BNUDC: A Two-Branched Deep Neural Network for Restoring Images from Under-Display Cameras

Jaihyun Koh[1,2], Jangho Lee[1] and Sungroh Yoon[1,3,4*]
[1] Department of Electrical and Computer Engineering, Seoul National University
[2] Samsung Display Corporation [3] Interdisciplinary Program in AI, Seoul National University
[4] AIIS, ASRI, INMC and ISRC, Seoul National University
{satyricon, ubuntu, sryoon}@snu.ac.kr

Abstract

The images captured by under-display cameras (UDCs) are degraded by the screen in front of them. We model this degradation in terms of a) diffraction by the pixel grid, which attenuates high-spatial-frequency components of the image; and b) diffuse intensity and color changes caused by the multiple thin-film layers in an OLED, which modulate the low-spatial-frequency components of the image. We introduce a deep neural network with two branches to reverse each type of degradation, which is more effective than performing both restorations in a single forward network. We also propose an affine transform connection to replace the skip connection used in most existing DNNs for restoring UDC images. Confining the solution space to the linear transform domain reduces the blurring caused by convolution; and any gross color shift in the training images is eliminated by inverse color filtering. Trained on three datasets of UDC images, our network outperformed existing methods in terms of measures of distortion and of perceived image quality.

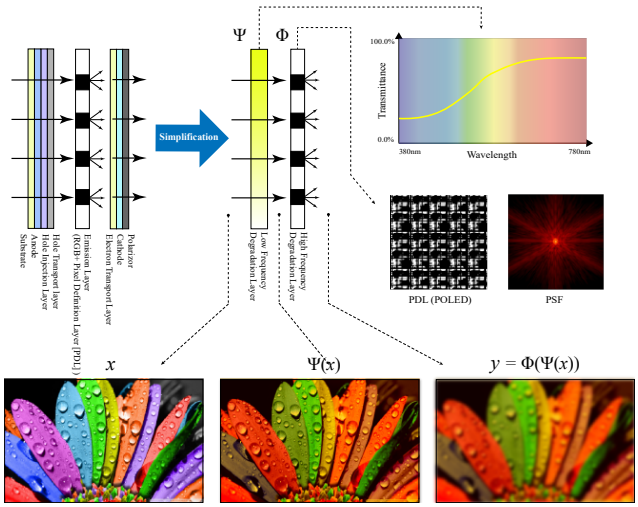

Figure 1. Physical model of the degradation of the image received by a UDC, simplified by two degradation processes: (1) color filtering and spatially variant attenuation Ψ by the thin-film layers of the OLEDs, and (2) diffraction Φ by the pixel definition layer (PDL), which can be represented by a point spread function (PSF). A UDC image y can be simulated by compositing the effect of Ψ and Φ on a latent image x.

1. Introduction

Under-display cameras (UDCs) are a key technology for realizing full-screen smartphones. Unfortunately, the quality of the images captured by a UDC is considerably reduced by the light loss and diffraction introduced by the display panel which is in front of it [8, 19, 40]. One way of addressing this problem is to increase the proportion of the panel which is transparent by reducing the pixel density in the region of the display immediately above the camera, and by modifying the layout of the RGB sub-pixels [26, 32, 37]. Increasing the transparent area reduces the resolution of the screen, and therefore distortion of the light reaching the UDC is unavoidable with a screen of any acceptable resolution; and the presence of the wires required to drive the pixels make this distortion worse. The natural limits on the extent to which an OLED display can be rearranged mean that restoration is required to make the images received by a UDC look like the images that would be received if the display was not in front of the UDC. Methods have recently been proposed based on paired image datasets [8, 21, 40] and image restoration [8, 16, 19, 27, 29, 33, 36, 39] which use a learning approach with a neural network.

Image restoration methods using deep neural networks (DNNs) have progressed substantially over the years: specific DNN architectures have been developed for many

*Correspondence to: Sungroh Yoon (sryoon@snu.ac.kr).

restoration tasks, such as super-resolution [22], deblurring [18], dehazing [10], and deraining [20]. The UDC presents a new image restoration task [39], which has received considerable attention, but most of the methods introduced so far use networks developed for other restoration tasks. Some methods [8, 19] do consider the physical processes that affect the image received by a UDC, in which the angle of the incident light is included in the inference process. However, in the same way that incorrect kernel estimation produces a severe artifact in deblurring techniques [18], errors in predicting the angle of incidence make this approach to restoring UDC images problematic.

In pioneering work on this topic [8, 19, 40], UDC degradation was modeled in terms of the diffraction and reduced intensity caused by display pixels. This model is effective in addressing the process that dominates image degradation, it only represents the reduction in high-spatial-frequency components of the image by diffraction, and does not consider the different transmission of wavelengths by the thin-film layers of an OLED, and long-range degradation caused by the non-uniformity of those layers, which is associated with the modulation of low-spatial-frequency components (see Fig. 1). We propose a physical model of UDC image degradation which includes low-spatial-frequency degradation processes, such as color shifts and spatial attenuation, and we introduce a DNN architecture to reverse the changes in the image predicted by our model. Whereas existing methods of UDC image restoration [8, 16, 19, 27, 29, 33, 36, 39] mainly involve a network performing a single deblurring task, we separate this task into high- and low-spatial-frequency reconstruction with two network branches. We induce each branch to deal with a difference range of frequencies. This branched network for UDC image restoration (BNUDC) effectively removes high-spatial-frequency noise such as 'flare' [8] and degradation with low spatial frequency, such as color shift. We also propose an affine transform connection which reduces over-smoothing by the convolution operation and preserves the structure of the image by constraining the solution space to the linear transform domain of the input image. In addition, we introduce inverse color filtering, which is a pre-processing technique that improves the color fidelity of the restoration of images with a severe color shift, such as those in the POLED dataset [40]. Simplifying the effect of the stacked thin films to that of a single color filter (see Fig. 1), allows inverse color filtering to be performed easily by inverting in the CIE XYZ color space. This is equivalent to the data normalization processes widely used in deep learning.

