This WACV 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.

Dropout the High-rate Downsampling: A Novel Design Paradigm for UHD Image Restoration

Chen Wu[1] Ling Wang[2] Long Peng[1] Dianjie Lu[3] Zhuoran Zheng[4] *
[1]University of Science and Technology of China
[2]The Hong Kong University of Science and Technology (Guangzhou)
[3]Shandong Normal University
[4]Sun Yat-sen University
wuchen5X@mail.ustc.edu.cn

Abstract

With the popularization of high-end mobile devices, Ultra-high-definition (UHD) images have become ubiquitous in our lives. The restoration of UHD images is a highly challenging problem due to the exaggerated pixel count, which often leads to memory overflow during processing. Existing methods either downsample UHD images at a high rate before processing or split them into multiple patches for separate processing. However, high-rate downsampling leads to significant information loss, while patch-based approaches inevitably introduce boundary artifacts. In this paper, we propose a novel design paradigm to solve the UHD image restoration problem, called D2Net. D2Net enables direct full-resolution inference on UHD images without the need for high-rate downsampling or dividing the images into several patches. Specifically, we ingeniously utilize the characteristics of the frequency domain to establish long-range dependencies of features. Taking into account the richer local patterns in UHD images, we also design a multi-scale convolutional group to capture local features. Additionally, during the decoding stage, we dynamically incorporate features from the encoding stage to reduce the flow of irrelevant information. Extensive experiments on three UHD image restoration tasks, including low-light image enhancement, image dehazing, and image deblurring, show that our model achieves better quantitative and qualitative results than state-of-the-art methods.

1. Introduction

With the widespread adoption of high-end mobile devices, such as smartphones, high-resolution images have become almost mainstream, capable of capturing more detailed visual content. However, various factors often lead to poor image quality during the imaging process, such as insufficient lighting conditions and unfavorable weather conditions. Image restoration constitutes a significant and demanding undertaking in the field of computer vision, aiming to recover degraded images into clear, realistic, and clean images. Although a series of promising methods have emerged for the task of image restoration, achieving satisfactory results, especially the learning-based methods [1, 5, 6, 17, 22–24, 33, 35, 38, 40–43, 47], these methods face challenges when it comes to transferring them to Ultra-high-definition (UHD) images. The high pixel count of UHD images (*i.e.*, 3840 × 2160) often leads to memory overflow issues during the processing of these methods.

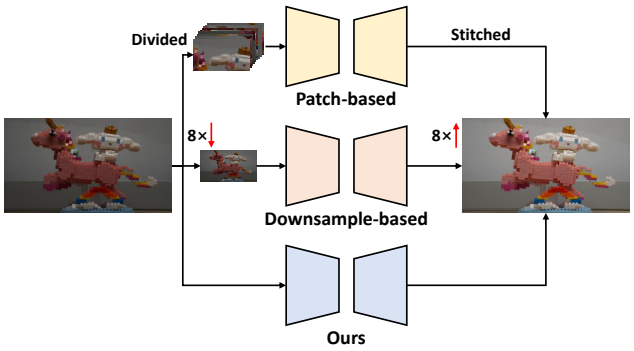

Figure 1. Comparison between previous methods and our proposed method for UHD image restoration. Due to hardware limitations, previous methods had to resort to patch-based or downsample-based approaches to enable consumer-grade GPUs to process UHD images. However, patch-based methods [32] introduce boundary artifacts during subsequent stitching processes, while downsample-based methods [18, 46] result in significant information loss, both of which can affect the quality of image enhancement. In contrast, our D2Net allows for direct full-resolution inference on UHD images.

*Corresponding author

To tackle this issue, there are currently two primary processing methods: patch-based methods [32] and downsample-based methods [18, 37, 46], as illustrated in Fig. 1. Patch-based methods refer to the approach of dividing UHD images into non-overlapping patches during processing. These patches are individually processed and then stitched to form the final restored image. However, this method inevitably introduces boundary artifacts during the stitching process. Additionally, the process of dividing and stitching patches incurs additional time overhead. Downsample-based methods involve high-rate downsampling of UHD images before processing. After completing the image restoration steps, the images are then upsampled to restore their original resolution. Undoubtedly, high-rate downsampling leads to a significant loss of valuable information, which impacts the quality of the final enhanced image.

In this paper, we design a novel framework to solve the UHD image restoration problem called D2Net, which allows for direct full-resolution inference of UHD images. This represents a new design paradigm that completely discards the conventional approaches to UHD image restoration (patch-based and downsample-based). Specifically, we delve deeper into the characteristics of the Fourier domain, as previous works primarily focus on the Fourier transform's capability to some extent in disentangling image degradation and content components. We find that when an image is transformed from the spatial domain to the frequency domain, the features in the frequency domain naturally exhibit interactions. In light of this, we design the Fourier-based Global Feature Extraction (FGFE) Module to capture long-range dependencies of features. In addition, UHD images, due to their high resolution, contain richer and more complex patterns. Therefore, we design the Multi-scale Local Feature Extraction (MLFE) Module to capture the multi-scale local features of the UHD images. It is worth noting that the convolution kernel sizes in the MLFE module are relatively large. This is because, considering the high-resolution characteristics of UHD images, the commonly used 1×1 convolution or 3×3 convolution are often unable to capture high-quality local features. Moreover, we also design an Adaptive Feature Modulation Module (AFMM) that dynamically adjusts the fusion of encoded and decoded features. Unlike the simple stacking of features used in previous approaches, this method effectively suppresses the flow of irrelevant information, leading to enhanced performance. Extensive experiments across three datasets demonstrate that D2Net surpasses prior arts in terms of restoration quality and generalization ability.

In summary, our main contributions are three-fold:

- To the best of our knowledge, we propose the first UHD restoration method that enables direct full-resolution inference on consumer-grade GPUs, without the need for high-rate downsampling or dividing the images into several patches.

- We design a novel UHD restoration framework named D2Net. D2Net leverages the FGFE module to achieve long-range modeling and utilizes the MLFE module to capture multi-scale local features. In addition, we design the AFMM to dynamically adjust the fusion of encoded and decoded features.

