FSI: Frequency and Spatial Interactive Learning for Image Restoration in Under-Display Cameras

Chengxu Liu[1,3], Xuan Wang[2], Shuai Li[2], Yuzhi Wang[2], Xueming Qian[1,3]
[1]Xi'an Jiaotong University [2]MEGVII Technology
[3]Shaanxi Yulan Jiuzhou Intelligent Optoelectronic Technology Co., Ltd
liuchx97@gmail.com, {wangxuan02,lishuai,wangyuzhi}@megvii.com, qianxm@mail.xjtu.edu.cn

Abstract

Under-display camera (UDC) systems remove the screen notch for bezel-free displays and provide a better interactive experience. The main challenge is that the pixel array of light-emitting diodes used for display diffracts and attenuates the incident light, leading to complex degradation. Existing models eliminate spatial diffraction by maximizing model capacity through complex design and ignore the periodic distribution of diffraction in the frequency domain, which prevents these approaches from satisfactory results. In this paper, we introduce a new perspective to handle various diffraction in UDC images by jointly exploring the feature restoration in the frequency and spatial domains, and present a Frequency and Spatial Interactive Learning Network (FSI). It consists of a series of well-designed Frequency-Spatial Joint (FSJ) modules for feature learning and a color transform module for color enhancement. In particular, in the FSJ module, a frequency learning block uses the Fourier transform to eliminate spectral bias, a spatial learning block uses a multi-distillation structure to supplement the absence of local details, and a dual transfer unit to facilitate the interactive learning between features of different domains. Experimental results demonstrate the superiority of the proposed FSI over state-of-the-art models, through extensive quantitative and qualitative evaluations in three widely-used UDC benchmarks.

1. Introduction

Under-display camera (UDC) systems are the foundation of full-screen display devices where the lens mounts under the display. It removes screen notches and provides a higher screen-to-body ratio without disrupting the screen's integrity [25, 38]. Recently, rising demand for full-screen devices enable the study of UDC to attract attention [8, 44].

To implement the full-screen display, it is essential to densely arrange some organic light-emitting diode (OLED) in the display area above the camera. However, this design also brings various image degradation. Specifically, the degradations are mainly due to 1) diffraction artifacts generated by the periodic gaps between the pixel grids as apertures[1], illustrated by Fig. 2(a), and 2) color shift from multiple thin-film layers in an OLED [7, 14, 45]. Besides, the regions with different diffraction intensities in the image cause different degrees of degradation, bringing challenges for diffraction removal. Typically, in Fig. 2(b), around the light source, diffraction causes flares that saturate one or more channels of the image, resulting in content loss.

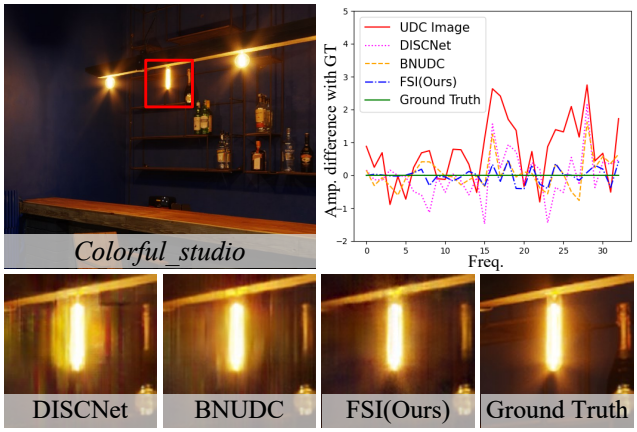

Figure 1. A comparison between FSI and other SOTA methods (DISCNet [7], BNUDC [14]) on SYNTH dataset. The upper right shows the amplitude-frequency curves horizontally oriented on clipped frequency bands of 0 to 35. The vertical axis indicates the amplitude difference between different methods and ground truth. The results of the proposed FSI are superior to others to approximate the curve of ground truth (best viewed in color).

To solve these challenges, recent years have witnessed an increasing number of UDC image restoration approaches, which can be categorized into two paradigms. The former makes attempts to leverage the prior knowledge of the diffraction blur kernel, *i.e.*, point spread function (PSF),

[1]Light diffracts as it propagates through obstacles with similar sizes to its wavelength.

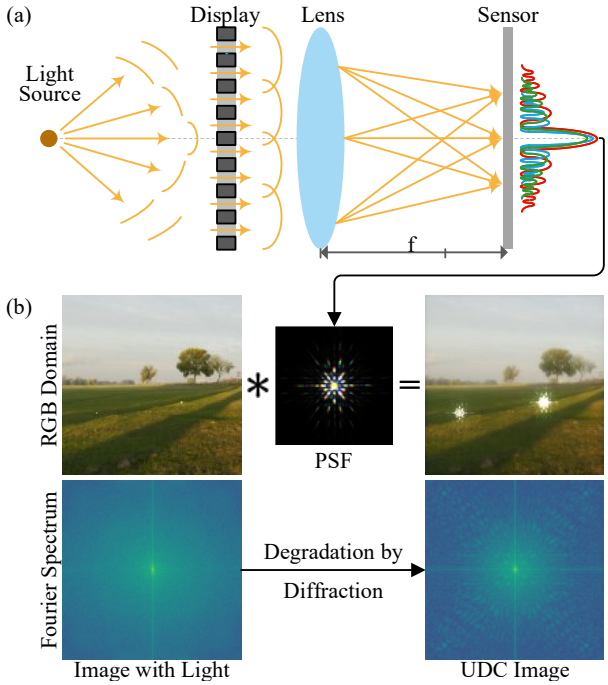

Figure 2. (a) illustrates the formation of the point spread function (PSF) caused by diffraction in the UDC system. The light emitted from light source crosses a display with arranged OLEDs and a lens before it is finally captured by the sensor. (b) is the generation of the UDC image and its Fourier spectrum, where the low-frequency and high-frequency signals are concentrated in the center and the edges, respectively. The UDC image after diffraction produces distinct high-frequency streaks in the Fourier spectrum.

as input and guide the removal of diffraction [7, 15, 20]. However, such methods rely heavily on the accuracy of the measured PSFs and have limitations in their application. The latter learns degradation patterns of UDC images directly by maximizing network capacity through complex design [14, 20, 45, 46]. The latest BNUDC [14] decouples the degradation of UDC images into high-frequency diffraction and low-frequency spatially attenuation for learning separately, achieving superior performance. Nonetheless, these methods only follow the spatial domain learning approaches, while ignoring the diffraction characteristics in frequency domain, thus producing artifacts around light sources, as shown in Fig. 1.