Overall, our contributions can be summarized as follows:

- We present a new model of the UDC degradation found in images captured by a UDC, which is specific to the optical properties of OLEDs. The model includes changes with a low spatial frequency, such as color shift and spatially variant attenuation. We propose a DNN architecture with one branch to restore high-frequency component, and a second branch to restore low-frequency components.
- We propose an affine transformation connection as an alternative to residual learning [11], This connection is specific to our model of image degradation. It removes noise introduced by restoration while preserving the structure of the image from the UDC.
- Our network achieves a-state-of-the-art performance in terms of both numerical and perceptual distortion metrics on three public datasets [8, 40]. Our network does not require a point-spread function (PSF) as a prior; but nevertheless it outperforms existing methods which are conditioned by a PSF.

2. Related Work

2.1. Restoring UDC images

UDC degradation model Zhou *et al.* [40] established the following model of the degradation of the image captured by a UDC:

$$y = \gamma(x * k) + e, \tag{1}$$

where x is the latent clean image, and y is the degraded image from the UDC, k is a blur kernel which represents the PSF by the pixel grid of displays, γ is the factor by which the intensity of the light is reduced, e is additive noise, and $*$ is the convolution operator. A similar model has been used in subsequent work [8, 19], except that the PSF varies with the incident angle of the light. Models of this sort represent the degradation of an image from a UDC in the simplest way, but a single scaling factor γ cannot model complicated forms of degradation with low spatial frequency, such as the spatial variation due to the OLED thin films. This requires a 2D transformation and a color shift is expressed as a 3D transformation in the color space.

Methods of restoring UDC images Supervised learning is the technique used for most restoration of UDC images, because datasets with paired observed and ground-truth images, either synthesized from a measured PSF [8] or created by capturing the same scene with a UDC and a surface mounted camera [40] are available. The first DNN architecture trained on such datasets was the scale-separated U-net architecture proposed by Zhou *et al.* [40]. Subsequent architectures include neural guided filter [29], latent image inference in the wavelet domain [27], residual dense networks [33], a weight-sharing multi-scale ResNet with channel attention, and a camera shading estimation module [39]. These methods use network modules commonly employed for image reconstruction, which are not specifically designed for UDC images. Kwon *et al.* [19] and Feng

et al. [8] noted that the profile of a long-tailed PSF varies with the angle of the incident light. To deal with this, the authors introduced a method of image restoration which can cope with multiple PSFs by conditioning the corresponding PSF on UDC images. The PSF prediction is a hard task, therefore a restored image may still be of low quality due to errors in predicting the PSF. We address this issue by getting the network to infer the latent PSF implicitly from the degraded image, rather than trying to predict it.

Separation of high- and low-frequency processes Methods of interpreting and reconstructing an image in the frequency domain are common in image processing [30]. In contrast, low-level machine vision algorithm, techniques based on learning mainly operate in the intensity domain. Methods which operate in the intensity domain, but on a number of frequency bands, represent a compromise. Examples of this approach include: multi-scale processing [24] in which several degraded images of different resolution are restored to approximate the same latent image. The feature pyramid [7] and octave convolution [2] use cross-frequency feature aggregation, which is another technique that operates in a 'pseudo frequency domain'. We take a similar approach to the restoration of UDC images.

Per-pixel transformations The simplest per-pixel transformation is residual learning using skip connections [11], and per-pixel filtering in which a latent image is estimated by filtering the observed image with a learned per-pixel kernel. This approach is used in classification [31] and image deblurring [38]. In per-pixel filtering, the DNN estimates a pseudo inverse kernel for each pixel, which allows it to deal with spatially variant blur. This method searches the dependencies of adjacent pixel, but this is unnecessary in the UDC image problem, because the PDL pattern is not spatially variant. We therefore propose a per-pixel affine transformation to deal with the coexisting high- and low-frequency degradation which is formed in a UDC image.

3. Method

3.1. Modeling the Degradation in UDC Images

We now put forward a model of the degradation of the image captured by a UDC in terms of the optical properties of the OLED display in front of the camera. As shown in Fig. 1, the layers of an OLED cell include transparent electrodes, the injection and transport layers of the hole and electron, and a light-emission layer which contains organic materials. Current OLED screens designed to accommodate for UDCs have a relatively large transparent area in which light loss is reduced [8,32], but all the layers are still present in the transparent area, except the organic materials which emit light, and the spectrum of the incident light is modulated by these layers. In addition, there are metal wires for electrical signals which are covered by black regions of the pixel definition layer (PDL). The PDL has a repeating pattern and causes diffraction. We model the spectrum modulation and diffraction separately. The effect of the thin-film layers, excluding PDL, is simplified to a single spatially variant color filter. Its effect is expressed as follows:

$$i_{\mathrm{cfl}}(m,n,\lambda) = i_{\mathrm{org}}(m,n,\lambda)\phi(m,n,\lambda), \tag{2}$$

