# Lightweight Neural Network for Enhancing Imaging Performance of Under-Display Camera

Yuenan Li, *Senior Member, IEEE*, Jin Wu, *Student Member, IEEE*, and Zetao Shi

***Abstract*— Under-Display Camera (UDC) is an emerging feature of cellphone. This technology makes full-screen cellphones possible by hiding the front-facing camera below the display panel, which is in contrast to the conventional designs that place the camera in a bezel or punch-hole on the screen border. However, this novel imaging paradigm also causes degradation. The display panel attenuates and diffracts incoming light, so the images captured by UDC contain multiple artifacts, such as blurring, color shift, and low intensities. This paper proposes a lightweight deep learning approach to restore UDC images in a blind setting. The restoration network uses cross-scale modulation to exploit complementary information from multi-scale representations and capture the self-similarity across scales, aiming to find the cues for recovering distortion-free images. To facilitate the deployment of this scheme across mobile devices, especially on those with limited memory space and computing power, we compress the restoration network by reducing architectural redundancy. An adaptive distillation algorithm is designed to exploit knowledge from a pre-trained full-size model. The proposed work also interprets the behavior of the neural network in utilizing local and non-local information to restore UDC images. The proposed algorithm is evaluated over three datasets of the images captured by the cameras below different types of display panels. The results of comparative experiments demonstrate that our algorithm shows comparable or superior performance to the competing ones that are much heavier in parameter amount and computational complexities.**

***Index Terms*— Blind image restoration, lightweight neural network, under-display camera.**

## I. Introduction

THE functions of cellphone have gone far beyond making voice calls and text messaging. With the rapid growth of computing power, cellphone has become an all-in-one portable terminal for entertainment, mobile payment, photography, fitness tracking, etc. Cellphone manufacturers have been exploring innovative designs to increase the size of the display panel (i.e., screen), but the front-facing camera is an obstacle to full-screen cellphone. Some designs make the camera less obtrusive by placing it in a punch-hole or notch on top of the front panel or using a pop-up camera. Under-display camera (UDC) is a cutting-edge technology making full-screen design possible. UDC mounts the front-facing camera below a display panel that is made of high-transparency organic light-emitting diode (OLED) materials. As a result, the screen can directly reach the borders of the frame without the disruption of punch-holes or notch. In addition to maximizing the aesthetic continuity of the front panel, UDC also improves the experience of eye contact during video chatting since a user can look directly through the camera.

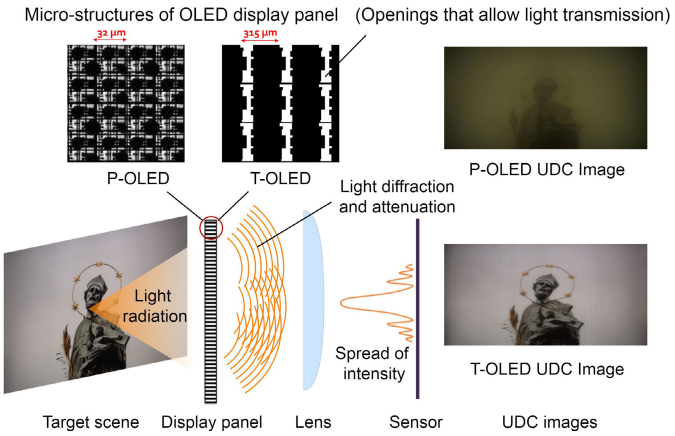



Fig. 1. Illustration of the imaging via under-display camera. The magnified images of the micro-structures of OLED display panels are adopted from [1].

OLED display panel is comprised of stacked organic layers. The display panel over UDC uses densely located micro-openings to allow light transmit through the stacked layers to reach the camera. Fig. 1 shows a schematic illustration of the UDC-based imaging system. The magnified micro-structures of the 4K Transparent-OLED (T-OLED) and Pentile-OLED (P-OLED) display panels, as well as the images captured by the cameras hidden below the two types of display panels are also shown in this figure. Apparently, the interference of the display panel with light degrades imaging quality. The micro-openings on the display panel (i.e., the white parts in the zoom-in views) that enable light transmission have complex structures, and their sizes are close to the wavelength of light, as shown in Fig. 1. As a result, the diffraction caused by the openings results in image blurring. In addition, the yellow substrate of the P-OLED display panel incurs color shift [1]. Most importantly, the opaque parts of the display panel attenuate the radiation sensed by the camera. For instance, the P-OLED display panel has a transmittance of

Manuscript received 9 November 2022; revised 11 May 2023; accepted 27 May 2023. Date of publication 9 June 2023; date of current version 8 January 2024. This work was supported by the National Natural Science Foundation of China under Grant 61972281 and Grant 61572352. This article was recommended by Associate Editor J. Liu. *(Corresponding author: Yuenan Li.)*

The authors are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: ynli@tju.edu.cn; wujin_tju@163.com; shizetao10@163.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3285014.

Digital Object Identifier 10.1109/TCSVT.2023.3285014

only 3% [1], so the raw images captured by P-OLED UDC are much darker than those captured by conventional cameras and have very low visibility (see the example presented in Fig. 1). Some UDC designs reduce the sizes of display pixels to create more space to let light pass through. Although enlarging the openings on a display panel increases transmittance, this is at the cost of lowering the resolution and brightness of the contents displayed on the screen. Designing UDC devices needs to achieve a balance between the imaging quality of the hidden camera and the displaying quality of the OLED panel. However, the cost of developing an OLED display panel is quite high, while re-designing the physical structure of the display panel (e.g., optimizing the spatial layouts of openings) cannot completely eliminate imaging artifacts. A more economical solution is to leverage computational approaches to restore the images captured by UDC.

Deep learning has demonstrated outstanding capability in low-level vision tasks, such as low-light image enhancement [2], [3], [4], deblurring [5], [6], denoising [7], [8], dehazing [9], super-resolution [10], [11], [12], and reflection removal [13]. It is worth noting that UDC image restoration is distinct from these conventional tasks. First, the degradation associated with UDC is more complex. As mentioned above, UDC distortion is a mixture of low illumination, blurring, flare, color shift, noise, etc, making it difficult to use a single model to formulate imaging distortion. Second, the point spread functions (PSFs) of display panels are not spatially invariant [14]. Third, the restoration of UDC images needs to be implemented on mobile devices, requiring the restoration algorithm to be lightweight and energy-efficient, so the models that are heavy in parameters and computation are not applicable to UDC image restoration.

In this paper, we propose a deep learning-based algorithm for the blind restoration of UDC images in the wavelet domain. The restoration network uses a parallel multi-scale architecture and cross-scale interactions to expand the receptive field of feature learning, aiming to accommodate the spatial supports of the PSFs of the display panel while constraining the parameter and computational budgets. To fit the UDC devices with less memory and computational resources, we further reduce the architecture redundancies of the restoration network and develop a wavelet-domain distillation algorithm. The distillation algorithm trains the compressed network by transferring knowledge from another pre-trained network (without accessing the ground-truth distortion-free images). This work also leverages the gradient-based approach to interpret the behavior of the restoration network in tackling this new type of image restoration task. We demonstrate how the restoration network exploits local and non-local dependencies among pixels to recover the clear-scene image. Quantitative and qualitative performance are assessed on three sets of UDC images captured through the display panels with different materials and microstructures. The comparison shows that the images restored by the proposed algorithm exhibit the highest quality scores, while the parameter amount and computational complexity are lower than the competing algorithms. To summarize, the contributions of this work are as follows.

1) We devise a lightweight neural network that can restore UDC images in a blind manner. The algorithm harnesses the multi-orientation and multi-resolution representation capabilities of wavelet decomposition and the multi-scale feature fusion capability of neural network to alleviate the mixed distortions introduced by UDC.

2) In order to save memory consumption, we present a lightweight variant of the restoration model by reducing the architectural redundancy in learning multi-scale features. We also propose an adaptive distillation algorithm for bridging the performance gap between the lightweight and full-size models.

3) We also attempt to understand the mechanisms of UDC restoration made by deep neural network. The results of model interpretation reveal the local and non-local cues that are helpful for inferring distortion-free images.

The rest parts of this paper are organized as follows. A brief literature review is presented in Section II. Section III describes the architecture of the restoration network, the compression and knowledge distillation schemes for constructing the lightweight variant, and the results of model interpretation. Experimental results are reported and discussed in Section IV. Finally, conclusions are summarized in Section V.

## II. Related Work

In light of the promising future of full-screen devices in the consumer electronics market, UDC imaging system has received increasing attention from both the academia and industry. Some experimental studies examined the optical characteristics of display panels [15], [16], [17], aiming to understand their impacts on imaging. The work in [15] measured the blurring effect caused by the light diffraction in transparent OLED and modeled the degree of blurriness using Gaussian kernels. To find out the origin of image blurring in UDC systems, Tang et al. made a component-wise analysis of OLED display panel [16]. The relationship between diffraction and the micro-structures of OLED display panel was studied in [18]. This work also describes a method for simulating distorted images based on diffraction theory. The following-up work in [19] demonstrates the impact of micro-structures on the subjective qualities of blurred images.