From the perspective of the frequency domain, light emitted from a light source across a display with arranged OLEDs is diffracted, which will produce the periodic decreasing spectral biases captured by the sensor [26, 38] , as illustrated in Fig. 2. UDC image restoration aims to eliminate the redundant spectral bias and reconstruct the loss of textures in the region of high diffraction intensity. In contrast, existing spatial domain models [7, 14] are hard to perceive the entire spectrum and thus have the bias for learning different frequency components. Besides, frequency learning [27, 36] has made significant progress in various low-level tasks recently [16, 32, 37, 40, 47]. This demonstrates that learning in the frequency domain not only helps to eliminate textures produced by the diffraction, but also enriches the global representation of features to reconstruct the image content [22, 32, 36, 47]. Therefore, a more promising solution is to explore proper ways of utilizing frequency learning to eliminate diffraction and reconstruct textures in UDC image restoration.

In this paper, we propose a novel **F**requency and **S**patial **I**nteractive learning network (FSI) for UDC image restoration. The key insight of FSI is to explore the properties of diffraction in the frequency domain, eliminate the diffraction-producing frequency components and reconstruct the loss of textures by joint learning in the frequency and spatial domains. The overview is shown in Fig. 3, FSI consists of a series of carefully designed Frequency-Spatial Joint (FSJ) modules and a Color Transform (CT) module. Specifically, in the FSJ module, a frequency learning block formulates image features as horizontal and vertical spectral components by Fourier transform to eliminate diffraction from both directions, a spatial learning block extracts hierarchical features step-by-step with a multi-distillation structure to further supplement the absence of local details, and a dual transfer unit enables the two parts to interact selectively and facilitates the joint learning. The CT module predicts a set of coefficients for blending different color spaces to adjust the color temperature.

Our contributions are summarized as follows:

- We propose a novel frequency and spatial interactive learning network (FSI), which is the first work to introduce frequency learning into UDC image restoration. By learning in the frequency domain, our method effectively eliminates the various diffraction and reconstructs the textures.
- We propose a frequency-spatial joint (FSJ) module, which introduces a new perspective to explore the union of information in frequency and spatial domains, providing inspiration for other PSF-conditional tasks.
- Extensive experiments demonstrate that the proposed method can significantly outperform existing SOTA methods in three widely-used UDC benchmarks.

2. Related Work

2.1. Image Restoration for UDC

Image restoration of the under-display camera (UDC) is a relatively new topic in low-level vision. To model the complex degradation of UDC systems, MSUNet [45] first analyzes the optical imaging process of real UDC and collects the diffraction kernel, *i.e.*, point spread function (PSF),

and two paired datasets, *i.e.*, transparent-organic LED (T-OLED) and pentile-organic LED (P-OLED), by mounting a display on top of a traditional digital camera lens. DSIC-Net [7] generates a larger dataset, based on the measured PSFs produced from the display layout patterns. They are also used by the ECCV challenge [8, 44] to compare.

The existing UDC image restoration algorithms treat them as an inversion problem for the measured PSFs, which can be divided into two paradigms: PSF-related and PSF-free. The PSF-related methods [7, 15, 20] consider the diversity of PSFs in various situations and propose frameworks for using PSF as priori information to guide the diffraction removal. However, in practice, the factors affecting measured PSF are complex, and inaccurate PSF will degrade the model performance. The PSF-free methods [14, 23, 31, 39, 45, 46] attempt to design more powerful feature learning networks that directly learn various degradations in UDC images. Typically, MSUNet [45] first proposes a learning-based restoration network based on U-Net [29]. Subsequent RDUNet [39], PDCRN [23], and DAGF [31] further propose residual dense network, wavelet decomposition CNN, and deep atrous guided filter for image restoration, respectively. Recently, BNUDC [14] uses a dual-stream network to decouple the diffraction and diffuse intensity in the UDC for learning separately. However, most methods focus on the design of complex modules, ignoring the frequency characteristics of diffraction [26, 38] and the effects of diffraction with different intensities. Therefore, we propose a frequency-spatial joint module, which is not limited to the specific PSFs and handles the various diffraction through joint learning in different domains.

2.2. Frequency Learning in Low-level Vision

Low-level vision always a attractive area to research [28, 40, 18, 19]. Recent years have witnessed an increasing number of studies on low-level vision tasks exploring image restoration from a frequency learning perspective [24, 30], such as super-resolution [16, 28], low-light enhancement [37], inpainting [32], and dehaze [40]. Typically, based on the Fast Fourier Transform (FFT), LaMa [32] obtains frequency features that the receptive field covers the entire image to inpainting with higher input resolution. FSDGN [40] finds the degradation property induced by haze is manifested in the amplitude spectrum and learns in frequency to image dehaze. DeepRFT [22] applies the ReLU in the frequency to extract the kernel-level information and integrates it into the ResBlock for deblurring. In general, frequency learning can be well-used for degraded reconstruction in many low-level vision tasks.

In UDC images, according to the characteristics that the spatial distribution of diffraction artifacts has regularity, we propose a more promising solution that eliminates diffraction in the frequency domain.

3. Methodology

3.1. Motivation

Our main inspiration comes from the frequency spectral properties of diffraction in UDC images. As shown in Fig. 2(b), the UDC image produces distinct streaks with a periodic distribution in the Fourier spectrum [26, 38]. In contrast to other restoration tasks, it is more reasonable to learn in the frequency domain to eliminate the spectral bias caused by PSF. Besides, the regions in the image with stronger diffraction intensity cause more severe degradation, even leading to content loss. Our second insight is that the joint learning of local spatial features and global frequency features can effectively recover the textures near the diffraction center, improving visual quality [22, 32, 36].

Therefore, we propose the Frequency and Spatial Interactive Learning Network (FSI) to handle the diffraction and color shift problems, in which the core design Frequency-Spatial Joint (FSJ) module enables effective feature representation learning for UDC image restoration.

