Modular Degradation Simulation and Restoration for Under-Display Camera

Yang Zhou† , Yuda Song†, and Xin Du ✉

Zhejiang University, Hangzhou, China
{yang_zhou,syd,duxin}@zju.edu.cn

Abstract. Under-display camera (UDC) provides an elegant solution for full-screen smartphones. However, UDC captured images suffer from severe degradation since sensors lie under the display. Although this issue can be tackled by image restoration networks, these networks require large-scale image pairs for training. To this end, we propose a modular network dubbed MPGNet trained using the generative adversarial network (GAN) framework for simulating UDC imaging. Specifically, we note that the UDC imaging degradation process contains brightness attenuation, blurring, and noise corruption. Thus we model each degradation with a characteristic-related modular network, and all modular networks are cascaded to form the generator. Together with a pixel-wise discriminator and supervised loss, we can train the generator to simulate the UDC imaging degradation process. Furthermore, we present a Transformer-style network named DWFormer for UDC image restoration. For practical purposes, we use depth-wise convolution instead of the multi-head self-attention to aggregate local spatial information. Moreover, we propose a novel channel attention module to aggregate global information, which is critical for brightness recovery. We conduct evaluations on the UDC benchmark, and our method surpasses the previous state-of-the-art models by 1.23 dB on the P-OLED track and 0.71 dB on the T-OLED track, respectively. Code is available at Github.

1 Introduction

Driven by the strong demand for full-screen mobile phones, the under-display camera (UDC) increasingly draws researchers' attention. UDC technology can deliver a higher screen-to-body ratio without disrupting the screen's integrity and introducing additional mechanics. However, UDC provides a better user experience at the expense of image quality. Since the sensor is mounted behind the display, the UDC images inevitably suffer severe degradation. Such image degradation is mainly caused by low light transmittance, undesirable light diffraction, and high-level noise, resulting in dark, blurred, and noisy images.

Following the prior work [1], the UDC imaging degradation process can be formulated as:

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

† Equal contribution

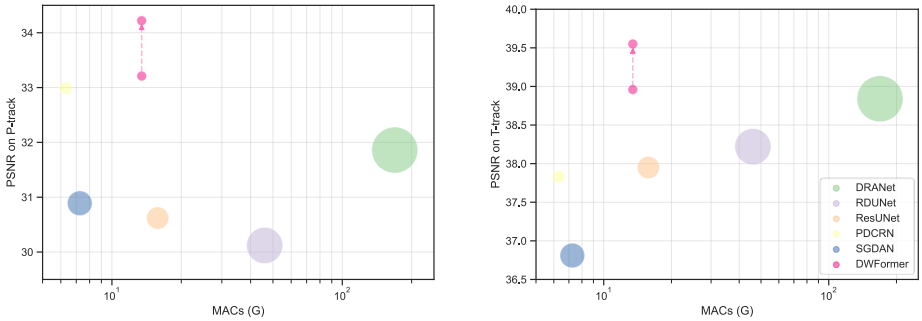


Fig. 1. Comparison of DWFormer with other UDC image restoration methods. The size of the dots indicates the Params of the method, and MACs are shown with the logarithmic axis. The arrows indicate that we use the generated data to improve the restoration model's performance further.

where $\cdot$ and $\otimes$ are multiplication and convolution operations, respectively, γ is the luminance scaling factor under the current gain setting and display type, k donates the point spread function (PSF) [2], and n is the zero-mean signal-dependent noise. Considering the attenuation of light transmission is wavelength-related [3], the luminance scaling factor γ should be different for each channel. Also, the PSF is spatially varying due to the different angles of the incident light [4]. And the signal-dependent noise n consists of shot and read noise [5] which can be modeled by heteroscedastic Gaussian distribution [6].

In recent years, many learning-based methods [7,8,9,10] have been introduced to improve the quality of UDC images and made significant advancements as they can learn strong priors from large-scale datasets. However, such methods require large amounts of data, and collecting aligned image pairs is labor-intensive. Facing insufficient data, generating realistic datasets can be a promising solution.

This paper proposes a modular pipeline generative network named MPGNet to simulate the UDC imaging degradation process and then use it to generate realistic image pairs. Unlike other end-to-end generation methods [11,12,13], we replace each degradation process with a subnetwork to preserve the physical properties of the imaging process. Specifically, we treat the image degradation process as three sequential steps, *i.e.,* brightness attenuation, non-uniform blurring, and noise corruption. And all modular subnetworks form the UDC imaging pipeline network. Besides the supervised learning, we employ GAN framework [14] to enhance the realism of generated images.

Based on the large amount of data generated by MPGNet, we can obtain a restoration model with well-generalized performance. However, designing a network suitable for the UDC image restoration task is not trivial. Recently, Transformers have attracted researchers' interest in the computer vision since ViT's [15] success on the high-level vision tasks. Thus we hope to build an efficient and effective UDC image restoration network based on the vision Transformer. Considering that MetaFormer [16] reveals the general architecture of the

transformers is more critical than the multi-head self-attention, we use depth-wise convolution and channel attention module to aggregate global information. Finally, we build a U-Net-like restoration network dubbed DWFormer for UDC image restoration.

We conduct evaluations on the UDC benchmark to verify the effectiveness of our MPGNet and DWFormer. Fig. 1 compares DWFormer with other UDC image restoration methods. Without synthetic datasets generated by MPGNet, DWFormer still achieves 33.21 dB on the P-OLED track and 38.96 dB on the T-OLED track, which surpasses the previous state-of-the-art models by 0.22 dB and 0.12 dB, respectively. Furthermore, if we use both generated and real data to train our DWFormer, it achieves 34.22 dB on the P-OLED track and 39.55 dB on the T-OLED track, 1.01 dB and 0.59 dB higher than the DWFormer trained with only real data. These results indicate that MPGNet can precisely model the UDC imaging process, and DWFormer can restore UDC images effectively. We hope our work can promote the application of UDC on full-screen phones.