The imaging performance of UDC system can be improved through optical and computational approaches. The Fourier analysis in [20] indicates that the periodic patterns and the shapes of the openings on the display panel are crucial factors affecting the blurring effect. The authors proposed to use the random tiling of openings to alleviate the blurring artifacts. However, breaking the periodicity of openings poses challenges to the fabrication of OLED display panels. Most of the UDC image restoration schemes in the literature are computationally based. To the best of our knowledge, the work presented in [21] should be one of the earliest explorations on this problem. This work was designed to enable more natural eye contact in video conferences. The teleconference system comprises a camera placed under the screen and a color+depth camera placed at the screen border. The information sensed by the color+depth camera is used to guide

the restoration of UDC image. Since this work relies on an auxiliary sensor, it is not applicable to current UDC mobile devices. The first data-driven UDC restoration method can be traced to Zhou et al.'s pilot study in [1]. Based on a systematic investigation of the UDC imaging system, including the light transmission rate, PSF, and the modulation transfer function of T-OLED and P-OLED display panels, the authors devised a Monitor-Camera Imaging System (MCIS) for capturing UDC images and their distortion-free correspondences. This work demonstrates the feasibility of restoring UDC images via data-driven approach, and a convolutional neural network (CNN) with encoder-decoder architecture was developed to suppress UDC degradation. In [14], Feng et al. described a non-blind restoration algorithm for removing the diffraction flare in UDC images. The measured PSF is taken as an auxiliary input of the restoration network, and the neural network uses dynamic skip connections to accommodate the spatial variability of PSF. The algorithm in [22] restores UDC images from the perspective of contrast enhancement, where a high-order curve for correcting the hue and saturation components of a UDC image is estimated via neural network. Reference [23] presents a restoration algorithm that can achieve controllable UDC restoration. The algorithm balances the levels of noise removal and de-blurring according to user preference. Similar to [14], it also requires measuring PSF in advance.

UDC restoration must be conducted on mobile platforms since restored images need to be presented to users immediately after the shot. Hence, handling heterogeneous and spatially-variant distortions while constraining the consumption of memory and computational resource are the key challenges of UDC restoration. Despite the initial success of deep learning in UDC restoration, there is much room for reducing resource consumption. Some of the prior approaches are heavy in computation and memory space. To pave the ways for the applications of deep neural networks in mobile and Internet-of-Things environments, increasing endeavors are being devoted to finding lightweight yet effective neural network architectures. One line of research leverages the strategies such as depth-wise separable convolution, group-wise convolution, and network architecture search to design compact network architectures. Some representative work includes the Inception series [24], [25], [26], Xception [27], MobileNets [28]. Another line of research reduces computational complexity and model size via weights pruning [29], quantization [30], or transferring knowledge from a larger network to a smaller one [31]. In image restoration, some recent studies explored the lightweight implementations of deep learning [10], [11], [32], [33], [34], [35]. For example, Hui et al. proposed a lightweight image super-resolution network that is composed of enhancement and compression modules [32]. Based on a systematic study on the efficiency of some essential components in low-level vision tasks, Lahiri et al. designed several efficient implementations of $3 \times 3$ convolution, dilated convolution, and upsampling [33]. Reference [10] presents an information multi-distillation network to accelerate image super-resolution, where channel splitting and hierarchical feature refinement are used in conjunction to progressively recover fine details of the input image while reducing the data loads of convolutional layers. The work in [10] optimizes this architecture using a more compact and flexible alternative to the channel splitting operation. Reference [11] makes a comprehensive analysis of the information distillation mechanism, and a more lightweight form was developed by refining the channel splitting operation in feature distillation. Nie et al. proposed a lightweight image super-resolution algorithm that first identifies a set of intrinsic features from a pre-trained model, and redundant features are then derived from the intrinsic ones using computationally-cheap operations [34].

## III. Methods

### A. Architecture of UDC Restoration Network

The proposed algorithm takes a single UDC image as input, requiring no prior knowledge about the device. The reason for adopting blind approach lies in the high variability of PSF and the difficulty of accurate PSF measurement. The PSF of UDC is determined by multiple factors, including the geometric structures of the micro-openings on the OLED display panel, the wavelength of light, the parameters of the camera, etc. Due to manufacturing imperfection, the structures of the openings are not congruent in general, so the PSF of UDC is spatially variant. In addition, the PSF also varies with the wavelength of light [19]. The studies in [23] reveal that the oblique incidence of light and conical diffraction also complicate the measurement of PSF. Therefore, a fixed PSF may not accurately describe the degradation patterns of UDC at all pixel coordinates, and some assumptions in deriving PSF do not hold in practice.

The architecture of our lightweight UDC restoration network is illustrated in Fig. 2. UDC images are restored in the wavelet domain due to the following advantages of discrete wavelet transform (DWT). Firstly, DWT can ease image restoration by separating the complicated imaging degradations of UDC into different sub-bands. The low-frequency sub-band is sensitive to the low-light artifact caused by light attenuation and the color shift caused by wavelength-selective light absorption, and the high-frequency sub-bands are sensitive to blurring artifacts. The proposed network uses two branches to estimate the low-frequency and high-frequency sub-bands for reconstructing the output image. Due to the de-interleaving effect of DWT, each branch can focus on relatively simple degradation patterns. Secondly, DWT and IDWT are free of learnable parameters. They also play the roles of down-sampling and up-sampling, which can enlarge the receptive fields of subsequent convolutions, while the invertibility of DWT ensures that there is no information loss. Finally, DWT allows us to allocate parameter and computational budgets to different parts of the restoration network according to the significance of the corresponding sub-bands.

As Fig. 2 shows, the input and output of the restoration network are bridged by a residual connection [36], so the intermediate layers of the network estimate the residual between the input UDC image and the clear-scene target in the wavelet domain. We first apply a double-level Haar wavelet transform on the input image. The restoration network has two parallel branches that process the wavelet coefficients output by the

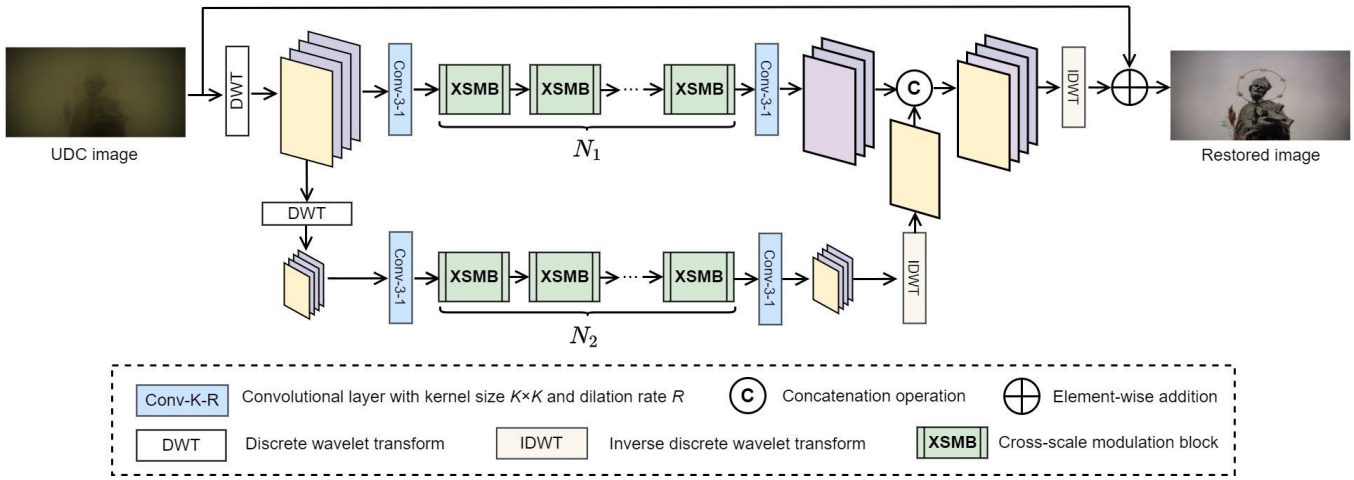



Fig. 2. Overall architecture of the restoration network. Details of XSMB will be shown in Fig. 3.

first and second levels of DWT, respectively. The two branches contain a stack of cross-scale modulation blocks (XSMBs).

We observe that UDC distortion is more prominent in the regions with edges and complex textures. The reason is that edges and textures usually associate with drastic variations of pixel intensities and are thus more susceptible to the blurring artifact. Accordingly, the contextual correlations among pixels are essential for analyzing the blurring artifact and modeling the PSF of the display panel. As will show later, besides local contexts, deep neural network also utilizes non-local information in UDC restoration, since self-similarity frequently recurs in natural images [37]. Hence, to model the local and non-local content dependencies, XSMBs need to ensure the diversity and spatial coverage of receptive fields.

Each XSMB first applies a $3 \times 3$ convolution on the input, giving a feature map $\boldsymbol{X} \in \mathbb{R}^{C \times H \times W}$. The feature map then goes through three parallel pipelines to obtain multi-scale representations of the input. Each pipeline consists of a dilated convolutional layer and a $1 \times 1$ convolutional layer, as shown in the upper part of Fig. 3. The dilated convolution layers are of dilation rates $R_L$, $R_M$ and $R_H$, respectively, where $R_L < R_M < R_H$. To save parameters, each dilated convolution is conducted in a grouped manner, and the number of groups is set to $G = 3$. Denote the outputs of the three branches by $\boldsymbol{F}_L$, $\boldsymbol{F}_M$, and $\boldsymbol{F}_H$, respectively, all of the resolution $C \times H \times W$. Along the bottom-up direction, the feature maps describe the characteristics of DWT coefficients within increasingly larger receptive fields. For the feature map extracted at each scale, we modulate it using the mask learned from those extracted at other scales. This enables the neural network to model the mutual dependencies among image representations and exploit complementary information in the scale space.