where $i_{\mathrm{org}}(m,n,\lambda)$ is the spectrum of the incident light at the location (m,n) on the sensor of the UDC, $\phi(m,n,\lambda)$ is the transmittance of wavelength λ by the color filter layer, and $i_{\mathrm{cfl}}(m,n,\lambda)$ is the spectrum of the light transmitted through the color filter in the simplified model of the display layers. Secondly, the diffraction of light depends on the PSF which is determined by the pattern of the PDL [32,37]. Assuming that the PSF does not change with wavelength, it can be expressed as the convolution of the spectral intensity and a kernel representing the PSF:

$$i_{\mathrm{pdl}}(m,n,\lambda) = \iint i_{\mathrm{cfl}}(m-\tau_1, n-\tau_2, \lambda)\psi(\tau_1,\tau_2)d\tau_1 d\tau_2, \tag{3}$$

where $i_{\mathrm{pdl}}(\cdot)$ is the spectrum of the light transmitted through the PDL, and $\psi(\cdot)$ is a 2D kernel representing the PSF. The spectrum of the light passing through the display is converted to the CIE XYZ color space [6] as follows:

$$\begin{aligned} X(m,n) &= \int i_{\mathrm{pdl}}(m,n,\lambda)S_X(\lambda)d\lambda, \\ Y(m,n) &= \int i_{\mathrm{pdl}}(m,n,\lambda)S_Y(\lambda)d\lambda, \\ Z(m,n) &= \int i_{\mathrm{pdl}}(m,n,\lambda)S_Z(\lambda)d\lambda, \end{aligned} \tag{4}$$

where $X, Y, Z(m,n)$ are the color coordinates in the XYZ color space corresponding the pixel (m,n), and S_X, S_Y and S_Z are color-matching functions [6]. The expected RGB color coordinates at each pixel of the image $y(m,n)$ captured by the camera can now be obtained using the linear transformation matrix $\mathbf{M} \in \mathbb{R}^{3\times 3}$ [15] as $y(m,n) = \mathbf{M}[X,Y,Z]^{\mathrm{T}}_{m,n}$.

We can now formulate a simplified model of the degradation processes described above as $y = (\Phi \circ \Psi)(x) + e$, where x is the latent clean image, the function $\Psi(\cdot)$ represents equation (2), the function $\Phi(\cdot)$ expresses the cumulative effect of equations (3)-(4). Given a degraded image y, the problem of finding an estimated latent clean image $\hat{x}$ requires the inversion of Ψ and Φ, as follows:

$$\hat{x} = (\Psi^{-1} \circ \Phi^{-1})(y). \tag{5}$$

We will now examine the nature of Ψ^{-1} and Φ^{-1}, and introduce a neural network architecture to approximate them.

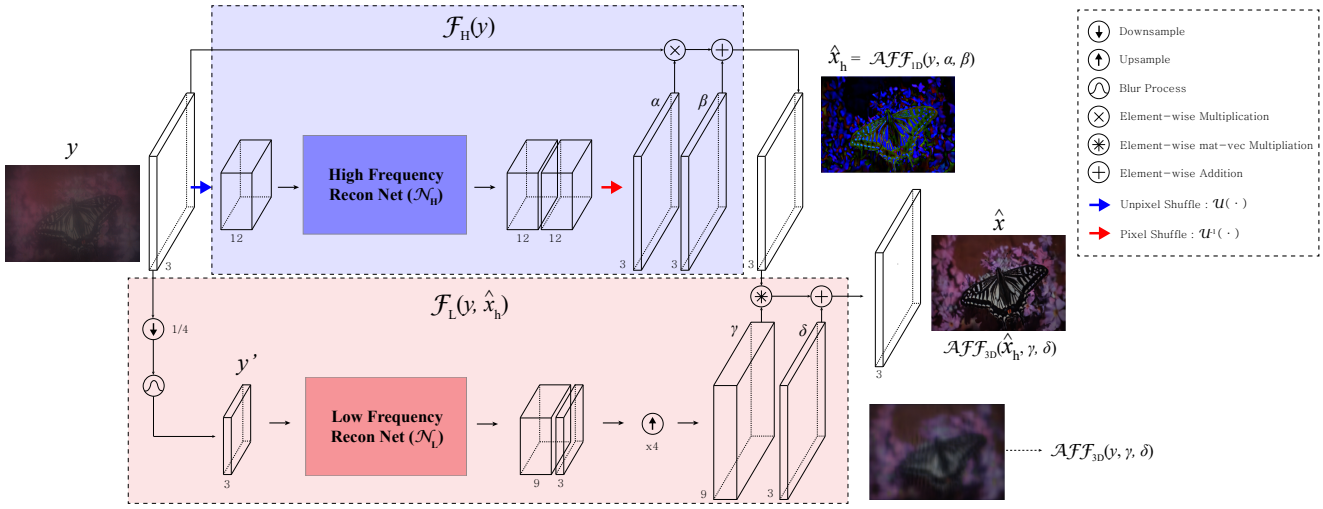

Figure 2. The proposed BNUDC is composed of (1) a branch $\mathcal{F}_{\mathrm{H}}$ that reverses high-spatial-frequency degradation by the PDL layer Φ and (2) a branch $\mathcal{F}_{\mathrm{L}}$ that reverses low-spatial-frequency degradation by other thin films Ψ . The two networks are connected in series by the affine transformation connection.

