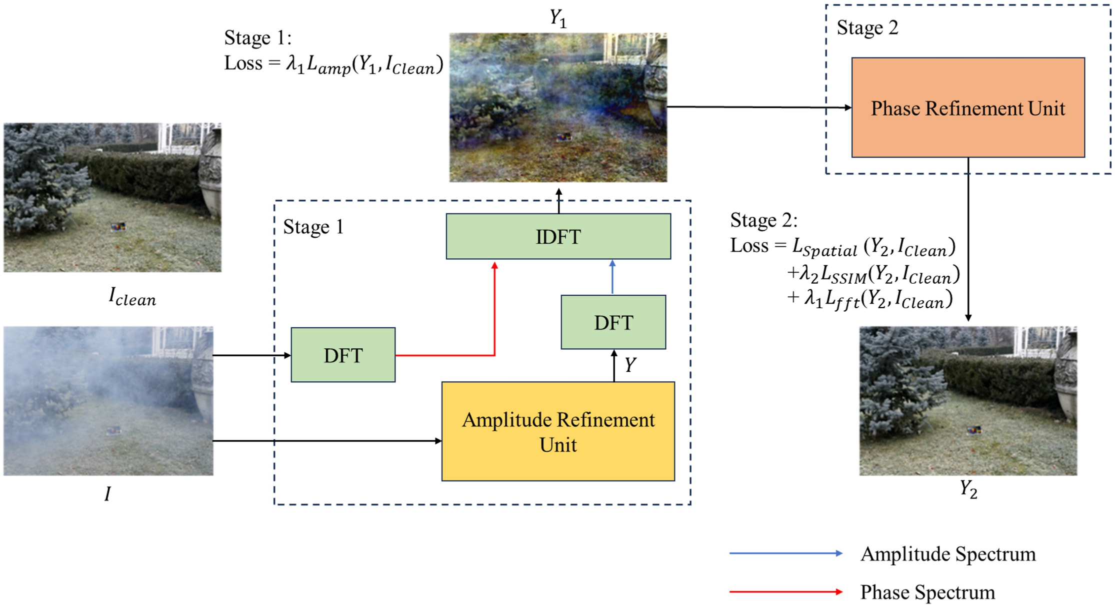

Stage 1:
Loss = $\lambda_1 L_{amp}(Y_1, I_{Clean})$
Y_1
Stage 2
Phase Refinement Unit
I_{clean}
Stage 1
IDFT
DFT
DFT
Y
Amplitude Refinement Unit
I
Stage 2:
Loss = $L_{Spatial}(Y_2, I_{Clean})$
$+\lambda_2 L_{SSIM}(Y_2, I_{Clean})$
$+\lambda_1 L_{fft}(Y_2, I_{Clean})$
Y_2
Amplitude Spectrum
Phase Spectrum

Figure 2. Two-stage architecture of FAPRNet comprising Amplitude Refinement followed by Phase Refinement stages for image restoration.

restored images with respect to ground truth.

- Extensive experiments on image deraining, dehazing, and low-light enhancement datasets demonstrate superior performance than CNN-based and Transformer-based state-of-the-art image restoration architectures.

2. Related Works

The process of restoring images entails eliminating noise, distortions, or artefacts from them in order to improve their clarity and overall appearance. Conventional image restoration techniques, like dark channel prior, and histogram-based prior, mainly rely on hand-crafted priors however, they lacked in generalization capabilities. Image restoration has experienced a dramatic paradigm shift due to the introduction of CNNs and Transformers and demonstrated remarkable performance in tasks like deraining, dehazing, and low-light enhancement.

2.1. CNN based restoration

Contrasting with traditional methods that relied on hand-crafted features or explicit degradation models, CNN-based image restoration techniques have advanced significantly in recent years, showcasing a plethora of architectures [8, 18, 23, 35]. Sophisticated modules like the multi-stage paradigm [12, 22], encoder-decoder architecture [8, 17], multi-patch learning, and attention module [23, 25, 28], which highlights pertinent information, have been integrated into these frameworks to improve their efficacy and performance. Convolution operations are localised, which restricts the ability to perceive global contextual information that is necessary for image restoration. To overcome this, techniques utilising various CNN-based attention modules have become popular in handling vision-related issues by focusing on important information in images while preserving spatial and inter-channel pixel connections [28].