Take $\boldsymbol{F}_L$ for example, we concatenate $\boldsymbol{F}_M$ and $\boldsymbol{F}_H$ along the channel direction, and the concatenated feature map is passed to two convolutional layers to learn a mask $\boldsymbol{M} \in \mathbb{R}^{C \times H \times W}$ (see the bottom part of Fig. 3):

$$\boldsymbol{M} = \text{Sigmoid}[\boldsymbol{W}_2 \cdot \text{ReLU}(\boldsymbol{W}_1 \cdot [\boldsymbol{F}_M, \boldsymbol{F}_H])], \tag{1}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are the weights of the two convolutional layers, respectively, and $[\cdot, \cdot]$ denotes the concatenation operation. The mask is then multiplied with $\boldsymbol{F}_L$ in an element-wise manner as $\boldsymbol{F}_L \leftarrow \boldsymbol{F}_L \otimes \boldsymbol{M}$, where $\otimes$ denotes element-wise multiplication. In this way, the elements of $\boldsymbol{F}_L$ are re-calibrated based on the contextual characteristics extracted from larger backgrounds. The cross-scale modulation provides a mechanism for simultaneously harnessing the fine-granular characteristics of local variation and the longer-range contextual features. Likewise, the largest-scale feature $\boldsymbol{F}_H$ is modulated by referencing the finer-detail information described by $\boldsymbol{F}_L$ and $\boldsymbol{F}_M$.

The cross-scale modulation offers the restoration network a high sensitivity to the local variation of UDC degradations. As mentioned above, the PSF of UDC can vary from pixel to pixel. Unlike the shift-invariant kernel of 2D convolution, the modulation weight assigned to each position is determined by the local characteristics extracted from its neighborhoods using different receptive fields. The diffraction effect of the display panel increases the correlation among neighboring pixels, and the multi-scale contextual features for computing the weights can characterize the degradation patterns at each position. Therefore, the fine granularity of XSMB helps the restoration network to handle the spatial variability of degradation patterns. In addition, the multiple receptive fields formed by parallel dilated convolutional layers can cover the PSFs with different spatial supports. The study on the statistical priors of natural images shows that some visual patterns repeat at multiple scales [37]. XSMB also allows utilizing the complementary information in different scales to restore such patterns in UDC images. The results of model interpretation also corroborate that the restoration network is able to utilize non-local self-similarity to restore UDC images, as will be discussed later in Section III-C.

Two-level inverse discrete wavelet transform (IDWT) is applied on the outputs of the first and second branches of the restoration network to generate the residual signal for restoring the input UDC image, as shown in Fig. 2. Due to the parallelization of the two branches and the multi-resolution representation of DWT, the proposed algorithm can make scalable restoration. For the scenarios that need low latency, such as video chatting and facial image-based de-lock, the clear-scene image can be directly recovered from the outputs of the second branch.

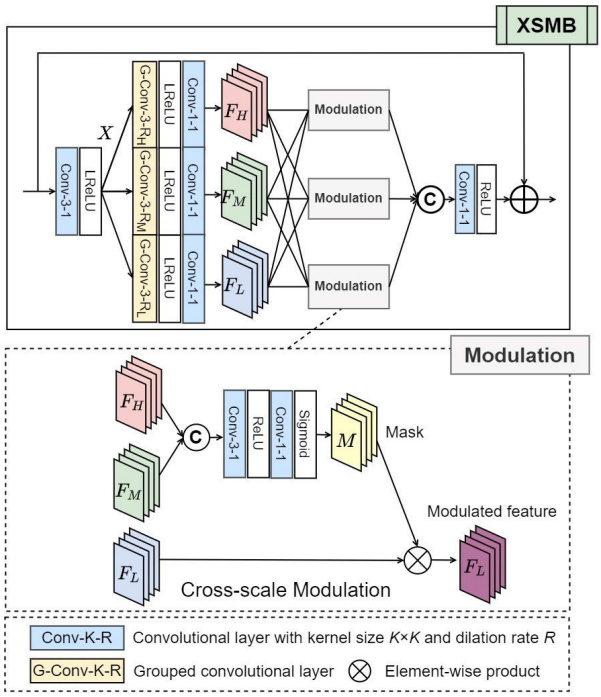



Fig. 3. Architecture of the cross-scale modulation block (XSMB).

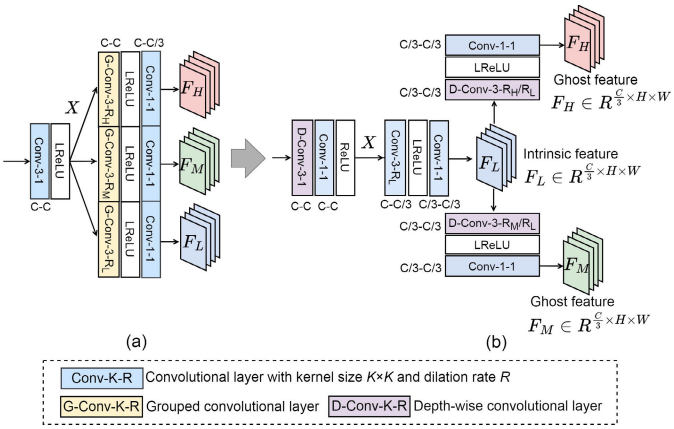



Fig. 4. Lightweight implementation of multi-scale feature extraction in XSMB. (a) The parallel pipelines in the original XSMB shown in Fig. 3, (b) Lightweight variant of (a). The numbers next to each convolutional layer indicate the channel numbers of the input and output.

The restoration network is trained to minimize the $L_1$ norm between the restored image and the ground-truth image:

$$L = |R(\boldsymbol{U}) - \boldsymbol{I}|_1, \tag{2}$$

where $R(\cdot)$, $\boldsymbol{U}$, and $\boldsymbol{I}$ denote the restoration network, input UDC image, and the ground-truth image, respectively.

### *B. Lite-XSMB and Knowledge Distillation*

The restoration network has less than 0.5M parameters and can beat the comparative models which have several times more parameters and computational complexity. To ease the deployment on cellphones, we also design a lightweight variant of the restoration network and the corresponding distillation algorithm. The lightweight network is designed by compressing the parameter-heavy components in the original network. Nearly 98% of the parameters are from cascaded XSMBs, and the parallel convolutional layers take a substantial amount of the parameters in each XSMB.

The function of the parallel pipelines in each XSMB is to capture the characteristics of the input at different scales. It is worth noting that larger-scale representations of the input (i.e., $\boldsymbol{F}_M$ and $\boldsymbol{F}_H$) can also be derived from the smallest-scale one (i.e., $\boldsymbol{F}_L$), since consecutive convolutions can progressively expand the receptive field of feature learning, which has similar effects as increasing dilation rates. If $\boldsymbol{F}_L$ is more compact than the input $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$, deriving $\boldsymbol{F}_M$ and $\boldsymbol{F}_H$ from $\boldsymbol{F}_L$ requires less parameters and computation than from $\boldsymbol{X}$. To this end, the lightweight XSMB first outputs a compact feature map $\boldsymbol{F}_L$ by applying depth-wise dilated convolutions (with the lowest dilation rate $R_L$) on $\boldsymbol{X}$, and then a $1\times 1$ convolutional layer reduces the channel number to $\frac{C}{3}$. Similar to [38], we take $\boldsymbol{F}_L$ as the intrinsic feature. Considering the redundancy among multi-scale features, $\boldsymbol{F}_M$ and $\boldsymbol{F}_H$ can be viewed as the ghost representations of $\boldsymbol{F}_L$. In contrast to the parallel pipelines shown in Fig. 3, the lightweight XSMB infers $\boldsymbol{F}_M$ and $\boldsymbol{F}_H$ from $\boldsymbol{F}_L$ using more efficient transformations. Specifically, two depth-wise dilated convolutional layers with dilation rates $\frac{R_M}{R_L}$ and $\frac{R_H}{R_L}$, respectively, followed by $1\times 1$ convolutional layers, are independently applied on $\boldsymbol{F}_L$ to generate $\boldsymbol{F}_M \in \mathbb{R}^{\frac{C}{3}\times H\times W}$ and $\boldsymbol{F}_H \in \mathbb{R}^{\frac{C}{3}\times H\times W}$, as illustrated in Fig. 4(b). Apart from the parallel pipelines, the lightweight architecture also use less parameters to compute $\boldsymbol{X}$, and the first $3\times 3$ convolutions with $C$-channel input and output are replaced by depth-wise convolutions and $1\times 1$ convolutions [see Fig. 4(b)].

