Image Restoration for Under-Display Camera

Yuqian Zhou[1], David Ren[2], Neil Emerton[3], Sehoon Lim[3], Timothy Large[3]
[1]IFP, UIUC, [2]CIL, UC Berkeley, [3]Microsoft

Abstract

The new trend of full-screen devices encourages us to position a camera behind a screen. Removing the bezel and centralizing the camera under the screen brings larger display-to-body ratio and enhances eye contact in video chat, but also causes image degradation. In this paper, we focus on a newly-defined Under-Display Camera (UDC), as a novel real-world single image restoration problem. First, we take a 4k Transparent OLED (T-OLED) and a phone Pentile OLED (P-OLED) and analyze their optical systems to understand the degradation. Second, we design a Monitor-Camera Imaging System (MCIS) for easier real pair data acquisition, and a model-based data synthesizing pipeline to generate Point Spread Function (PSF) and UDC data only from display pattern and camera measurements. Finally, we resolve the complicated degradation using deconvolution-based pipeline and learning-based methods. Our model demonstrates a real-time high-quality restoration. The presented methods and results reveal the promising research values and directions of UDC.

1. Introduction

Under-display Camera (UDC) is a new imaging system that mounts display screen on top of a traditional digital camera lens, as shown in Figure 1. Such a system has mainly two advantages. First, it brings a new product trend of full-screen devices [11] with larger screen-to-body ratio, which can provide better user perceptive and intelligent experience [12]. Without seeing the bezel and extra buttons, users can easily access more functions by directly touching the screen. Second, it provides better human computer interaction. By putting the camera in the center of the display, it enhances teleconferencing experiences with perfect gaze tracking, and it is increasingly relevant for larger display devices such as laptops and TVs.

Unlike pressure or fingerprint sensors that can be easily integrated into a display, it is relatively difficult to retain full functionality of an imaging sensor after mounting it behind a display. The imaging quality of a camera will be severely degraded due to lower light transmission rate and diffraction effects. As a result, images captured will be noisy and blurry. Therefore, while bringing better user experience and interaction, UDC may sacrifice the quality of photography, face processing [35] and other downstream vision tasks. Restoring and enhancing the images captured by UDC system will be desired.

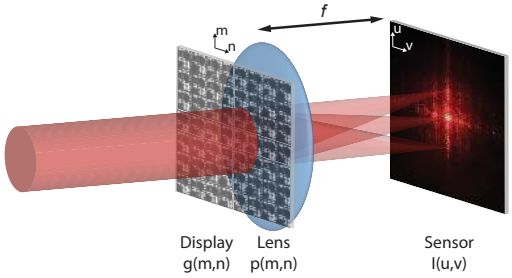


Figure 1: The newly proposed imaging system named Under-Display Camera (UDC). We mount display screen on top of a traditional digital camera lens. The design brings new trend of full-screen devices.

Traditional image restoration approaches form the task as an inverse problem or an optimization problem like Maximum-a-Posterior (MAP). For the UDC problem, for practical purposes, the proposed image restoration algorithm and system are expected to work in real-time. Therefore, deconvolutional-based methods like Wiener Filter [14] should be preferred. Deconvolution is the inverse process of convolution and recovers the original signal from the point-spread-function (PSF)-convolved image. The fidelity of the deconvolution process is dependent on the space-invariance of the PSF over the image field of-view (FOV) and on a low condition number for the inverse of the PSF [19]. For strongly non-delta-function-like PSFs such as those encountered when imaging through a display, the value of condition number can be large. For such PSFs an additional denoising step may be essential.

Another option is the emerging discriminative learning-based image restoration model. Data-driven discriminative learning-based image restoration models usually outperform traditional methods in specific tasks like image denoising [3, 26, 42, 43, 47, 48], de-bluring [21, 28],de-raining [39, 40], de-hazing [13, 33], super-resolution [22, 37], and

light-enhancement [9]. However, working on synthesis data with single degradation type, existing models can be hardly utilized to enhance real-world low-quality images with complicated or combined degradation types. To address complicated real degradation like the UDC problem, directly collecting real paired data or synthesizing near-realistic data after fully understanding the degradation model is necessary.

In this paper, we present the first study to define and analyze the novel Under-Display Camera (UDC) system from both optics and image restoration viewpoints. For optics, we parse the optical system of the UDC pipeline and analyze the characteristics of light transmission. Then we relate the obtained intuitions and measurements to an image restoration pipeline, and propose two ways of resolving the single-image restoration: A deconvolution-based Wiener Filter [29] pipeline (DeP) and a data-driven learning-based approach. Specifically, we regard UDC restoration as a combination of tasks such as low-light enhancement, deblurring, and de-noising.

Without loss of generality, our analysis focuses on **two types of displays**, a 4K Transparent Organic Light-Emitting Diode (T-OLED) and a phone Pentile OLED (P-OLED), and **a single camera type,** a 2K FLIR RGB Point Grey research camera. To obtain the real imaging data and measure the optical factors of the system, we also propose a data acquisition system using the above optical elements.

In summary, the main contributions of our paper are: (1) A brand new imaging system named Under-Display Camera (UDC) is defined, measured and analyzed. Extensive experiments reveal the image degradation process of the system, inspiring better approaches for restoring the captured images. (2) As a baseline, two practical and potential solutions are proposed, including conventional Wiener Filter and the recent learning-based method. (3) Adopting the newly-assembled image acquisition system, we collect the first Under-Display Camera (UDC) dataset which will be released and evaluated by the public.

2. Related Work

Real-world Image Reconstruction and Restoration Image restoration for UDC [23, 24, 46, 49] can be categorized into the problem of Real-world restoration [3, 45]. It is becoming a new concept in low-level vision. In the past decades, low-level vision works on synthetic data (denoising on AWGN and SR on Bicubic), but the models are not efficient for images with real degradation such as real noises or blur kernels. Making models perform better on real-world inputs usually requires new problem analysis and a more challenging data collection. Recently, researchers also worked on challenging cases like lensless imaging problems [20, 27, 30], or integrating optics theory with High Dynamic Range imaging [34]. Previously, there has been two common ways to prepare adaptive training data for real-world problems: real data collection and near-realistic data synthesis.

