Decoupling Degradations with Recurrent Network for Video Restoration in Under-Display Camera

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

Abstract

Under-display camera (UDC) systems are the foundation of full-screen display devices in which the lens mounts under the display. The pixel array of light-emitting diodes used for display diffracts and attenuates incident light, causing various degradations as the light intensity changes. Unlike general video restoration which recovers video by treating different degradation factors equally, video restoration for UDC systems is more challenging that concerns removing diverse degradation over time while preserving temporal consistency. In this paper, we introduce a novel video restoration network, called D^{2}RNet, specifically designed for UDC systems. It employs a set of Decoupling Attention Modules (DAM) that effectively separate the various video degradation factors. More specifically, a soft mask generation function is proposed to formulate each frame into flare and haze based on the diffraction arising from incident light of different intensities, followed by the proposed flare and haze removal components that leverage long- and short-term feature learning to handle the respective degradations. Such a design offers an targeted and effective solution to eliminating various types of degradation in UDC systems. We further extend our design into multi-scale to overcome the scale-changing of degradation that often occur in long-range videos. To demonstrate the superiority of D^{2}RNet, we propose a large-scale UDC video benchmark by gathering HDR videos and generating realistically degraded videos using the point spread function measured by a commercial UDC system. Extensive quantitative and qualitative evaluations demonstrate the superiority of D^{2}RNet compared to other state-of-the-art video restoration and UDC image restoration methods.

Introduction

The rising popularity of full-screen mobile devices has driven the development of under-display camera (UDC) systems. While research on UDC primarily focuses on single image restoration (Feng et al. 2022; Zhou et al. 2020), few works are available on video restoration, which impedes the popularity of devices with UDC systems. UDC system is an imaging system where the lens is mounted under the display. It can eliminate the screen notch of the traditional front camera in mobile devices, providing a bezel-less display without disrupting the screen's integrity (Qin et al. 2021).

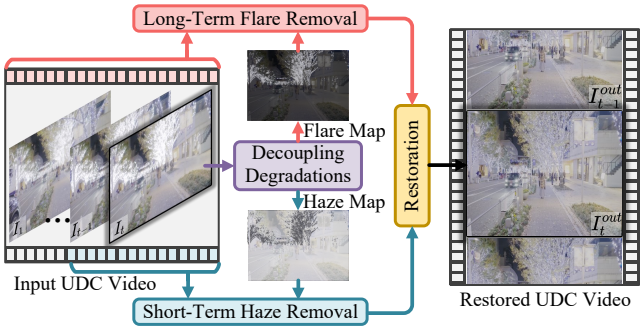

Figure 1: Method illustration. In UDC systems, the degree of degradation is positively correlated with the intensity of incident light. Our method decouples the degradation into brighter flare and darker haze, which are handled using information from long and short distances, respectively.

In contrast to conventional cameras, during UDC imaging, the incident light will cross the densely arranged organic light-emitting diodes (OLEDs) used for display before arriving at the lens. It implies that incident light is diffracted when propagating the aperture between the OLEDs, especially when the wavelength of the light is similar to the gaps between the obstacles (Zhou et al. 2021) (illustrated by Fig. 2(a)). Besides, the degree of degradation arising from diffraction is positively correlated with the intensity of incident light (Kwon et al. 2021). As depicted in Fig. 1, in brighter regions close to the light source, diffraction causes **flare** that saturates one or more channels of the image, resulting in content loss. In contrast, in other darker regions, diffraction causes **haze** that makes the content fuzzy.

To solve these challenges, many efforts have been devoted to handle image restoration for UDC through deep learning-based models. These works can be categorized into two paradigms. Some attempts to leverage the prior knowledge of the diffraction blur kernel, *i.e.*, point spread function (PSF) illustrated in Fig. 2(b), to guide the removal of diffraction (Feng et al. 2021; Kwon et al. 2021; Liu et al. 2022c). Another part directly learns diffraction removal through complex network design (Feng et al. 2023; Koh, Lee, and Yoon 2022; Liu et al. 2023a). Unfortunately, unlike UDC images, diffraction will change dynamically with motion in

UDC video. Therefore, existing methods for images are unable to take advantage of the strong temporal coherence of diffraction over time, leading to poor performance.