2.2. Transformers based restoration

Vision Transformer (ViT) emerge, a visionary approach that treats images as sequences of patches or tokens, similar to words in Natural Language Processing (NLP), thus providing a novel framework for understanding images holistically. Their innate ability to recognize complex visual patterns, coupled with the dynamic flexibility of the self-attention mechanism, puts ViT at the forefront of computer vision innovation. However, the original ViT suffers from limited inductive bias and quadratic computational costs with increase in image size. In an effort to address these issues, ViTs have undergone recent iterations [7, 27] with other strategies to include depthwise convolution in the feed-forward network [29, 31]. As a result of these developments, numerous image restoration methods have been created, utilising the global modelling capacity and adaptability to a wide range of input information. However, compared to CNNs, the Transformer's self-attention ability to represent local invariant properties is weaker. A hybrid architecture that combines CNNs for local feature extraction and Transformers for global feature capture has been proposed as a solution to these drawbacks.

3. Method

This section commences with a review of the fundamental properties of the Discrete Fourier Transform (DFT) in the context of images. Subsequently, the architectural details of the proposed FAPRNet architecture are meticulously presented, including a thorough discussion of the core building block ARU and PRU. The section concludes by detailing the loss function used for training the architecture.

3.1. Preliminary

The DFT is widely utilised for image analysis in the frequency domain. The DFT, denoted by F, takes an image $x \in R^{H \times W \times C}$, where H, W, and C represent image dimensions and transforms each colour channel into the frequency domain as a complex-valued representation, $F(x)(u,v)$. This is mathematically expressed as:

$$F(x)(u,v) = \frac{1}{HW} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h,w) e^{-j2\pi(\frac{h}{H}u + \frac{w}{W}v)} \tag{1}$$

Here, u and v represent the coordinates in the frequency domain. The amplitude, $A(x)(u,v)$, and phase component $P(x)(u,v)$, are expressed from the real, $F_R(x)(u,v)$, and imaginary, $F_I(x)(u,v)$, parts of $F(x)$ as follows:

$$A(x)(u,v) = \sqrt[2]{F_R^2(x)(u,v) + F_I^2(x)(u,v)} \tag{2}$$

$$P(x)(u,v) = tan^{-1}\left(\frac{F_I(x)(u,v)}{F_R(x)(u,v)}\right) \tag{3}$$

The inverse DFT (IDFT), denoted by F^{-1}, transforms the image back from the frequency domain to the spatial domain. The DFT has been instrumental in performing detailed frequency analysis for image restoration tasks. Notably, by analyzing the phase and amplitude components in the frequency domain, it has been observed that image degradation primarily affects the amplitude. This observation becomes evident when interchanging the amplitude and phase components of a degraded image to its ground truth, as demonstrated in Figure 1. This observation suggests that the DFT effectively separates image degradation from the underlying content information. This insight paves the way for utilizing the DFT as a prior for image degradation within image restoration frameworks where both amplitude and phase spectrum can be refined separately.