We now calculate the reduction of parameters. Consider the original XSMB module discussed in Section III-A. Assume that all the dilated convolutional layers use $K\times K$ kernels, and their outputs have $C$ channels, as shown in Fig. 4(a). Since each pipeline adopts a $G$-group dilated convolution, the parameter amount of three parallel ones is:

$$P_{\text{org}} = 3\left[\underbrace{G\left(\frac{C}{G}\right)^2 K^2}_{\text{grouped dilated conv.}} + \underbrace{\frac{C}{3}\cdot C}_{1\times 1\text{ conv.}}\right] = \left(\frac{3K^2}{G}+1\right)C^2$$
$$= (K^2+1)C^2 \quad (\forall G = 3). \tag{3}$$

While for the lightweight version shown in Fig. 4(b), there is only one stream for generating the $\frac{C}{3}$-channel intrinsic feature $\boldsymbol{F}_L$ from $\boldsymbol{X}$ and two streams for deriving the ghost representations from $\boldsymbol{F}_L$. The parameter amount reduces to:

$$P_{\text{lite}} = \underbrace{\left[\underbrace{\frac{C}{3}\cdot C\cdot K^2}_{K\times K\text{ conv.}} + \underbrace{\left(\frac{C}{3}\right)^2}_{1\times 1\text{ conv.}}\right]}_{\text{Compute } \boldsymbol{F}_L \text{ from } \boldsymbol{X}} + 2\underbrace{\left[\underbrace{\frac{C}{3}K^2}_{\text{depthwise conv.}} + \underbrace{\left(\frac{C}{3}\right)^2}_{1\times 1\text{ conv.}}\right]}_{\text{Compute } \boldsymbol{F}_M \text{ and } \boldsymbol{F}_H \text{ from } \boldsymbol{F}_L}$$
$$= \frac{(K^2+1)C^2}{3} + \frac{2K^2C}{3}. \tag{4}$$

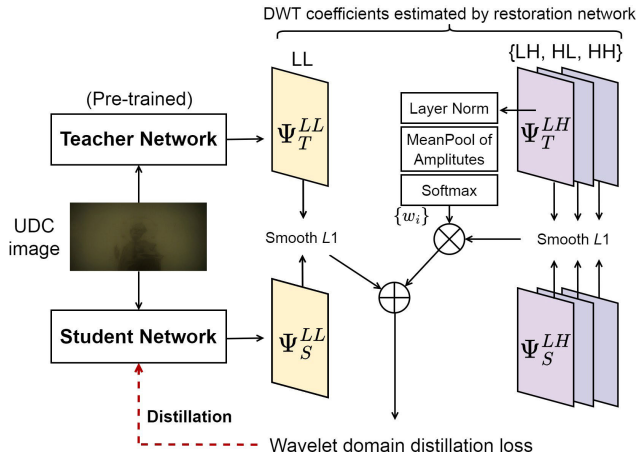



Fig. 5. Teacher-student distillation in the wavelet domain.

For the parameter settings used in this work: $K = 3$ and $C = 48$, we have $P_{\text{lite}} \approx \frac{1}{3} P_{\text{org}}$.

It is desirable that the reduction of parameters does not cause notable performance degradation of UDC restoration, so a distillation algorithm is designed to train the lightweight network (i.e., student network) by transferring knowledge from a pre-trained full-size network (i.e., teacher network). This work addresses the scenario where only a pre-trained teacher network and a set of UDC images are available for training the lightweight student network. The motivation for using unpaired training data is to enhance the flexibility of the distillation algorithm. The cost of collecting the distortion-free images for training UDC restoration models is quite high, since they need to align with the UDC images pixel by pixel. Capturing the same scene using two different types of cameras while maintaining pixel-level alignment requires dedicated tuning of imaging conditions. Due to the high cost of data acquisition, the owner of a training set may not always release the ground-truth images. We expect that the distillation algorithm works even without access to the ground-truth images.

We compute distillation loss from the first-level wavelet coefficients (i.e., the inputs to the last IDWT), as shown in Fig. 5. The coefficients represent the residual between the raw UDC image and its distortion-free correspondence, so their amplitudes are small, especially those in the high-frequency sub-bands. To give more penalty to the small-value discrepancy and prevent outliers from dominating the loss, we use the smooth $L_1$ loss [39] to define distillation loss. Since the statistical distributions of the wavelet coefficients in different sub-brands vary a lot, the discrepancy between the outputs of the teacher and student networks in low- and high-frequency sub-bands are measured in different ways. For the low-frequency sub-band, the distillation loss is computed as:

$$\begin{aligned} L_1 &= \text{Smooth}L_1(\boldsymbol{\Psi}_T^{LL}, \boldsymbol{\Psi}_S^{LL}) \\ &= \sum_{m,n} l(\boldsymbol{\Psi}_T^{LL}[m,n] - \boldsymbol{\Psi}_S^{LL}[m,n]) \end{aligned} \tag{5}$$

where $\boldsymbol{\Psi}_T^{LL}$ and $\boldsymbol{\Psi}_S^{LL}$ are the wavelet coefficients in the LL sub-band estimated by the teacher and student networks, respectively. The $l(\cdot)$ in (5) is defined as follow [39]:

$$l(x) = \begin{cases} \dfrac{x^2}{2\beta}, & \text{if } |x| < \beta; \\ |x| - \dfrac{\beta}{2}, & \text{otherwise,} \end{cases} \tag{6}$$

where the const $\beta$ is set to 1.

The high-frequency sub-bands do not exhibit equal importance in UDC restoration, so the distillation loss is computed according to their significance. Take an image with rich vertical and horizontal edges for example, strong responses can be observed in the LH and HL sub-bands, while complex textures are more salient in the HH sub-band. In light of the directional sensitivity of DWT, we propose an adaptive distillation loss that emphasizes the information with higher visual significance. The DWT coefficients estimated by the teacher network can largely reflect the characteristics of the target scene and UDC artifacts. As mentioned above, high-frequency coefficients are sparse and quite weak in amplitudes. To highlight the informative components, we first apply layer normalization (LN) on the high-frequency sub-bands estimated by the teacher network. Subsequently, the amplitudes of the coefficients within each sub-band are aggregated via mean-pooling. The statistics are fed to a softmax layer to generate sub-band weights. The distance between the estimated DWT coefficients in the high-frequency sub-bands is measured by weighted smooth $L_1$ loss:

$$L_2 = \sum_i w_i \cdot \text{Smooth}L_1(\boldsymbol{\Psi}_T^i, \boldsymbol{\Psi}_S^i), \ \forall i \in \{\text{LH, HL, HH}\}. \tag{7}$$

where $w_i$ is the weight for the $i$-th sub-band inferred from statistics:

$$w_i = \text{Softmax}\{\text{MeanPool}\left[|\text{LN}(\boldsymbol{\Psi}_T^i)|\right]\} \tag{8}$$

The distillation loss combines the distances calculated in the wavelet and spatial domains:

$$L_{\text{distill}} = L_1 + \lambda L_2 + |\hat{\boldsymbol{I}}_S - \hat{\boldsymbol{I}}_T|_1. \tag{9}$$

where $\lambda$ is weight, $\hat{\boldsymbol{I}}_S$ and $\hat{\boldsymbol{I}}_T$ are the images restored by the teacher and student networks, respectively.

### C. Interpretation of UDC Restoration Model

Deep learning approaches restore images in an end-to-end manner, making it difficult to understand the internal behavior of the restoration model. For this new image restoration task, we would like to gain more insights besides algorithm design, so this work also tries to interpret how the deep neural network restores the mixed distortion brought by UDC.

The diffraction effect of the openings on OLED display panel increases the correlation among neighboring pixels in UDC images. Thus, the restoration network needs to leverage the contextual information within the neighborhood of a pixel to estimate its intensity in the clear-scene image. It would be interesting to explore how neighboring pixels contribute to the recovery of a specific region and to what extent the influence of neighboring pixels lasts, especially for dealing the regions with

rich textures. Moreover, we are also curious if the non-local patterns that are similar to those in a given region aid the restoration of this region. Answering these questions requires attributing the output of the restoration network to every pixel of the input image. We use the integrated gradient (IG) based attribution algorithm [40] to interpret the behavior of the neural network in restoring UDC images. The algorithm quantifies the contribution of a given dimension of the input based on the gradient of the output with respect to it. Let us denote the input UDC image and the restored image as $\boldsymbol{U}$ and $\hat{\boldsymbol{I}}$, respectively, and $\hat{\boldsymbol{I}} = R(\boldsymbol{U})$, where $R(\cdot)$ is the restoration network. Without the loss of generality, let us consider two arbitrary pixels at coordinates $(m, n)$ and $(i, j)$ of the input and output images, respectively. The contribution given by the input pixel $\boldsymbol{U}[m, n]$ in reconstructing the output pixel $\hat{\boldsymbol{I}}[i, j]$ can be measured by the following IG value [40]:

$$\mathrm{IG}^{i,j}[m, n] = \boldsymbol{U}[m, n] \int_0^1 \frac{\partial(\hat{\boldsymbol{I}}_\alpha[i, j])}{\partial \boldsymbol{U}[m, n]} d\alpha. \tag{10}$$

where $\hat{\boldsymbol{I}}_\alpha = R(\alpha \boldsymbol{U}),\ \alpha \in [0, 1]$. The IG value examines the gradients of the restoration results with respect to $\boldsymbol{U}[m, n]$ along a linear path where $\alpha$ increases from 0 to 1. According to the property of IG, summing up $\mathrm{IG}^{i,j}[m, n]$ over the input image gives

$$\hat{\boldsymbol{I}}[i, j] = \sum_{(m,n)\in \boldsymbol{U}} \mathrm{IG}^{i,j}[m, n]. \tag{11}$$

Hence, $\mathrm{IG}^{i,j}[m, n]$ shows the importance of $\boldsymbol{U}[m, n]$ in estimating the intensity of $\hat{\boldsymbol{I}}[i, j]$. We sample a small patch $\boldsymbol{P}$ with complex details from a reconstructed image $\hat{\boldsymbol{I}}$ and identify the most influential pixels for restoring the patch. For an arbitrary pixel $(i, j)$ within the patch, we compute its IG values with respect to every pixel in the input image $\boldsymbol{U}$, giving a heat map

$$\boldsymbol{h}^{i,j} = \{|\mathrm{IG}^{i,j}[m, n]|,\ (m, n) \in \boldsymbol{U}\},\ \forall (i, j) \in \boldsymbol{P} \subset \hat{\boldsymbol{I}}. \tag{12}$$