2 Related Works

2.1 UDC Imaging

Several previous works [1,4,17,18] have constructed the optical system of under-display camera (UDC) imaging and analyzed its degradation components as well as the causes. While these works provided good insights into the UDC imaging system, their modeling approaches simplify the actual degradation process. Thus the derived degradation images differ significantly from the real ones. Therefore, how to generate realistic degradation images is still a problem to be solved, and this is one of the focuses of our work. We found that some work has been done to study the degradation module individually.

Blur Modeling. The blurring process can be modeled as a blurring kernel performing a convolution operation on a sharp image [19,20]. Many methods [21,22,23] estimate the blur kernel by assuming the characteristics of the blur kernel, and other methods [1,4] models blur by a point spread function (PSF). However, the blur in UDC imaging is blind and spatially varying, increasing the difficulty of accurately estimating the blur kernel. Unlike previous works, we try to model the blur directly using a convolutional neural network.

Noise Modeling. The noise is usually modeled as Poissonian-Gaussian noise [24] or heteroscedastic Gaussian [25]. While the heteroscedastic Gaussian noise model can provide a proper approximation of the realistic noise [26] to some extent, several studies [27,28] have demonstrated that the real-world cases appear to be much more complicated. To this end, GCBD [11] proposed a generation-based method to generate realistic blind noise. C2N [29] adopted a new generator architecture to represent the signal-dependent and spatially correlated noise distribution. Compared to C2N, our proposed generator has a receptive field limited by the demosaicing method [30] and considers quantization noise.

2.2 Generative Adversarial Network (GAN)

GAN [14] was proposed to estimate the generative model via simultaneous optimization of the generator and discriminator. And many researchers have leveraged GANs for image-to-image translation [31,32,33,34,35], whose goal is to translate an input image from one domain to another domain. However, GAN may suffer from gradient vanishing or exploding during training, and several works [36,37,38] have proposed proper loss functions to stabilize training. We build our GAN framework using supervised loss and adversarial loss, thus stabilizing training and achieving promising results.

2.3 Image Restoration Architecture

A popular solution for recovering degraded images is to use CNN-based U-Net-like networks to capture hierarchical information for various image restoration tasks, including image denoising [39,26], deblurring [40,41] and low-light enhancement [42,43]. Recently, Transformer achieved great advancements in high-level vision problem [15,44,16] and has also been introduced for image restoration [45,46]. IPT [45] uses standard Transformer to build a general backbone for various restoration problems. However, IPT is quite huge and requires large-scale datasets. And Uformer [46] proposed a U-Net-like architecture based on the Swin Transformer [44] for noise and blur removal. Motivated by MetaFormer [16], we use depth-wise convolution instead of multi-head self-attention to aggregate spatial information. The most similar work to ours is NAFNet [47], which also uses depth-wise convolution and channel attention module to build a MetaFormer-like network. However, we choose FrozenBN [48] instead of LN as the normalization layer. We also propose a novel channel attention module dubbed ACA, which increases the computational cost slightly compared to the SE modules [49] but can aggregate global information more effectively.

3 Method

3.1 Overall Network Architecture

Our method consists of a new generative network called MPGNet for modeling the UDC imaging degradation process and a U-Net-like network called DWFormer for UDC image restoration.

For MPGNet, we adopt GAN framework [14] to improve the realism of generated images. As shown in Fig. 2, our generate network architecture consists of a degradation generator and a pixel-wise discriminator. The degradation generator MPGNet comprises three parts, *i.e.*, brightness attenuation module, blurring module, and noise module. The three parts correspond to channel scaling, diffraction blurring, and Poisson-Gaussian noise corruption in the Statistical Generation Method (SGM) [1]. Since the generation process can be considered an image-to-image translation task, we employ a pixel-wise U-Net-like discriminator [50] for better performance. DWFormer is a U-Net-like network, as shown

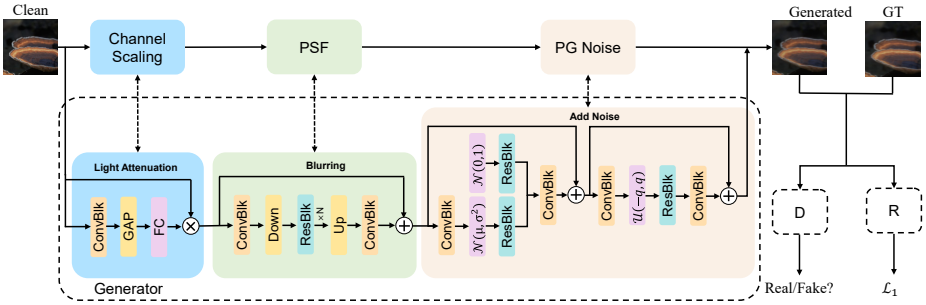


Fig. 2. Generator network architecture of our proposed MPGNet. We replace each degradation process with a characteristic-related subnetwork to simulate the UDC imaging pipeline and use a GAN-based training scheme for more realistic results.

in Fig. 3, built with our proposed DWFormer Block (DWB). The DWB evolves from the standard Transformer, and we replace the multi-head self-attention with a depth-wise convolution and channel attention module.

3.2 MPGNet

Brightness Attenuation. The brightness attenuation occurs due to organic light-emitting diode (OLED) displays [51] absorbing part of the light. Since the attenuation is wavelength-related [3], the attenuation coefficients to be estimated should be channel-dependent. Besides, brightness is a global statistic and therefore requires global information to be aggregated. To this end, we use several convolution blocks H_{FE} to extract features, a global average pooling H_{GAP} to aggregate information, and a multi-layer perceptron (MLP) to encode attenuation coefficients [52]. We multiply attenuation coefficients of size $1 \times 1 \times C$ with the clean image to obtain the dark image:

$$\begin{aligned} F_{BF} &= H_{FE}(x_{clean}), \\ F_G &= H_{GAP}(F_{BF}), \\ x_{dark} &= x_{clean} \cdot \sigma(H_D(\delta(H_U(F_G)))), \end{aligned} \tag{2}$$

where δ is ReLU and σ is Sigmoid function, H_D and H_U are the channel reduction and channel upsampling operators, respectively [49].