From a methodology perspective, unlike image restoration which only learns on spatial dimensions, video restoration pays more attention to exploiting temporal information. Existing video restoration methods either align features of adjacent frames (*e.g.*, 5 or 7 frames) through a sliding window input mechanism (Liang et al. 2022a), or learn features from the more distant frame through a recurrent mechanism (Liang et al. 2022b). Among them, benefiting from the long-term modeling capability of recurrent mechanisms, significant progress has been made in video super-resolution (Liu et al. 2022a), deblurring (Zhong et al. 2023), and denoise (Tassano, Delon, and Veit 2020) tasks. For UDC videos with complex degradations, the distant frames have more content differences but help recover lost content through the recurrent mechanism (Liu et al. 2022a; Chan et al. 2022; Liu et al. 2023c) (*i.e.*, eliminate flare). Exploiting more similar scene patterns from adjacent frames in spatio-temporal neighborhood is essential for recovering clear content (Wang et al. 2022; Lin et al. 2022; Zhang, Xie, and Yao 2022) (*i.e.*, eliminate haze). Therefore, a more promising solution is to explore proper network with long- and short-term video representation learning to effectively and pertinently eliminate various degradations in UDC video.

In this paper, we propose a novel UDC video restoration network to enable effective video representation learning (D^2RNet). The key idea of D^2RNet is to decouple the degradations in UDC videos while recovering them separately with different features pertinently (as shown in Fig 1). To achieve this, we propose a decoupling attention module (DAM) in conjunction with a globally multi-scale bidirectional recurrent framework. In particular, a soft mask generation function is used to partition each frame into flare and haze regions, which are produced by the diffraction of strongly and weakly incident light, respectively. For the flare region, a flare removal component learns long-term features to recover the content loss. For the haze region, a haze removal component learns short-term features to recover content fuzzy. DAM is extended to three scales to overcome the scale-changing of degradation in long-range UDC videos. Besides, for evaluation, we establish a large-scale UDC video restoration benchmark, dubbed VidUDC33K. It contains 677 paired videos of length 50 with 1080p resolution, covering various challenging scenarios.

Our contributions are summarized as follows:

- We propose a novel network with long- and short-term video representation learning by decoupling video degradations for the UDC video restoration task (D^2RNet), which is the first work to address UDC video degradation. The core decoupling attention module (DAM) enables a tailored solution to the degradation caused by different incident light intensities in the video.
- We propose a large-scale UDC video restoration dataset (VidUDC33K), which includes numerous challenging scenarios. To the best of our knowledge, this is the first dataset for UDC video restoration.
- Extensive quantitative and qualitative evaluations demonstrate the superiority of D^2RNet. In the proposed VidUDC33K dataset, D^2RNet gains 1.02db PSNR improvements more than other restoration methods.

Related Work

UDC Restoration. Recently, UDC restoration based on deep learning has made significant progress and become an increasingly promising research topic. Existing benchmarks are mainly studies based on images. Typically, MSUNet (Zhou et al. 2021) proposes a first UDC image restoration benchmark by analyzing the optical imaging process of real UDC for ECCV20 challenge (Zhou et al. 2020). It includes diffraction kernels, *i.e.*, point spread function (PSF), and two paired datasets, *i.e.*, transparent-organic LED (T-OLED) and pentile-organic LED (P-OLED), by mounting a display on top of a traditional digital camera lens. However, they are unaligned with the real UDC degradation due to the lack of high dynamic range (HDR). Based on the measured PSF in real devices, DSICNet (Feng et al. 2021) generates a larger benchmark with the proposed model-based data synthesis pipeline for ECCV22 challenge (Feng et al. 2022). In addition, since multiple artificial lights at night may introduce different diffraction patterns, nighttime flare removal (Dai et al. 2022) and haze removal (Liu et al. 2023d) are also partially similar to the UDC restoration.

Existing works treat UDC image restoration as an inversion problem for the measured PSFs (*i.e.*, diffraction templates). To eliminate the degradation arising from the PSF, some PSF-related methods (Kwon et al. 2021; Liu et al. 2022c) use the PSF as a priori to guide the diffraction removal. Further, to avoid the PSF diversity caused by multiple external factors, the PSF-free methods (Panikkasseril Sethumadhavan et al. 2020; Zhou et al. 2021) directly learn various degradations in UDC images through complex network design. Typically, MSUNet (Zhou et al. 2021), DAGF (Feng et al. 2021), and BNUDC (Koh, Lee, and Yoon 2022) propose U-Net (Ronneberger, Fischer, and Brox 2015) framework, deep atrous guided filter, and dual-stream framework for UDC restoration, respectively. Recently, AlignFormer (Feng et al. 2023) proposes the first reference-based framework for non-aligned UDC image restoration.