3.2. Overall Architecture

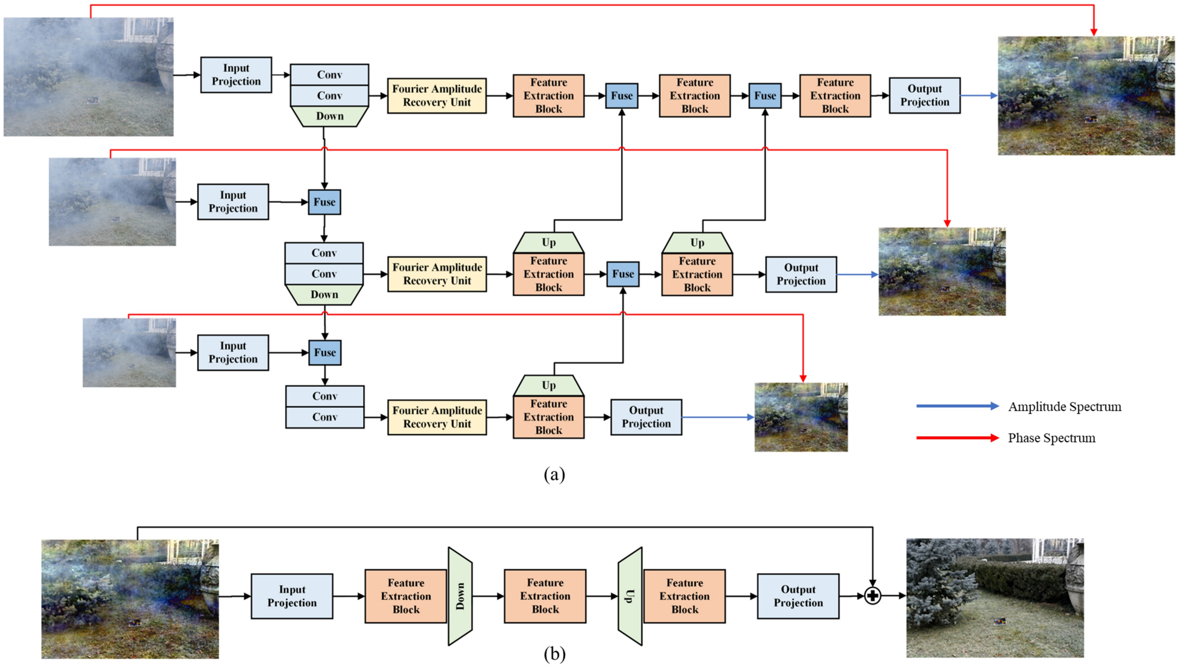
Leveraging the Fourier prior, this work proposes a novel two-stage architecture illustrated in Figure 2. The FAPRNet architecture consists of an amplitude refinement stage and a phase refinement stage. The first stage aims to restore the amplitude of degraded images to resemble amplitude of ground truth by amplitude refinement unit (ARU). To achieve this, the network utilizes the inverse Fourier transform applied to the clean image's amplitude and the degraded image's phase given by $F^{-1}(A(I_{clean}), P(I))$, where I_{clean} and I represents ground truth and degraded images respectively, as the supervisory signal. This approach preserves the phase information of degraded images, that is crucial for retaining background structures. The second stage receives the output from the first stage $Y_1 = (F^{-1}(A(Y), P(I)))$ as input. The second stage then refines the phase information to recover fine-grained background details using the ground truth clean image as the supervisory signal through PRU.

The ARU processes a multi-scale input ($I \in R^{H \times W \times 3}$) and its downsampled versions ($I_1 \in R^{\frac{H}{2} \times \frac{W}{2} \times 3}$) and ($I_2 \in R^{\frac{H}{4} \times \frac{W}{4} \times 3}$). It extracts shallow features through 3×3 convolution operations. These feature maps undergo further processing through CNN blocks for feature extraction, with generated feature maps being downsampled after each stage and expanding their channels. Feature maps from the previous stage are fused with the downsampled input features using 1×1 convolution operations. These multi-scale high-level features are utilized for restoring the amplitude of degraded images by the Fourier Amplitude Recovery Unit (FARU), which operates on the amplitude spectrum of the feature maps. The multi-scale features obtained from FARU are fused with a series of Feature Extractor Blocks (FEB) as shown in figure 3(a). Upsampling of feature maps is performed using transpose convolution operations. Finally, the output images for stage one is obtained by combining the amplitude spectrum of the generated image with the phase of degraded images. Architecture for PRU in second stage targeting of refining phase spectrum consist of series of three FEBs with downsampling and upsamplig after first and second stage respectively as illusrated in figure 3(b). The residual generated is added with the input image of stage two to finally get the restored output.

3.3. Fourier Amplitude Recovery Unit

Since convolution usually takes place in the spatial domain, we use a residual network to refine the amplitude spectrum and preserve spatial features. The input features are first passed through a 1×1 convolution layer to generate F_{res}, as shown in Figure 4, before performing the DFT. The amplitude spectrum is then fed into a Transformer block, followed by two 1×1 convolution layers to generate $\hat{F_{res}}$. Subsequently, the calculation of F_{fft} can be accomplished by employing the IDFT, given by $F_{fft} = F^{-1}(A(\hat{F_{res}}), P(F_{res}))$. The main branch consist of series of 3×3 convolution layers to generate F_{main}. Finally output from the FARU (F_{out}) is obtained by summing results from all three branches ie. F_{res}, F_{main} and F_{fft}.