Blurring. The blurring in UDC imaging is caused by light diffraction as the size of the openings in the pixel layout is on the order of the wavelength of visible light. Although the blur kernel of diffraction could be accurately estimated by computing the Fourier transform of the wavefront function [53], it is too complicated, especially when light enters from all directions and forms spatially varying PSFs. For practical purposes, we choose residual blocks [54] as the basis to build the characteristic-related subnetwork to model the blurring process. Furthermore, since the blur kernel size is not fixed [4], to better

cover the blur kernel of various sizes, we use convolution to downsampling and use sub-pixel convolution [55] to upsampling the feature maps. Finally, we use residual connection to fuse the input and output:

$$x_{blur} = H_B(x_{dark}) + x_{dark}, \tag{3}$$

where H_B is the blurring module, x_{dark} is the output of the previous module, and x_{blur} is the blurred and dark image.

Noise. The noise consists of read and shot noise [5], which are usually formulated as heteroscedastic Gaussian distribution [6]. However, noise in the real world is more complicated and spatially correlated, making the heteroscedastic Gaussian distribution model inaccurate. Inspired by C2N [56], we generate realistic noise by combining signal-independent and signal-dependent noise. First, we use a residual block H_{R_1} to transform the noise $n_{s_1} \in \mathbb{R}^{h\times w\times d}$ (d=32 is the feature dimension) sampled from the standard normal distribution $\mathcal{N}(0,1)$ to noise n_i with a more complicated distribution:

$$n_i = H_{R_1}(n_{s_1}) \ . \tag{4}$$

Second, the mean and variance of signal-dependent noise should be highly related to the image signal. We use a convolutional block to encode the pixel-wise mean $\mu \in \mathbb{R}^{h\times w\times d}$ and variance $\sigma \in \mathbb{R}^{h\times w\times d}$ from x_{blur}. Since the sampling is not differentiable, we use the reparameterization trick to transform the noise. Specifically, we sample the noise $n_{s_2} \in \mathbb{R}^{h\times w\times d}$ from $\mathcal{N}(0,1)$, and transform it to noise $n_d \sim \mathcal{N}(\mu, \sigma^2)$ via

$$n_d = H_{R_2}(n_{s_2} \cdot \sigma + \mu), \tag{5}$$

where H_{R_2} is also a residual block to transform the distribution of signal-dependent noise to a more complicated distribution. Considering that the noise is spatially correlated, we use residual blocks with two pixel-wise convolutions and one 3×3 convolution for both noises. After mapping these noises from the initial noise to noise with the target distribution, we take 1×1 convolution H_M to reduce the dimension to the color space and add them to output x_{blur} of the blurring module:

$$x_{noisy} = x_{blur} + H_M(n_i + n_d) \ . \tag{6}$$

Moreover, recent work [57] shows that the quantification noise significantly impact on low-light imaging. And following the ISP pipeline, the quantization noise should be signal-dependent and added after other noise. Thus we use a convolutional block to encode the pixel-wise quantization noise interval $q \in \mathbb{R}^{h\times w\times d}$ from x_{noisy}. Also, we use a residual block H_{R_3} to transform the quantization noise $n_{s_3} \in \mathbb{R}^{h\times w\times d}$ sampled from the uniform distribution $\mathcal{U}(-q, q)$ to more realistic noise n_q:

$$n_q = H_{R_3}(n_{s_3}) \ . \tag{7}$$

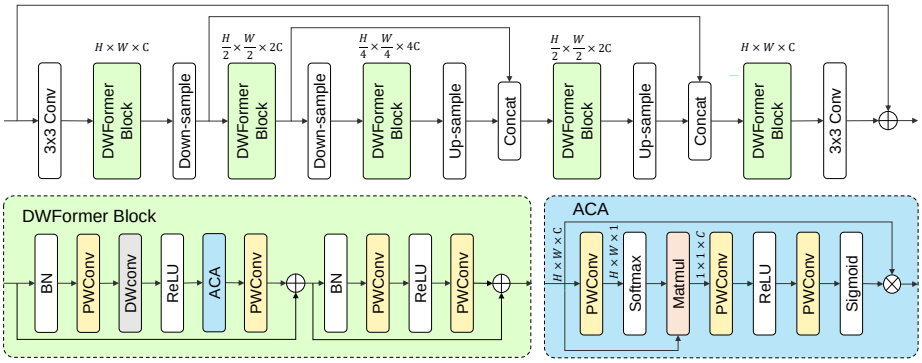


Fig. 3. Restoration network architecture of our proposed DWFormer.

After transforming n_q to the color space by a pixel-wise convolution, we add the quantization noise to the previous noisy image x_{noisy} by a residual connection:

$$x_{final} = H_{N_q} + x_{noisy}, \tag{8}$$

where H_{N_q} is the quantization noise module, and x_{final} is the final degraded image.

3.3 DWFormer

Overall Pipeline. The overall structure of the proposed DWFormer is a U-Net-like hierarchical network with skip connections. Specifically, given a degraded image $x \in \mathbb{R}^{h\times w\times 3}$, DWFormer firstly applied a 3×3 convolution to extract image features $F_0 \in \mathbb{R}^{h\times w\times c}$. Then, the feature map F_0 passes through all stages, and each stage contains a stack of proposed DWBs and a downsampling layer or upsampling layer. We use convolution operation to downsampling and sub-pixel convolution to upsampling, respectively. Finally, we also use a 3×3 convolution to get the residual image $r \in \mathbb{R}^{h\times w\times 3}$ and the restored image is obtained by $\hat{x} = y + r$.