Recently, more real noise datasets such as DND [31], SIDD [2, 28], and RENOIR [5], have been introduced to address practical denoising problems. Abdelrahman et al. [3] proposed to estimate ground truth from captured smartphone noise images, and utilized the paired data to train and evaluate the real denoising algorithms. In addition to noise, Chen et al. first introduced the SID dataset [9] to resolve extreme low-light imaging. In the area of Single Image Super Resolution (SISR), researchers considered collecting optical zoom data [10, 45] to learn better computational zoom. Other restoration problems including reflection removal [32, 36] also follow the trend of real data acquisition. Collecting real data suffers from limitation of scene variety since most previous models acquire images of postcards, static objects or color boards. In this paper, we propose a novel monitor-camera imaging system, to add real degradation to the existing natural image datasets like DIV2K [4].

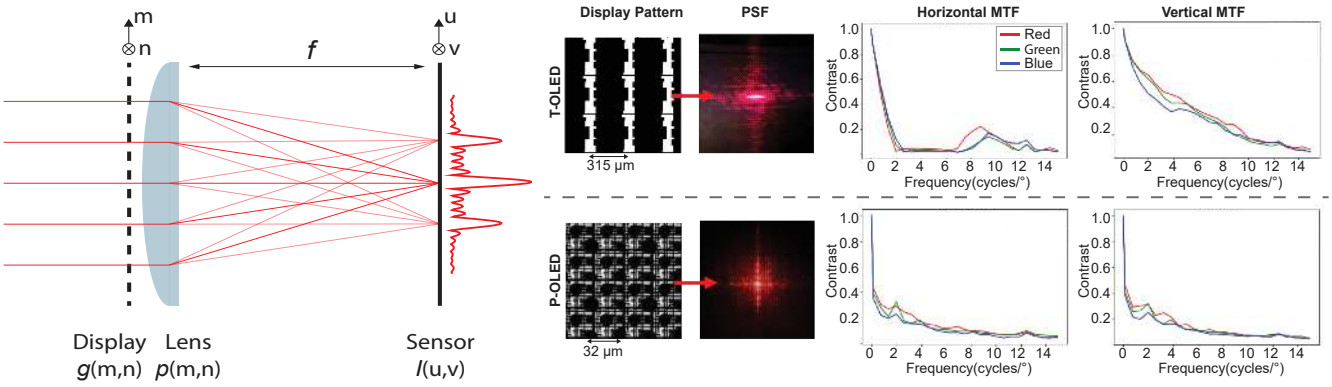
A realistic dataset can be synthesized if the degradation model is fully understood and resolved. One good practice of data synthesis is generating real noises on raw sensors or RGB images. CBDNet [17] and Tim et al. [8] synthesized realistic noise by unfolding the in-camera pipeline, and Abdelhamed et al. [1] better fitted the real noise distribution with flow-based generative models. Zhou et al. [48] adapted the AWGN-RVIN noise into real RGB noise by analyzing the demosacing process. Other physics-based synthesis was also explored in blur [7] or hazing [6]. For the UDC problem in this paper, we either collected real paired data, or synthesized near-realistic data from model simulation. In particular, we applied the theory of Fourier optics to simulate the diffraction effects, and further adjusted the data with other camera measurements. Our data synthesizing pipeline demonstrates a promising performance for addressing real complicated degradation problem.

3. Formulation

In this section, we discuss the optical system and image formation process of the proposed UDC imaging system. We analyze the degradation type, light transmission rate and visualize the Point Spread Function (PSF). Moreover, we formulate the image formation pipeline to compute simulated PSF from measurements.

3.1. Optical System Analysis

Optical Elements. In our experiments, we focus on the Organic Light-Emitting Diode (OLED) displays [38] as they have superior optical properties compared to the traditional LCDs (Liquid Crystal Display). Due to confidentiality reasons it is often difficult to obtain the sample materials used for demos from commercial companies. In this



(a) (b)

Figure 2: Image formation pipeline of under-display camera (UDC) problem. (a) Image Formation Pipeline. (b)Optics characteristics of UDC. The structure of the 4K T-OLED has a grating-like pixel layout. P-OLED differs from T-OLED in sub-pixel design. From left to right: Micrography of display patterns, PSFs (red light only) and MTFs (red, green, and blue).

Table 1: Comparison of two displays in terms of light transmission rate and physical pixel layout and open areas.

Metrics	T-OLED	P-OLED
Pixel Layout Type	Stripe	Pentile
Open Area	21%	23%
Transmission Rate	20%	2.9%
Major Degradation	Blur, Noise	Low-light, Color Shift, Noise

case, we select the displays with different transparencies to improve the generalization. Note that all the displays are **non-active** in our experiments, since in real scenario, the display can be turned off locally by setting black pixels on local regions of the OLED display when the camera is in operation to 1) reduce unnecessary difficulty from display contents while not affecting user experience and 2) provide users with the status of the device and thus ensure privacy.

Owing to transparent materials being used in OLED display panels, visible lights can be better transmitted through the OLEDs than LCDs. In the meantime, pixels are also arranged such that open area is maximized. In particular, we focus on 4k Transparent OLED (T-OLED) and a phone Pentile OLED (P-OLED). Figure 2 is a micrograph illustration of the pixel layout in the two types of OLED displays. The structure of the 4K T-OLED has a grating-like pixel layout. P-OLED differs from T-OLED in sub-pixel design. It follows the basic structure of RGBG matrix.