The heat maps corresponding to all pixels within the patch are summed up as $\boldsymbol{H} = \sum_{(i,j)\in \boldsymbol{P}} \boldsymbol{h}^{i,j}$, and $\boldsymbol{H}$ is overlaid over the restored image $\hat{\boldsymbol{I}}$, as illustrated in Fig. 6. The aggregated heat map $\boldsymbol{H}$ visualizes the pixel-wise contribution in reconstructing the patch, and red parts highlight those giving the most significant contribution. It is obvious that a large amount of the information for reconstructing the patch comes from its internal pixels. In particular, we find that the edges and corners, as can be seen from the pixels colored in red in the IG map, contribute most. This result agrees with the fact that abrupt intensity variations are more sensitive to the light diffraction and attenuation effects of OLED display panel. Accordingly, these regions provide more information for estimating the distortion and PSF of the display panel, and the faithful recovery of the salient parts is also vital for the visual qualities of restored images. In contrast, the contribution given by the black squares, which have nearly uniform intensities, is much lower. It is worth noting that the information for restoring the patch is not solely from its internal pixels, while non-local pixels also contribute information, as can be seen from the pixels outside the patch and marked in blue. From the zoom-in figure, the facade of the building contains several shadow-like vertical strips that repeatedly appear in the image. The IG values imply that the restoration of the given patch also depends on distant vertical stripes. We can conclude that neural network can exploit non-local information when restoring some repetitive image patterns.

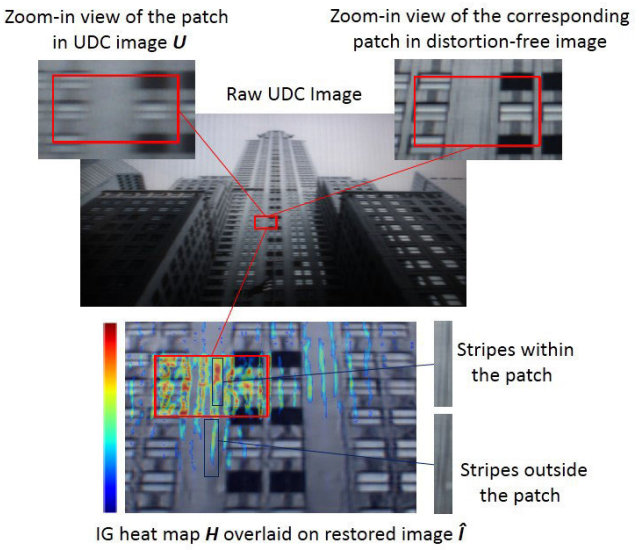



Fig. 6. Visualization of model interpretation results.

## IV. Experimental Results and Discussions

### A. Implementation Details

The proposed restoration network was implemented and tested in PyTorch on a workstation with NVIDIA RTX 2080Ti GPUs and 64G RAM. The network was trained for 500 epochs using the Adam optimizer ($\beta_1$=0.95, $\beta_2$=0.999). The initial learning rate was set to $3\times10^{-4}$ and then gradually declined to $3\times10^{-5}$ through a cosine annealing scheduler. The first and second branches of the restoration network contain five and one XSMBs, respectively (i.e., $N_1 = 1$ and $N_2 = 5$ in Fig. 2). We also tested the restoration performance, model size, and complexity of the proposed algorithm under different combinations of $N_1$ and $N_2$, and the results will be discussed in an individual subsection. The channel numbers of the input and output of each XSMB were set to $C = 48$. All the dilated convolutional layers use $3 \times 3$ kernels, and the dilation rates were set to $R_L = 6$, $R_M = 12$, and $R_H = 18$.

### B. Datasets

To examine the versatility of the proposed algorithm to different types of UDC devices, we experimented with the images shot by the cameras mounted below T-OLED, P-OLED, and ZTE Axon 20's display panels, respectively.

Quantitative performance comparisons were carried out on the datasets developed by Zhou et al. [1]. Per our knowledge, these are the only public datasets covering multiple materials of display panels and supporting the quantitative assessment of blind UDC restoration with pixel-wise aligned distortion-free ground truths. The UDC images were captured by placing T-OLED and P-OLED display panels over a 2K FLIR RGB

TABLE I
QUANTITATIVE COMPARISON RESULTS OF DIFFERENT METHODS ON P-OLED AND T-OLED DATASETS (THE BEST AND SECOND-BEST RESULTS ARE COLORED IN RED AND BLUE, RESPECTIVELY)

| Methods | MACs (G)↓ | Params (M)↓ | P-OLED | | | T-OLED | | |
|---|---|---|---|---|---|---|---|---|
| | | | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| PANet | 25.88 | 6.06 | 28.05 | 0.930 | 0.302 | 35.23 | 0.966 | 0.165 |
| FFA | 287.55 | 4.46 | 28.32 | 0.929 | 0.276 | 35.95 | 0.972 | 0.136 |
| DEUNet | 17.15 | 8.94 | 30.09 | 0.938 | 0.234 | 35.72 | 0.967 | 0.143 |
| IMDN | 44.90 | 0.69 | 26.99 | 0.916 | 0.309 | 35.14 | 0.965 | 0.145 |
| RFDN | 42.18 | 0.68 | 27.30 | 0.919 | 0.306 | 35.34 | 0.967 | 0.168 |
| DAGF | 4.57 | 1.12 | 33.23 | 0.954 | 0.236 | 37.46 | 0.973 | 0.147 |
| Proposed | 3.00 | 0.48 | 33.46 | 0.957 | 0.219 | 38.23 | 0.977 | 0.122 |

camera. The testing images are all of resolution $1024 \times 2048$, and there are 240 pairs of images in both the P-OLED and T-OLED datasets, respectively.

The proposed algorithm was also tested on the high dynamic range UDC images captured by the ZTE Axon 20 UDC cellphone in [14]. The testing images were obtained from real scenes. The image of each scene is obtained by fusing three images captured using different exposures, and some scenes contain strong light sources. The dataset provides paired synthetic UDC images and ground-truth images for training. The testing set contains 30 images, but the ground-truth states of the testing images are unavailable, so this dataset was only used for qualitative evaluation. We trained the restoration algorithms on the synthetic image pairs in the training set and compared their performance using the real-world testing images.

### C. Quantitative and Qualitative Evaluation on T-OLED and P-OLED Datasets

UDC is a recently emerging imaging apparatus, so the research on the related imaging restoration problems is far less than those on conventional low-level vision problems. Moreover, most of the algorithms in the literature are non-blind and require the knowledge about PSFs. For a comprehensive comparison, besides the state-of-the-art blind UDC restoration algorithm in [1], we also included several representative blind image restoration models as baselines, especially the lightweight ones. The proposed work was compared with six algorithms, including the Dual-Encoder-UNet (DEUNet) for blind UDC restoration [1], the Feature Fusion Attention Network (FFA) [41], the Pyramid Attention Network (PANet) [42], the Information Multi-Distillation Network (IMDN) [10], the Residual Feature Distillation Network (RFDN) [11], and the Deep Atrous Guided Filter (DAGF) [43]. The proposed and comparative algorithms were trained using the training examples provided by [1]. The visual qualities of restored images were measured by three metrics, including Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS).

Table I compares the restoration performance of the proposed algorithm and competing methods. Our algorithm constantly achieves the best results on both datasets in terms of all quality metrics. We note that our algorithm outperforms DAGF, which is the second-best, by 0.23dB and 0.77dB in PSNR on the P-OLED and T-OLED datasets, respectively, and outperforms the rest methods by a large margin. It is apparent from the table that restoring the images captured by the UDC under P-OLED display panel is more challenging, where the PSNR values measured on the P-OLED dataset are at least 4dB lower than those measured on the T-OLED dataset. The reason is that the P-OLED panel has a much lower transmission rate, resulting in notable information loss in UDC images. Moreover, the substrate of the P-OLED display panel also introduces severe color distortion, so the colors in the UDC images are very monotonous (most are yellow and black, see the input image in Fig. 7).

The qualitative comparisons of restoration results are presented in Fig. 7 and Fig. 8. The testing images were shot under the P-OLED and T-OLED display panels, respectively, and representative patches are displayed in zoom-in view to demonstrate the reconstruction of fine details. The ground truths are shown in the last row and last column for comparison. We can see that the images restored by the proposed method are closer to the ground truths. As can be seen from the zoom-in views, the algorithm can faithfully recover fine details. For example, the fur of the panther in the reconstructed image shown in Fig. 8 is quite clear and sharp, and there are no aliasing or blurring artifacts.

We also evaluated the model compactness and computational complexities of the testing algorithms. Computational complexity was quantified by the number of multiply-accumulate operations (MACs) required for restoring a UDC image of resolution $256 \times 256$. The parameter amounts and MACs are shown in Table I. The proposed algorithm achieves the best restoration performance with the least memory consumption and computational load. The restoration network has 0.48M parameters. Among the competing algorithms, DAGF has the closest quality scores and MACs as the proposed one, while it has twice the parameter amount. IMDN and RFDN are another two algorithms with less than 1M parameters besides the proposed one, but their MACs are at least 14 times of ours. In Section III-B, we also described a more compact and efficient variant of the proposed model. The proposed compressing scheme can further save nearly half of the parameters and computational complexity. Detailed results will be presented in Section IV-H.