DWFormer Block. Transformer [58] conducts global self-attention, which leads to quadratic complexity with respect to the number of tokens and has much less inductive bias than CNNs, which may be helpful for low-level tasks. Thus, we modify it to be more suitable for the UDC image restoration task. Specifically, considering local information is more favorable for noise and blur removal, we use a depth-wise convolution H_D to achieve spatial aggregation and thus significantly reduce the computational cost. Further, since brightness correction requires global information for each color channel, we propose an augmented channel attention (ACA) module H_{ACA} to capture global information and place it after the depth-wise convolution. Unlike SENet [49], which uses global average pooling to get the global feature descriptor, our proposed ACA starts with a

pixel-wise convolution followed by reshape and softmax operations to obtain the weights $W_p \in \mathbb{R}^{1\times1\times HW}$ of each position of the feature maps $F_{in} \in \mathbb{R}^{H\times W\times C}$. Then we also reshape the F_{in} to $\mathbb{R}^{1\times HW\times C}$ and and use matrix multiplication with W_p to obtain the global feature descriptor $F_C \in \mathbb{R}^{1\times1\times C}$. Also, we use a pixel-wise convolution with ReLU activation and a pixel-wise convolution with Sigmoid activation to fully capture channel-wise dependencies. Finally, we use the descriptor to rescale the input feature maps to obtain the augmented feature maps. Besides, we found that BatchNorm (BN) performs better than LayerNorm (LN) in UDC image restoration task when batch size exceeds 16 on a single GPU. Therefore, we choose BN as the normalization layer, and the whole DWFormer block is computed as:

$$\begin{aligned} \hat{z}^l &= H_{P_2}(H_{ACA}(\delta(H_D(H_{P_1}(\mathrm{BN}(z^l)))))) + z^l, \\ z^{l+1} &= \mathrm{MLP}(\mathrm{BN}(\hat{z}^l)) + \hat{z}^l, \end{aligned} \tag{9}$$

where δ is ReLU and z^l and z^{l+1} denote the input feature maps and output feature maps of DWB respectively. Note that we use the FrozenBN [48] to avoid the inconsistency in training and testing, *i.e.*, we first train with minibatch statistics in the early training period and use fixed population statistics in the later training period.

3.4 Training

Our degradation generator network G is designed to generate a realistic degraded image $\hat{y} \in \mathbb{R}^{h\times w\times3}$ from its clean version $x \in \mathbb{R}^{h\times w\times3}$ and our discriminator network D is designed to distinguish each pixel's real probability. Similar to other GAN applications [31,59], the two networks G and D can be simultaneously optimized in an adversarial way [14] with the least-squares loss function [38]:

$$\begin{aligned} \min_G \max_D \mathcal{L}_{adv}(G,D) = \mathbb{E}_{y\sim p_{real}(y)}[logD(y)] \\ + \mathbb{E}_{x\sim p_{real}(x)}[1 - logD(G(x))], \end{aligned} \tag{10}$$

where y is the real degraded image and x is the corresponding clean image. Here we concat the degraded images with clean images as inputs to D. And since there have been real paired data [60,31], we can also use supervised algorithms to optimize our model. However, the noise is a distribution-unknown random variable, and if we directly adopted $\mathcal{L}_1$ loss on generated and real degraded images, the noise will be eliminated while constructing the image alignment. To this end, we feed generated degraded images $G(x)$ and real degraded images y into a pre-trained restoration model R to obtain their restored versions and then perform $\mathcal{L}_1$ loss between them [61].

$$\mathcal{L}_{sup} = ||R(G(x)) - R(y)||_1 \ . \tag{11}$$

We use a hyperparameter λ to balance the supervised loss and adversarial loss, and the final loss is:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda\mathcal{L}_{sup} \ . \tag{12}$$

For convenience, we set the λ to 10 for both the T-OLED and P-OLED tracks. For DWFormer, we just use $\mathcal{L}_1$ to train it.

3.5 Implementation Details

We train MPGNet for the UDC image generation task and DWFormer for the UDC image restoration task. Thus the training settings are different. MPGNet is trained with Adam [62] optimizer ($\beta_1 = 0.5$, and $\beta_2 = 0.999$) for 6×10^4 iterations. We update the generator once and the discriminator three times in each iteration. The initial learning rates are set to 1×10^{-4} and 1×10^{-3} for generator and discriminator, respectively. We employ the cosine annealing strategy [63] to steadily decrease the learning rate from an initial value to 1×10^{-6} and 1×10^{-5} during training. The batch size is set to 8, and we randomly perform horizontal and vertical flips for data augmentation. And DWFormer has five stages, and the number of DWB in each layer is $\{8, 8, 8, 6, 6\}$, respectively. The model is also trained with Adam optimize ($\beta_1 = 0.9$, and $\beta_2 = 0.999$) for 2×10^5 iterations. The initial learning rate is set to 1×10^{-4} and the cosine annealing strategy is adopted to steadily decrease the learning rate to 1×10^{-6}. The batch size is 64, and we achieve data augmentation by randomly performing flips and rotations.