3.2. The Inverse Transformation Functions

The transform Φ in (5) can be regarded as the blurring of an original image by a long-tailed PSF; thus the computation of Φ^{-1} is effectively a deblurring problem. Deblurring is ill-posed and the reconstruction of sharp edges and high-frequency components in the Fourier domain is challenging [18]. The transformation Ψ can be regarded as a color filtering operation, thus Ψ^{-1} is an inverse color filter. If the filter function Ψ is spatially uniform, then inverse filtering is trivial. However, Ψ is spatially variant due to the non-uniform color of the display. This non-uniform degradation of color can be viewed as a corruption of the low-spatial-frequency components of the image.

3.3. Branched Network Architecture

Our BNUDC approximate $\Psi^{-1} \circ \Phi^{-1}$ in (5) by means of two branches: a high-frequency reconstruction (HFR) branch $\mathcal{F}_{\mathrm{H}}$ models Φ^{-1}, and a low-frequency reconstruction (LFR) branch $\mathcal{F}_{\mathrm{L}}$ models Ψ^{-1}. The branches $\mathcal{F}_{\mathrm{H}}$ and $\mathcal{F}_{\mathrm{L}}$ are cascaded as follows:

$$\hat{x} = \mathcal{F}_{\mathrm{L}}(y, \mathcal{F}_{\mathrm{H}}(y)). \tag{6}$$

Connecting the LFR and HFR networks in series, as represented by $\hat{x} = \mathcal{F}_{\mathrm{L}}(\mathcal{F}_{\mathrm{H}}(y))$, is the most straightforward and direct implementation of (5). However, in this formulation, the outer function $\mathcal{F}_{\mathrm{L}}$ does not receive the input image y, but an intermediate image from the inner function $\mathcal{F}_{\mathrm{H}}$. This does not allow the LFR network used to realize the outer function $\mathcal{F}_{\mathrm{L}}$ to operate on low spatial frequencies. Moreover, such an architecture makes it difficult to ensure that the HFR network deals with high-frequency components of the image, and the LFR network deals with low-frequency components because this architecture is eventually a feed-forward network with a single branch. This is one source of the poor performance often associated with sequentially connected networks. To avoid this pitfall, we arranged for both of the parallel network branches to receive the input image. The outputs of these networks are then connected in series to realize the sequential process described by (5). Fig. 2 shows this branched network, and its components are presented in more detail in the Supplementary material.

High-frequency reconstruction branch Reconstructing UDC images degraded by diffraction in the PDL requires a network that concentrates on the sharp edges remaining in the degraded image. To preserve the edge information in the degraded image, we use the flat network which maintains the resolution of the input image in the feature space [34] instead of using a bottleneck architecture such as that used for encoder-decoders, which down-scales the resolution of feature maps. The implementation of this architecture as a full-sized network would place a large computational burden on memory during training and inference. A pixel shuffle scheme [28] is therefore used to reduce the spatial dimension which reduces the use of memory without a loss of input information. As in the super-resolution process in which the restoration of sharp edges is the main task, the high-frequency network requires a wide receptive field. We therefore use a parallel dilated convolution residual block [1, 29].

A CNN with a large number of layers, performing a long series of convolution operations, which is likely to cause blurring of edge features. When this occurs at the beginning

Table 1. Average PSNR, SSIM, LPIPS, and DISTS for the images in the POLED dataset: the best and second-best result for each metric is highlighted in blue and red respectively. Up-arrows mean that higher values are better, and down-arrow means the opposite.

POLED	PARAM	IT(s)	TEST SET				VALIDATION SET			
			PSNR↑	SSIM↑	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
MSUNET [40]	8.9M	0.08	29.17	0.9393	0.2239	0.1746	29.96	0.9343	0.2281	0.1774
DAGF [29]	1.1M	1.12	32.29	0.9509	0.2163	0.1913	33.79	0.9580	0.2250	0.1942
PDCRN [27]	4.7M	0.08	32.99	0.9578	0.2102	0.2075	33.58	0.9593	0.2188	0.2121
BNUDC	4.6M	0.08	33.39	0.9610	0.1748	0.1511	34.39	0.9634	0.1871	0.1612

Table 2. Average PSNR, SSIM, LPIPS, and DISTS for the images in the TOLED dataset: the best and second-best result for each metric is highlighted in blue and red respectively. Up-arrows mean that higher values are better, and down-arrow means the opposite. The † symbol indicates that the figure quoted is the number of parameters that we required to reproduce the scores reported in the original paper.

TOLED	PARAM	IT(s)	TEST SET				VALIDATION SET			
			PSNR↑	SSIM↑	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
MSUNET [40]	8.9M	0.08	37.40	0.9756	0.1093	0.1052	38.25	0.9772	0.1174	0.1155
IPIUer [39]	24.7M†	0.10	38.18	0.9796	0.1128	0.1050	39.03	0.9813	0.1210	0.1145
BAIDU [39]	20.0M†	0.18	38.23	0.9803	0.1026	0.0966	39.06	0.9812	0.1108	0.1057
BNUDC	4.6M	0.08	38.22	0.9798	0.0988	0.0964	39.09	0.9814	0.1072	0.1052
BNUDC-1D	4.6M	0.08	38.26	0.9800	0.1007	0.0942	39.12	0.9814	0.1086	0.1034

Table 3. Average PSNR, SSIM, LPIPS, and DISTS for synthetic images in SYNTH dataset. BNUDC-S is a version of our BNUDC with fewer parameters, to permit a more direct comparison with SFTMD and DISC.