### D. Qualitative Evaluation on ZTE UDC Dataset

As mentioned earlier, the testing images in this dataset have no ground-truth images, so this dataset was adopted to

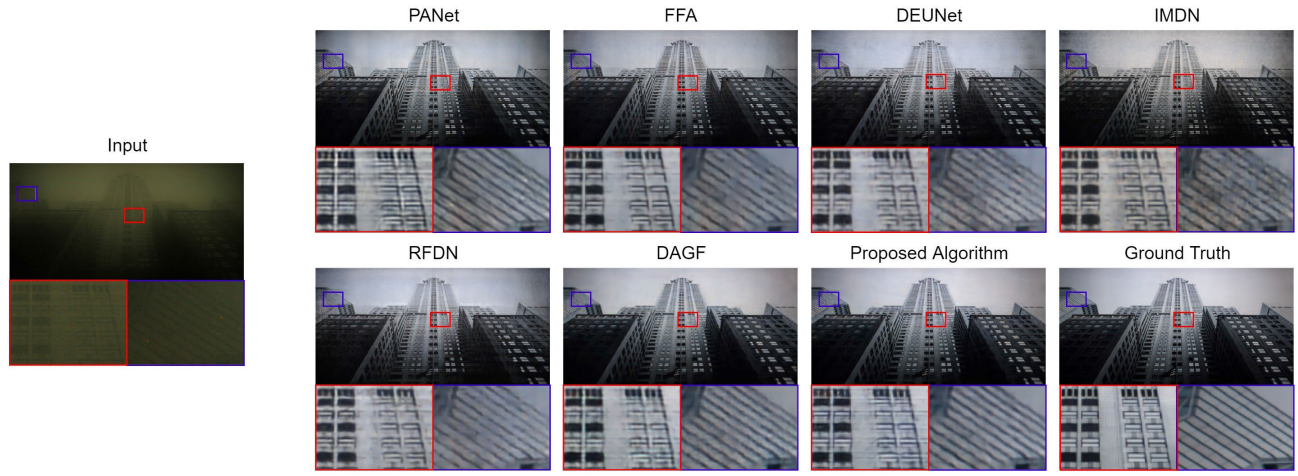



Fig. 7. Visualization of the restoration results of an image captured by a UDC below P-OLED display panel.


Fig. 8. Visualization of the restoration results of an image captured by a UDC below T-OLED display panel.

qualitatively compare restoration results. All the algorithms were trained on the synthetic UDC images provided by the dataset. The training and testing processes do not access to the PSF data. The input images and the restoration results are displayed in Fig. 9. Since no image signal processing (ISP) was applied to the raw UDC images, the restored images are post-processed using the same procedures as [14] for better visualization, including color correction and scaling, and contrast enhancement. Some testing images contain light sources that are sensitive to diffraction. The major artifacts in testing images are haze, blur, glare, and noise, so the qualities of restored images are judged by checking the naturalness of color and saturation, the suppression of glare and haze, the sharpness of fine details, etc. Compared with the raw inputs, the restored images show brighter colors and are less hazy. Take the second image for example, the characters on the bright yellow plate can be clearly distinguished from the background after restoration, and colors have been recovered from the saturated regions. The proposed algorithm does not introduce notable artifacts. In contrast, noisy patterns appear in the outputs of IMDN and RFDN (see the zoom-in images in the fourth and last rows). Slight ringing artifacts can be observed in the output of DEUNet, especially in the regions near the borders of the lamp (see the zoom-in images in the last row), while the regions in our result are more natural. However, as stated in [14], even for non-blind algorithms, it is still very challenging to correct the diffraction artifacts of light source, so remaining flares can be observed in the zoom-in views of output images.

### *E. Comparison Among Different Configurations of XSMBs*

The restoration performance, parameter amount, and computational complexity of the proposed network depend on the numbers of XSMBs in the two branches (denoted by $N_1$ and $N_2$, respectively). In this subsection, we demonstrate the variation of these factors with the configurations of XSMBs. The experiments tested the restoration networks with five and

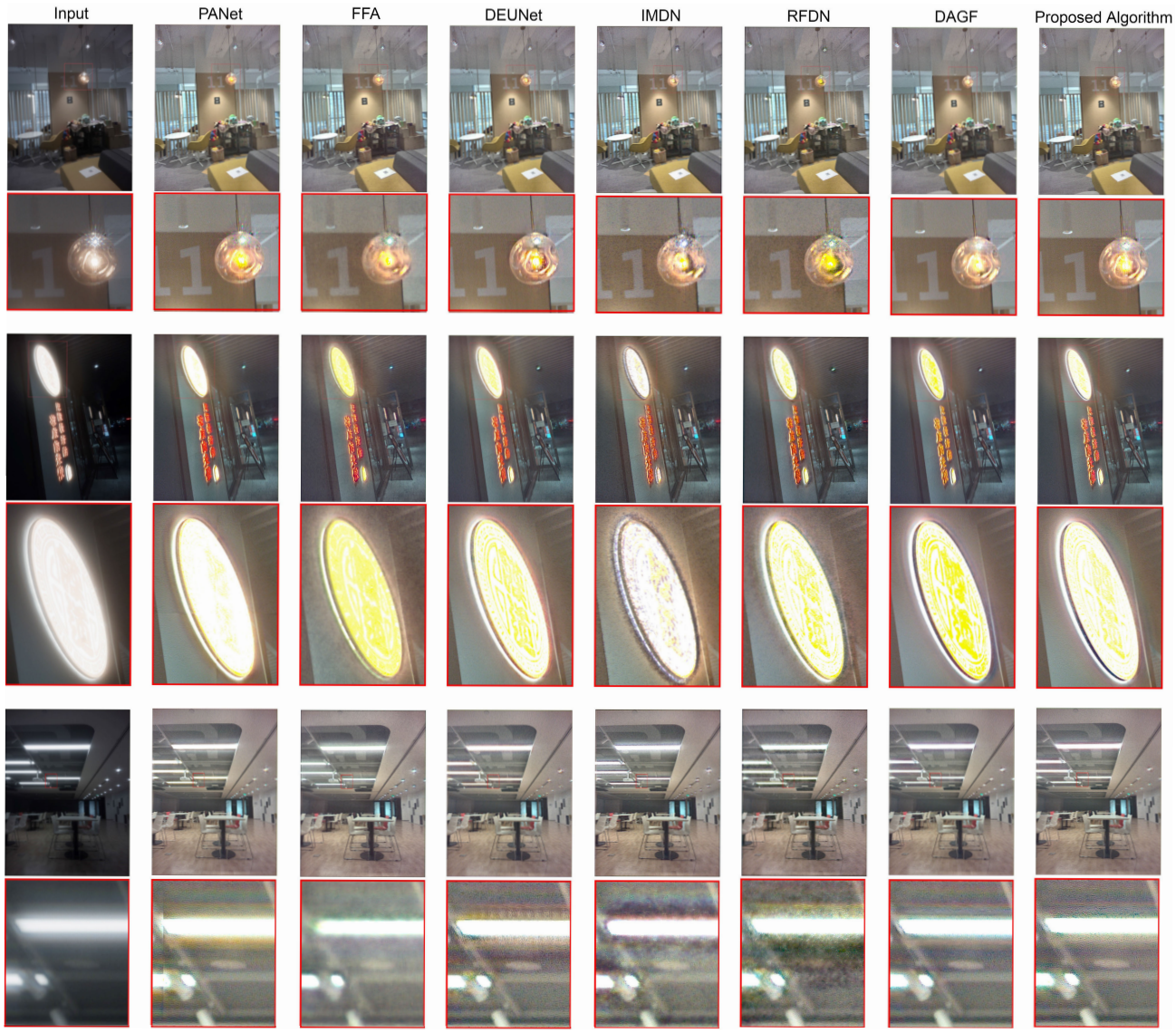



Fig. 9. Visualization of the restoration results of the images captured by a ZTE Axon 20 UDC cellphone.

TABLE II
COMPARISON AMONG DIFFERENT ARRANGEMENTS OF XSMBS IN THE TWO BRANCHES

| $N_1$ | $N_2$ | MACs (G)↓ | Params (M)↓ | P-OLED | | | T-OLED | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| 1 | 5 | 3.00 | 0.48 | 33.46 | 0.957 | 0.219 | 38.23 | 0.977 | 0.122 |
| 2 | 4 | 3.93 | 0.48 | 33.27 | 0.956 | 0.216 | 38.19 | 0.977 | 0.121 |
| 3 | 3 | 4.86 | 0.48 | 33.25 | 0.956 | 0.215 | 38.19 | 0.977 | 0.119 |
| 4 | 2 | 5.79 | 0.48 | 32.73 | 0.953 | 0.216 | 38.09 | 0.977 | 0.118 |
| 5 | 1 | 6.71 | 0.48 | 32.97 | 0.955 | 0.215 | 38.27 | 0.978 | 0.117 |
| 1 | 4 | 2.69 | 0.40 | 33.28 | 0.956 | 0.220 | 38.06 | 0.977 | 0.125 |
| 2 | 3 | 3.62 | 0.40 | 33.08 | 0.955 | 0.217 | 38.03 | 0.976 | 0.125 |
| 3 | 2 | 4.55 | 0.40 | 32.67 | 0.953 | 0.218 | 38.01 | 0.977 | 0.119 |
| 4 | 1 | 5.48 | 0.40 | 32.37 | 0.951 | 0.224 | 38.14 | 0.978 | 0.117 |

six XSMBs in total. Table II lists the results obtained using all possible combinations of $N_1$ and $N_2$.