3.2. Problem Formulation

We follow the existing works [14, 45] to define the UDC image restoration as the diffraction removal and color correction problems. It can be formulated as:

$$\hat{y} = f(\gamma \cdot x * k + n), \tag{1}$$

where x and $\hat{y}$ are the clean image and degraded image from UDC, respectively. k is the diffraction blur kernel, which is generated by the pixel grid of display and is commonly denoted as the point spread function (PSF). γ is the intensity scaling factor, which is degraded by multiple thin-film layers and usually produces color shift. $*$ denotes the convolution operator, and n denotes the additive noise of the UDC camera. $f(\cdot)$ denotes the clamp function used to model the saturation of pixels that are beyond the limited range. Here we omit the non-linear mapping for brevity.

3.3. Frequency and Spatial Interactive Learning Network

As illustrated in Fig. 3, the proposed FSI takes the UDC image $3 \times H \times W$ as input. To accelerate the inference process, we use pixel-unshuffle to reduce the spatial resolution by four times for feature learning. After that, the features are fed into frequency and spatial domains for interactive learning, which means that the frequency and spatial features can propagate and enhance each other during learning. Then, a pixel-shuffle outputs two $3 \times H \times W$ images, one preserves the important information in UDC input by multiplication, which is then added to the other one as residuals. It improves the content consistency of the output and input. Finally, the restored image is outputted through a color

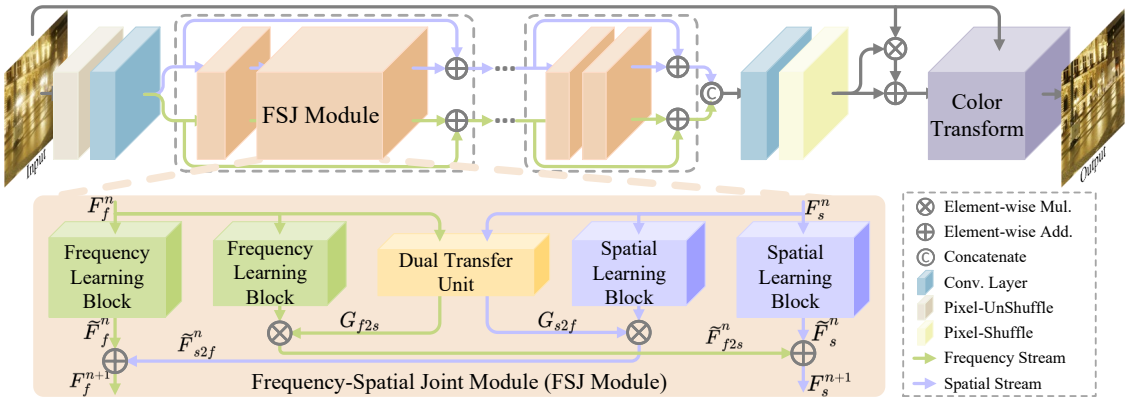

Figure 3. Overview of proposed Frequency and Spatial Interactive Learning Network (FSI). FSI generates restored images through a series of stacked Frequency-Spatial Joint (FSJ) modules and a Color Transform module. FSJ module consists of frequency learning blocks, a spatial learning block, and a dual transfer unit. F and G denote the features and transfer gates of the intermediate process, respectively.

transform module. Besides, we also group two FSJ modules and use skip connection [10] to improve capacity and accelerate convergence.

3.4. Frequency-Spatial Joint Module

The FSJ module is used to eliminate the redundant frequencies and reconstruct the textures. Benefiting from the capability of promoting each other to learn complementary features, the paradigm of the dual-stream network has achieved promising applications on various restoration tasks [3, 14, 41]. We take one step further and introduce this design of dual-stream paradigm into the FSJ. Specifically, as shown in Fig. 3, a FSJ module consists of two frequency learning blocks, two spatial learning blocks, and a dual transfer unit. It learns the features of both frequency stream and spatial stream. In each stream, one block is used to learn domain-specific features exclusively, and the other is used to learn cross-domain features.

In terms of formula, we use $\mathrm{FLB}(\cdot)$, $\mathrm{SLB}(\cdot)$, and $\mathrm{DTU}(\cdot)$ to denote the frequency learning block, spatial learning block, and dual transfer unit, respectively. Take the n^{th} FSJ module for example, where $F_f^n, F_s^n \in R^{C\times H\times W}$ denote the input features in frequency and spatial domain. H, W, and C represent the height, width, and channel of the feature maps, respectively. Then, the domain-specific features $\widetilde{F}_f^n, \widetilde{F}_s^n \in R^{C\times H\times W}$ learned in frequency and spatial can be formulated as:

$$\widetilde{F}_f^n = \mathrm{FLB}(F_f^n), \quad \widetilde{F}_s^n = \mathrm{SLB}(F_s^n). \tag{2}$$

The cross-domain features $\widetilde{F}_{f2s}^n, \widetilde{F}_{s2f}^n \in R^{C\times H\times W}$ transferred from frequency and spatial stream, formulated as:

$$\begin{aligned} \widetilde{F}_{f2s}^n &= \mathrm{FLB}(F_f^n) \otimes G_{f2s}, \\ \widetilde{F}_{s2f}^n &= \mathrm{SLB}(F_s^n) \otimes G_{s2f}, \end{aligned} \tag{3}$$

where $\otimes$ denotes the element-wise multiplication. $G_{f2s}, G_{s2f} \in R^{1\times H\times W}$ are the transfer gates for selecting interactive features, obtained by:

$$G_{f2s}, G_{s2f} = \mathrm{DTU}(F_f^n, F_s^n). \tag{4}$$

Finally, the output frequency and spatial features denoted as $F_f^{n+1}, F_s^{n+1} \in R^{C\times H\times W}$, can be formulated as:

$$F_f^{n+1} = \widetilde{F}_f^n \oplus \widetilde{F}_{s2f}^n, \quad F_s^{n+1} = \widetilde{F}_s^n \oplus \widetilde{F}_{f2s}^n, \tag{5}$$

where $\oplus$ denotes the element-wise addition. We empirically set the number of feature channels and FSJ modules to 96 and 6, respectively. Following will focus on the structure of the three parts $\mathrm{FLB}(\cdot)$, $\mathrm{SLB}(\cdot)$, and $\mathrm{DTU}(\cdot)$ in detail.

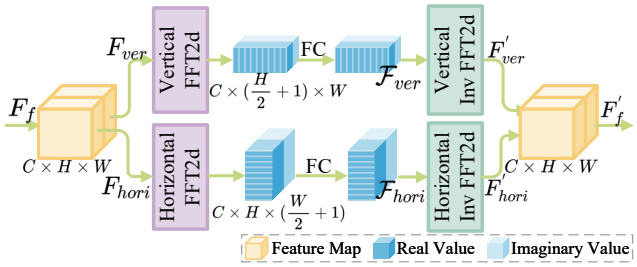

Figure 4. The structure of Frequency Learning Block (FLB).