The Transformer blocks efficiently extract long-range



(a)

(b)

Figure 3. Architectures for (a) amplitude refinement unit (ARU) and (b) phase refinement unit (PRU) of FAPRNet.

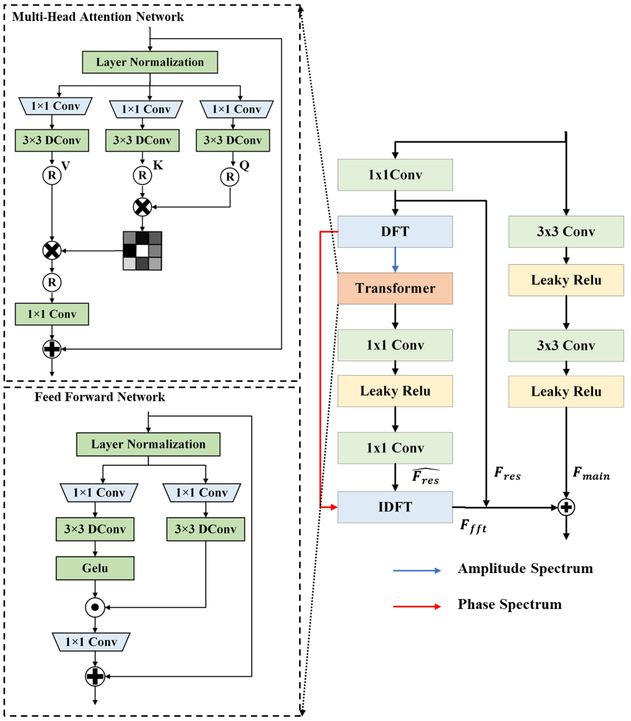


Figure 4. Architectures of Fourier amplitude recovery unit

dependencies and global features, overcoming computational challenges associated with traditional self-attention layers. By applying self-attention across channels rather than spatial dimensions, the computation of cross-covariance across channels generate attention maps encapsulating global context. Multi-head attention modules efficiently generate query (Q), key (K), and value (V) projections through 1×1 convolutions and 3×3 depth-wise convolutions. Reshaping of query and key projections enables their dot-product interaction, yielding a transposed-attention map A_t. Attention is evaluated as: $A_t = V.Softmax(\frac{K.Q}{\alpha})$. The Feed Forward Network (FFN) incorporates a gating mechanism and depth-wise convolutions, managing information flow across hierarchical levels in the architecture. Expansion factor (γ) is utilized to reduce computational complexity by expanding and then reducing the number of channels within the FFN.

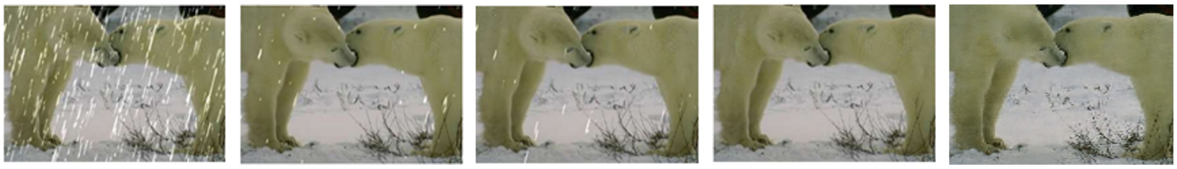
Architecture for FEB is similar to FARU with a difference that residual branch operates in spatial domain i.e F_{res} is directly given as input to Transformer block.

3.4. Loss Function

The Mean Square Error (MSE) often leads to an over-smoothed image due to its squared penalty. To strike a balance between preserving the details and removing degradation Mean Absolute Error (MAE) is utilised. In addition to pixel-level loss functions that guarantee faithful reconstruction, our method incorporates a frequency-domain loss function computed via the DFT to ensure the recovery of global image information. Y_1 and Y_2 are outputs for ampli-

Table 1. Comparison of image deraining results on Test100, Rain100H, Rain100L and Test1200 datasets