- We conduct comprehensive quantitative and qualitative evaluations of our method, which demonstrate that D2Net achieves state-of-the-art results across multiple UHD image restoration tasks.

2. Related Work

2.1. UHD Image Restoration

In recent years, data-driven CNN [5, 17, 22–24, 33, 43] or Transformer [1–3, 6, 19, 35, 39, 42, 47] architectures have been shown to out perform conventional restoration approaches. However, these methods often focus on processing low-resolution images (400×600), and when faced with high-resolution images (3840×2160), they often encounter problems such as memory overflow.

Learning-based UHD image restoration has received some attention in recent years. Zheng *et al*. [48] incorporate bilateral learning into deep learning by using a network to learn local affine coefficients of the bilateral grid from low-resolution images and apply them to high-resolution images. Subsequently, several methods based on deep bilateral learning have emerged [49, 50]. LLFormer [32] divides the UHD image into multiple non-overlapping patches and processes them individually, treating the UHD image as multiple low-resolution images. Li *et al*. [18] discover that a high-resolution image and its corresponding low-resolution version exhibit similar amplitude patterns. Based on this observation, they propose UHDFour which restores the low-light UHD images in the frequency domain. Similarly, the UHD-SFNet [36] also attempts to enhance underwater UHD images in the frequency domain. UHDformer [30] designs two branches to handle UHD images, placing the global feature modeling on the low-resolution branch to avoid memory overflow, significantly reducing the computational complexity of the model. However, these methods introduce boundary artifacts and incur additional time overhead when patch-based strategies are employed, while downsample-based strategies result in a significant loss of useful information. These factors adversely affect the quality of the restored images. Our method performs full-resolution inference directly on UHD images, thereby avoiding the problems mentioned above.

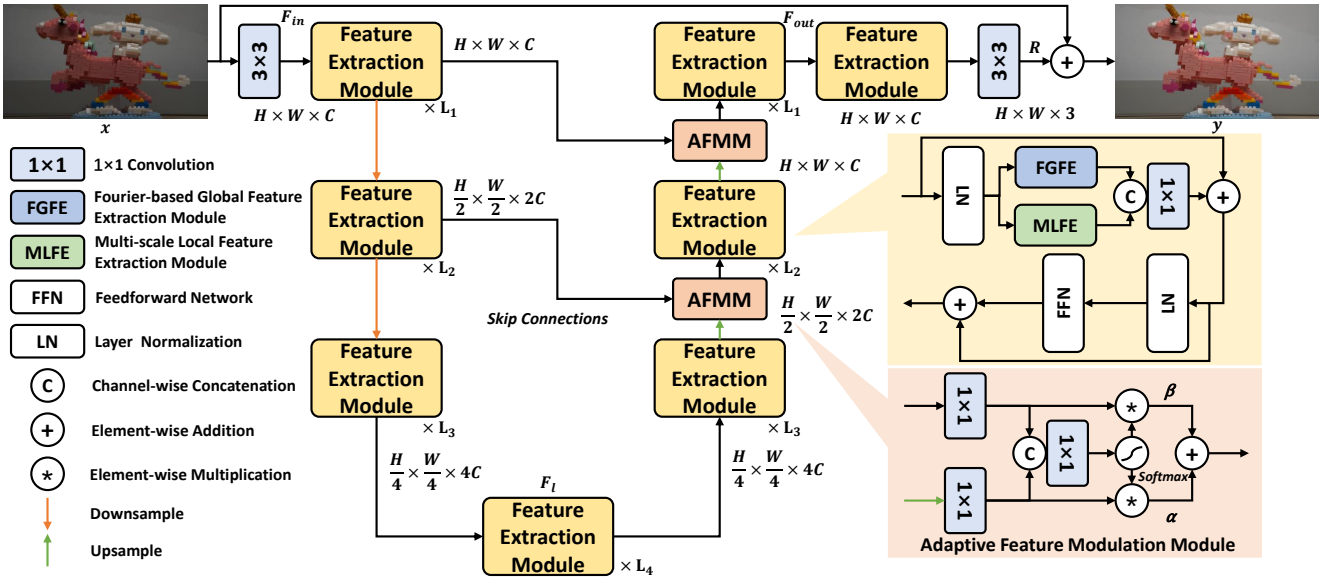

Figure 2. An overview of the network architecture of our D2Net. The degraded image is forwarded to a 4-level hierarchical encoder-decoder Unet-like structure to get a normal-light image. D2Net primarily consists of several Feature Extraction Modules (FEMs) and an Adaptive Feature Modulation Modules (AFMMs). FEM is responsible for feature extraction through global and local feature modeling. AFMM dynamically adjusts the fusion of encoded and decoded features to suppress the flow of irrelevant information.

2.2. Image Restoration in Frequency Domain

Learning with frequency domain has become increasingly common in low-level tasks. DWSR [9] is trained in the wavelet domain with four input and output channels respectively for super-resolution tasks. Compared to learning with the wavelet domain, learning with the Fourier domain is more commonly used in image restoration tasks. Fourier transform has been widely employed for tasks such as image de-raining [10], image dehazing [44], low-light image enhancement [18, 31], image deblurring [15, 21], image artifact removal [43] and image warping [41]. These methods are mostly based on the observation that tshe Fourier transform can effectively decouple degraded information from image content. For instance, in low-light image enhancement tasks, a significant portion of illumination information is captured in the amplitude components, while the phase component contains most of the structure representation. In contrast to these methods, we focus on the excellent global modeling capability of the Fourier domain. When an image undergoes Fourier transformation, the features in the frequency domain inherently exhibit interactions. This greatly encourage us to try using Fourier transform to capture the long-range dependencies between features.

3. Method

In this section, we first take a brief review of the Fourier transform, and then we provide the overall pipeline of our method and further details on the critical components of our approach below.