Light Transmission Rate. We measure the transmission efficiency of the OLEDs by using a spectrophotometer and white light source. Table 1 compares the light transmission rate of the two displays. For T-OLED, the open area occupies about 21%, and the light transmission rate is around 20%. For P-OLED, although the open area can be as large as 23%, the light transmission rate is only 2.9%.

The loss of photons can be attributed mainly to the structure of P-OLED. First, the P-OLED has a finer pixel pitch, so photos are scattered to higher angles comparing to the T-OLED. As a result, high angle photons are not collected by the lens. Second, P-OLED is a flexible/bendable display, which has a poly-amide substrate on which the OLED is formed. Such a substrate has relatively low transmission efficiency, causing photons to be absorbed. The absorbed light with certain wavelengths may make the images captured through a polyamide-containing display panel by a UDC appear yellow. As a result, imaging through a P-OLED results in lower signal-to-noise ratio (SNR) comparing to using a T-OLED, and has a color shift issue. One real imaging example is shown in Figure 4.

Diffraction Pattern and Point Spread Function (PSF). Light diffracts as it propagates through obstacles with sizes that are similar to its wavelength. Unfortunately, the size of the openings in the pixel layout is on the order of wavelength of visible light, and images formed will be degraded due to diffraction.

Here we characterize our system by measuring the point spread function (PSF). We do so by pointing a collimated red laser beam ($\lambda = 650$nm) at the display panel and recording the image formed on the sensor, as demonstrated in Figure 1 and 2. An ideal PSF shall resemble a delta function, which then forms a perfect image of the scene. However, light greatly spreads out in UDC. For T-OLED, light spreads mostly across the horizontal direction due to its nearly one dimensional structure in the pixel layout, while for P-OLED, light is more equally distributed as the pixel layout is complex. Therefore, images captured by UDC are either blurry (T-OLED) or hazy (P-OLED).

Modulation Transfer Function (MTF) Modulation Transfer Function (MTF) is another important metric for an imaging system, as it considers the effect of finite lens aperture, lens performance, finite pixel size, noise, nonlinearities, quantization (spatial and bit depth), and diffraction in our systems. We characterize the MTF of our systems by recording sinusoidal patterns with increasing frequency in both lateral dimensions, and we report them in

Figure 2. For T-OLED, contrasts along the horizontal direction are mostly lost in the mid-band frequency due to diffraction. This phenomenon is due to the nearly one-dimensional pixel layout of the T-OLED. Figure 4 shows severe smearing horizontally when putting T-OLED in front of the camera. While for P-OLED, the MTF is almost identical to that of display-free camera, except with severe contrast loss. Fortunately, however, nulls have not been observed in any particular frequencies.

3.2. Image Formation Pipeline

In this section, we derive the image formation process of UDC based on the analysis in the previous sections. Given a calibrated pixel layout and measurements using a specific camera, degraded images can be simulated from a scene. From the forward model, we can compute the ideal PSF and consequently synthesize datasets from ground truth images.

Given an object in the scene $\mathbf{x}$, the degraded observation $\mathbf{y}$ can be modeled by a convolution process,

$$\mathbf{y} = (\gamma \mathbf{x}) \otimes \mathbf{k} + \mathbf{n}, \tag{1}$$

where γ is the intensity scaling factor under the current gain setting and display type, $\mathbf{k}$ is the PSF, and $\mathbf{n}$ is the zero-mean signal-dependent noise. Notice that this is a simple noise model that approximately resembles the combination of shot noise and readout noise of the camera sensor, and it will be discussed in a later section.

Intensity Scaling Factor (γ) The intensity scaling factor measures the changing ratio of the average pixel values after covering the camera with a display. It simultaneously relates to the physical light transmission rate of the display, as well as the digital gain δ setting of the camera. γ can be computed from the ratio of δ-gain amplified average intensity values $I_d(\delta, s)$ at position s captured by UDC, to the 0-gain average intensity values $I_{nd}(0, s)$ by naked camera within an enclosed region S. It is represented by,

$$\gamma = \frac{\int_S I_d(\delta, s) ds}{\int_S I_{nd}(0, s) ds} \tag{2}$$

Diffraction Model We approximate the blur kernel $\mathbf{k}$, which is the Point Spread Function (PSF) of the UDC. As shown in Figure 1, in our model, we assume the display panel is at the principle plane of the lens. We also assume the input light is monochromatic plane wave with wavelength λ (i.e. perfectly coherent), or equivalently light from a distance object with unit amplitude. Let the display pattern represented by transparency with complex amplitude transmittance be $g(m, n)$ at the Cartesian co-ordinate (m, n), and let the camera aperture/pupil function $p(m, n)$ be 1 if (m, n) lies inside the lens aperture region and 0 otherwise, then the display pattern inside the aperture range $g_p(m, n)$ becomes,

$$g_p(m, n) = g(m, n)p(m, n). \tag{3}$$

At the focal plane of the lens (i.e. 1 focal length away from the principle plane), the image measured is the intensity distribution of the complex field, which is proportional to the Fourier transform of the electric field at the principle plane [16]:

$$I(u, v) \propto \left| \iint_{-\infty}^{\infty} g_p(m, n) \exp\left[-j\frac{2\pi}{\lambda f}(mu + nv)\right] \mathrm{d}m\mathrm{d}n \right|^2. \tag{4}$$

Suppose $G_p(v_m, v_n) = \mathscr{F}(g_p(m, n))$, where $\mathscr{F}(\cdot)$ is the Fourier transform operator, then

$$I(u, v) \propto |G_p(v_m, v_n)|^2 = \left| G_p(\frac{u}{\lambda f}, \frac{v}{\lambda f}) \right|^2, \tag{5}$$

which performs proper scaling on the Fourier transform of the display pattern on the focal plane.