Model	Network	Test100		Rain100H		Rain100L		Test1200		Params (M)	GFLOPS
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM		
CNN	DerainNet[11]	22.77	0.810	14.92	0.592	27.03	0.884	23.38	0.835	0.058	1.453
	RESCAN[20]	25.00	0.835	26.36	0.786	29.80	0.881	30.55	0.882	1.040	20.361
	MSPFN[15]	27.50	0.876	28.66	0.860	32.40	0.933	32.39	0.916	13.220	604.700
	MPRNet[35]	30.27	0.897	30.41	0.890	36.40	0.965	32.91	0.916	3.640	141.280
	HINet[5]	30.29	0.906	30.65	0.894	37.28	0.970	33.05	0.919	3.720	170.710
	Fourmer[37]	30.54	0.911	30.76	0.896	37.47	0.970	33.05	0.919	0.400	16.753
Transformer	Restormer[34]	32.00	0.923	31.46	0.904	38.99	0.978	33.19	0.926	26.120	140.990
Ours	FAPRNet	**33.92**	**0.948**	**33.47**	**0.940**	**41.84**	**0.980**	**34.29**	**0.951**	5.220	25.080



Input MPRNet DerainNet Ours Ground Truth

Figure 5. Visual comparison of image deraining results with prior works on Rain100L dataset

tude and phase refinement stage respectively. For amplitude refinement, we minimize the loss function expressed as:

$$L_{amp} = \sum_{i=0}^{i=2} ||A(I_{clean_i}) - A(Y_{1_i})|| \tag{4}$$

In the second stage, we enhance the phase spectrum of predicted restored images by employing loss function in both spatial and frequency domain:

$$L_{fft} = ||A(I_{clean}) - A(Y_2)|| + ||P(I_{clean}) - P(Y_2)|| \tag{5}$$

$$L_{spatial} = ||I_{clean} - Y_2|| \tag{6}$$

Furthermore, to ensure predicted images are perceptually similar to the ground truth we add Structural Similarity Index Measure (SSIM) loss to the total loss function. The total loss function for training the architecture is expressed as:

$$L_{total} = L_{spatial} + \lambda_1(L_{amp} + L_{fft}) + \lambda_2(L_{SSIM}) \tag{7}$$

In this equation λ_1 and λ_2 are set to 0.1. The value of λ_1 and λ_2 are determined through experimentation and validation to balance the contribution of losses in dual domain while training.

4. Experiments

The proposed FAPRNet architecture's effectiveness was evaluated on diverse image restoration tasks, encompassing deraining, enhancement, and dehazing. This section presents the experimental results, including ablation studies, and compares our method to state-of-the-art approaches. We employ established image quality metrics Peak Signal-to-Noise Ratio (PSNR) and SSIM for evaluation alongside model parameters (Params) and computational complexity measured in Giga Floating-Point Operations (GFLOPs).

4.1. Image Deraining

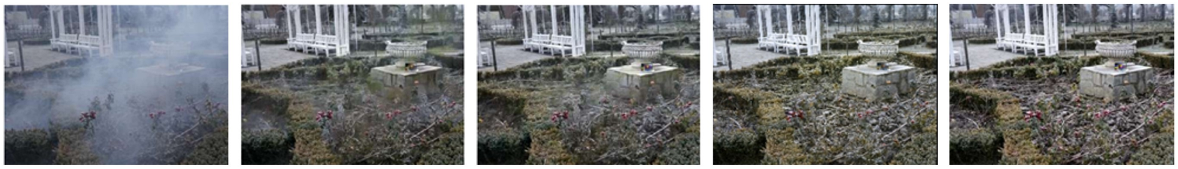
Training for the image deraining application was conducted using the Rain13k dataset, and the architecture's performance was subsequently evaluated on the Test100 [36], Rain100H [32], Rain100L [32], and Test1200 [36] datasets. Comparative results for the FAPRNet architecture, alongside CNN and Transformer architectures, are presented in Table 1. FAPRNet demonstrates a 2.925dB improvement in PSNR across all datasets compared to the CNN-based Fourmer architecture [37], while necessitating $1.49\times$ more computations. Conversely, when compared with the Transformer architecture Restormer [34], the proposed architecture achieves a 1.97dB improvement in PSNR across all datasets, with a substantial reduction of $5\times$ parameters and $5.62\times$ computations, respectively. Thus, the proposed architecture strikes a balance between performance and computational complexity for image deraining applications. A qualitative comparison of the deraining application results is depicted in Figure 5.