SYNTH	PARAM	TEST SET			
		PSNR↑	SSIM↑	LPIPS↓	DISTS↓
SFTMD [9]	3.9M	42.35	0.9863	0.0123	-
DISC (w/ PSF) [8]	3.8M	43.27	0.9877	0.0108	0.0182
DISC (w/o PSF)	-	42.77	0.9870	-	-
BNUDC (w/o PSF)	4.6M	45.78	0.9942	0.0106	0.0150
BNUDC-S	3.5M	45.33	0.9939	0.0108	0.0158

of training, it is difficult to obtain a sharp image. We therefore introduce a pixel-wise affine transform connection. The HFR network can now be expressed as follows:

$$\alpha, \beta = \mathcal{U}^{-1}(\mathcal{N}_{\mathrm{H}}(\mathcal{U}(y))), \tag{7}$$

where $y \in \mathbb{R}^{h\times w\times 3}$ is a UDC image in the RGB color space which has $h \times w$ resolution. $\alpha \in \mathbb{R}^{h\times w\times 3}$ and $\beta \in \mathbb{R}^{h\times w\times 3}$ are the element-wise gain and bias inferred by the HFR network $\mathcal{N}_{\mathrm{H}}(\cdot)$. The functions $\mathcal{U}(\cdot)$ and $\mathcal{U}^{-1}(\cdot)$ express a unpixel shuffle [28] and its inverse. The 1D affine transform $\mathcal{AFF}_{1\mathrm{D}}$ is expressed by the following pixel-wise linear transform: $\hat{x}_{\mathrm{h}} = \mathcal{AFF}_{1\mathrm{D}}(y, \alpha, \beta) = y \cdot \alpha + \beta$, where $\hat{x}_{\mathrm{h}} \in \mathbb{R}^{h\times w\times 3}$ is an intermediate result from the HFR branch, and $\cdot$ is element-wise multiplication. The HFR network $\mathcal{N}_{\mathrm{H}}$ infers pixel-wise gain and bias, and the output image is estimated by a pixel-wise linear transformation of the degraded image y using these gain and bias values.

Low-frequency reconstruction branch $\mathcal{F}_{\mathrm{L}}$ is a data pipeline that reverses low-frequency degradation. It consists of an encoder-decoder network which has a bottleneck structure so that it only compresses the relevant low-frequency features. First, the UDC image y is down-scaled by a factor of four to remove high-frequency features. High-frequency components that remain in this reduced-resolution image are further removed by blurring with a 3x3 box filter. The resulting image y' lacks edges and other high-frequency components, allowing the network to focus on the remaining low-frequency features. The proposed LFR network $\mathcal{N}_{\mathrm{L}}$ is shown in Fig. 2, and its effect can be expressed as follows:

$$\gamma, \delta = \mathcal{N}_{\mathrm{L}}(y') \uparrow, \tag{8}$$

where $\gamma \in \mathbb{R}^{h\times w\times 9}$ and $\delta \in \mathbb{R}^{h\times w\times 3}$ are inferred pixel-wise matrix elements and translator for 3D affine transform, and the up-arrow expresses up-scaling. A 3D affine transfrom of the intermediate result $\hat{x}_{\mathrm{h}}$ with γ and δ is performed to correct color shift and other low-frequency degradation as $\hat{x} = \mathcal{AFF}_{3\mathrm{D}}(\hat{x}_{\mathrm{h}}, \gamma, \delta)$, where $\hat{x}$ is the transformed image, and the 3D affine transform $\mathcal{AFF}_{3\mathrm{D}}$ is expressed by pixel-wise matrix-vector multiplication and translation, as follows: $\hat{\mathbf{x}}_{m,n} = \mathbf{G}_{m,n}\mathbf{x}_{m,n} + \mathbf{d}_{m,n}$, where $\mathbf{x}_{m,n} \in \mathbb{R}^3$ and $\hat{\mathbf{x}}_{m,n} \in \mathbb{R}^3$ are color vectors in RGB space corresponding to the pixel (m, n), in the image $\hat{x}_{\mathrm{h}}$ and $\hat{x}$ respectively. $\mathbf{G}_{m,n} \in \mathbb{R}^{3\times 3}$ is a 3x3 matrix which is a rearrangement of $\gamma(m, n, :) \in \mathbb{R}^9$, and $\mathbf{d}_{m,n} \in \mathbb{R}^3$ is a vector of $\delta(m, n, :)$. This transformation can be viewed as a regularizer which induces the network to perform the color reconstruction. Experimental results are presented in the ablation study.

Figure 3. Two example images from the POLED dataset restored by four different networks. The first column contains the original UDC images. The next four columns shows the reconstructed images obtained using MSUNET [40], DAGF [29], PDCRN [27] and our BNUDC. The last column contains the ground truth images.

Figure 4. Two example images from the TOLED dataset restored by four networks. The first column contains the UDC images. Subsequent columns shows the reconstructed images obtained using MSUNET [40], IPIUer [39], BAIDU [39] and our BNUDC.

Considering high-frequency information in terms of intensity alone makes it more likely that the HFR network will produce an accurate reconstruction of that information. The LFR network takes a 'broad-brush' low-frequency approach to restoring the chrominance information. This separation accords with the relative sensitivity of the human eye to intensity and color [4, 13].