In the case of P-OLED, the highest quality scores are achieved when the first branch has one XSMB and all the remaining ones go to the second branch. This is because the P-OLED display panel heavily distorts all frequency sub-bands. As can be seen from the image shown in Fig. 7, the coarse structures are nearly invisible in the image captured by UDC, suggesting that the low-frequency sub-band has severe distortion. Hence, allocating more blocks to the second branch helps recover the most visually significant LL sub-band, so that the visibility of the coarse structures of the restored image can be guaranteed. Another trend observed from Table II is that MACs increase with $N_1$, since the sub-bands in the first-level

TABLE III
ROBUSTNESS OF RESTORATION PERFORMANCE AGAINST NOISE

| $\sigma^2$ | P-OLED | | T-OLED | |
|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ |
| 0 | 33.46 | 0.957 | 38.23 | 0.977 |
| 1 | 33.20 | 0.954 | 37.91 | 0.975 |
| 2 | 32.71 | 0.946 | 37.42 | 0.972 |

TABLE IV
RESULTS OF ABLATION EXPERIMENTS

| Networks | P-OLED | | T-OLED | |
|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ |
| w/o dilated convolution | 29.08 | 0.934 | 37.73 | 0.976 |
| w/o cross-scale modulation | 32.95 | 0.953 | 37.85 | 0.975 |
| w/o DWT and IDWT | 33.07 | 0.951 | 37.62 | 0.973 |
| Original network | 33.46 | 0.957 | 38.23 | 0.977 |

DWT have larger resolutions than those in the second one. Hence, a small $N_1$ benefits both the quality of restored image and the efficiency of inference.

When it comes to T-OLED, the optimal configuration is quite different from the one for P-OLED, which highlights the distinction between the degradation patterns associated with the two types of display panels. For both $N_1 + N_2 = 5$ and $N_1 + N_2 = 6$, the best quality scores are observed when $N_2 = 1$, indicating that the restoration of high-frequency has higher priority. From Fig. 8, the artifacts in the images captured through the T-OLED screen are mainly blurring and some periodic strips, while the low-frequency components are less affected. As a result, more emphasis should be placed on restoring the high-frequency components. However, as discussed above, larger $N_1$ leads to higher computational loads. It should be noted that unlike the case in P-OLED, different settings of $N_1$ and $N_2$ do not show remarkable differences in terms of quality scores. When there are five XSMBs, the maximum gap between the PSNRs corresponding to different settings is only 0.13dB. Hence, considering computational complexity, small values of $N_1$ are also feasible choices in this case.

### F. Robustness of UDC Image Restoration

We also evaluated the robustness of the proposed algorithm. Gaussian noise (with zero mean and variance $\sigma^2$) was added to the testing images in the T-OLED and P-OLED datasets, and performance of the pre-trained restoration network was measured under the noisy setting. This is to simulate the scenario where the target UDC introduces more noise than the one for capturing training images. Table III shows the variation of quality scores after noise addition. The restoration performance is stable in the presence of noise. When the variance of Gaussian noise increases to two from one, the average PSNR of original UDC images decreases by approximately 3 dB. In comparison, the decrease in the average PSNR of restored images is around 0.5 dB.

### G. Ablation Studies

*1) Effect of Expanding Receptive Fields:* The measurements in [44] show that the PSFs of OLED display panels, especially those of the P-OLED one, have large spatial supports. This requires the restoration network to have sufficiently large receptive fields to capture the spatial dependencies brought by light diffraction. To this end, the proposed work integrates parallel dilated convolutions into XSMBs. To verify the effectiveness of this design, we conducted ablation experiments by replacing the dilated convolutions with the regular ones (i.e., using a dilation rate of one). As shown in Table IV, the ablated network shows PSNR values that are 4.38dB and 0.5dB lower than those of the original network on the P-OLED and T-OLED datasets, respectively.

*2) Effect of Cross-Scale Modulation:* Another ablated network was constructed to examine the effect of cross-scale modulation, where the outputs of the parallel pipelines in each XSMB are directly concatenated without being mutually modulated. The results of ablation experiments imply that the dependencies in the scale space can be harnessed to boost the performance of UDC image restoration. The average gap between the PSNR values achieved by the networks with and without the cross-scale modulation is 0.45dB. The decline of restoration performance after discarding XSMBs indicates that simply combining the outputs of multiple dilated convolutional layers is not able to fully exploit the information provided by multi-scale features. The cross-scale modulation scheme compliments the feature learned at a specific scale with those from neighboring scales, so it overcomes the limitation that the output of a single convolutional layer can only represent the information at one scale. From Fig. 3, the cross-scale modulation also diversifies the paths of data flow between the input and output of an XSMB, which increases the expressive power of the restoration network.

*3) Effects of DWT and IDWT:* As discussed in Section III-A, DWT offers several benefits to UDC image restoration. Ablation experiments were also conducted to demonstrate the advantages of wavelet-domain restoration over the one in spatial domain. We retrained the restoration network by replacing the double-level DWT and IDWT operations with $3 \times 3$ convolution, down-sampling, and up-sampling operations, and the results of ablation experiments are listed in the third row of Table IV. The ablated network exhibits inferior performance on both datasets. Meanwhile, using convolutional layers instead of DWT and IDWT introduces additional 648 parameters.

### H. Performance of Lightweight Restoration Model

In this sub-section, we demonstrate the performance of the lightweight architecture described in Section III-B and the capability of the distillation algorithm in balancing restoration performance, model compactness, and computational loads. The lightweight restoration model was trained in a teacher-student mode by taking the knowledge distilled from the full-size model tested in Section III-A as the supervision signals, and the training process requires no ground-truth clear-scene targets. The weight $\lambda$ in (9) was set to 160 and 500 for the T-OLED and P-OLED datasets, respectively.

The first two rows of Table V compare the quantitative scores, complexity, and parameter amounts of the full-size and

TABLE V
PERFORMANCE OF THE LIGHTWEIGHT NETWORKS TRAINED WITH DIFFERENT STRATEGIES

| Networks | MACs (G)↓ | Params (M)↓ | P-OLED | | | T-OLED | | |
|---|---|---|---|---|---|---|---|---|
| | | | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| Full-size | 3.00 | 0.48 | 33.46 | 0.957 | 0.219 | 38.23 | 0.977 | 0.122 |
| Lightweight (w/o using GTs for training) | 1.69 | 0.26 | 32.53 | 0.950 | 0.251 | 37.57 | 0.974 | 0.139 |
| Lightweight (use GTs for training) | 1.69 | 0.26 | 32.35 | 0.950 | 0.248 | 37.61 | 0.975 | 0.136 |

lightweight networks. The compression scheme reduces nearly half of the network parameters and computational complexity. The lightweight network has less than 0.3M parameters, which is within the capacity of most mobile devices. Compared with the competing algorithms listed in Table I, the lightweight network has only $\frac{1}{2} \sim \frac{1}{170}$ parameters and $\frac{1}{2} \sim \frac{1}{34}$ MACs. Despite the reduction of resource consumption, the performance of the lightweight variant still ranks first and second on the T-OLED and P-OLED datasets, respectively. We find that the restoration task for P-OLED UDC is more susceptible to model compression. As shown earlier, the images captured by the camera under the P-OLED display panel have much lower visibility. In contrast, the T-OLED panel mainly affects fine details. Accordingly, the model for P-OLED UDC image restoration needs to recover a larger amount of the information about the original scene, so the impact of model compression becomes more remarkable.

To demonstrate the effect of the distillation algorithm, we re-trained the lightweight restoration network using paired UDC images and the corresponding ground-truth images without using the pre-trained teacher network. The performance of two networks trained using ground truths and the pre-trained teacher network, respectively, are compared in the second and third rows of Table V. The results suggest that the knowledge distilled from the pre-trained network has almost identical effects as the ground truths in training the lightweight network. For the T-OLED dataset, the gap between the PSNRs of the two networks is only 0.04dB. It is interesting to note that on the P-OLED dataset, the network trained under the guidance of the pre-trained teacher network even outperforms the one trained using ground truths (by 0.18dB in PSNR). The reason is that the distortions introduced by the P-OLED display panel spread over the whole spectrum of the wavelet domain and vary across sub-bands. The proposed distillation algorithm separately exploits knowledge from low-frequency and high-frequency sub-bands. We conjecture that the wavelet-domain distillation strategy makes it easier for the lightweight network to model the heterogeneous sources of distortion. Another advantage of the distillation algorithm is that it can reduce the cost of collecting the pixel-wise aligned distortion-free counterparts of UDC images for training. This offers the flexibility of accomodating a pre-trained restoration model to the UDC devices with different hardware configurations.

### *I. Runtime on Mobile Device*

Since UDC image restoration needs to be conducted on mobile devices, we experimented with the mobile-end deployment of restoration algorithms. The proposed work, including the full-size and lightweight networks, and the competing algorithms except PANet were all deployed on a Razer-Phone2 cellphone. The memory consumption of PANet exceeds the capability of the cellphone. The cellphone is equipped with Qualcomm Snapdragon 845 processor and has 8GB RAM.