4.2. Image Dehazing

The proposed FAPRNet architecture is evaluated for image dehazing using synthetic and real-world datasets. The RESIDE dataset [19] introduces synthetic haze in both indoor

Table 2. Comparison of image dehazing results on NH-Haze, Dense-Haze, and RESIDE datasets.

Method	Network	NH-Haze		Dense-Haze		RESIDE		Params(M)	GFLOPs
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM		
CNN	DCP[13]	10.57	0.519	10.06	0.387	15.09	0.765	-	-
	DehazeNet[3]	16.62	0.524	13.84	0.425	20.64	0.779	0.010	0.58
	AOD-Net[18]	15.40	0.564	13.14	0.414	19.82	0.818	0.002	0.11
	FFA-Net[23]	19.87	0.691	14.39	0.452	36.39	0.988	4.680	288.10
	Fourmer[37]	19.91	0.721	15.95	0.492	37.32	0.990	1.290	20.60
	IRNext[9]	20.55	0.813	17.60	0.659	40.19	**0.996**	5.460	-
Transformer	ITB-Dehaze[21]	21.44	0.710	16.31	0.561	-	-	110.000	-
Ours	FAPRNet	**25.26**	**0.880**	**21.39**	**0.770**	**41.43**	0.995	5.220	25.08



Input Dehammer IRNext Ours Ground Truth

Figure 6. Visual comparison of image dehazing results with prior works on NH-haze dataset

and outdoor scenes, while real-world datasets NH-Haze [2] and Dense-Haze [1] assess the model's performance in realistic scenarios. The proposed FAPRNet architecture outperforms all CNN-based architectures in terms of PSNR and SSIM on synthetic as well as real-world datasets as evident in Table 2. The proposed architecture results in higher PSNR across all datasets. In comparison with recent algorithms Fourmer [37] and IRNext [9], the proposed FAPRNet provides 5.35dB and 4.71dB of improvement in PSNR respectively, on NH-Haze dataset while maintaining computational complexity. Furthermore, FAPRNet also outperforms the Transformer based architecture ITB-Dehaze [21] by 3.82dB in terms of PSNR while requiring 21× fewer operations. The visual comparison of results is shown in Figure 6.

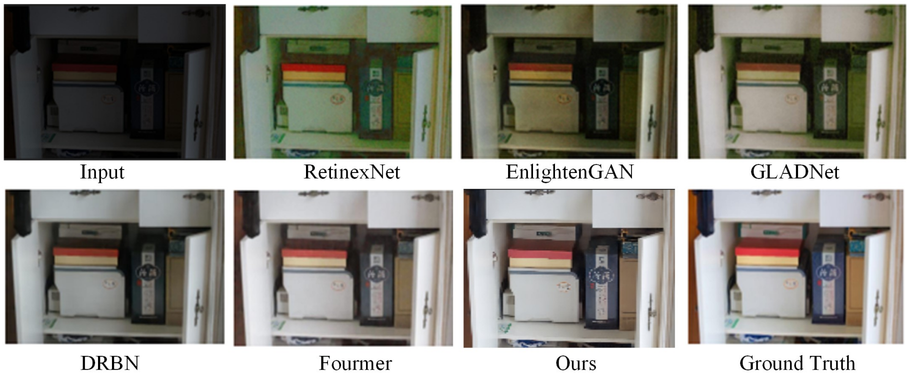
4.3. Low Light Image Enhancement

The proposed architecture was evaluated for low light image enhancement task on LOL dataset [6]. It outperforms all previously proposed CNN-based architectures. Results are compared with previously proposed architectures as given in Table 3. The proposed architecture achieves a 6.35dB improvement in PSNR over Fourmer architecture [37] with a $5\times$ increase in FLOPs. However, compared to Retinexformer [4], FAPRNet significantly improves the performance in terms of PSNR and SSIM with increase in computations. Figure 7 depicts that the result of the proposed architecture is visually closer to the ground truth. FAPRNet can achieve state-of-the-art performance with an increase in computational complexity.