4. Experiments

4.1. Datasets

We evaluated the performance of our network after training on three public datasets: POLED and TOLED [40], which are realistic datasets captured using an RGBG PenTile [5] and a transparent OLED respectively, and SYNTH [8] which is a set of synthetic images generated by blur kernels based on the measured PSFs of a commercial UDC smartphone. The UDC images in the SYNTH dataset are labeled with ground-truth PSFs. Therefore DISCnet [8] trained on this dataset uses the PSFs as an external prior, whereas our network does not require a PSF prior to make it robust against PSF prediction errors. The images in the SYNTH dataset are degraded only based on the measured PSFs, without color shifts but with other low-frequency degradation, therefore we use this dataset to test the LFR branch in such a case.

Image pre-processing The severe color shift found in the POLED dataset can be normalized by a uniform inverse filter η_{xyz}, which can be obtained by comparing the average of the UDC and ground-truth images in XYZ space. This pre-processing can be expressed by

$$\eta_{\text{xyz}} = \frac{1}{N}\sum_{i=1}^{N}\frac{y^i_{\text{xyz}}}{x^i_{\text{xyz}}}, \quad \tilde{y}^i_{\text{xyz}} = \frac{y^i_{\text{xyz}}}{\eta_{\text{xyz}}}, \tag{9}$$

where $\{x^i_{\text{xyz}}, y^i_{\text{xyz}}\}_{i=1}^{N}$ are pairs of training images in XYZ color space, and $\tilde{y}^i_{\text{xyz}}$ is the i^{th} pre-processed image. This global color filtering can reverse a significant color distortion: the average PSNR of images in the training dataset is improved from 15.59dB to 20.02dB. Example images are shown in the first row in Fig. 6. Refer to the Supplementary material for more details.

4.2. Training Procedure

Data augmentation is performed by random horizontal and vertical flipping and Gaussian noise with a standard de-

Figure 5. Three example images from the SYNTH dataset. The first column contains the original UDC images. The next two columns show the reconstructed images restored using DISCnet [8] and our BNUDC.

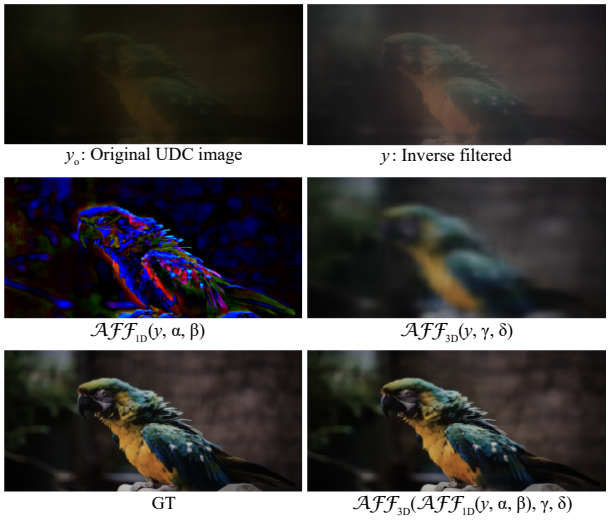

Figure 6. Visualizations of intermediate images from the HFR and LFR branches on an example from the POLED dataset. The top row shows the captured image (left) and the same image after preprocessing (right); the second row show the results from the HFR (left) and the LFR network (right); and the last row contains the ground-truth image (left) and the restored image (right).

viation of 1×10^{-3}. Flipping is not applied to the SYNTH dataset to preserve the variance of the PSF. The logarithm of squared error is used as the loss function for the POLED and TOLED datasets to penalize small errors, and L1 with perceptual loss [14] is used for the SYNTH dataset. For training we used the Adam optimizer [17] with β_1 set to 0.9 and β_2 set to 0.999, and with an initial learning-rate of 2×10^{-4}, which is reduced to 1×10^{-6} using gradual warm-up cosine annealing [23]. All the networks were trained on

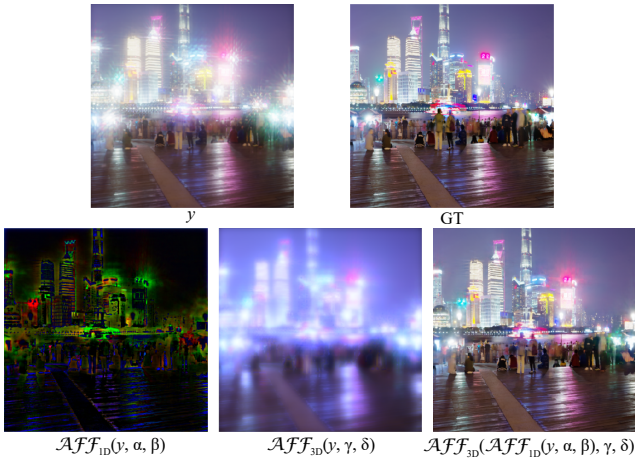

Figure 7. Visualizations of intermediate images from the HFR and LFR branches on SYNTH dataset. The top row shows the UDC image (left) and the ground-truth image (right), the second row contains the results from the HFR (left), the LFR network (middle), and the restored image (right).

patches taken from full-resolution images with a batch size of 2 over 3×10^5 iterations. We implemented our BNUDC in the PyTorch framework [25] and trained the model using two NVIDIA V100 GPUs.

4.3. Results

We used PSNR to measure pixel-wise distance and SSIM to measure structural similarity [12]. We also used LPIPS [35] and DISTS [3] to measure perceptual difference. We performed comparative experiments using the publicly available DAGF [29], DISCnet [8] codes, and re-implemented models for PDCRN [27], IPIUer, and BAIDU [39], which are not publicly available. The MSUNET [40] used images of 16-bit color depth to train, but the images in the public dataset have a color depth of 8 bits. We therefore implemented a network model which can use the 8bits images.