4 Experiment

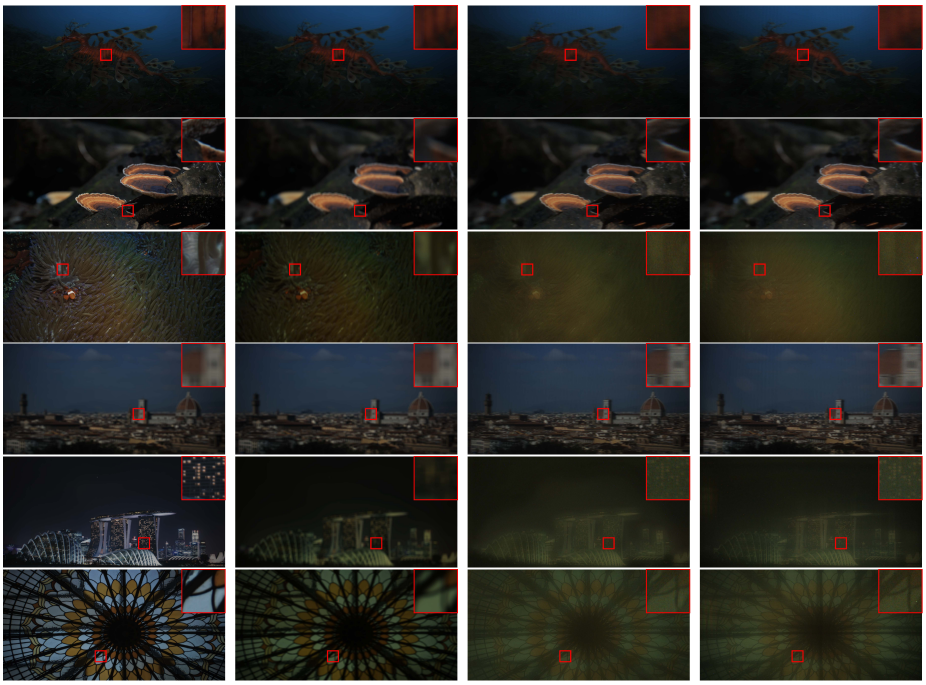
4.1 Dataset

The real image pairs used in the experiments are the P-OLED and T-OLED datasets [1] provided by UDC 2020 Image Restoration Challenge. Both datasets have 300 clean-degraded image pairs of size 1024×2048. Also, we leverage the high-resolution DIV2K dataset [64] captured under real environments to generate realistic degraded images. The DIV2K dataset provides 900 clean images. Similar to prior work [1], we optionally rotate and resize the clean images and generate clean-degraded image pairs with the resolution of 1024×2048. The real and generated image pairs are cropped into patches of size 256×256 and are randomly sampled to gather training mini-batches.

4.2 Comparison of Generators

To holistically evaluate the quality of MPGNet-generated images, we employ several tactics. First, we present several qualitative examples for perceptual studies in Fig. 4. The results demonstrate that the SGM-generated [1] degraded images cannot accurately estimate the blur and noise distribution. In contrast, the MPGNet-generated results are closer to the ground truths and preserve the degradation's diversity.

Second, we use different generated datasets to train the restoration models and evaluate them on the UDC benchmark. Fig. 5 shows that the model trained with the SGM-generated dataset yields still blurry and brightness-unaligned



(a) Clean (b) SGM (c) Ours (d) GT

Fig. 4. Visual comparison of the generated degraded image samples on the UDC dataset. (a) A clean image. (b-e) Generated degraded image of SGM and Ours, respectively. (f) The ground truth degraded image. The top half of the figure is on the T-OLED track, and the bottom half of the figure is on the P-OLED track. We amplify the images on the P-OLED track for comparison.

results, while the model trained with the MPGNet-generated dataset produces results closer to the ground truth. Also, Table 1 illustrates that our method outperforms SGM by 3.31 dB on the P-OLED track and 2.49 dB on the T-OLED track by only using the synthetic dataset. We can further improve the restoration model's performance by using both generated and real datasets for training, implying that our generated data can complement the real data and thus enhance the model's generalization.

Intuitively, using a single model as a generator is more common. Thus we replace the entire generation model with an end-to-end U-Net [64], which has a competitive number of parameters and computation cost to MPGNet. However, the U-Net performs poorly. We believe this is mainly because a single network tends to confuse multiple degradation processes, leading to model convergence to a poor local optimum. We assign the degradation process to multiple characteristic-related sub-networks, which dramatically avoids this local optimum, showing that physical constraints work. Also, from the results, we find that SGM performs worse than U-Net, which may be due to the inaccurate

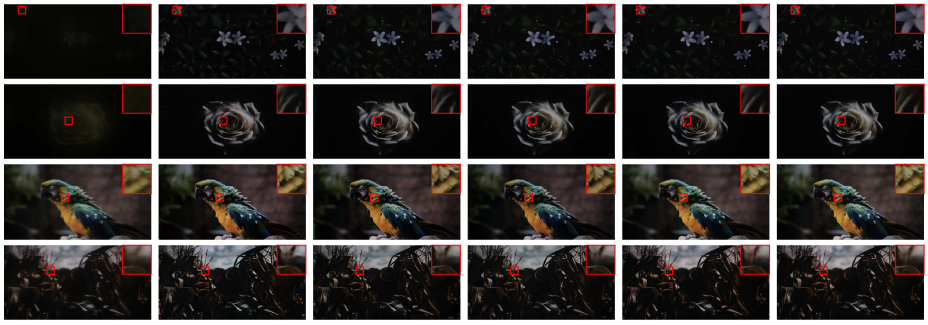


Fig. 5. Comparison between restoration results on the UDC dataset. (a) The degraded image. (b-c) The clean images are restored using the models trained by the different datasets. (f) The ground-truth. The top two are from the T-OLED track, and the bottom two are from the P-OLED track.

Table 1. Performance comparison between other methods and our MPGNet.

Methods	P-OLED		T-OLED	
	PSNR	SSIM	PSNR	SSIM
Real	33.21	0.960	38.96	0.984
SGM	$28.61_{\downarrow 4.60}$	$0.911_{\downarrow 0.049}$	$34.82_{\downarrow 4.14}$	$0.950_{\downarrow 0.034}$
SGM + Real	$32.56_{\downarrow 0.65}$	$0.957_{\downarrow 0.003}$	$38.42_{\downarrow 0.54}$	$0.980_{\downarrow 0.004}$
U-Net	$30.41_{\downarrow 2.80}$	$0.912_{\downarrow 0.048}$	$36.20_{\downarrow 2.76}$	$0.958_{\downarrow 0.026}$
MPGNet	$31.95_{\downarrow 1.26}$	$0.941_{\downarrow 0.019}$	$37.33_{\downarrow 1.63}$	$0.970_{\downarrow 0.014}$
MPGNet + Real	$34.22_{\uparrow 1.01}$	$0.964_{\uparrow 0.004}$	$39.55_{\uparrow 0.59}$	$0.986_{\uparrow 0.002}$

parameters from the manual statistics. We argue that our MPGNet provides a solution to avoid it.

4.3 Comparison of Restoration Models

We compare the performance of the proposed DWFormer with other learning-based approaches on the UDC benchmark, and the quantitative comparisons in terms of PSNR and SSIM metrics are summarized in Table 2. The results show that our method achieves state-of-the-art results for both the P-OLED track and T-OLED track. In particular, our DWFormer achieves 0.22 and 0.12 dB PSNR improvements over the previous best methods PDCRN [7] and DRANet [9] on the P-OLED and T-OLED tracks, respectively. Using both generated and real data to train our model, DWFormer can improve 1.01 dB on the P-OLED track and 0.59 dB on the T-OLED track over the previous one.