3.1. Preliminary

The Fourier transform is an extensively employed methodology for analyzing the frequency characteristics of an image [8]. In the case of images possessing multiple color channels, the Fourier transform is individually applied to each channel. Given an image $\boldsymbol{x} \in \mathbb{R}^{\mathrm{H} \times \mathrm{W} \times \mathrm{C}}$, the Fourier transform $\mathcal{F}$ converts it from the spatial domain to Fourier domain as the complex component $\mathcal{F}(\boldsymbol{x})$, which is expressed as:

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

where $\boldsymbol{u}$ and $\boldsymbol{v}$ and the coordinates of the Fourier domain. $\mathcal{F}^{-1}(\cdot)$ defines the inverse Fourier transform. The amplitude component $\mathcal{A}(\boldsymbol{x})(\boldsymbol{u}, \boldsymbol{v})$ and phase component $\mathcal{P}(\boldsymbol{x})(\boldsymbol{u}, \boldsymbol{v})$ are expressed as:

$$\begin{aligned} \mathcal{A}(\boldsymbol{x})(\boldsymbol{u}, \boldsymbol{v})) &= \sqrt{\mathcal{R}^2(\boldsymbol{x})(\boldsymbol{u}, \boldsymbol{v})) + \mathcal{I}^2(\boldsymbol{x})(\boldsymbol{u}, \boldsymbol{v}))}, \\ \mathcal{P}(\boldsymbol{x})(\boldsymbol{u}, \boldsymbol{v})) &= \arctan[\frac{\mathcal{I}(\boldsymbol{x})(\boldsymbol{u}, \boldsymbol{v}))}{\mathcal{R}(\boldsymbol{x})(\boldsymbol{u}, \boldsymbol{v}))}], \end{aligned} \tag{2}$$

where $\mathcal{R}(\boldsymbol{x})(\boldsymbol{u}, \boldsymbol{v})$ and $\mathcal{I}(\boldsymbol{x}(u, \boldsymbol{v})$ represent the real and imaginary parts respectively.

By examining Eqs. 1 and 2, it becomes evident that once an image is mapped to the Fourier domain, inherent correlations exist among individual pixels. This motivated us to

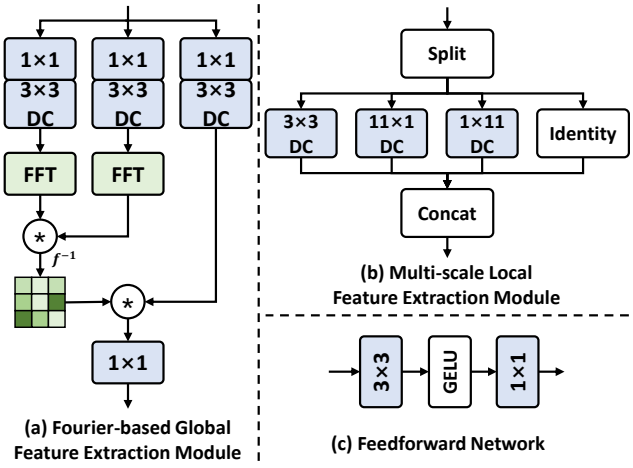

Figure 3. The illustration of (a) Fourier-based Global Feature Extraction (FGFE) Module, (b) Multi-scale Local Feature Extraction (MLFE) Module and (c) Feedforward Network (FFN).

develop a method for modeling long-range feature dependencies in the Fourier domain.

3.2. Overview

An overview of D2Net is shown in Fig. 2. For a given degraded image $\boldsymbol{x} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times 3}$, we first map it to the feature space through a 3×3 convolution to obtain the deep features $\boldsymbol{F_{in}} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$; where $\mathrm{H}\times\mathrm{W}$ is the spatial resolution and C denotes the channel. Subsequently, deep features $\boldsymbol{F_{in}}$ will pass through an encoder to obtain latent features $\boldsymbol{F}_l \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 4C}$. The latent features will be combined with encoded features for restoration, resulting in the final output $\boldsymbol{F_{out}} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$. The overall architecture of D2Net is a 4-level hierarchical encoder-decoder UNet-like [25] structure. During the encoding stage, the feature maps progressively reduce the spatial resolution while increasing the channel capacity. Conversely, during the decoding stage, the process is reversed, where the spatial resolution is increased while the channel capacity decreases. Unlike the simple stacking of features used in previous UNet-like approaches, we design the Adaptive Feature Modulation Module to adjust the fusion of encoded and decoded features dynamically. Lastly, the final features $\boldsymbol{F_{out}}$ are refined to obtain the residual image $\boldsymbol{R}$ to which degraded image is added to obtain the normal-light image: $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{R}$.

3.3. Fourier-based Global Feature Extraction Module

FGFE module aims to capture long-range dependencies of features. For UHD images, modeling long-range dependencies between features is a highly challenging task due to the high-resolution nature of UHD images and this often leads to memory overflow issues. This motivates us to develop an efficient global modeling approach to address the modeling of long-range dependencies between features. We observe that when an image is transformed from the spatial domain to the Fourier domain, the features in the Fourier domain inherently exhibit interactions. Therefore, leveraging the Fourier transform, we design an efficient global modeling module, and the architecture is shown in Fig. 3 (a). Following the design principles of self-attention [29], we first map the feature maps through a convolutional group to obtain $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$. Next, we use Fourier transform $\mathcal{F}(\cdot)$ to convert the spatial domain representation of $\boldsymbol{Q}$ and $\boldsymbol{K}$ into the frequency domain. We compute the frequency attention map $\boldsymbol{M}$ in the Fourier domain and then we apply inverse Fourier transform $\mathcal{F}^{-1}(\cdot)$ to return from the frequency domain to the spatial domain and compute the final result by combining the transformed attention map with $\boldsymbol{V}$. Given the input feature maps $\boldsymbol{F_{in}}$, this process is summarized as:

$$\boldsymbol{Q} = \boldsymbol{F_{in}} \cdot \boldsymbol{W_Q}, \boldsymbol{K} = \boldsymbol{F_{in}} \cdot \boldsymbol{W_K}, \boldsymbol{V} = \boldsymbol{F_{in}} \cdot \boldsymbol{W_V}, \quad (3)$$

$$\boldsymbol{F_{out}} = \mathrm{Conv}(\boldsymbol{V} \cdot \mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{Q}) \cdot \mathcal{F}(\boldsymbol{k}))), \quad (4)$$

where $\boldsymbol{W_Q}$, $\boldsymbol{W_K}$ and $\boldsymbol{W_V}$ are different projection matrices. $\mathrm{Conv}(\cdot)$ stands for the 1×1 convolution layer. A convolutional group consists of a 1×1 depth-wise convolution followed by a 3×3 convolution.