Frequency learning block. According to the theory of Fourier transform, 1) the Fourier transform of a 2D image can be decomposed into x-axis and y-axis parts, which reflect diffraction in the horizontal and vertical directions, respectively, and 2) a point-wise update in the frequency spectrum will affect the global features of the image [2, 4, 32].

Therefore, we construct the frequency learning block to deal with the global diffraction by using the horizontal and vertical FFT in both directions. Specifically, as shown in Fig. 4, we transform the image features into two orthogonal Fourier spectrums, which are obtained from the Fast Fourier Transform (FFT) in the horizontal and vertical directions. For each direction of the spectrums, we use a fully connected (FC) layer to eliminate the spectral biases generated by diffraction and reconstruct the textures.

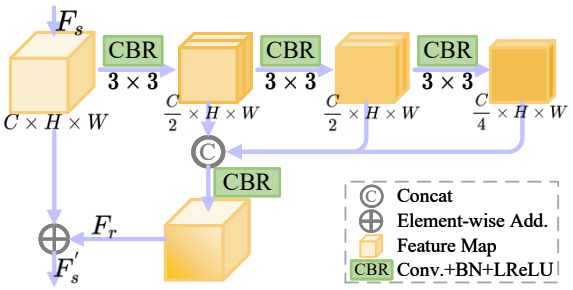

Figure 5. The structure of Spatial Learning Block (SLB).

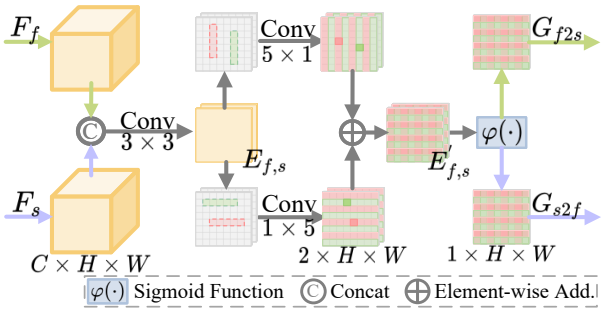

Figure 6. The structure of Dual Transfer Unit (DTU).

In terms of formula, we denote the vertical and horizontal Fast Fourier Transform as $\mathrm{FFT}_v(\cdot)$ and $\mathrm{FFT}_h(\cdot)$, respectively, and the corresponding inverse transforms as $\mathrm{iFFT}_v(\cdot)$ and $\mathrm{iFFT}_h(\cdot)$. We first partition the input frequency feature F_f into two parts F_{ver} and F_{hori} with the same channel. And the transformed features in the frequency domain can be formulated as:

$$\begin{aligned} \mathcal{F}_{ver}^{real}, \mathcal{F}_{ver}^{imag} &= \mathrm{FFT}_v(F_{ver}), \\ \mathcal{F}_{hori}^{real}, \mathcal{F}_{hori}^{imag} &= \mathrm{FFT}_h(F_{hori}), \end{aligned} \tag{6}$$

where $\mathcal{F}_{ver}^{real}, \mathcal{F}_{hori}^{real}$ and $\mathcal{F}_{ver}^{imag}, \mathcal{F}_{hori}^{imag}$ denote the real and imaginary parts of the spectrums, respectively. Notably, we only keep half of the spectrums to reduce the computational cost according to the conjugate symmetry in the Fourier transform theory. Then we concatenate the real and imaginary parts to reconstruct them with $\mathrm{FC}(\cdot)$, formulated as:

$$\begin{aligned} \mathcal{F}_{ver} &= \mathrm{FC}(Concat(\mathcal{F}_{ver}^{real}, \mathcal{F}_{ver}^{imag})), \\ \mathcal{F}_{hori} &= \mathrm{FC}(Concat(\mathcal{F}_{hori}^{real}, \mathcal{F}_{hori}^{imag})), \end{aligned} \tag{7}$$

where $Concat(\cdot)$ for concatenating the real and imaginary parts. Since the spectrum has a global receptive field, using the FC layer can learn the spectral changes at a small cost. We finally transform the reconstructed spectrums into feature maps $F_{ver}^{'}, F_{hori}^{'}$ by inverse FFT, formulated as:

$$\begin{aligned} F_{ver}^{'} &= \mathrm{iFFT}_v(Split(\mathcal{F}_{ver})), \\ F_{hori}^{'} &= \mathrm{iFFT}_h(Split(\mathcal{F}_{hori})), \end{aligned} \tag{8}$$

where $Split(\cdot)$ for dividing the spectrums into real and imaginary parts. The two features $F_{ver}^{'}$ and $F_{hori}^{'}$ are combined into the final output, formulated as:

$$F_f^{'} = Concat(F_{ver}^{'}, F_{hori}^{'}), \tag{9}$$

where $F_f^{'}$ is the final output of the frequency learning block.

Spatial learning block. Learning only in the frequency domain leads to the absence of local detailed features [32, 47]. Inspired by the success of multi-distillation mechanism in low-level vision tasks [11, 33, 42], we construct a spatial learning block to complement local fine-grained features that are neglected in frequency.

Specifically, as shown in Fig. 5, the spatial feature F_s is progressively refined through multiple cascaded convolution layers. Each layer retains half of the channel features for further refinement, and the rest are outputted after feature reconstruction. This process can be formulated as:

$$\begin{aligned} F_{r1}, F_{c1} &= Split(\mathrm{CBR}(F_s)), \\ F_{r2}, F_{c2} &= Split(\mathrm{CBR}(F_{c1})), \\ F_{r3} &= \mathrm{CBR}(F_{c2}), \\ F_r &= \mathrm{CBR}(Concat(F_{r1}, F_{r2}, F_{r3})), \end{aligned} \tag{10}$$

where $\mathrm{CBR}(\cdot)$ denotes the combination of convolution, batch normalization, and Leaky ReLU. F_{c1}, F_{c2} denote the coarse features for further progressive refinement. F_{r1}, F_{r2}, F_{r3} denote the refined features and F_r denotes the reconstructed feature after aggregating them. Finally, F_r is added to the spatial feature F_s as a residual, formulated as:

$$F_s^{'} = F_s \oplus F_r, \tag{11}$$

where $F_s^{'}$ is the final output of the spatial learning block.