Video Restoration. Video restoration aims to recover a high-quality video from a low-quality counterpart. Existing video restoration methods can be categorized into two kinds of paradigms: based on sliding-window structure (Kim et al. 2018; Li et al. 2021; Wang et al. 2019; Li et al. 2023) and based on recurrent structure (Huang, Wang, and Wang 2017; Isobe et al. 2020; Tao et al. 2018; Sajjadi, Vemulapalli, and Brown 2018). The methods based on sliding-window structure use adjacent frames within a sliding window as inputs to recover the high-quality frame (*e.g.*, 5 or 7 frames). They mainly use Transformer (Liang et al. 2022a) or deformable convolutions (Wang et al. 2019) to design advanced alignment modules and fuse useful features from adjacent frames. Nevertheless, multi-frame inputs often require higher computational costs, especially when using larger window sizes

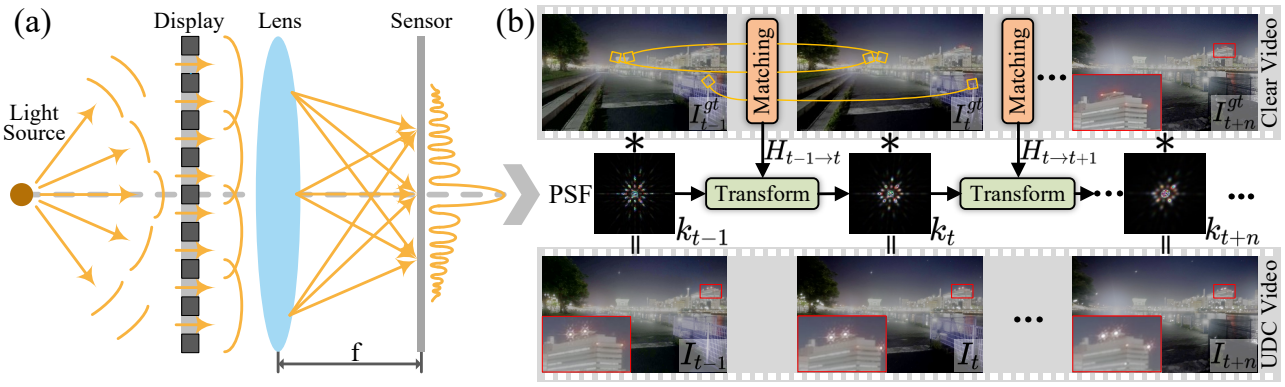

Figure 2: (a) illustrates the formation of the PSF in UDC systems. The light emitted from the light source crosses a display and a lens before it is finally captured by the sensor. (b) is the generation of UDC video, where the matching part computes the Homography matrix (*i.e.*, H) corresponding to inter-frame motion, and the transform part performs perspective warp on PSF.

to model frames at more distance. Rather than only aggregating information from adjacent frames, methods based on the recurrent structure can deliver the relevant features from past frames over time. These methods either devote their attention to designing advanced propagation methods for utilizing frames at longer distances (Chan et al. 2022; Liu et al. 2022a), or exploit powerful attention mechanisms to enhance feature extraction in the recurrent framework (Liang et al. 2022b; Zhong et al. 2023).

However, in contrast to degradation in super-resolution, deblurring, and other low-level tasks, variations in the intensity of incident light cause different degrees of degradation in UDC video. We propose a more promising solution to eliminate various degradations in UDC video restoration by taking full advantage of the recurrent framework.

Problem Formulation and Dataset

Problem Formulation. Inspired by the real-world imaging process of commercial UDC systems, we follow the existing UDC image restoration works (Koh, Lee, and Yoon 2022; Zhou et al. 2021) to define the UDC video restoration as the diffraction removal problem. In UDC video, the degradation model of the t^{th} frame can be formulated as:

$$I_t = f(\gamma \cdot I_t^{GT} * k_t + n), \tag{1}$$

where I_t^{GT} and I_t denote the clean and degraded frame, respectively. k_t is the diffraction kernel (*i.e.*, PSF), which is the primary factor affecting the visual quality. γ and n are the intensity scaling factor and additive noise, respectively. $*$ denotes the convolution operator. $f(\cdot)$ denotes the clamp function used to simulate the channel saturation. Here we omit the non-linear mapping for brevity.