4.4. Ablation Studies

Effectiveness of Fourier Prior Three different models are evaluated to determine the effectiveness of Fourier prior. Model 1 is trained by proving Y i.e output of ARU in stage 1 directly to stage 2 for phase refinement. Notably, this process does not incorporate the phase of the input degraded image I within stage 1. Therefore, the learning process for stage 1 can be denoted as $(I \rightarrow Y)$, and stage 2 as $(P(Y) \rightarrow P(I_{clean}))$. Model 2 performs phase refinement $(P(I) \rightarrow P(I_{clean}))$ at first stage followed by amplitude refinement $(A(I) \rightarrow A(I_{clean}))$. And finally last model is the proposed architecture performing amplitude refinement followed by phase refinement. Based on the results in Table 4, Model 2 performs the worst. This is likely because the training signal for this model, $F^{-1}(A(I), P(I_{clean}))$, contains degradation throughout the entire image. This extensive degradation makes it difficult for the first sub-network in Model 2 to learn the transformation effectively and thereby not utilizing first stage architecture efficiently. Furthermore, FAPRNet has improved performance as compared to Model 1, suggesting that preserving the phase information of the degraded image $(P(I))$ plays a crucial role in the Fourier prior.

Efficacy of Multi-scale Inputs in ARU Incorporating inputs at multiple scales is essential for capturing both coarse and fine-grained features, leading to improved model performance. The lack of multi-scale inputs leads to a de-


Input

RetinexNet

EnlightenGAN

GLADNet

DRBN

Fourmer

Ours

Ground Truth

Figure 7. Visual comparison of image low light enhancement results with prior works on LOL dataset

Table 3. Comparison of low light image enhancement application on LOL dataset.

Method	Network	PSNR	SSIM	Params (M)	GFLOPS
CNN	RetinexNet[6]	16.77	0.425	0.84	148.54
	GLADNet[26]	19.72	0.680	1.13	275.32
	EnlightenGAN[16]	17.48	0.647	8.37	72.61
	DRBN[33]	20.13	0.801	0.58	42.41
	Uretinex-Net[30]	21.31	0.835	1.23	68.37
	Fourmer[37]	25.61	0.840	0.08	5.03
Transformer	Retinexformer[4]	25.16	0.845	1.61	15.57
Ours	FAPRNet	**31.96**	**0.936**	5.22	25.08

Table 4. Effectiveness of Fourier Prior across various architecture configurations

Architecture	Test100		Dense-Haze		LOL	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Model 1	29.43	0.880	17.91	0.730	28.42	0.840
Model 2	27.26	0.813	16.43	0.520	25.19	0.781
FAPRNet	**33.92**	**0.948**	**21.39**	**0.770**	**31.96**	**0.936**

Table 5. Effectiveness of multi-scale inputs in the ARU

Architecture	Test100		Dense-Haze		LOL	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
w/o multi-resolution inputs	31.97	0.924	20.82	0.710	30.68	0.910
with multi-resolution inputs	**33.92**	**0.948**	**21.39**	**0.770**	**31.96**	**0.936**

Table 6. Efficacy of multi-domain loss function

Loss Function	Test100		Dense-Haze		LOL	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
$L_{spatial}$	28.67	0.891	18.47	0.690	28.41	0.878
$L_{spatial} + \lambda_1 L_{fft}$	31.53	0.913	20.72	0.710	29.82	0.890
$L_{spatial} + \lambda_1(L_{amp} + L_{fft})$ $+ \lambda_2 L_{SSIM}$	**33.92**	**0.948**	**21.39**	**0.770**	**31.96**	**0.936**

cline in all performance metrics, as indicated by the results in Table 5. These observations underscore the significance of multi-scale inputs in attaining optimal results through the learning of features at various scales.

Efficacy of Loss Function In order to comprehensively evaluate the efficacy of the implemented loss function, we examine different combinations of losses. These include first, solely spatial domain loss $(L_{spatial})$ between the final predicted and ground truth image. Second, spatial domain loss combined with frequency domain loss composed of MSE of amplitude and phase of restored image and ground truth image $(L_{spatial} + \lambda_1 L_{fft})$, and finally, a multi-scale integration involving amplitude spectrum from stage one and phase spectrum from stage two, along with final spatial domain and SSIM loss $(L_{spatial} + \lambda_1(L_{amp} + L_{fft}) + \lambda_2 L_{SSIM})$. The results indicate that the implemented multi-scale dual-domain loss function significantly improves model performance, as illustrated in Table 6.