Dual transfer unit. We present a dual transfer unit to generate transfer gates that serve to fuse features between different domains better. It enables selective interaction and joint learning of the frequency and spatial features.

Specifically, as shown in Fig. 6, we first compress two features into a lower dimensional embedding space $E_{f,s} \in R^{2 \times H \times W}$ by a convolutional layer, formulated as:

$$E_{f,s} = Conv_{3\times3}(Concat(F_f, F_s)). \tag{12}$$

And then the receptive field of the embedding space is expanded along the vertical and horizontal directions using 5×1 and 1×5 convolutional layers, respectively. Thereafter, an element-wise addition aggregates the embedding space in both directions, formulated as:

$$E_{f,s}^{'} = Conv_{5\times1}(E_{f,s}) \oplus Conv_{1\times5}(E_{f,s}). \tag{13}$$

The merit of $Conv_{5\times1}(\cdot), Conv_{1\times5}(\cdot)$ is that DTU could be focused on complementary features learning with less cost. Finally, the transfer gates G_{s2f}, G_{f2s} obtained by:

$$G_{s2f}, G_{f2s} = Split(\varphi(E_{f,s}^{'})), \tag{14}$$

where $\varphi(\cdot)$ denotes the sigmoid activation function used for normalization. This design allows to facilitate the learning of complementary features with little cost and is much simpler than the typical self-attention [34] as well.

3.5. Color Transform

In UDC, degradation caused by multiple thin-film layers in the display usually produces color shifts. Therefore, inspired by the solutions in white balance [1, 17], we use a lightweight U-Net [29] to predict a set of coefficients to adjust the color temperature by matrix transformation.

Specifically, we denote the lightweight network as $\Phi(\cdot)$, the coefficients $\mathcal{T} \in R^{12\times H\times W}$ can be obtained by:

$$\mathcal{T} = \Phi(\hat{y}), \tag{15}$$

where $\hat{y}$ is the input UDC image. The elements of each coordinate in $\mathcal{T}$ include 9-dimensional weight coefficients $(t_r^r, t_g^r, t_b^r, t_r^g, t_g^g, t_b^g, t_r^b, t_g^b, t_b^b)$ and 3-dimensional bias coefficients $(t_r^\theta, t_g^\theta, t_b^\theta)$. When the input is (r_{in}, g_{in}, b_{in}), the corresponding output $(r_{out}, g_{out}, b_{out})$ can be calculated as:

$$\begin{bmatrix} r_{out} \\ g_{out} \\ b_{out} \end{bmatrix} = \begin{bmatrix} t_r^r & t_g^r & t_b^r \\ t_r^g & t_g^g & t_b^g \\ t_r^b & t_g^b & t_b^b \end{bmatrix} \begin{bmatrix} r_{in} \\ g_{in} \\ b_{in} \end{bmatrix} + \begin{bmatrix} t_\delta^r \\ t_\delta^g \\ t_\delta^b \end{bmatrix}. \tag{16}$$

For each coordinate in the image, the weight coefficients can balance the luminance of different color spaces. The bias coefficients can adjust the image brightness. Such a design delivers superior color enhancement.

4. Experiments

4.1. Datasets and Metrics

We evaluate the proposed FSI and compare it with other SOTA approaches on three widely-used datasets: **P-OLED** [45], **T-OLED** [45], and **SYNTH** [7]. For **P-OLED** and **T-OLED** [45], there are published in the UDC 2020 challenge [44] and captured using an RGBG PenTile [6] and a transparent OLED, respectively. It contains a total of 300 images, in which 240 for training, 30 for validation, and 30 for testing. The resolution of each image is 1024×2048. For **SYNTH** [7], it contains 2,016 images for training and 360 for testing. This is a set of synthetic images generated by blur kernels based on the measured PSFs of a commercial UDC smartphone. Each image is degraded based on the measured PSFs with a resolution of 800×800. For fair comparisons, we keep the same evaluation metrics: 1) peak signal-to-noise ratio (PSNR), 2) structural similarity index (SSIM) [35], 3) human perceptual similarity (LPIPS) [43], and 4) deep image structure and texture similarity (DISTS) [5] as previous works [7, 14].

4.2. Training Details

During training, we use Cosine Annealing scheme [21] and Adam [13] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is reduced from the initial 2×10^{-4} to 1×10^{-6}. We set the batch size as 4 and the input size as full-resolution of image. For fair comparisons, we augment the training data with random horizontal flips and vertical flips. For P-OLED and T-OLED, the logarithm of the squared error is used as the loss function. For SYNTH, we follow existing work [7] using L_1 loss and perceptual loss [12]. The total number of iterations is 400K. All models are built with PyTorch and trained on 2 NVIDIA V100 GPUs.

Method	RT(s)	#P(M)	PSNR	SSIM	LPIPS	DISTS
SFTMD [9]	-	3.9	42.35	0.9863	0.0123	-
DISCNet [7]	0.21	3.8	43.27	0.9877	0.0108	0.0182
UDC-UNet [20]	0.11	5.7	45.37	0.9898	0.0162	0.0175
BNUDC [14]	0.03	4.6	45.78	0.9942	0.0106	0.0150
FSI(Ours)	0.03	3.8	46.14	0.9951	0.0101	0.0138

Table 1. Quantitative comparison (PSNR↑, SSIM↑, LPIPS↓, and DISTS↓) on the SYNTH [7] dataset. RT and #P indicate the runtimes and parameters, respectively. Red indicates the best and blue indicates the second best performance (best viewed in color).

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

We compare FSI with nine start-of-the-art methods. These methods can be summarized into two categories: PSF-related methods [7, 9, 20] and PSF-free methods [14, 23, 31, 39, 45, 46]. For fair comparisons, we obtain the performance from their original paper or reproduce results by authors' officially released models.

Quantitative comparison. The performance comparisons on SYNTH [7] are shown in Tab. 1. In comparison with the latest PSF-related method UDC-UNet [20], our method not only has faster runtimes and fewer parameters, but also achieves a higher performance. It is because FSI eliminates diffraction directly in the frequency domain and does not require additional cost to learn the PSFs. Besides, benefiting from the Fourier transform having global receptive fields, our FSI does not need convolution kernels with more parameters. Therefore, in comparison with the latest PSF-free method BNUDC [14] that learns features in the spatial domain, our FSI outperforms it by **0.36 dB** with fewer parameters. Such superior performances mainly benefit from interactive learning in frequency and spatial.