Therefore, to compute the PSF $\mathbf{k}$ for image $\mathbf{x}$, we start from computing Discrete Fourier Transform (DFT) with squared magnitude $M(a, b) = |\hat{G_p}(a, b)|^2$ of the $N \times N$ microscope transmission images $\hat{g_p}$ of the display pattern and re-scaling it. Then, the spatial down-sampling factor r (denoted by $\downarrow r$) becomes,

$$r = \frac{1}{\lambda f} \cdot \delta_N N \cdot \rho, \tag{6}$$

where δ_N is the pixel size of the $\hat{g_p}$ images, and ρ is the pixel size of the sensor. Finally, $\mathbf{k}$ can be represented as

$$k(i, j) = \frac{M_{\downarrow r}(i, j)}{\sum_{(\hat{i},\hat{j})} M_{\downarrow r}(\hat{i}, \hat{j})}. \tag{7}$$

k is a normalized form since we want to guarantee that it represents the density distribution of the intensity with diffraction effect. Note that only PSF for a single wavelength is computed for simplicity. However, scenes in the real-world are by no means monochromatic. Therefore, in order to calculate an accurate color image from such UDC systems, PSF for multiple wavelengths shall be computed. More details will be shown in Section 4.2.

Adding Noises We follow the commonly used shot-read noise model [8, 18, 25] to represent the real noise on the imaging sensor. Given the dark and blur signal $w = (\gamma \mathbf{x}) \otimes \mathbf{k}$, the shot and readout noise can be modeled by a heteroscedastic Gaussian,

$$\mathbf{n} \sim \mathcal{N}(\mu = 0, \sigma^2 = \lambda_{read} + \lambda_{shot} w), \tag{8}$$

where the variance σ is signal-dependent, and λ_{read} , λ_{shot} are determined by camera sensor and gain values.

4. Data Acquisition and Synthesis

We propose an image acquisition system called Monitor-Camera Imaging System (MCIS). In particular, we display

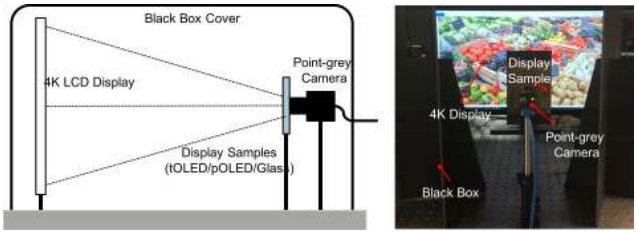
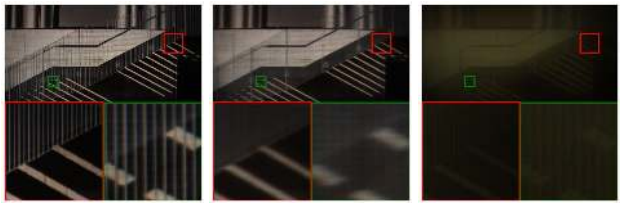


Figure 3: Monitor-Camera Imaging System (MCIS). MCIS consists of a 4K LCD monitor, the 2K FLIR RGB Point-Grey research camera, and a panel that is either T-OLED, P-OLED or Glass(i.e. no display). The camera is mounted on the center line of the 4K monitor, and adjusted to cover the full monitor range.



(a) Display-free (b) TOLED (c) POLED

Figure 4: Real samples collected by the proposed MCIS. Images captured by T-OLED are blur and noisy, while those captured by P-OLED are low-light, color-shifted and hazy.

natural images with rich textures on high-resolution monitor and capture them with a static camera. The method is more controllable, efficient, and automatic to capture a variety of scene contents than using mobile set-ups to capture limited static objects or real scenes.

4.1. Monitor-Camera Imaging System

The system architecture is shown in Figure 3. MCIS consists of a 4K LCD monitor, the 2K FLIR RGB Point-Grey research camera, and a panel that is either T-OLED, P-OLED or Glass(i.e. no display). The camera is mounted on the center line of the 4K monitor, and adjusted to cover the full monitor range. We calibrate the camera gain by measuring a 256×256 white square shown on the monitor and matching the RGB histogram. For fair comparison and simplicity, we adjust the focus and fix the aperture to f/1.8. It guarantees a reasonable pixel intensity range avoiding saturation while collecting data with no gain. Suppose we develop a real-time video system, the frame rate has to be higher than 8 fps. So the lowest shutter speed is 125 ms for the better image quality and the higher Signal-to-Noise Ratio (SNR).

We split 300 images from DIV2K dataset [4], and take turns displaying them on a 4K LCD in full screen mode. We either rotate or resize the images to maintain the Aspect Ratio. For training purposes, we capture two sets of images,

Table 2: Camera Settings for different set of collected data

Parameteres	No-Display	T-OLED	P-OLED
Aperture	f/1.8		
FPS/Shutter	8/125ms		
Brightness	0		
Gamma	1		
Gain	1	6	25(Full)
White-balance	Yes	None	None

Table 3: Measured parameters for data synthesis

Parameteres	T-OLED			P-OLED		
	R	G	B	R	G	B
γ	0.97	0.97	0.97	0.34	0.34	0.20
λ (nm)	640	520	450	640	520	450
r	2.41	2.98	3.44	2.41	2.98	3.44

which are the degraded images $\{y_i\}$, and the degradation-free set $\{x_i\}$.

To capture $\{x_i\}$, we first cover the camera with a thin glass panel which has the same thickness as a display panel. This process allows us to avoid the pixel misalignment issues caused by light refraction inside the panel. To eliminate the image noises in $\{x_i\}$, we average the 16 repeated captured frames. Then we replace the glass with a display panel (T-OLED or P-OLED), calibrate the specific gain value avoiding saturation, and capture $\{y_i\}$. For each set, we record both the 16-bit 1-channel linear RAW CMOS sensor data as well as the 8-bit 3-channel linear RGB data after in-camera pipeline that includes demosaicing. The collected pairs are naturally well spatially-aligned in pixel-level. They can be directly used for deep model training without further transformations.