TABLE VI
COMPARISON OF RUNTIME ON CELLPHONE (SECOND)

| FFA | DEUNet | IMDN | RFDN | DAGF | Full | Lightweight |
|---|---|---|---|---|---|---|
| 68.03 | 1.97 | 8.38 | 4.46 | 23.03 | 0.53 | 0.36 |

Table VI lists the comparison of the runtime measured on the cellphone, where 'Full' and 'Lightweight' denote the two variants of our work. The proposed algorithm ranks first in speed and is the only algorithm whose runtime is within one second. DEUNet is the second most efficient algorithm. For the testing images with $512 \times 512$ pixels, the speeds of the full-size and lightweight networks (without GPU acceleration) are 0.53 sec/image and 0.36 sec/image, respectively. In general, the computational budget is affordable for mobile devices. The latest Qualcomm Snapdragon 8 Gen2 platform for mobile devices can perform over 3800G floating-point operations per second. The full-size network requires about 12.0 G MACs to accomplish the above image restoration task, which accounts for a small portion of the computational resource. It is worth noting that runtime was measured without using GPU or software libraries for acceleration. Tools such as Qualcomm Neural Processing SDK can speed up the execution of deep neural networks on mobile platforms, taking advantage of which can enhance the efficiency of on-device image restoration. Besides saving multiply-accumulate operations, increasing the degree of parallelism can also speed up the inference. However, hardware-level performance optimization is beyond the scope of this study.

We also measured the CPU runtime of our algorithm on a PC with Intel i9-9900K CPU and 64GB RAM, and GPU acceleration was disabled to match the experimental setting on the mobile device. The full-size network takes 0.17s to restore an input image with $512 \times 512$ pixels, and the lightweight one takes 0.12s. A single NVIDIA 2080Ti GPU can speed up the inference speed by around ten times.

## V. CONCLUSION

We have proposed a deep learning-based algorithm to enhance the imaging quality of UDC devices. The neural network tackles the diverse distortions caused by OLED display panels at multiple scales in the wavelet domain, and the residual between the wavelet representations of the raw UDC image and its clear-scene state is inferred via cross-scale modulation. To cater to the mobile devices with constrained memory space, the proposed work further reduces the architectural redundancy

of the restoration network. An adaptive distillation algorithm has been developed to exploit knowledge from a pre-trained full model to guide the training of its lightweight counterpart. The performance comparison confirms the efficacy of the proposed algorithm. It demonstrates higher quality scores than state-of-the-arts with much lower consumption of memory and computational resources.

As an exploratory study, the proposed work still has several limitations, some of which also exist in prior work on UDC image restoration. Firstly, the on-device inference speed has space to improve. As mentioned above, the restoration model has not been fully optimized toward the hardware architectures of mobile devices. Secondly, the thorough removal of severe UDC artifacts remains challenging. For example, the strong flares near light sources cannot be completely removed (see Fig. 9). Finally, the training algorithm requires paired data. It would be more practical if restoration models could learn from partially paired UDC images.

We believe the following topics in UDC image restoration are worth further investigation: 1) algorithms with extremely low latency time, 2) semi-supervised training for UDC image restoration models, and 3) performance optimization for downstream applications on UDC devices (e.g., face verification).

## REFERENCES


[1] Y. Zhou, D. Ren, N. Emerton, S. Lim, and T. Large, "Image restoration for under-display camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9175–9184.

[2] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3291–3300.

[3] M. Lamba and K. Mitra, "Restoring extremely dark images in real time," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3486–3496.

[4] J. Li, X. Feng, and Z. Hua, "Low-light image enhancement via progressive-recursive network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4227–4240, Nov. 2021.

[5] S. Brehm, S. Scherer, and R. Lienhart, "High-resolution dual-stage multi-level feature aggregation for single image and video deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1872–1881.

[6] S. Wan et al., "Deep convolutional-neural-network-based channel attention for single image dynamic scene blind deblurring," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 2994–3009, Aug. 2021.

[7] Y. Kim, J. W. Soh, and N. I. Cho, "Adaptively tuning a convolutional neural network by gate process for image denoising," *IEEE Access*, vol. 7, pp. 63447–63456, 2019.

[8] B. Jiang, Y. Lu, J. Wang, G. Lu, and D. Zhang, "Deep image denoising with adaptive priors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5124–5136, Aug. 2022.

[9] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 154–169.

[10] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2024–2032.

[11] J. Liu, J. Tang, and G. Wu, "Residual feature distillation network for lightweight image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 41–55.

[12] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.

[13] Y. Li, Q. Yan, K. Zhang, and H. Xu, "Image reflection removal via contextual feature fusion pyramid and task-driven regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 553–565, Feb. 2022.

[14] R. Feng, C. Li, H. Chen, S. Li, C. C. Loy, and J. Gu, "Removing diffraction image artifacts in under-display camera via dynamic skip connection network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 662–671.

[15] H. Kwon, C. Yang, M. Kim, C. Kim, J. Ahn, and P. Kim, "Simulation of blur in transmitted image through transparent plastic for transparent OLEDs," *J. Display Technol.*, vol. 12, no. 8, pp. 851–858, Aug. 2016.

[16] Q. Tang, H. Jiang, X. Mei, S. Hou, G. Liu, and Z. Li, "Study of the image blur through FFS LCD panel caused by diffraction for camera under panel," in *SID Symp. Dig. Tech. Papers*, 2020, vol. 51, no. 1, pp. 406–409.

[17] N. Emerton, D. Ren, and T. Large, "Image capture through TFT arrays," in *SID Symp. Dig. Tech. Papers*, 2020, vol. 51, no. 1, pp. 402–405.

[18] Z. Qin, Y. Tsai, Y. Yeh, Y. Huang, and H. D. Shieh, "See-through image blurring of transparent organic light-emitting diodes display: Calculation method based on diffraction and analysis of pixel structures," *J. Display Technol.*, vol. 12, no. 11, pp. 1242–1249, Nov. 2016.

[19] Z. Qin, J. Xie, F. Lin, Y. Huang, and H. D. Shieh, "Evaluation of a transparent display's pixel structure regarding subjective quality of diffracted see-through images," *IEEE Photon. J.*, vol. 9, no. 4, pp. 1–14, Aug. 2017.

[20] A. Yang and A. C. Sankaranarayanan, "Designing display pixel layouts for under-panel cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2245–2256, Jul. 2021.

[21] B. Fu, M. Ye, R. Yang, and C. Zhang, "See-through image enhancement through sensor fusion," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 687–692.

[22] J. Luo, W. Ren, T. Wang, C. Li, and X. Cao, "Under-display camera image enhancement via cascaded curve estimation," *IEEE Trans. Image Process.*, vol. 31, pp. 4856–4868, 2022.

[23] K. Kwon et al., "Controllable image restoration for under-display camera in smartphones," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2073–2082.

[24] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.

[27] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[28] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[29] Z. Hou and S. Kung, "Efficient image super resolution via channel discriminative deep neural network pruning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3647–3651.

[30] Y. Ma, H. Xiong, Z. Hu, and L. Ma, "Efficient super resolution using binarized neural network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 694–703.

[31] S. Cho, S. Kim, S. Jung, and S. Ko, "Blur-robust object detection using feature-level deblurring via self-guided knowledge distillation," *IEEE Access*, vol. 10, pp. 79491–79501, 2022.

[32] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 723–731.

[33] A. Lahiri, S. Bairagya, S. Bera, S. Haldar, and P. K. Biswas, "Lightweight modules for efficient deep learning based image restoration," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1395–1410, Apr. 2021.

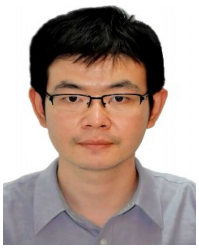
[34] Y. Nie, K. Han, Z. Liu, C. Liu, and Y. Wang, "GhostSR: Learning ghost features for efficient image super-resolution," 2021, *arXiv:2101.08525*.

[35] H. Ullah et al., "Light-DehazeNet: A novel lightweight CNN architecture for single image dehazing," *IEEE Trans. Image Process.*, vol. 30, pp. 8968–8982, 2021.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[37] M. Zontak and M. Irani, "Internal statistics of a single natural image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 977–984.

[38] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1577–1586.
[39] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[40] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.
[41] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "FFA-Net: Feature fusion attention network for single image dehazing," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11908–11915.
[42] Y. Mei et al., "Pyramid attention networks for image restoration," 2020, *arXiv:2004.13824*.
[43] V. Sundar, S. Hegde, D. Kothandaraman, and K. Mitra, "Deep atrous guided filter for image restoration in under display cameras," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 379–397.
[44] S. Lim, Y. Zhou, N. Emerton, T. Large, and S. Bathiche, "Image restoration for display-integrated camera," in *SID Symp. Dig. Tech. Papers*, 2020, vol. 51, no. 1, pp. 1102–1105.



**Yuenan Li** (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees in measurement technology and instrumentation and the Ph.D. degree in information and communication engineering from the Harbin Institute of Technology, China, in 2004, 2006, and 2010, respectively. From 2019 to 2020, he was a Visiting Researcher with the University of Maryland, College Park, MD, USA. He is currently an Associate Professor with the School of Electrical and Information Engineering, Tianjin University, China. His research interests include multimedia signal processing, information security and forensics, computer vision, and physiological signal processing.

**Jin Wu** (Student Member, IEEE) received the B.S. degree in electronics and information engineering from Tianjin University, China, in 2020, where he is currently pursuing the M.Eng. degree. His research interests include image processing and computer vision.

**Zetao Shi** received the B.S. degree in measurement technology and instrumentation from the Taiyuan University of Technology, China, in 2021. He is currently pursuing the M.Eng. degree with Tianjin University. His research interests include deep learning and image restoration.