We further validate the generalization capability of FSI on additional UDC datasets P-OLED and T-OLED [45]. As shown in Tab. 2, due to the well-designed FLB/SLB and the feature interactive learning mechanism, FSI achieves better results in two kinds of datasets, which outperforms other SOTA methods between **0.38 dB** to **0.46 dB**. This large margin demonstrates the power of FSI in feature restoration. Especially, on P-OLED with low light transmission rate, compared to T-OLED, FSI exceeds BNUDC [14] by **0.46 dB**. This is because FSI introduces the CT module to effectively adjust the color temperature and brightness. The performances verify that FSI has strong generalization ca-

Method	Runtimes(s)	#Params(M)	P-OLED				T-OLED			
			PSNR	SSIM	LPIPS	DISTS	PSNR	SSIM	LPIPS	DISTS
MSUNet [45]	0.08	8.9	29.17	0.9393	0.2239	0.1746	37.40	0.9756	0.1093	0.1052
DAGF [31]	1.12	1.1	32.29	0.9509	0.2163	0.1913	36.49	0.9716	0.1392	0.1217
PDCRN [23]	0.08	4.7	32.99	0.9578	0.2102	0.2075	37.83	0.9780	-	-
ResUNet [39]	0.43	16.5	31.39	0.9506	0.2129	0.1823	37.95	0.9790	-	-
RDUNet [39]	0.53	47.9	30.12	0.9410	-	-	38.13	0.9797	0.0992	0.0971
UDC-UNet [20]	0.30	5.7	33.12	0.9592	0.1924	0.1617	38.10	0.9796	0.1003	0.0971
BNUDC [14]	0.08	4.6	33.39	0.9610	0.1748	0.1511	38.22	0.9798	0.0988	0.0964
FSI(Ours)	0.07	3.8	33.85	0.9627	0.1738	0.1457	38.60	0.9805	0.0979	0.0915

Table 2. Quantitative comparison (PSNR↑, SSIM↑, LPIPS↓, and DISTS↓) on the POLED and TOLED [45] dataset. The results are tested on RGB channels. Red indicates the best and blue indicates the second best performance (best viewed in color).

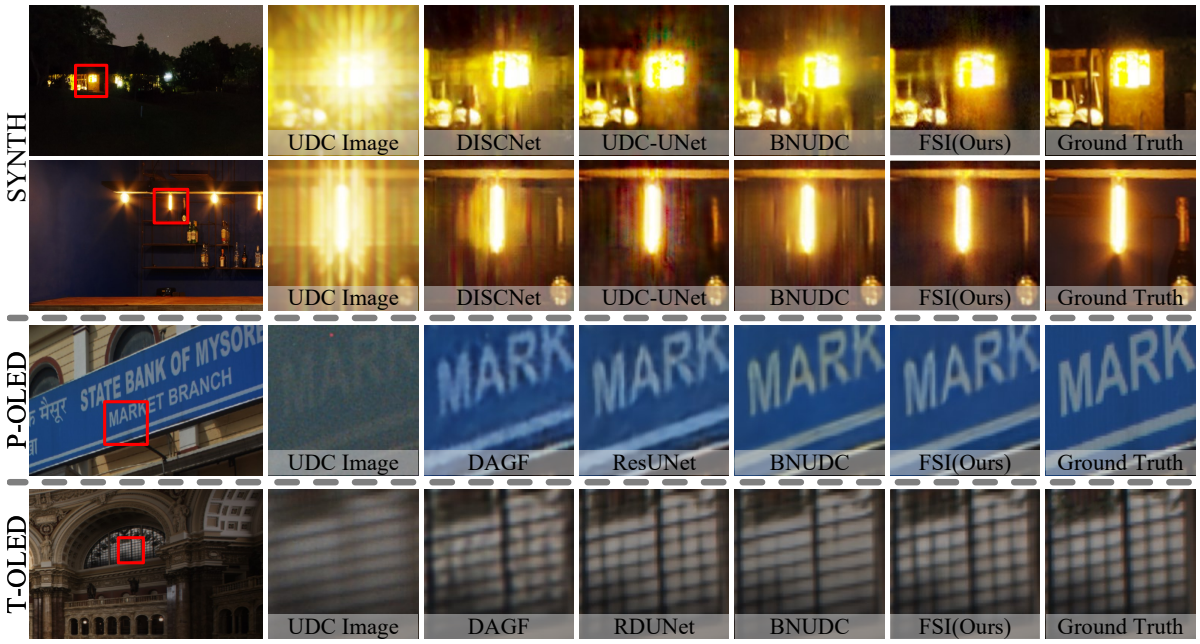

Figure 7. Visual results on SYNTH [7], P-OLED [45], and T-OLED [45] dataset. The method is shown at the bottom of each case. Zoom in to see better visualization.

pabilities under different UDC datasets.

Qualitative comparison. To further compare the visual qualities of different approaches, we show visual results restored by FSI and other SOTA methods on different datasets in Fig. 7. It can be observed that FSI has a great improvement in visual quality, especially in areas with flare and texture loss caused by diffraction. More visualizations can be found in the supplementary.

4.4. Ablation Study

In this section, we conduct ablation for each component and study the effect of the number of FSJ and channels.

Individual components. Based on our proposed model, we directly use ResBlock [10] to replace frequency and spatial learning blocks as the "Base" model and progressively add each component for comparisons. To ensure a fair comparison, we try to keep same parameters for each experiment by changing the channels. As shown in Tab. 3, with the addition of SLB and FLB, PSNR can be improved from 44.84 dB to 45.79 dB, which verifies their powerful ability for feature learning and spectral bias removal. When DTU is involved, the frequency and spatial features promote each other, and the performance is improved to 46.05 dB. This demonstrates the superiority of FSJ(SLB+FLB+DTU) for jointly learning spatial and frequency domain features.