Due to the yellow substrate inside the P-OLED, certain light colors, especially blue, are filtered out and changes the white balance significantly. We therefore did not further alter the white balance. The light transmission ratio of P-OLED is extremely low, so we set up the gain value to be the maximum (25) for higher signal values. All the detailed camera settings for the two display types are shown in Table 2. One real data sample is shown in Figure 4. As discussed and analyzed in Section 3.1, images captured by T-OLED are blur and noisy, while those captured by P-OLED are low-light, color-shifted and hazy.

4.2. Realistic Data Synthesis Pipeline

We follow the image formation pipeline to simulate the degradation on the collected $\{x_i\}$. A model-based data synthesis method will benefit concept understanding and further generalization. Note that all the camera settings are the same as the one while collecting real data. We first transform the 16-bit raw sensor data $\{x_i\}$ into four bayer channels x_r, x_{gr}, x_{gl}, and x_b. Then, we multiply the measured intensity scaling factor γ, compute the normalized and scaled PSF k, and add noises to the synthesize degraded data.

Measuring γ: To measure γ for each channel using the MCIS, we select the region of interest S to be a square re-

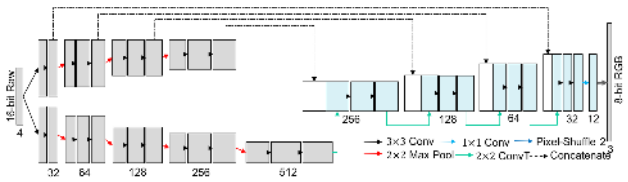


Figure 5: Network structure of the proposed UNet. It takes a 4-channel RAW sensor data observation y, and outputs the restored 3-channel RGB image x.

gion of size 256×256, and display the intensity value input from 0 to 255 with stride 10 on the monitor. We then record the average intensity both with and without the display for each discrete intensity value, and plot the relationship between display-covered values and no-display-covered ones. Using linear regression, we obtain the ratios of lines for different RGGB channel. For T-OLED, the measured γ is 0.97, same for all the channels. For P-OLED, $\gamma = 0.20$ for the blue channel, and $\gamma = 0.34$ for the other three channels.

Computing PSF: Following equation 3, we acquire the transmission microscope images of the display pattern and crop them with the approximated circular aperture shape with diameter $3333\mu m$, the size of the camera aperture. In equation 6, the $\delta_N N$ is $3333\mu m$. ρ equals to $1.55\mu m/pixel$ in Sony sensor. However, after re-arranging the raw image into four RGGB channels, ρ becomes 3.1 for each channel. The focal length is $6000\mu m$. $\lambda = (640, 520, 450)$ for R, G, B channel, which are the approximated center peaks of the R, G, B filters respectively on the sensor. It yields the down-sampling ratio $r = (2.41, 2.98, 3.44)$ for the R, G, and B channels.

Adding Noises: We measure λ_{read} and λ_{shot} to estimate the noise statistics. We display random patterns within the 256×256 window on the monitor, collect the paired noisy and noise-free RAW data, and compute their differences. For each of the RGGB channel, we linearly regress the function of noise variance to the intensity value, and obtain the ratio as the shot noise variance, and the y-intersection as the readout noise variance. We then repeat the process 100 times and collect pairs of data points. Finally, we estimate the distribution and randomly sample λ_{read} and λ_{shot}. All the measurements are listed in Table 3.

5. Image Restoration Baselines

We use the collected real paired data, synthetic paired data, simulated PSF, and all the necessary measurements to perform image restoration. We split the 300 pairs of images in the UDC dataset into 200 for training, 40 for validation and 60 images in the testing partition. All the images have a resolution of 1024×2048.

5.1. Deconvolution Pipeline (DeP)

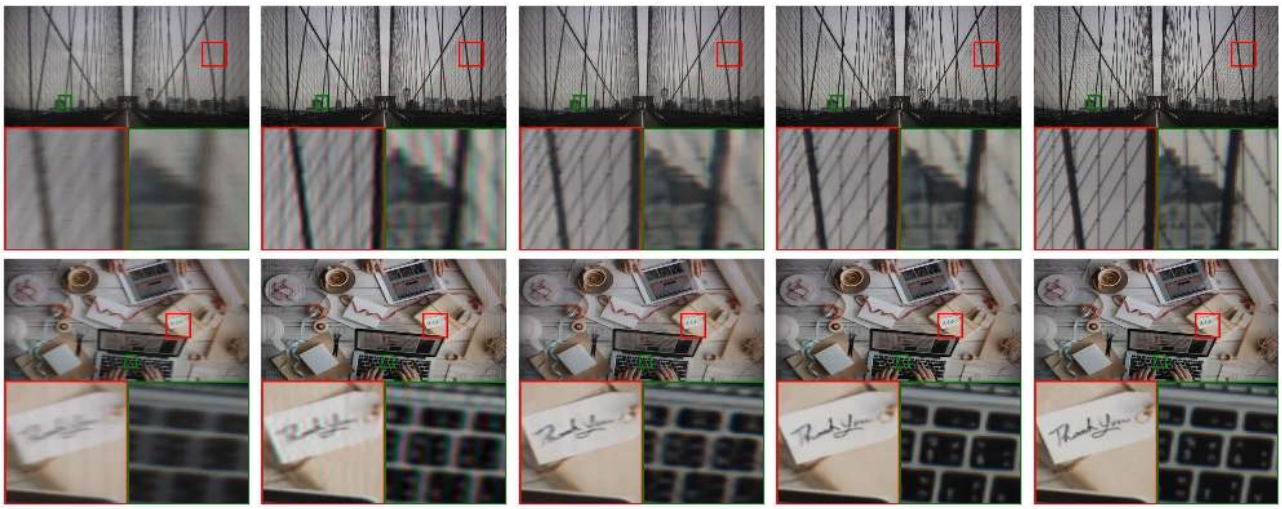
The DeP is a general-purpose conventional pipeline concatenating denoising and deconvolution (Wiener Filter), which is an inverse process of the analyzed image formation. To better utilize the unsupervised Wiener Filter (WF) [29], we first apply the BM3D denoiser to each RAW channel separately, afterwards we linearly divide the measured γ with the outputs for intensity scaling. After that, WF is applied to each channel given the pre-computed PSF $\mathbf{k}$. Finally, RAW images with bayer pattern are demosaiced by linear interpolation. The restored results are evaluated on the testing partition of the UDC dataset.