Table 2. Restore qualitative evaluation on the UDC benchmark. Note that † means that the models are trained with both generated and real datasets.

Methods	P-OLED		T-OLED		Overhead	
	PSNR	SSIM	PSNR	SSIM	Params	MACs
ResNet [1]	27.42	0.918	36.26	0.970	1.37M	40.71G
UNet [1]	29.45	0.934	36.71	0.971	8.94M	17.09G
SGDAN [65]	30.89	0.947	36.81	0.971	21.1M	7.25G
ResUNet [66]	30.62	0.945	37.95	0.979	16.50M	15.79G
DRANet [9]	31.86	0.949	38.84	0.983	79.01M	168.98G
PDCRN [7]	32.99	0.958	37.83	0.978	3.65M	6.31G
RDUNet [66]	30.12	0.941	38.22	0.980	47.93M	46.01G
DWFormer(Ours)	33.21	0.960	38.96	0.984	1.21M	13.46G
PDCRN†	34.02	0.962	38.75	0.982	3.65M	6.31G
RDUNet†	32.45	0.952	39.05	0.984	47.93M	46.01G
DWFormer†(Ours)	34.22	0.964	39.55	0.986	1.21M	13.46G

4.4 Ablation Studies

We first study the impact of each module in MPGNet on performance. Specifically, we replace brightness correction, blurring, and noise modules with corresponding modules in SGM. Further, we remove each noise component, *i.e.*, quantization noise n_q, signal-dependent noise n_d and signal-independent noise n_i to explore their effects. We use these modified networks to generate datasets that are used to train restoration models. And we evaluate the restoration models on the UDC benchmark as shown in Table 3.

As we can see, the blurring module has the most significant impact on the model performance, indicating that diffraction caused by hardware structure is the most critical degradation factor in the UDC imaging process. And ablation studies of noise and brightness modules show that different materials' main degradation components differ. Surprisingly, quantization noise plays such a significant role in the noise module. It may be due to the brightness attenuation during the imaging process and the low number of image bits. Also, such a phenomenon indicates that the actual noise module of the UDC images is quite complex and challenging to simulate using only Poisson-Gaussian distribution.

To verify the effectiveness of the DWFormer modules, we performed ablation studies for the normalization, depth-wise convolution, and attention mechanism module, and Table 3 shows the results. We notice that the removal of ACA causes a significant performance drop, implying that global information is indispensable. And our ACA is much better than SE with negligible additional computation cost. Also, the performance decreased if we used Swin Transformer's block, indicating that the locality properties provided by depth-wise convolution work for the UDC restoration task. Besides, BatchNorm surpasses LayerNorm. It is worth

Table 3. Different modules' effects on the MPGNet's performance. The performance is evaluated on the generated datasets using our DWFormer.

Methods	P-OLED		T-OLED	
	PSNR	SSIM	PSNR	SSIM
MPGNet	31.95	0.941	37.33	0.970
MPGNet w/ SGM-Light	$30.81_{\downarrow 1.14}$	$0.930_{\downarrow 0.011}$	$36.96_{\downarrow 0.37}$	$0.965_{\downarrow 0.005}$
MPGNet w/ SGM-Blur	$30.10_{\downarrow 1.85}$	$0.915_{\downarrow 0.026}$	$35.87_{\downarrow 1.46}$	$0.959_{\downarrow 0.011}$
MPGNet w/ SGM-Noise	$31.16_{\downarrow 0.79}$	$0.934_{\downarrow 0.007}$	$36.30_{\downarrow 1.03}$	$0.962_{\downarrow 0.008}$
MPGNet w/o n_q	$31.36_{\downarrow 0.59}$	$0.935_{\downarrow 0.006}$	$36.87_{\downarrow 0.46}$	$0.964_{\downarrow 0.006}$
MPGNet w/o n_d	$31.26_{\downarrow 0.69}$	$0.935_{\downarrow 0.006}$	$36.55_{\downarrow 0.78}$	$0.963_{\downarrow 0.007}$
MPGNet w/o n_i	$31.64_{\downarrow 0.31}$	$0.937_{\downarrow 0.004}$	$37.04_{\downarrow 0.29}$	$0.967_{\downarrow 0.003}$

Table 4. Different modules' effects on the DWFormer's performance.

Methods	P-OLED		T-OLED		MACs
	PSNR	SSIM	PSNR	SSIM	
DWFormer	33.21	0.960	38.96	0.984	13.46G
ACA $\rightarrow$ None	$32.62_{\downarrow 0.59}$	$0.951_{\downarrow 0.009}$	$38.45_{\downarrow 0.51}$	$0.980_{\downarrow 0.004}$	13.42G
ACA $\rightarrow$ SE	$33.00_{\downarrow 0.21}$	$0.958_{\downarrow 0.002}$	$38.72_{\downarrow 0.24}$	$0.982_{\downarrow 0.002}$	13.43G
DWB $\rightarrow$ Swin	$33.07_{\downarrow 0.14}$	$0.959_{\downarrow 0.001}$	$38.84_{\downarrow 0.12}$	$0.982_{\downarrow 0.002}$	16.22G
BN $\rightarrow$ LN	$33.11_{\downarrow 0.10}$	$0.957_{\downarrow 0.003}$	$38.90_{\downarrow 0.06}$	$0.982_{\downarrow 0.002}$	13.46G

noting that BatchNorm can be fused into a convolutional layer when inferring, which makes the network run faster than the network with LayerNorm.

Moreover, we explore the effect of $\mathcal{L}_{sup}$ and $\mathcal{L}_{adv}$ and use them for training MPGNet. Fig. 6 shows the generated results. $\mathcal{L}_{sup}$ alone leads to reasonable but noiseless results. $\mathcal{L}_{adv}$ alone gives much blurrier results. It is because the $\mathcal{L}_{sup}$ loss will still eliminate the noise while constructing the image alignment though we have attempted to alleviate it, and the weak constraint of GAN causes the content restoration to be more difficult.