3.4. Multi-scale Local Feature Extraction Module

Due to the high-resolution nature of UHD images, they contain richer and more detailed patterns compared to images with regular resolutions. In addition, smaller convolution kernel sizes would make it difficult to capture high-quality local features in UHD images, as their high-resolution characteristics require a larger receptive field. Therefore, we design the MLFE module and the architecture is shown in Fig. 3 (b). To reduce computational complexity, we employ depth-wise convolutions in this module and selectively process only a subset of channels [45]. For input $\boldsymbol{F_{in}}$, we split it into four groups along the channel dimension,

$$\begin{aligned}\boldsymbol{F_{hw}}, \boldsymbol{F_w}, \boldsymbol{F_h}, \boldsymbol{F_{id}} &= \mathrm{Split}(\boldsymbol{F_{in}}) \\ &= \boldsymbol{F_{in:\!:g}}, \boldsymbol{F_{in:g:2g}}, \boldsymbol{F_{in:2g:3g}}, \boldsymbol{F_{in:3g:}},\end{aligned} \quad (5)$$

where $\boldsymbol{g}$ is the channel numbers of convolution branches. We set a ratio $\boldsymbol{r_g}$ to determine the branch channel numbers by $\boldsymbol{g} = \boldsymbol{r_g} \cdot \boldsymbol{C}$. $\mathrm{Split}(\cdot)$ is the feature splitting operation. Next, the splitting inputs are fed into different parallel branches, this process can be formulated as:

$$\begin{aligned}\boldsymbol{F'_{hw}} &= \mathrm{DWConv}^{g\to g}_{k_s\times k_s} g(\boldsymbol{F_{hw}}), \\ \boldsymbol{F'_{w}} &= \mathrm{DWConv}^{g\to g}_{1\times k_b} g(\boldsymbol{F_{w}}), \\ \boldsymbol{F'_{h}} &= \mathrm{DWConv}^{g\to g}_{k_b\times 1} g(\boldsymbol{F_{h}}), \\ \boldsymbol{F'_{id}} &= \boldsymbol{F_{id}},\end{aligned} \quad (6)$$

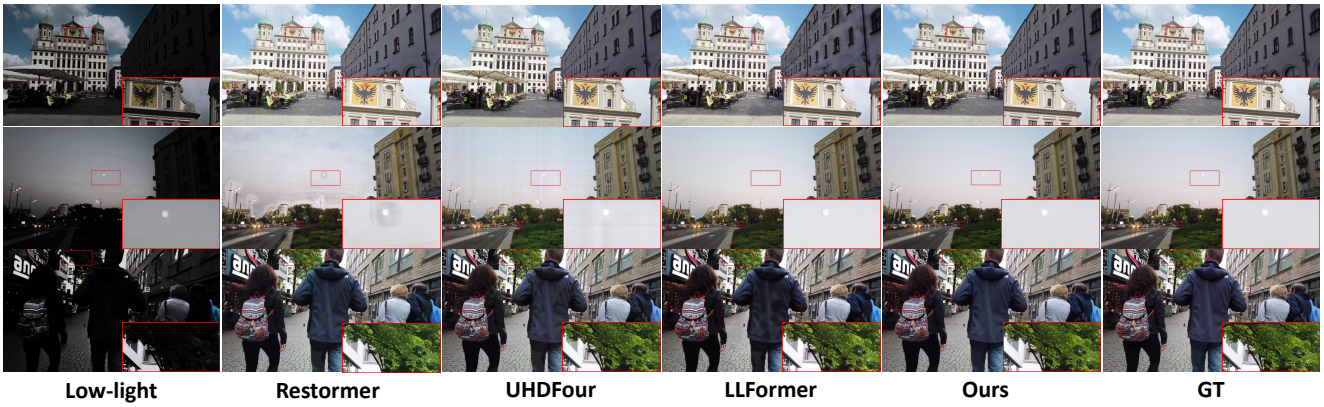

Figure 4. Visual quality comparisons with state-of-the-art methods on UHD-LOL4K dataset. Please zoom in for details.

where k_s denotes the small square kernel size and k_b represents the band kernel size. $\mathrm{DWConv}(\cdot)$ is the depth-wise convolution. Finally, the outputs from each branch are concatenated,

$$F_{out} = \mathrm{Concat}(F'_{\mathbf{hw}}, F'_{\mathbf{w}}, F'_{\mathbf{h}}, F'_{\mathbf{id}}), \tag{7}$$

where $\mathrm{Concat}(\cdot)$ denotes the feature concatenating operation.

3.5. Adaptive Feature Modulation Module

Previous Unet-like methods typically employ a simple concatenation of encoded and decoded features for processing, which we consider to be a crude approach. In the process of obtaining encoded features at each level in the Unet-like structure, there is a progression from coarse to fine-grained for the feature maps. This means that the feature maps obtained in the initial few steps contain a lot of irrelevant information. Therefore, we design the AFMM to reduce the flow of irrelevant information. The encoded and decoded features at the same level are first aggregated together to generate interactions. Subsequently, a softmax operation is applied to obtain the corresponding weight maps. Based on the weight maps, the encoded and decoded features adaptively adjust their feature maps and are then summed together to form the final feature representations. Given the encoded features F_{en} and decoded features F_{de} at the same level, this process is summarized as:

$$M = \mathrm{Conv}_2(\mathrm{Concat}(\mathrm{Conv}_0(F_{en}), \mathrm{Conv}_1(F_{de}))), \tag{8}$$
$$F_{out} = \mathrm{Conv}_0(F_{en}) \cdot \mathrm{S}_0(M) + \mathrm{Conv}_1 \cdot \mathrm{S}_1(M), \tag{9}$$

where $\mathrm{Conv}(\cdot)$ denotes the 1×1 convolution and $\mathrm{S}(\cdot)$ is the Softmax operation.