5.2. Learning-based Methods

UNet. We propose a learning-based restoration network baseline as shown in Figure 5. The proposed model takes a 4-channel RAW sensor data observation y, and outputs the restored 3-channel RGB image x. The model conducts denoising, debluring, white-balancing, intensity scaling, and demosaicing in a single network, whose structure is basically a UNet. We split the encoder into two sub-encoders, one of which is for computing residual details to add, and the other one learns content encoding from degraded images. By splitting the encoder, compared with doubling the width of each layer, we will have fewer parameters, and make the inference and learning more efficient. To train the model from paired images, we apply the L_1 loss, which will at large guarantee the temporal stability compared with adversarial loss [15]. Besides, we also apply $SSIM$ and Perception Loss (VGG Loss) for ablation study.

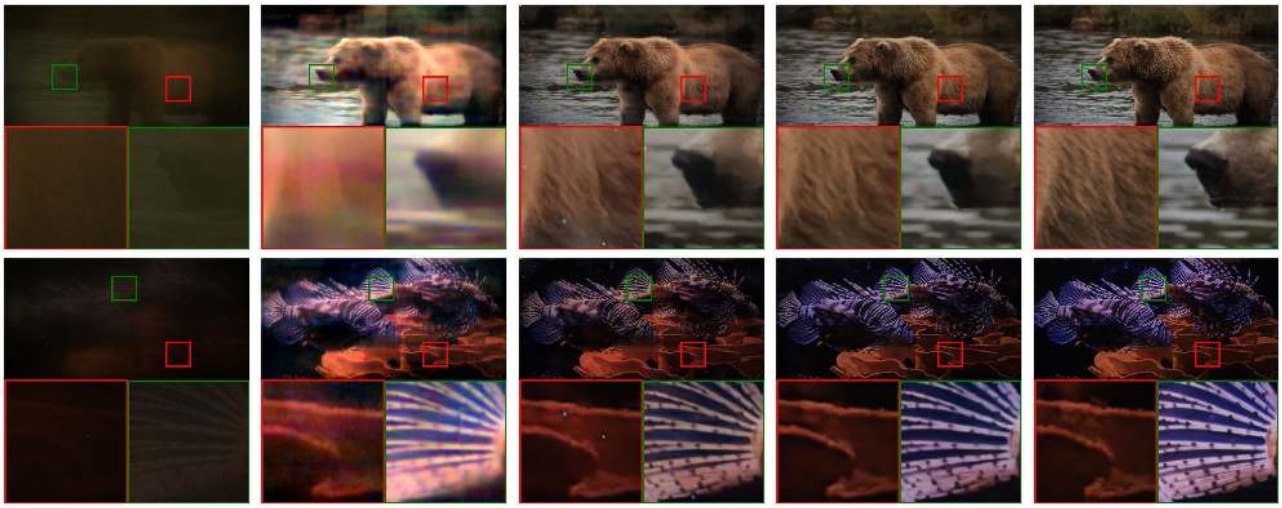
We crop patches of 256×256, and augment the training data using the raw image augmentation [26] while preserving the RGGB bayer pattern. We train the model for 400 epochs using Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$) with learning rate 10^{-4} and decay factor 0.5 after 200 epoches. We also train the same structure using the synthetic data (denoted as **UNet(Syn)**) generated by the pipeline proposed in section 4.2.

ResNet. Additionally, a data-driven ResNet trained with the same data is utilized for evaluation. To our knowledge, UNet and ResNet-based structures are two widely-used deep models for image restoration. We use 16 residual blocks with a feature width of 64 for our ResNet architecture, just as Lim *et al.* do for EDSR [22]. The model also takes 4-channel RAW data, and outputs 3-channel RGB images. The data-driven models cannot be directly adaptive to UDC inputs if only trained with bi-cubic degradation. We did not compare with their model structures because model novelty is not our main claim, and the presented two methods are the most general ways which can achieve real-time inference as the baselines. Other model variants can be further explored in future work.



(a) T-OLED (b) DeP (c) UNet(Syn) (d) UNet (e) GT

Figure 6: Restoration Results Comparison for T-OLED. GT: Ground Truth.



(a) P-OLED (b) DeP (c) UNet(Syn) (d) UNet (e) GT

Figure 7: Restoration Results Comparison for P-OLED. GT: Ground Truth.

6. Experimental Results

6.1. Qualitative and Quantitative Comparisons

The qualitative restoration results are shown in Figure 6 and 7. As shown, image Deconvolution Pipeline (DeP) successfully recovers image details but still introduces some artifacts, and suffers from the inaccuracy of the computed ideal PSF. The UNet-based model achieves better visual quality and denoising performance. The results of UNet trained with the synthetic data are visually better than DeP.