POLED dataset Table 1 compares the performance of four trained networks on the restoration of 2k resolution images from this dataset. Our network outperforms the other those, especially in terms of perceptual quality, as illustrated by Fig. 3. Fig. 6 shows intermediate results from the HFR and LFR branches. As expected (see Section 3.3), the HFR branch restores the edges, largely in terms of luminance, while the LFR network restores color in a more general fashion. The 1D and 3D affine transforms that we use differ from those in a standard CNN architecture; but any increase in computation time is insignificant due to the parallel tensor processing using the GPUs.

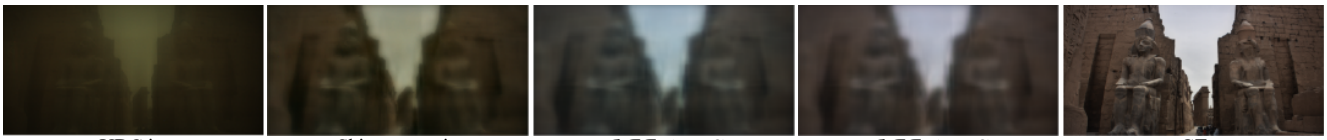

Figure 8. Task separability. The first column is the UDC image, the second shows images obtained with a skip connection in the LFR branch (corresponding to the third row in Table 4); the third shows images obtained using a 1D affine transform (corresponding to the fourth row in Table 4); and the fourth shows images obtained using a 3D affine transform (corresponding to the fifth row in Table 4).

Table 4. Results of the ablation study: the first row shows PSNR, SSIM, LPIPS and DISTS values for a single-branch network with a skip connection. This is then augmented by a 1D affine transform, with the results shown in the second row. The third row shows results for a dual-branch network with a skip connection for the LFR and a 1D affine transform in the HFR branch. This is then augmented by a 1D affine transform or by a 3D affine transform for the LFR branch, with the results shown in the fourth and fifth rows. The final row shows results for the dual-branch and 3D affine transform configuration with inverse color filtering of the images in the training dataset.

POLED	TEST SET			
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
Single Branch (skip)	32.68	0.9530	0.1961	0.1857
+ 1D Affine ($\mathcal{N}_H$)	32.70	0.9546	0.1932	0.1815
Dual Branch (skip)	33.19	0.9606	0.1773	0.1571
+ 1D Affine ($\mathcal{N}_L$)	33.27	0.9608	0.1767	0.1547
+ 3D Affine ($\mathcal{N}_L$)	33.32	0.9608	0.1752	0.1531
+ Inverse color filter	33.39	0.9610	0.1748	0.1511

TOLED dataset Table 2 compares the performance of four trained networks on the restoration of 2k resolution images from this dataset. The performance of the BNUDC-1D network which uses a 1D affine transform in its LFR branch is slightly better than that of the basic BNUDC network in terms of image distortion. This can be attributed to a smaller color shift in the TOLED images compared to those from POLED images.

SYNTH dataset The images in this dataset have no color shift because they have been synthesized with PSFs considering only luminances. However, these images do exhibit flare, which is a significant issue with UDCs, because high-frequency noise is typically introduced during its elimination. Our network is effective on this task, because the LFR branch is able to compensate for the noise introduced by the HFR brach, resulting in perceptually pleasing images. Quantitative results are presented in Table 3 and example images are shown in Fig. 5. Fig. 7 shows intermediate results on images from the SYNTH dataset. In this case, the LFR branch does not seem to try to match the color as it does on POLED images, Instead, the LFR network eliminates the effect of over-estimation of the flare artifact by the HFR branch. Our network does not need prior PSFs and thus potential errors in predicting the PSFs are eliminated.

4.4. Ablation

We analyzed the effectiveness of each component in the BNUDC against a baseline configuration of a single branch network with a skip connection (1st row in Table 4). In the single branch networks, the inter-feature depth is increased so that it has a similar number of parameters to the dual branch network. The dual-branch network increases PSNR by +0.49dB and the 3D affine transformation adds a further +0.23dB. As shown in Fig. 8, the use of a 3D affine transform produces more accurate colors in the LFR branch. This suggests that the use of a 3D affine transform in the LFR branch allows the two branches to play more distinct roles. Based on these results, either a 1D or 3D affine transform could be used in the LFR branch, depending on the color shift produced by the display panel.

5. Conclusions

We have proposed a two-branch network that effectively restores images captured by an under-display camera. This branched neural network architecture achieves state-of-the-art performance on this restoration task. Restoration of natural images from the SYNTH dataset produces result with 46dB of PSNR, so that the restored images cannot be visually distinguished from the ground-truth images. However, high-frequency artifacts, which are perceptually annoying, remain in restored images in which there is flare caused by strong lights. Also, the restoration of images from the POLED dataset, which exhibit large shift in colors and luminance, lack detail. These problem images require further work.

Acknowledgements: This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)], AIRS Company in Hyundai Motor and Kia through HMC/KIA-SNU AI Consortium Fund, and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2022.