3.6. Feedforward Network

FFN is designed to map features into a more compact representation [27, 37]. It consists of a 3×3 convolution, a 1×1 convolution, and the GULE activation function. FFN initially increases the number of channels in the feature maps by several times. After applying the GULE activation function, the number of channels is then adjusted back to the original size. The entire process can be represented as follows:

$$F_{out} = \mathrm{Conv}_{1\times1}(\mathrm{GELU}(\mathrm{Conv}_{3\times3}(F_{in}))), \tag{10}$$

where F_{in} and F_{out} are the input features and output features, respectively.

Table 1. Comparison of quantitative results on UHD-LOL4K dataset. Best and second best values are indicated with **bold** text and underlined text respectively.

Methods	Type	Venue	PSNR	SSIM	Param
Z_DCE++	non-UHD	TPAMI'21	15.58	0.934	79.42K
RUAS		CVPR'21	14.68	0.757	3.44K
ELGAN		TIP'21	18.36	0.864	-
Uformer		CVPR'22	29.98	0.980	20.63M
Restormer		CVPR'22	36.90	0.988	26.11M
NSEN	UHD	MM'23	29.49	0.980	2.67M
UHDFour		ICLR'23	36.12	0.990	17.54M
LLFormer		AAAI'23	37.33	0.988	24.52M
Ours		-	**37.73**	**0.992**	5.22M

4. Experiment

4.1. Implementation Details and Datasets

Implementation Details. We conduct experiments in PyTorch on eight NVIDIA GeForce RTX 3090 GPUs. The model is trained with the Adam optimizer [14] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for 300k iterations. The learning rate is initially set to 2×10^{-4} and employs the cosine annealing scheme during the training process. We randomly crop the full-resolution 4K image to a resolution of 512×512 as the input with a batch size of 16. To augment the training data, random horizontal and vertical flips are applied to the input images. The architecture of our method consists of a

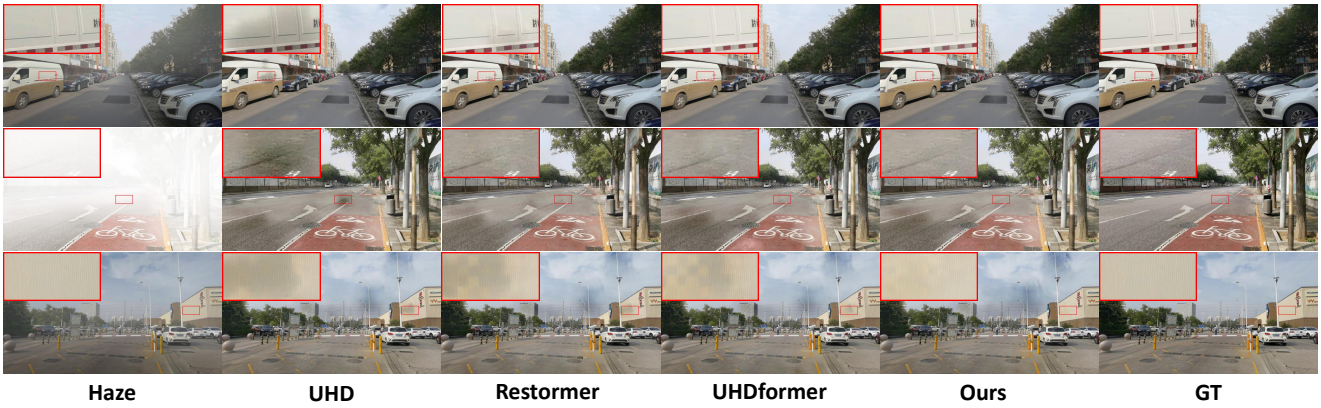

Figure 5. Visual quality comparisons with state-of-the-art methods on UHD-Haze dataset. Please zoom in for details.

4-level encoder-decoder, with varying numbers of FEMs at each level, specifically [2, 4, 4, 6] from level-1 to level-4. The refinement stage contains 2 blocks and feature channels set to 24. For the FGFE module, we focus on a patch size of 8×8 pixels when computing the frequency attention map. The training objective is to minimize the mean absolute error (MAE, *i.e.*, L1 loss) between the restored image and ground truth.

Datasets. To verify the effectiveness of D2Net, we evaluate it on the UHD-LOL dataset [32], UHD-Haze dataset and UHD-Blur dataset. The UHD-LOL dataset serves as a comprehensive benchmark for evaluating the performance of D2Net in low-light conditions. It consists of two subsets, namely UHD-LOL4K and UHD-LOL8K, which contain UHD images with resolutions of 4K and 8K, respectively. In this study, we focus on the UHD-LOL4K subset to validate the effectiveness of D2Net. The UHD-LOL4K subset comprises a total of 8,099 image pairs, with 5,999 pairs allocated for training purposes and 2,100 pairs designated for testing. UHD-Haze and UHD-Blur datasets are respectively re-collected from the datasets of Zheng *et al*. [48] and Deng *et al*. [7] by Wang *et al*. [30]. Among them, the UHD-Haze dataset has 2,290 haze-clear image pairs for training and 231 pairs for testing. The UHD-Blur dataset has 1,964 blur-clear image pairs for training and 300 pairs for testing.

4.2. Comparisons with the State-of-the-art Methods

Evaluation. Two well-known metrics, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [34], are employed for quantitative comparisons. Higher values of these metrics indicate superior performance of the methods. PSNR and SSIM are calculated along the RGB channels. We also report the trainable parameters (Param). For methods that cannot perform full-resolution inference, we have adopted the previous patch-based or downsample-based approaches to handle them.

Low-Light UHD Image Enhancement Results. We evaluate UHD low-light image enhancement results on UHD-LOK4K dataset with, and compare our method with approaches Z_DCE++ [16], RUAS [20], ELGAN [13], Uformer [35], Restormer [47], NSEN [46], UHDFour [18] and LLFormer [32]. As shown in Tab. 1, our method outperforms the current state-of-the-art method, LLFormer, by 0.4 dB in PSNR. And you can find the visualization results in Fig. 4.

Table 2. Comparison of quantitative results on UHD-Haze dataset. Best and second best values are indicated with **bold** text and underlined text respectively.