From (Zhou et al. 2021), PSF is determined by the screen pattern $p(x, y)$ in the view of the light source. Different from the image restoration that keeps the PSF constant, when the light source changed during video shooting, PSF changes accordingly (Kwon et al. 2021). As illustrated by Fig. 2(b), we follow existing works (Babbar and Bajaj 2022; Ye et al. 2021; Liu et al. 2022b) to simulate the dynamic changes of k_t during the motion by computing the inter-frame Homography matrix $H_{t-1\to t}$, formulated as:

$$\begin{aligned} k_t &= \mathcal{T}(k_{t-1}, H_{t-1\to t}) \\ &= \left|\mathcal{F}(H^{-1}_{t-1\to t}(\mathcal{F}^{-1}(sqrt(k_{t-1})))\right|^2, \\ H_{t-1\to t} &= \mathcal{M}(I^{GT}_{t-1}, I^{GT}_t), \end{aligned} \tag{2}$$

where $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ are the Fourier transform and its inverse transform, respectively. $H^{-1}_{t-1\to t}$ is the inverse matrix of $H_{t-1\to t}$. $\mathcal{T}(\cdot)$ is the transform function that uses the $H^{-1}_{t-1\to t}$ to perspective warp the PSF of the previous frame k_{t-1}. $\mathcal{M}(\cdot)$ is the matching part used to compute the Homography matrix between frames.

Simulated Data. To keep the high dynamic range and high resolution of UDC video, we collected a total of 677 HDR videos from YouTube covering various scenarios present in HDRi Haven (*e.g.*, Outdoor, Skies, Urban, Night, Nature, and so on) and measured the PSF using a commercial ZTE Axon 20 device. Each video consists of 50 frames with a resolution of 1080×1920. For each video, we simulated the corresponding degraded video using Eq. (1), where the PSF k_t is dynamically changed by Eq. (2). To simulate the exhibit of structured flares near strong light sources, brightness augmentation is also applied in each frame. Finally, 627 videos are selected for training, and the remaining 50 videos are for testing randomly.

Real Data. To verify the effectiveness of the D^{2}RNet in real world, we captured 10 raw videos of different scenarios using the same ZTE Axon 20. We keep a high dynamic range and the same resolution with the simulated data.

Methodology

Overview of D^{2}RNet

By analyzing the properties of different degradations (*i.e.*, flare and haze) caused by variations in incident light intensity, we introduce valuable insight into handling UDC video restoration by decoupling the degradations. The overall structure of the proposed D^{2}RNet is shown in Fig. 3.

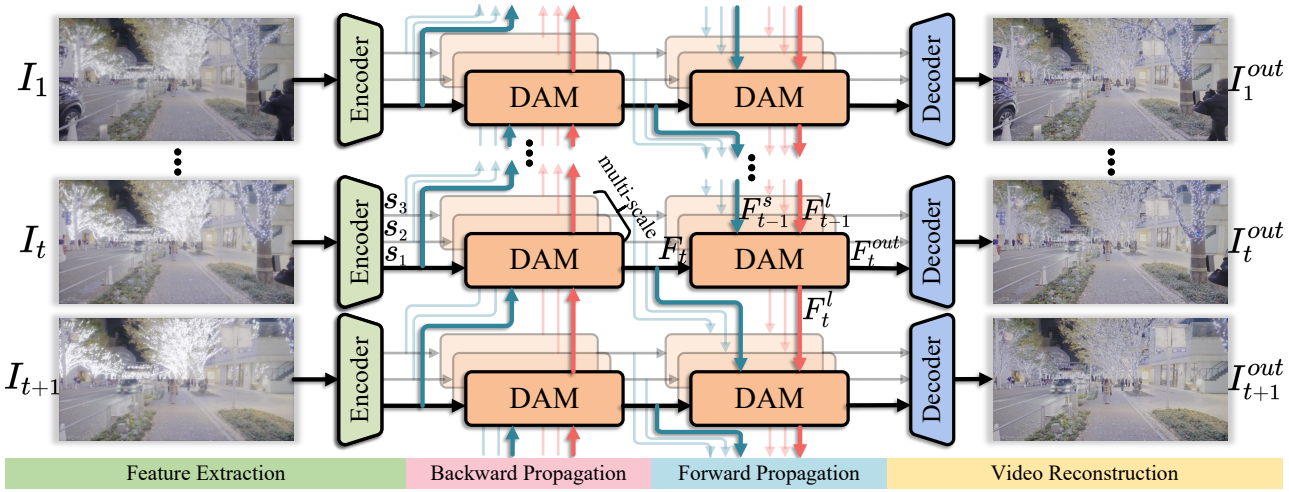

Figure 3: Overview of D^{2}RNet. It adopts a multi-scale bilateral recurrent architecture. Where an encoder and decoder are used to extract frame features and reconstruct the output frame, respectively. The proposed decoupling attention modules (DAM, see details in Fig. 4) is used to refine the features in both backward and forward propagation, which is supervised at multi-scale.