Base	FSJ Module			CT	PSNR	SSIM	LPIPS	DISTS
	SLB	FLB	DTU					
✓					44.84	0.9918	0.0136	0.0219
✓	✓				45.02	0.9929	0.0122	0.0198
✓		✓			45.45	0.9930	0.0122	0.0190
✓	✓	✓			45.79	0.9935	0.0116	0.0183
✓	✓	✓	✓		46.05	0.9938	0.0109	0.0147
✓	✓	✓	✓	✓	**46.14**	**0.9951**	**0.0101**	**0.0138**

Table 3. Ablation study on the SYNTH [7] dataset. SLB: spatial learning block. FLB: frequency learning block. DTU: dual transfer unit. CT: color transform module. Our proposed FSI can be interpreted as "Base+FSJ(SLB+FLB+DTU)+CT"

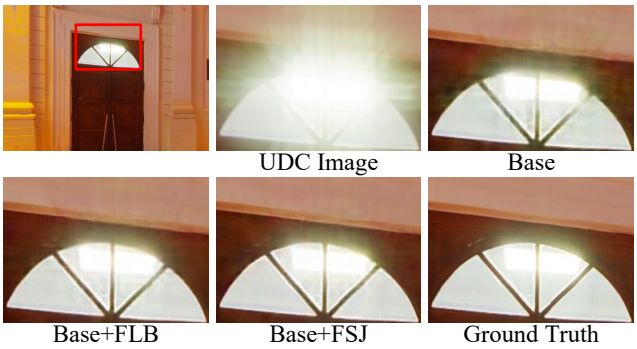

Figure 8. Ablation study on the frequency learning block (FLB) and frequency-spatial joint (FSJ) module on the SYNTH [7] dataset. FSJ can be interpreted as "FLB+SLB+DTU".

FSJ Num.	PSNR	SSIM	LPIPS	DISTS
2	40.48	0.9870	0.0277	0.0399
4	43.36	0.9889	0.0208	0.0299
6	46.14	0.9951	0.0101	0.0138
8	46.41	0.9956	0.0100	0.0134

Table 4. Ablation study results of the number of FSJ modules used on the SYNTH [7] dataset.

When the CT module is progressively added, the color temperature is optimized further, and the performance is improved to 46.14 dB. This demonstrates the superiority of each part in FSI. We further explore the visual differences as shown in Fig. 8. FLB can eliminate the diffraction-induced flare, and FSJ can interact frequency features with spatial features to produce clearer textures.

Number of FSJ modules. As shown in Tab. 4, we experiment with different numbers of FSJ in intervals of 2. The performance is positively correlated with the FSJ number. It demonstrates the effectiveness of the FSJ module for joint learning in different domains. However, the performance gain gradually decreases when the FSJ number is more than 6. After a trade-off between performance and parameters, the final number is set to 6.

Number of channels. As shown in Tab. 5, we progressively increase the channel number in intervals of 16 to explore their impact on performance. The performance bottleneck appears at the channel count of 96 and leads to the redundancy of features. In our model, we empirically choose 96 as the channel number.

4.5. Evaluation on Real UDC Image

Besides the generated SYNTH dataset, we follow the settings in DISCNet [7] and conduct a comparison on real images. Since the real data is collected without ISP, we use the same post-processing as in [7] for better visualization. As shown in Fig. 9, for diffraction-induced flare and texture loss, our FSI can produce clearer textures. This demonstrates the robustness of our FSI. It is noteworthy that, the same as [14], our method is not designed to eliminate noise.

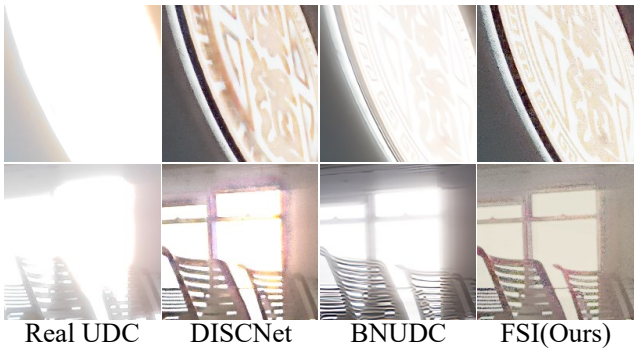

Figure 9. Visual comparison on a real input UDC image.

Channel Num.	PSNR	SSIM	LPIPS	DISTS
72	45.49	0.9929	0.0126	0.0201
80	45.73	0.9934	0.0116	0.0188
96	46.14	0.9951	0.0101	0.0138
112	46.23	0.9950	0.0100	0.0135

Table 5. Ablation study results of the number of channels used on the SYNTH [7] dataset.

5. Limitations

When the lens is closer to the light source, the diffraction-induced flare covers much of the image content. Our work is superior in recovering regular textures (*e.g.*, the fourth row of Fig. 7) through frequency domain learning. Nevertheless, for the recovery of the irregular texture loss in large areas, there are still some limitations. Besides, our method restores the UDC image only in the RGB domain. While processing in the RAW domain is beyond the scope of this paper and requires further exploration.

6. Conclusion

In this paper, we study UDC image restoration in the frequency domain and introduce a new perspective to handle them by jointly exploring the information in the frequency and spatial domains. In particular, we propose a novel frequency and spatial interactive learning network (FSI), which includes a series of frequency-spatial joint (FSJ) modules to handle diffraction and a color transform module to handle the color shift. Such a design can efficiently eliminate the diffraction-produced spectral biases and recover the textures near the center of high-intensity diffraction. Experimental results show significant performance improvements and clear visual margins between the proposed FSI and existing SOTA models. In the future, we will focus on evaluating and extending our method in other low-level vision tasks through more explorations.

Acknowledgement. This work was supported in part by the NSFC under Grant 62272380 and 62103317, the Science and Technology Program of Xi'an, China under Grant 21RGZN0017, and MEGVII Technology.

References

[1] Mahmoud Afifi and Michael S Brown. Deep white-balance editing. In *CVPR*, pages 1397–1406, 2020.

[2] E Oran Brigham and RE Morrow. The fast fourier transform. *IEEE spectrum*, 4(12):63–70, 1967.

[3] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *CVPR*, pages 182–192, 2021.

[4] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *NeurIPS*, 33:4479–4488, 2020.

[5] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 44(5):2567–2581, 2020.

[6] CH Brown Elliott, TL Credelle, S Han, MH Im, MF Higgins, and P Higgins. Development of the pentile matrix™ color amlcd subpixel architecture and rendering algorithms. *Journal of the Society for Information Display*, 11(1):89–98, 2003.

[7] Ruicheng Feng, Chongyi Li, Huaijin Chen, Shuai Li, Chen Change Loy, and Jinwei Gu. Removing diffraction image artifacts in under-display camera via dynamic skip connection network. In *CVPR*, pages 662–671, 2021.