Methods	Type	Venue	PSNR	SSIM	Param
Restormer	non-UHD	CVPR'22	12.72	0.693	26.11M
Uformer		CVPR'22	19.83	0.737	20.63M
DehazeFormer		TIP'23	15.37	0.725	2.5M
UHD	UHD	ICCV'21	18.04	0.811	34.5M
UHDformer		AAAI'24	22.59	0.943	0.34M
Ours		-	**24.88**	**0.944**	5.22M

UHD Image Dehazing Results. For the UHD image dehazing task, we compare our method to recent approaches such as Restormer [47], Uformer [35], DehazeFormer [26], UHD [49] and UHDformer [30]. And the results are shown in Tab. 2. Compared to the latest method, UHDformer, our approach achieved a 2.29 dB improvement in PSNR, which is undoubtedly a significant improvement. Compared to previous methods, D2Net can better achieve color correction, as you can see from the qualitative comparison results in Fig. 5.

UHD Image Deblurring Results. In the UHD image deblurring task, we evaluate our proposed D2Net against existing deblurring approaches, including MIMO-Unet++ [4], Restormer [47], Uformer [35], Stripformer [28], FFTformer [15] and UHDformer [30]. As shown in Tab. 3, our method achieved a 1.64 dB improvement in PSNR compared to the current state-of-the-art method, UHDformer. Additionally, you can find the visual comparison results in Fig. 6.

User Study Score. We conduct a user study to quantify the

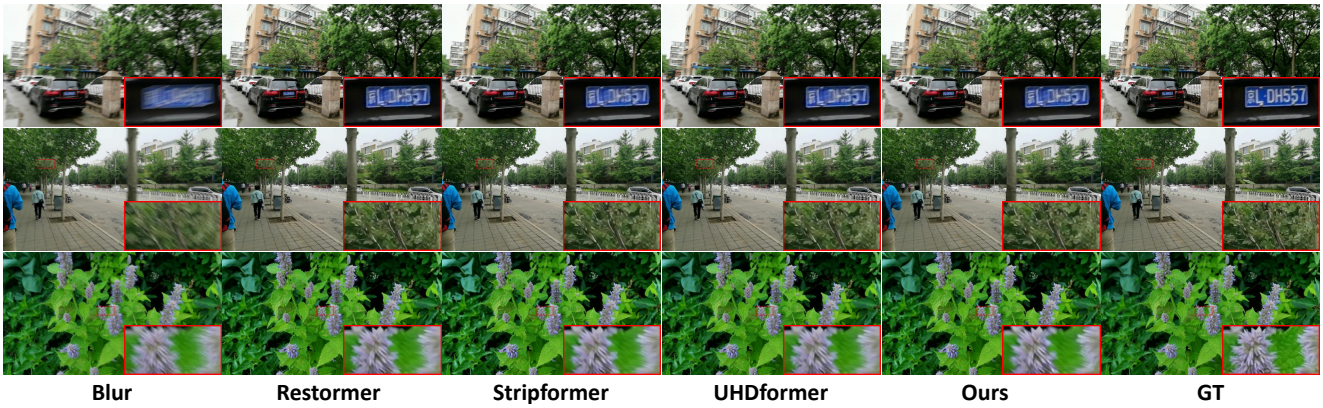

Figure 6. Visual quality comparisons with state-of-the-art methods on UHD-Blur dataset. Please zoom in for details.

Table 3. Comparison of quantitative results on UHD-Blur dataset. Best and second best values are indicated with **bold** text and underlined text respectively.

Methods	Type	Venue	PSNR	SSIM	Param
MIMO-Unet++	non-UHD	ICCV'21	25.03	0.752	16.1M
Restormer		CVPR'22	25.21	0.752	26.1M
Uformer		CVPR'22	25.27	0.752	20.6M
Stripformer		ECCV'22	25.05	0.750	19.7M
FFTformer		CVPR'23	25.41	0.757	16.6M
UHDformer	UHD	AAAI'24	28.82	0.844	0.34M
Ours		-	**30.46**	**0.872**	5.22M

human subjective visual perception quality of the restored UHD images from the three datasets. 5 human subjects are invited to score the visual quality of 15 restored images, independently. Specifically, we allocate 5 restored images per task, which consist of the results of 4 different methods and the ground truth. These subjects are instructed to subjectively rank the quality of these 5 images, and The scores range from 1 (worst) to 5 (best). From Fig. 7, we can see the results of user study. From the results, we can draw two conclusions: 1) Our approach has a more realistic and precise visual quality compared to other methods. 2) The gap between the current methods' restored images and the GT is still relatively large, and further improvement is needed.

4.3. Ablation Study

We further conduct extensive ablation studies to understand better and evaluate each component in the proposed D2Net. To ensure a fair comparison, all experiments are conducted on the UHD-LOL4K dataset. In order to demonstrate the effectiveness of the proposed module, we replace it with the Residual Block (ResBlock) [11] of comparable parameters.

Effectiveness of Multi-scale Local Feature Extraction Modules. To demonstrate the gains of MLFE modules, we repalce it with the ResBlock of comparable parameters. The results shown in Tab. 4 demonstrate the effectiveness of the MLFE modules. Multi-scale feature representation to provide performance gains is already a common practice [12, 45]. However, existing UHD image restoration methods have overlooked this aspect. For UHD images, it is even more important, as at higher resolutions, convolutions with different receptive fields can more easily capture different patterns.

Effectiveness of Fourier-based Global Feature Extraction Modules. To demonstrate the gains of FGFE modules, we repalce it with the ResBlock of comparable parameters. From the results in Tab. 4, we can see that the FGFE modules bring significant performance gains, which are greater than the gains brought by the MLFE modules. This suggests that the performance gains from non-local features are substantial, which is likely why researchers have been trying to apply self-attention to various tasks. Furthermore, we will discuss the relationship between self-attention and our FGFE modules.