The quantitative results are listed in Table 4. We report the performance in PSNR, SSIM, a perceptual metric LPIPS [44], inference time T (ms/MPixels) and GFLOPs. The inference time is tested with one single Titan X, and the GFLOPs is computed by input size of $512 \times 1024 \times 4$. ResNet achieves a comparable performance to UNet, but it requires more computation operations and longer inference time. The proposed UNet-based structure is efficient and effective, which can therefore be deployed for real-time inference for high-resolution inputs with a single GPU. In Table 4, we demonstrate that synthetic data still has gaps with the real data, though it has already greatly out-performed the DeP for the two display types. The domain gap mainly comes from the following aspects. First, due to the existing distances between display and lens, in real data there appears visible patterns of the display on the image plane. We recall in the assumption of the diffraction model, the display panel is exactly at the principle plane of the lens system. The cause of the visible bands are illustrated in the supplementary material. Second, the approximated light transmis-

Table 4: Pipeline Comparison

				4K T-OLED		P-OLED	
Pipeline Structure	#P ↓	GFLOPs ↓	T ↓	PSNR/SSIM ↑	LPIPS ↓	PSNR/SSIM ↑	LPIPS ↓
DeP	-	-	-	28.50/0.9117	0.4219	16.97/0.7084	0.6306
ResNet	**1.37M**	721.76	92.92	36.26/0.9703	0.1214	27.42/0.9176	0.2500
UNet(Syn)	8.93M	**124.36**	**21.37**	32.42/0.9343	0.1739	25.88/0.9006	0.3089
UNet	8.93M	**124.36**	**21.37**	**36.71/0.9713**	**0.1209**	**30.45/0.9427**	**0.2219**

Table 5: Ablation Study on UNet alternatives.

Alternatives				4K T-OLED		P-OLED	
	#P ↓	GFLOPs ↓	T ↓	PSNR/SSIM ↑	LPIPS ↓	PSNR/SSIM ↑	LPIPS ↓
UNet Basseline	8.93M	124.36	21.37	36.71/0.9713	0.1209	30.45/0.9427	0.2219
Double Width	31.03M	386.37	40.42	37.00/0.9730	0.1171	30.37/0.9425	0.2044
Single Encoder	7.76M	97.09	15.85	36.47/0.9704	0.1288	30.26/0.9387	0.2318
$L_1 \rightarrow L_1 + SSIM$	8.93M	124.36	21.37	36.69/0.9714	0.1246	30.37/0.9403	0.2131
$L_1 \rightarrow L_1 + VGG$	8.93M	124.36	21.37	36.31/0.9711	0.1130	30.37/0.9403	0.2130

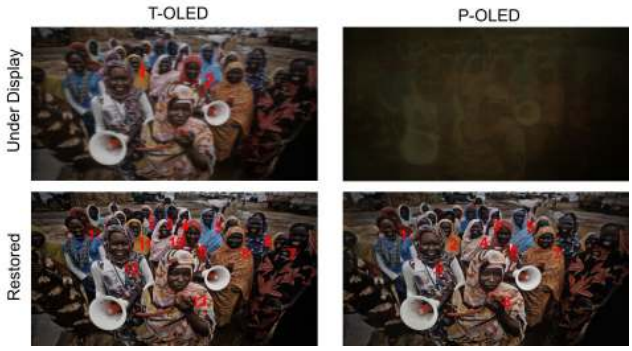


Figure 8: Face detection performance before and after applying restoration. Without display, the original face recall rate is 60%. Covering the camera with T-OLED or P-OLED will decrease the recall rate to 8% and 0%. After image restoration, the recall rates recovered back to 56% and 39%.

sion rate may not be accurate, the measured values may be influenced by other environment light sources. Third, impulse noise caused by dead pixels or over-exposure in the camera sensors widely exist in the real dataset. Those factors provide more improvement space for this work.

6.2. Ablation Study

For the best-performed UNet structure, we compare different UNet alternatives in Table 5. We increase the parameter size by splitting the original encoders into two sub-encoders, so the performance is also increased. The increment parameter size and inference time is far less than doubling the width of each layer of UNet, but the performance improvement is comparable (T-OLED), even better (P-OLED). We claim that the proposed UNet structure will both maintain a small number of parameters and operations, and achieve a real-time high-quality inference. To try alternative loss functions, we add $SSIM$ or VGG loss in additional to L_1 loss with 1:1 ratio. However, the performance gains on either $SSIM$ or perceptual metric LPIPS are not significant enough, and are not visually distinctive. Adversarial loss is not implemented due to its temporal instability of GAN-based training.

6.3. Downstream Applications

The proposed image restoration also enhances the performance of downstream applications including face detection. Figure 8 shows an example of detecting faces using MTCNN [41]. Without display, the original face recall rate is 60%. Covering the camera with T-OLED or P-OLED will decrease the recall rate to 8% and 0%. After image restoration, the recall rates are recovered to 56% and 39%.

7. Conclusion and Limitations

This paper defined and presented a novel imaging system named Under-Display-Camera (UDC). Deploying UDC to full-screen devices improves the user interaction as well as teleconferencing experience, but does harm to imaging quality and other downstream vision applications. We systematically analyzed the optical systems and modelled the image formation pipeline of UDC, and both collected real data using a novel acquisition system and synthesized realistic data and the PSF of the system using optical model. We then proposed to address the image restoration of UDC using a Deconvolution-based Pipeline (DeP) and data-driven learning-based methods. Our experiments showed that the former achieved basic restoration and the latter demonstrated an efficient high-quality restoration. The model trained with synthetic data also achieved a remarkable performance indicating the potential generalization ability.

UDC problem has its promising research values in complicated degradation analysis. In real-world applications, other factors like an active display, reflection, lens flare *etc.* are still very challenging and complicated. Future work can be exploring UDC-specific restoration models and working with aperture and display researchers to analyze the influential factors of image degradation. It will make the restoration model generalized better for mass production, or helpful for down-stream tasks, as an ultimate goal.