[8] Ruicheng Feng, Chongyi Li, Shangchen Zhou, Wenxiu Sun, Qingpeng Zhu, Jun Jiang, Qingyu Yang, Chen Change Loy, and Jinwei Gu. Mipi 2022 challenge on under-display camera image restoration: Methods and results. *arXiv preprint arXiv:2209.07052*, 2022.

[9] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *CVPR*, pages 1604–1613, 2019.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[11] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *ACM MM*, pages 2024–2032, 2019.

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

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

[14] Jaihyun Koh, Jangho Lee, and Sungroh Yoon. Bnudc: A two-branched deep neural network for restoring images from under-display cameras. In *CVPR*, pages 1950–1959, 2022.

[15] Kinam Kwon, Eunhee Kang, Sangwon Lee, Su-Jin Lee, Hyong-Euk Lee, ByungIn Yoo, and Jae-Joon Han. Controllable image restoration for under-display camera in smartphones. In *CVPR*, pages 2073–2082, 2021.

[16] Xin Li, Xin Jin, Tao Yu, Simeng Sun, Yingxue Pang, Zhizheng Zhang, and Zhibo Chen. Learning omni-frequency region-adaptive representations for real image super-resolution. In *AAAI*, volume 35, pages 1975–1983, 2021.

[17] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. 4D LUT: Learnable context-aware 4d lookup table for image enhancement. *arXiv preprint arXiv:2209.01749*, 2022.

[18] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *CVPR*, pages 5687–5696, 2022.

[19] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. TTVFI: Learning trajectory-aware transformer for video frame interpolation. *arXiv preprint arXiv:2207.09048*, 2022.

[20] Xina Liu, Jinfan Hu, Xiangyu Chen, and Chao Dong. Udc-unet: Under-display camera image restoration via u-shape dynamic network. *arXiv preprint arXiv:2209.01809*, 2022.

[21] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[22] Xintian Mao, Yiming Liu, Fengze Liu, Qingli Li, Wei Shen, and Yan Wang. Intriguing findings of frequency selection for image deblurring. In *AAAI*, 2023.

[23] Hrishikesh Panikkasseril Sethumadhavan, Densen Puthussery, Melvin Kuriakose, and Jiji Charangatt Victor. Transform domain pyramidal dilated convolution networks for restoration of under display camera images. In *ECCVW*, pages 364–378. Springer, 2020.

[24] Fengqing Qin, Chaorong Li, Lilan Cao, Lihong Zhu, Xuyan Zou, Xiaomei Li, Tianqi Zhang, and Yilan Xue. Blind image restoration with defocus blur by estimating point spread function in frequency domain. In *2021 5th International Conference on Advances in Image Processing (ICAIP)*, pages 62–67, 2021.

[25] Zong Qin, Ruijin Qiu, Minyi Li, Xinyi Yu, and Bo-Ru Yang. P-78: Simulator-based efficient panel design and image retrieval for under-display cameras. In *SID Symposium Digest of Technical Papers*, volume 52, pages 1372–1375. Wiley Online Library, 2021.

[26] Zong Qin, Yu-Hsiang Tsai, Yen-Wei Yeh, Yi-Pai Huang, and Han-Ping David Shieh. See-through image blurring of transparent organic light-emitting diodes display: calculation method based on diffraction and analysis of pixel structures. *Journal of Display Technology*, 12(11):1242–1249, 2016.

[27] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. FcaNet: Frequency channel attention networks. In *ICCV*, pages 783–792, 2021.

[28] Zhongwei Qiu, Huan Yang, Jianlong Fu, and Dongmei Fu. Learning spatiotemporal frequency-transformer for compressed video super-resolution. In *ECCV*, pages 257–273. Springer, 2022.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[30] Mingwen Shao, Yuanjian Qiao, Deyu Meng, and Wangmeng Zuo. Uncertainty-guided hierarchical frequency domain transformer for image restoration. *KBS*, page 110306, 2023.

[31] Varun Sundar, Sumanth Hegde, Divya Kothandaraman, and Kaushik Mitra. Deep atrous guided filter for image restoration in under display cameras. In *ECCVW*, pages 379–397. Springer, 2020.

[32] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor

Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, pages 2149–2159, 2022.

[33] Wei Tu, Yong Yang, Shuying Huang, Weiguo Wan, Lixin Gan, and Hangyuan Lu. Mmdn: Multi-scale and multi-distillation dilated network for pansharpening. *IEEE TGRS*, 60:1–14, 2022.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

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

[36] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *CVPR*, pages 1740–1749, 2020.

[37] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *CVPR*, pages 2281–2290, 2020.

[38] Anqi Yang and Aswin C Sankaranarayanan. Designing display pixel layouts for under-panel cameras. *IEEE TPAMI*, 43(7):2245–2256, 2021.

[39] Qirui Yang, Yihao Liu, Jigang Tang, and Tao Ku. Residual and dense unet for under-display camera restoration. In *ECCVW*, pages 398–408. Springer, 2021.

[40] Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao. Frequency and spatial dual guidance for image dehazing. In *ECCV*, pages 181–198. Springer, 2022.

[41] 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*, pages 14821–14831, 2021.

[42] Qianyi Zhang, Zhixin Zeng, Yiming Liu, Kang Tang, and Ji Wang. Dynamic scene deblurring using enhanced feature fusion and multi-distillation mechanism. In *IJCNN*, pages 1–8. IEEE, 2021.

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[44] Yuqian Zhou, Michael Kwan, Kyle Tolentino, Neil Emerton, Sehoon Lim, Tim Large, Lijiang Fu, Zhihong Pan, Baopu Li, Qirui Yang, et al. Udc 2020 challenge on image restoration of under-display camera: Methods and results. In *ECCVW*, pages 337–351. Springer, 2020.

[45] Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, and Timothy Large. Image restoration for under-display camera. In *CVPR*, pages 9179–9188, 2021.

[46] Yang Zhou, Yuda Song, and Xin Du. Modular degradation simulation and restoration for under-display camera. In *ACCV*, pages 265–282, 2022.

[47] Yunliang Zhuang, Zhuoran Zheng, and Chen Lyu. DPFNet: A dual-branch dilated network with phase-aware fourier convolution for low-light image enhancement. *arXiv preprint arXiv:2209.07937*, 2022.