Both self-attention and the FGFE modules aim to establish long-range feature dependencies. So why do we not try using self-attention and instead take a different approach by modeling global features in the frequency domain? LLFormer [32] has already tried using self-attention to solve the UHD image restoration problem, but encountered issues with memory overflow, forcing them to divide the image into patches and process them individually. The main factor causing the memory overflow is the storage of the attention map. For self-attention, the spatial complexity is generally $O(HW \times HW)$. And our FGFE modules only require $O(HW)$, which is sufficient to model the long-range feature dependencies, thereby successfully avoiding the memory overflow issue. There are also methods that attempt to calculate the attention map along the channel dimension to reduce memory consumption [47]. However, these methods often require stacking multiple attention blocks to achieve performance comparable to the methods that compute attention maps along the spatial dimension. This approach is also not suitable for UHD images.

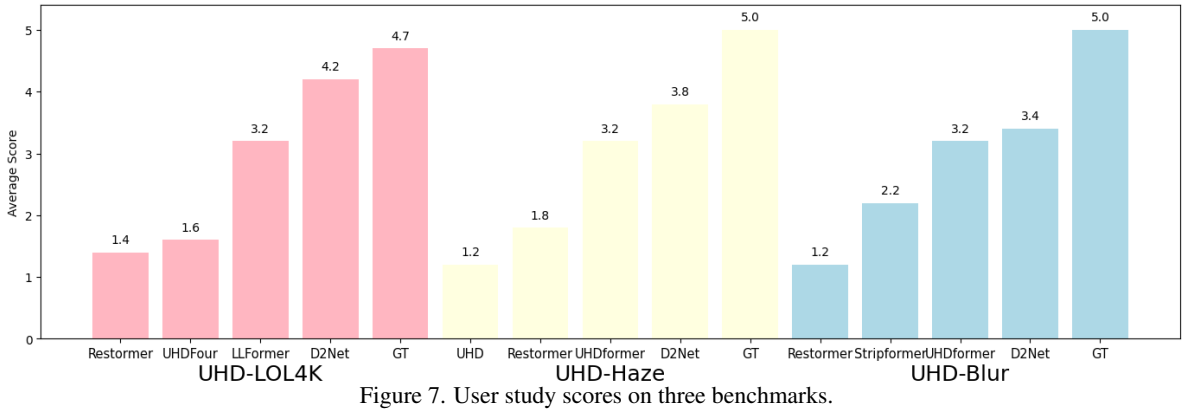

Figure 7. User study scores on three benchmarks.

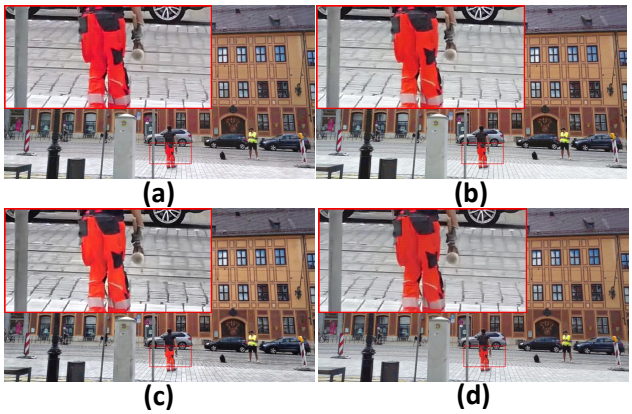

Figure 8. Visual quality comparisons of ablation study. Please zoom in for details. The labels in the figure correspond to the labels in Tab. 4.

Effectiveness of Adaptive Feature Modulation Modules. We compare our proposed AFMM with the commonly used direct feature concatenation approach in the UNet structure. The results are shown in Tab. 5. Based on the observed results, our proposed AFMM is clearly superior to the simple approach of stacking features together. The effectiveness of AFMM, in our view, may be attributed to the inherent redundancy in image data, where the same patterns can appear multiple times in a single image. The encoded features may contain some irrelevant information, and the flow of such information can degrade the model's performance. Therefore, the adaptive selection of features in AFMM proves to be effective in mitigating this issue.

5. Conclusion

In this paper, we proposed a novel UHD image restoration framework, named D2Net. D2Net enables direct full-resolution inference on consumer-grade GPUs, without the need for high-rate downsampling or dividing the images into several patches. Specifically, we ingeniously leveraged the characteristics of the Fourier transform to achieve long-range dependency modeling of features in the frequency domain. We also designed an MLFE module to extract multi-scale local features from the images. Additionally, we introduced the AFMM to reduce the flow of irrelevant information in the Unet-like structure. Extensive experiments showed the superiority of D2Net on several UHD image restoration tasks.

Table 4. Ablation study of proposed blocks. In order to ensure consistency in parameter magnitude, we gradually replace the proposed module with Residual Block of comparable parameters.

#	FGFE	MLFE	PSNR	SSIM
(a)			36.88	0.984
(b)	✓		37.47	0.988
(c)		✓	37.13	0.984
(d)	✓	✓	37.73	0.992

Table 5. Ablation study of Adaptive Feature Modulation Module.

	PSNR	SSIM
Concat	36.97	0.989
AFMM	37.73	0.992

Table 6. Trainable parameters,running time and inference memory cost of some UHD image restoration methods.

	UHD	UHDFour	UHDformer	Ours
Param	34.5M	17.54M	0.34M	5.22M
RT	1.43s	1.67s	5.90s	1.63s
Mem	4.50G	2.02G	11.46G	10.97G

Limitations. Full-resolution inference can improve the quality of the restored images, but it also results in a significant memory overhead. In Tab. 6, we report the training parameters, inference time, and memory overhead required for recent UHD image restoration methods to process 3840×2160 resolution images on a single NVIDIA GeForce RTX 3060 GPU. From the results, we can see that our method, like most previous approaches, incurs a relatively large inference memory cost when processing a single UHD image. Looking ahead, we will focus on optimizing the inference memory overhead of our method.