We further use these generated datasets to train our DWFormer and evaluate them on the UDC benchmark for quantitative assessment, and the results are shown in Table 5. Note that $\mathcal{L}_1$ means that we directly compute the $\mathcal{L}_1$ loss on the generated image. In contrast, $\mathcal{L}_{sup}$ means we take the generated images through a restoration network and then calculate the loss. The $\mathcal{L}_1$ is worse than the $\mathcal{L}_{sup}$, implying that the latter can preserve the random noise to some extent. Note that if we do not feed clean images as extra inputs to D, MPGNet fails to generate valid degraded images and falls into a mode collapse.

Here, we use both $\mathcal{L}_{sup}$ and $\mathcal{L}_{adv}$ to boost the model performance. And We further evaluate it with different ratios (changing the value of λ). It is found

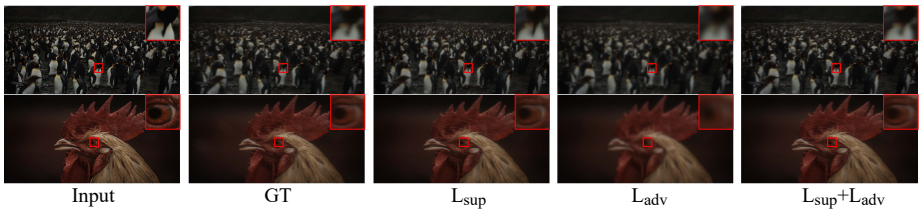


Fig. 6. Different losses lead to different generation results. Each column shows the results of training at different losses.

Table 5. Different loss terms for training MPGNets. The performance is evaluated on the generated datasets using our DWFormer.

Methods	P-OLED		T-OLED	
	PSNR	SSIM	PSNR	SSIM
$\mathcal{L}_1$	31.13	0.908	35.98	0.951
$\mathcal{L}_{sup}$	31.23	0.911	36.08	0.966
$\mathcal{L}_{adv}$	31.55	0.945	36.02	0.957
$\mathcal{L}_{adv} + \mathcal{L}_{sup}$	33.22	0.960	38.80	0.981
$\mathcal{L}_{adv} + 10\mathcal{L}_{sup}$	33.21	0.960	38.96	0.984
$\mathcal{L}_{adv} + 100\mathcal{L}_{sup}$	33.03	0.956	38.98	0.984

that the optimal ratio is highly correlated with the data itself, and different datasets hold different optimal ratios. Experiments show that $\mathcal{L}_{adv}$ plays a more significant role on the P-OLED track because the dataset is severely degraded. And the degradation is diminished on the T-OLED track, so the weight of $\mathcal{L}_1$ needs to be increased to ensure content consistency.

5 CONCLUSION

In this paper, we start from the degradation pipeline of the UDC imaging and replace each degradation process with a subnetwork, which forms our MPGNet. Further, the GAN framework is adopted to generate more realistic degraded images. Based on the analysis of UDC image degradation, we propose a novel block modified from Transformer and use it to build a U-Net-like image restoration network named DWFormer. Experiments show that our MPGNet generates more realistic degraded images than previous work, and DWFormer achieves superior performance. Finally, We use both generated and real datasets to train DWFormer, and further boost its performance, showing our generated data can be complementary to real data.

References

1. Zhou, Y., Ren, D., Emerton, N., Lim, S., Large, T.: Image restoration for under-display camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 9179–9188 1, 3, 4, 9, 12
2. Heath, M.T.: Scientific computing: an introductory survey, revised second edition. SIAM (2018) 2
3. Nayar, S.K., Narasimhan, S.G.: Vision in bad weather. In: Proceedings of the seventh IEEE international conference on computer vision. Volume 2., IEEE (1999) 820–827 2, 5
4. Kwon, K., Kang, E., Lee, S., Lee, S.J., Lee, H.E., Yoo, B., Han, J.J.: Controllable image restoration for under-display camera in smartphones. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 2073–2082 2, 3, 5
5. Hasinoff, S.W.: Photon, poisson noise. (2014) 2, 6
6. Kersting, K., Plagemann, C., Pfaff, P., Burgard, W.: Most likely heteroscedastic gaussian process regression. In: Proceedings of the 24th international conference on Machine learning. (2007) 393–400 2, 6
7. Panikkasseril Sethumadhavan, H., Puthussery, D., Kuriakose, M., Charangatt Victor, J.: Transform domain pyramidal dilated convolution networks for restoration of under display camera images. In: European Conference on Computer Vision, Springer (2020) 364–378 2, 11, 12
8. Sundar, V., Hegde, S., Kothandaraman, D., Mitra, K.: Deep atrous guided filter for image restoration in under display cameras. In: European Conference on Computer Vision, Springer (2020) 379–397 2
9. Nie, S., Ma, C., Chen, D., Yin, S., Wang, H., Jiao, L., Liu, F.: A dual residual network with channel attention for image restoration. In: European Conference on Computer Vision, Springer (2020) 352–363 2, 11, 12
10. Feng, R., Li, C., Chen, H., Li, S., Loy, C.C., Gu, J.: Removing diffraction image artifacts in under-display camera via dynamic skip connection network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 662–671 2
11. Chen, J., Chen, J., Chao, H., Yang, M.: Image blind denoising with generative adversarial network based noise modeling. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 3155–3164 2, 3
12. Chang, K.C., Wang, R., Lin, H.J., Liu, Y.L., Chen, C.P., Chang, Y.L., Chen, H.T.: Learning camera-aware noise models. In: European Conference on Computer Vision, Springer (2020) 343–358 2
13. Kim, D.W., Ryun Chung, J., Jung, S.W.: Grdn: Grouped residual dense network for real image denoising and gan-based real-world noise modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. (2019) 0–0 2
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014) 2, 4, 8
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 2, 4

16. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. arXiv preprint arXiv:2111.11418 (2021) 2, 4
17. Kwon, H.J., Yang, C.M., Kim, M.C., Kim, C.W., Ahn, J.Y., Kim, P.R.: Modeling of luminance transition curve of transparent plastics on transparent oled displays. Electronic Imaging **2016** (2016) 1–4 3
18. Zong, Qin, Wei-Yuan, Cheng, Han-Ping, David, Shieh, Yi-Pai, Huang, and, Y.H.: See-through image blurring of transparent oled display: Diffraction analysis and oled pixel optimization. SID International Symposium: Digest of Technology Papers **47** (2016) 393–396 3
19. Whyte, O., Sivic, J., Zisserman, A., Ponce, J.: Non-uniform deblurring for shaken images. International journal of computer vision **98** (2012) 168–186 3
20. Gupta, A., Joshi, N., Lawrence Zitnick, C., Cohen, M., Curless, B.: Single image deblurring using motion density functions. In: European conference on computer vision, Springer (2010) 171–184 3
21. Sun, J., Cao, W., Xu, Z., Ponce, J.: Learning a convolutional neural network for non-uniform motion blur removal. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 769–777 3
22. Xu, L., Ren, J.S., Liu, C., Jia, J.: Deep convolutional neural network for image deconvolution. Advances in neural information processing systems **27** (2014) 3
23. Chakrabarti, A.: A neural approach to blind motion deblurring. In: European conference on computer vision, Springer (2016) 221–235 3
24. Foi, A., Trimeche, M., Katkovnik, V., Egiazarian, K.: Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. IEEE Transactions on Image Processing **17** (2008) 1737–1754 3
25. Hasinoff, S.W., Durand, F., Freeman, W.T.: Noise-optimal capture for high dynamic range photography. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE (2010) 553–560 3
26. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward convolutional blind denoising of real photographs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 1712–1722 3, 4
27. Wei, K., Fu, Y., Yang, J., Huang, H.: A physics-based noise formation model for extreme low-light raw denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 2758–2767 3
28. Zhang, Y., Qin, H., Wang, X., Li, H.: Rethinking noise synthesis and modeling in raw denoising. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 4593–4601 3
29. Hong, Z., Fan, X., Jiang, T., Feng, J.: End-to-end unpaired image denoising with conditional adversarial networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 34. (2020) 4140–4149 3
30. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE transactions on image processing **26** (2017) 3142–3155 3
31. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 1125–1134 4, 8
32. Dong, H., Yu, S., Wu, C., Guo, Y.: Semantic image synthesis via adversarial learning. In: Proceedings of the IEEE international conference on computer vision. (2017) 5706–5714 4
33. Kaneko, T., Hiramatsu, K., Kashino, K.: Generative attribute controller with conditional filtered generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 6089–6098 4

34. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 4681–4690 4
35. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 2536–2544 4
36. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. Advances in neural information processing systems **30** (2017) 4
37. Lim, J.H., Ye, J.C.: Geometric gan. arXiv preprint arXiv:1705.02894 (2017) 4
38. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. (2017) 2794–2802 4, 8
39. Yue, Z., Zhao, Q., Zhang, L., Meng, D.: Dual adversarial network: Toward real-world noise removal and noise generation. In: European Conference on Computer Vision, Springer (2020) 41–58 4
40. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: Deblurgan: Blind motion deblurring using conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8183–8192 4
41. Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 8878–8887 4
42. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 3291–3300 4
43. Xia, Z., Gharbi, M., Perazzi, F., Sunkavalli, K., Chakrabarti, A.: Deep denoising of flash and no-flash pairs for photography in low-light environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 2063–2072 4
44. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 10012–10022 4
45. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 12299–12310 4
46. Wang, Z., Cun, X., Bao, J., Liu, J.: Uformer: A general u-shaped transformer for image restoration. arXiv preprint arXiv:2106.03106 (2021) 4
47. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. arXiv preprint arXiv:2204.04676 (2022) 4
48. Wu, Y., Johnson, J.: Rethinking" batch" in batchnorm. arXiv preprint arXiv:2105.07576 (2021) 4, 8
49. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 7132–7141 4, 5, 7
50. Schonfeld, E., Schiele, B., Khoreva, A.: A u-net based discriminator for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 8207–8216 4
51. Li, D., Zhang, H., Wang, Y.: Four-coordinate organoboron compounds for organic light-emitting diodes (oleds). Chemical Society Reviews **42** (2013) 8416–8433 5

52. Fu, Q., Di, X., Zhang, Y.: Learning an adaptive model for extreme low-light raw image processing. arXiv preprint arXiv:2004.10447 (2020) 5
53. Voelz, D.G.: Computational fourier optics: a MATLAB tutorial. SPIE press Bellingham, Washington (2011) 5
54. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778 5
55. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 1874–1883 6
56. Jang, G., Lee, W., Son, S., Lee, K.M.: C2n: Practical generative noise modeling for real-world denoising. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 2350–2359 6
57. Monakhova, K., Richter, S.R., Waller, L., Koltun, V.: Dancing under the stars: video denoising in starlight. arXiv preprint arXiv:2204.04210 (2022) 6
58. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 7
59. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8798–8807 8
60. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. (2017) 2223–2232 8
61. Cai, Y., Hu, X., Wang, H., Zhang, Y., Pfister, H., Wei, D.: Learning to generate realistic noisy images via pixel-level noise-aware adversarial training. Advances in Neural Information Processing Systems **34** (2021) 8
62. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9
63. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) 9
64. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. (2017) 126–135 9, 10
65. Zhou, Y., Kwan, M., Tolentino, K., Emerton, N., Lim, S., Large, T., Fu, L., Pan, Z., Li, B., Yang, Q., et al.: Udc 2020 challenge on image restoration of under-display camera: Methods and results. In: European Conference on Computer Vision, Springer (2020) 337–351 12
66. Yang, Q., Liu, Y., Tang, J., Ku, T.: Residual and dense unet for under-display camera restoration. In: European Conference on Computer Vision, Springer (2020) 398–408 12