5. Conclusion

This work presents a Fourier prior for image restoration applications, leveraging the observation that the amplitude in the frequency domain holds crucial information about degradation, while the phase retains background information in the given image. We propose a two-stage architecture that independently refines the amplitude and phase spectrum of degraded images that effectively separates image degradation from content, leading to improved restoration results. The proposed architecture demonstrates superior performance over recently proposed architectures in various image restoration tasks, including deraining, dehazing, and low-light enhancement. The FAPRNet architecture not only overcomes the computational overhead associated with Transformers but also exhibits superior performance compared to standalone CNN or Transformer models.

References

[1] Codruta Orniana Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In *Proceedings*

of the IEEE International Conference on Image Processing, pages 1014–1018, 2019.

[2] Codruta Orniana Ancuti, Cosmin Ancuti, and Radu Timofte. Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1798–1805, 2020.

[3] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016.

[4] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12504–12513, 2023.

[5] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 182–192, 2021.

[6] Wei Chen, Wang Wenjing, Yang Wenhan, and Liu Jiaying. Deep retinex decomposition for low-light enhancement. In *Proceedings of the British Machine Vision Conference*, 2018.

[7] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, and Q. Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 569–578, 2021.

[8] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2021.

[9] Yuning Cui, Wenqi Ren, Sining Yang, Xiaochun Cao, and Alois Knoll. Irnext: Rethinking convolutional network design for image restoration. In *Proceedings of the International Conference on Machine Learning*, pages 6545–6564, 2023.

[10] AA Dixit and AC Phadke. Image de-noising by non-local means algorithm. In *International Conference on Signal Processing, Image Processing and Pattern Recognition*, pages 275–277, 2013.

[11] Xueyang Fu, Jiabin Huang, Xinghao Ding, Yinghao Liao, and John Paisley. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Transactions on Image Processing*, 26(6):2944–2956, 2017.

[12] Xin Guo, Xueyang Fu, Man Zhou, Zhen Huang, Jialun Peng, and Zheng-Jun Zha. Exploring fourier prior for single image rain removal. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 935–941, 2022.

[13] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2010.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[15] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8343–8352, 2020.

[16] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Chen, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021.

[17] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2034–2042, 2021.

[18] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4770–4778, 2017.

[19] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018.

[20] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European Conference on Computer Vision*, pages 254–269, 2018.

[21] Yangyi Liu, Huan Liu, Liangyan Li, Zijun Wu, and Jun Chen. A data-centric solution to nonhomogeneous dehazing via vision transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1406–1415, 2023.

[22] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 257–265, 2017.

[23] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11908–11915, 2020.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention*, pages 234–241, 2015.

[25] Chull Hwan Song, Hye Joo Han, and Yannis Avrithis. All the attention you need: Global-local, spatial-channel attention for image retrieval. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 2754–2763, 2022.

[26] Wenjing Wang, Chen Wei, Wenhan Yang, and Jiaying Liu. Gladnet: low-light enhancement network with global awareness. In *Proceedings of the International Conference on Automatic Face Gesture Recognition*, pages 751–755, 2018.

[27] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE In-*

ternational Conference on Computer Vision, pages 568–578, 2021.

[28] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, pages 3–19, 2018.

[29] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 22–31, 2021.

[30] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5891–5900, 2022.

[31] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal attention for long-range interactions in vision transformers. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 30008–30022, 2021.

[32] Wenhan Yang, Robby T. Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1694, 2017.

[33] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3060–3069, 2020.

[34] S. Zamir, A. Arora, S. Khan, M. Hayat, F. Khan, and M. Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5718–5729, 2022.

[35] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14821–14831, 2021.

[36] He Zhang, Vishwanath Sindagi, and Vishal M. Patel. Image de-raining using a conditional generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):3943–3956, 2020.

[37] Man Zhou, Jie Huang, Chun-Le Guo, and Chongyi Li. Fourmer: An efficient global modeling paradigm for image restoration. In *Proceedings of the International Conference on Machine Learning*, pages 42589–42601, 2023.