References

[1] 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/CVF International Conference on Computer Vision*, pages 12504–12513, 2023. 1, 2

[2] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5896–5905, 2023. 2

[3] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Cross aggregation transformer for image restoration. *Advances in Neural Information Processing Systems*, 35:25478–25490, 2022. 2

[4] 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/CVF international conference on computer vision*, pages 4641–4650, 2021. 6

[5] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Revitalizing convolutional network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2

[6] Yuning Cui, Yi Tao, Wenqi Ren, and Alois Knoll. Dual-domain attention for image deblurring. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 479–487, 2023. 1, 2

[7] Senyou Deng, Wenqi Ren, Yanyang Yan, Tao Wang, Fenglong Song, and Xiaochun Cao. Multi-scale separable network for ultra-high-definition video deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14030–14039, 2021. 6

[8] Matteo Frigo and Steven G Johnson. Fftw: An adaptive software architecture for the fft. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 3, pages 1381–1384. IEEE, 1998. 3

[9] Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga. Deep wavelet prediction for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 104–113, 2017. 3

[10] Xin Guo, Xueyang Fu, Man Zhou, Zhen Huang, Jialun Peng, and Zheng-Jun Zha. Exploring fourier prior for single image rain removal. In *IJCAI*, pages 935–941, 2022. 3

[11] 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. 7

[12] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th acm international conference on multimedia*, pages 2024–2032, 2019. 7

[13] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, 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. 6

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[15] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5886–5895, 2023. 3, 6

[16] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4225–4238, 2021. 6

[17] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE transactions on image processing*, 29:4376–4389, 2019. 1, 2

[18] Chongyi Li, Chun-Le Guo, Man Zhou, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Embedding fourier for ultra-high-definition low-light image enhancement. *arXiv preprint arXiv:2302.11831*, 2023. 1, 2, 3, 6

[19] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 2

[20] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10561–10570, 2021. 6

[21] Xintian Mao, Yiming Liu, Fengze Liu, Qingli Li, Wei Shen, and Yan Wang. Intriguing findings of frequency selection for image deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1905–1913, 2023. 3

[22] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. Low-light image enhancement via a deep hybrid network. *IEEE Transactions on Image Processing*, 28(9):4364–4375, 2019. 1, 2

[23] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 154–169. Springer, 2016. 1, 2

[24] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3253–3261, 2018. 1, 2

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted*

intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015. 4

[26] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. 6

[27] Long Sun, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Spatially-adaptive feature modulation for efficient image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13190–13199, 2023. 5

[28] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *European conference on computer vision*, pages 146–162. Springer, 2022. 6

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[30] Cong Wang, Jinshan Pan, Wei Wang, Gang Fu, Siyuan Liang, Mengzhu Wang, Xiao-Ming Wu, and Jun Liu. Correlation matching transformation transformers for uhd image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5336–5344, 2024. 2, 6

[31] Chenxi Wang, Hongjun Wu, and Zhi Jin. Fourllie: Boosting low-light image enhancement by fourier frequency information. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7459–7469, 2023. 3

[32] Tao Wang, Kaihao Zhang, Tianrun Shen, Wenhan Luo, Bjorn Stenger, and Tong Lu. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2654–2662, 2023. 1, 2, 6, 7

[33] Yang Wang, Long Peng, Liang Li, Yang Cao, and Zheng-Jun Zha. Decoupling-and-aggregating for image exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18115–18124, 2023. 1, 2

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[35] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022. 1, 2, 6

[36] Yiwen Wei, Zhuoran Zheng, and Xiuyi Jia. Uhd underwater image enhancement via frequency-spatial domain aware network. In *Proceedings of the Asian Conference on Computer Vision*, pages 299–314, 2022. 2

[37] Chen Wu, Zhuoran Zheng, Xiuyi Jia, and Wenqi Ren. Mixnet: Towards effective and efficient uhd low-light image enhancement. *arXiv preprint arXiv:2401.10666*, 2024. 2, 5

[38] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25456–25467, 2024. 1

[39] Jie Xiao, Xueyang Fu, Feng Wu, and Zheng-Jun Zha. Stochastic window transformer for image restoration. *Advances in Neural Information Processing Systems*, 35:9315–9329, 2022. 2

[40] Jun Xiao, Xinyang Jiang, Ningxin Zheng, Huan Yang, Yifan Yang, Yuqing Yang, Dongsheng Li, and Kin-Man Lam. Online video super-resolution with convolutional kernel bypass grafts. *IEEE Transactions on Multimedia*, 25:8972–8987, 2023. 1

[41] Jun Xiao, Zihang Lyu, Cong Zhang, Yakun Ju, Changjian Shui, and Kin-Man Lam. Towards progressive multi-frequency representation for image warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2995–3004, 2024. 1, 3

[42] Qiuhai Yan, Aiwen Jiang, Kang Chen, Long Peng, Qiaosi Yi, and Chunjie Zhang. Textual prompt guided image restoration. *arXiv preprint arXiv:2312.06162*, 2023. 1, 2

[43] Guang Yang, Lu Lin, Chen Wu, and Feng Wang. Dual-domain learning for jpeg artifacts removal. In *International Conference on Neural Information Processing*, pages 556–568. Springer, 2023. 1, 2, 3

[44] Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao. Frequency and spatial dual guidance for image dehazing. In *European Conference on Computer Vision*, pages 181–198. Springer, 2022. 3

[45] Weihao Yu, Pan Zhou, Shuicheng Yan, and Xinchao Wang. Inceptionnext: When inception meets convnext. *arXiv preprint arXiv:2303.16900*, 2023. 4, 7

[46] Wei Yu, Qi Zhu, Naishan Zheng, Jie Huang, Man Zhou, and Feng Zhao. Learning non-uniform-sampling for ultra-high-definition image enhancement. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1412–1421, 2023. 1, 2, 6

[47] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 1, 2, 6, 7

[48] Zhuoran Zheng, Wenqi Ren, Xiaochun Cao, Xiaobin Hu, Tao Wang, Fenglong Song, and Xiuyi Jia. Ultra-high-definition image dehazing via multi-guided bilateral learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16180–16189. IEEE, 2021. 2, 6

[49] Zhuoran Zheng, Wenqi Ren, Xiaochun Cao, Tao Wang, and Xiuyi Jia. Ultra-high-definition image hdr reconstruction via collaborative bilateral learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4449–4458, 2021. 2, 6

[50] Yijie Zhou, Chao Li, Jin Liang, Tianyi Xu, Xin Liu, and Jun Xu. 4k-resolution photo exposure correction at 125 fps with˜ 8k parameters. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1587–1597, 2024. 2