References

[1] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1375–1383. IEEE, 2019. 4

[2] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3435–3444, 2019. 3

[3] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728*, 2020. 7

[4] Emil Dumic, Mario Mustra, Sonja Grgic, and Goran Gvozden. Image quality of 4 2 2 and 4 2 0 chroma subsampling formats. In *2009 International Symposium ELMAR*, pages 19–24. IEEE, 2009. 6

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

[6] Mark D Fairchild. *Color appearance models*. John Wiley & Sons, 2013. 3

[7] Yuchen Fan, Jiahui Yu, Ding Liu, and Thomas S Huang. Scale-wise convolution for image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10770–10777, 2020. 3

[8] Ruicheng Feng, Chongyi Li, Huaijin Chen, Shuai Li, Chen Change Loy, and Jinwei Gu. Removing diffraction image artifacts in under-display camera via dynamic skip connection network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 662–671, 2021. 1, 2, 3, 5, 6, 7

[9] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019. 5

[10] Jie Gui, Xiaofeng Cong, Yuan Cao, Wenqi Ren, Jun Zhang, Jing Zhang, and Dacheng Tao. A comprehensive survey on image dehazing based on deep learning. *arXiv preprint arXiv:2106.03323*, 2021. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3

[12] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 7

[13] David H Hubel. *Eye, brain, and vision*. Scientific American Library/Scientific American Books, 1995. 6

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 7

[15] Henry R Kang. *Computational color technology*. SPIE press Bellingham, WA, 2006. 3

[16] Daewook Kim, Jahoon Koo, Jewon Yoo, Hyun-joo Hwang, Sujin Choi, Eunkyung Koh, Hyunguk Cho, Seungin Baek, Yongjo Kim, and Jaebum Cho. 18-4: Fast image restoration and glare removal for under-display camera by guided filter. In *SID Symposium Digest of Technical Papers*, volume 52, pages 222–223. Wiley Online Library, 2021. 1, 2

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[18] Jaihyun Koh, Jangho Lee, and Sungroh Yoon. Single-image deblurring with neural networks: A comparative survey. *Computer Vision and Image Understanding*, 203:103134, 2021. 2, 4

[19] Kinam Kwon, Eunhee Kang, Sangwon Lee, Su-Jin Lee, Hyong-Euk Lee, ByungIn Yoo, and Jae-Joon Han. Controllable image restoration for under-display camera in smartphones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2073–2082, 2021. 1, 2

[20] Siyuan Li, Wenqi Ren, Feng Wang, Iago Breno Araujo, Eric K Tokuda, Roberto Hirata Junior, Roberto M Cesar-Jr, Zhangyang Wang, and Xiaochun Cao. A comprehensive benchmark analysis of single image deraining: Current challenges and future perspectives. *International Journal of Computer Vision*, 129(4):1301–1322, 2021. 2

[21] 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. 1

[22] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Blind image super-resolution: A survey and beyond. *arXiv preprint arXiv:2107.03055*, 2021. 2

[23] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 7

[24] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 3

[25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 7

[26] Zong Qin, Ruijin Qiu, Minyi Li, Xinyi Yu, and Bo-Ru Yang. P-78: Simulator-based efficient panel design and image retrieval for under-display cameras. In *SID Symposium Digest of Technical Papers*, volume 52, pages 1372–1375. Wiley Online Library, 2021. 1

[27] Hrishikesh Panikkasseril Sethumadhavan, Densen Puthussery, Melvin Kuriakose, and Jiji Charangatt Victor. Transform domain pyramidal dilated convolution networks for restoration of under display camera images. In *European Conference on Computer Vision*, pages 364–378. Springer, 2020. 1, 2, 5, 6, 7

[28] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 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*, pages 1874–1883, 2016. 4, 5

[29] Varun Sundar, Sumanth Hegde, Divya Kothandaraman, and Kaushik Mitra. Deep atrous guided filter for image restoration in under display cameras. In *European Conference on Computer Vision*, pages 379–397. Springer, 2020. 1, 2, 4, 5, 6, 7

[30] David CC Wang, Anthony H Vagnucci, and Ching-Chung Li. Digital image enhancement: a survey. *Computer vision, graphics, and image processing*, 24(3):363–381, 1983. 3

[31] Ze Wang, Zichen Miao, Jun Hu, and Qiang Qiu. Adaptive convolutions with per-pixel dynamic filter atom. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12302–12311, 2021. 3

[32] Anqi Yang and Aswin Sankaranarayanan. Designing display pixel layouts for under-panel cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 3

[33] Qirui Yang, Yihao Liu, Jigang Tang, and Tao Ku. Residual and dense unet for under-display camera restoration. In *European Conference on Computer Vision*, pages 398–408. Springer, 2020. 1, 2

[34] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14821–14831, 2021. 4

[35] 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

[36] Zhenhua Zhang. 74-3: Image deblurring of camera under display by deep learning. In *SID Symposium Digest of Technical Papers*, volume 51, pages 1109–1112. Wiley Online Library, 2020. 1, 2

[37] Zhenhua Zhang. 14.4: Diffraction simulation of camera under display. In *SID Symposium Digest of Technical Papers*, volume 52, pages 93–96. Wiley Online Library, 2021. 1, 3

[38] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2482–2491, 2019. 3

[39] Yuqian Zhou, Michael Kwan, Kyle Tolentino, Neil Emerton, Sehoon Lim, Tim Large, Lijiang Fu, Zhihong Pan, Baopu Li, Yang Qirui, Yihao Liu, Jigang Tang, Tao Ku, Shibin Ma, Bingnan Hu, Jiarong Wang, Densen Puthussery, Hrishikesh P S, Melvin Kuriakose, and Lianping Xing. Udc 2020 challenge on image restoration of under-display camera: Methods and results. 08 2020. 1, 2, 5, 6, 7

[40] Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, and Timothy Large. Image restoration for under-display camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9179–9188, 2021. 1, 2, 5, 6, 7