References

[1] Abdelrahman Abdelhamed, Marcus A Brubaker, and Michael S Brown. Noise flow: Noise modeling with conditional normalizing flows. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3165–3173, 2019. 2

[2] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2018. 2

[3] Abdelrahman Abdelhamed, Radu Timofte, and Michael S Brown. Ntire 2019 challenge on real image denoising: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2

[4] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 2, 5

[5] Josue Anaya and Adrian Barbu. Renoir–a dataset for real low-light image noise reduction. *Journal of Visual Communication and Image Representation*, 51:144–154, 2018. 2

[6] Codruta O Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense haze: A benchmark for image dehazing with dense-haze and haze-free images. *arXiv preprint arXiv:1904.02904*, 2019. 2

[7] Tim Brooks and Jonathan T Barron. Learning to synthesize motion blur. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6840–6848, 2019. 2

[8] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019. 2, 4

[9] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3291–3300, 2018. 2

[10] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1652–1660, 2019. 2

[11] Dong-Ming Chen, Bin Xiong, and Zhen-Yu Guo. Full-screen smartphone, Sept. 3 2019. US Patent App. 29/650,323. 1

[12] V David John Evans, Xinrui Jiang, Andrew E Rubin, Matthew Hershenson, and Xiaoyu Miao. Optical sensors disposed beneath the display of an electronic device, Oct. 17 2019. US Patent App. 16/450,727. 1

[13] Raanan Fattal. Single image dehazing. *ACM transactions on graphics (TOG)*, 27(3):72, 2008. 1

[14] J Scott Goldstein, Irving S Reed, and Louis L Scharf. A multistage representation of the wiener filter based on orthogonal projections. *IEEE Transactions on Information Theory*, 44(7):2943–2959, 1998. 1

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 6

[16] Joseph W Goodman. *Introduction to Fourier optics*. Roberts and Company Publishers, 2005. 4

[17] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1712–1722, 2019. 2

[18] Samuel W Hasinoff. Photon, poisson noise. *Computer Vision: A Reference Guide*, pages 608–610, 2014. 4

[19] Michael T Heath. *Scientific Computing: An Introductory Survey, Revised Second Edition*. SIAM, 2018. 1

[20] Salman S Khan, VR Adarsh, Vivek Boominathan, Jasper Tan, Ashok Veeraraghavan, and Kaushik Mitra. Towards photorealistic reconstruction of highly multiplexed lensless images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7860–7869, 2019. 2

[21] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8183–8192, 2018. 1

[22] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 1, 6

[23] Sehoon Lim, Yuqian Zhou, Neil Emerton, and Tim Large. Aperture design for learning-based image restoration. In *3D Image Acquisition and Display: Technology, Perception and Applications*, pages DF3A–2. Optical Society of America, 2020. 2

[24] Sehoon Lim, Yuqian Zhou, Neil Emerton, Tim Large, and Steven Bathiche. 74-1: Image restoration for display-integrated camera. In *SID Symposium Digest of Technical Papers*, volume 51, pages 1102–1105. Wiley Online Library, 2020. 2

[25] Ce Liu, Richard Szeliski, Sing Bing Kang, C Lawrence Zitnick, and William T Freeman. Automatic estimation and removal of noise from a single image. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):299–314, 2007. 4

[26] Jiaming Liu, Chi-Hao Wu, Yuzhi Wang, Qin Xu, Yuqian Zhou, Haibin Huang, Chuan Wang, Shaofan Cai, Yifan Ding, Haoqiang Fan, et al. Learning raw image denoising with bayer pattern unification and bayer preserving augmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 6

[27] Kristina Monakhova, Joshua Yurtsever, Grace Kuo, Nick Antipa, Kyrollos Yanny, and Laura Waller. Learned reconstructions for practical mask-based lensless imaging. *Optics express*, 27(20):28075–28090, 2019. 2

[28] Seungjun Nah, Radu Timofte, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, and Kyoung Mu Lee.

Ntire 2019 challenge on video deblurring: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2

[29] François Orieux, Jean-François Giovannelli, and Thomas Rodet. Bayesian estimation of regularization and point spread function parameters for wiener–hunt deconvolution. *JOSA A*, 27(7):1593–1607, 2010. 2, 6

[30] Yifan Peng, Qilin Sun, Xiong Dun, Gordon Wetzstein, Wolfgang Heidrich, and Felix Heide. Learned large field-of-view imaging with thin-plate optics. *ACM Trans. Graph.*, 38(6):219–1, 2019. 2

[31] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1586–1595, 2017. 2

[32] Abhijith Punnappurath and Michael S Brown. Reflection removal using a dual-pixel sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1556–1565, 2019. 2

[33] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *European conference on computer vision*, pages 154–169. Springer, 2016. 1

[34] Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1386–1396, 2020. 2

[35] Jasper Tan, Li Niu, Jesse K Adams, Vivek Boominathan, Jacob T Robinson, Richard G Baraniuk, and Ashok Veeraraghavan. Face detection and verification using lensless cameras. *IEEE Transactions on Computational Imaging*, 5(2):180–194, 2018. 1

[36] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Benchmarking single-image reflection removal algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3922–3930, 2017. 2

[37] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 1

[38] Ing G Wenke. Organic light emitting diode (oled). *Research gate*, 2016. 2

[39] He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 695–704, 2018. 1

[40] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *IEEE transactions on circuits and systems for video technology*, 2019. 1

[41] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 8

[42] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 1

[43] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 1

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 7

[45] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2019. 2

[46] Zhenhua Zhang. Image deblurring of camera under display by deep learning. In *SID Symposium Digest of Technical Papers*, volume 51, pages 43–46. Wiley Online Library, 2020. 2

[47] Yuqian Zhou, Jianbo Jiao, Haibin Huang, Jue Wang, and Thomas Huang. Adaptation strategies for applying awgn-based denoiser to realistic noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 10085–10086, 2019. 1

[48] Yuqian Zhou, Jianbo Jiao, Haibin Huang, Yang Wang, Jue Wang, Honghui Shi, and Thomas Huang. When awgn-based denoiser meets real noises. *arXiv preprint arXiv:1904.03485*, 2019. 1, 2

[49] 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 *European Conference on Computer Vision*, pages 337–351. Springer, 2020. 2