Specifically, given a degraded low-quality sequence $\mathbf{I}_{LQ} = \{I_t \in \mathbb{R}^{C\times H\times W}, t \in \{1,2,\dots,T\}\}$, the goal is to recover a high-quality version $\mathbf{I}_{HQ} = \{I_t^{out} \in \mathbb{R}^{C\times H\times W}, t \in \{1,2,\dots,T\}\}$. Where T, C,H, and W are the sequence length, channel, height, and width, respectively. The whole model adopts the recurrent architecture that combines multi-scale feature learning and bi-directional propagation. In which, the core decoupling attention module (DAM) refines features during backward and forward propagation.

Multi-scale Bi-directional Recurrent Architecture

Inspired by the success of bi-directional recurrent (Chan et al. 2022; Huang, Wang, and Wang 2015) and multi-scale fusion (Cho et al. 2021; Zamir et al. 2021) in low-level tasks, we combine them to enhance video representations. As shown in Fig. 3, from left to right, there is an encoder for extracting multi-scale features, backward and forward propagation for features learning, and a decoder for reconstructing the output frames, respectively.

Formally, take the restoration process of input I_t of the t^{th} frame as an example. First, during feature extraction, we use the contracting path (up-to-down) of UNet (Ronneberger, Fischer, and Brox 2015) as the structure of the encoder. This structure tailored for image restoration is broadly considered to enhance local details at different scales (Koh, Lee, and Yoon 2022; Zamir et al. 2021). Specifically, we denote the encoder as $\mathrm{E}(\cdot)$, the output can be obtained by:

$$F_t^{s_1}, F_t^{s_2}, F_t^{s_3} = \mathrm{E}(I_t), \tag{3}$$

where $F_t^{s_1} \in \mathbb{R}^{C_{s_1}\times\frac{H}{s_1}\times\frac{W}{s_1}}$, $F_t^{s_2} \in \mathbb{R}^{C_{s_2}\times\frac{H}{s_2}\times\frac{W}{s_2}}$, and $F_t^{s_3} \in \mathbb{R}^{C_{s_3}\times\frac{H}{s_3}\times\frac{W}{s_3}}$ indicate the obtained multi-scale features. C_{s_1}, C_{s_2}, and C_{s_3} are the number of feature channels at scales s_1, s_2, and s_3, respectively, which increases progressively as the spatial resolution of the feature decreases.

Then, during bi-directional propagation, the proposed DAM, denoted as $\mathrm{DAM}(\cdot)$, refines the features recurrently by inputting the current and the historical features. Take one of the forward propagation as an example (omitting the scale symbols for brevity). This process can be formulated as:

$$F_t^{out}, F_t^l = \mathrm{DAM}(F_t, F_{t-1}^s, F_{t-1}^l), \tag{4}$$

where $F_{t-1}^s = F_{t-1}$ denotes the short-term features from the previous frame, and F_{t-1}^l indicates the long-term features from the DAM output of the previous frame. Likewise, DAM is also used for backward propagation and multi-scale. Besides, to enable better learning of features, we multiplex the features by taking the output of backward propagation as the input of forward propagation.

Finally, we use the expanding path (down to up) of UNet (Ronneberger, Fischer, and Brox 2015) as the structure of the decoder, denoted as $\mathrm{D}(\cdot)$. The features after propagation are used to reconstruct the output, formulated as:

$$I_t^{out} = \mathrm{D}(F_t^{out,s_1}, F_t^{out,s_2}, F_t^{out,s_3}), \tag{5}$$

where I_t^{out} is the outut frame. The inputs F_t^{out,s_1}, F_t^{out,s_2}, and F_t^{out,s_3} indicate multi-scale features output from the bi-directional recurrent propagation.

Decoupling Attention Module

Along with the diffraction changes with video motion, the distant frame has some complementary features for recovering the lost content due to the flare. Conversely, the removal of haze in the current frame is less correlated with the content in the distant frames, and utilizing more similar features in the spatio-temporal neighborhood is more cost-effective for removing content haze. Therefore, we propose a tailored decoupling attention module (DAM) in Fig. 4. It includes a soft mask generation function to decouple the flare and haze, and flare and haze removal components handle the respective degradations. We omit scale symbols for brevity.

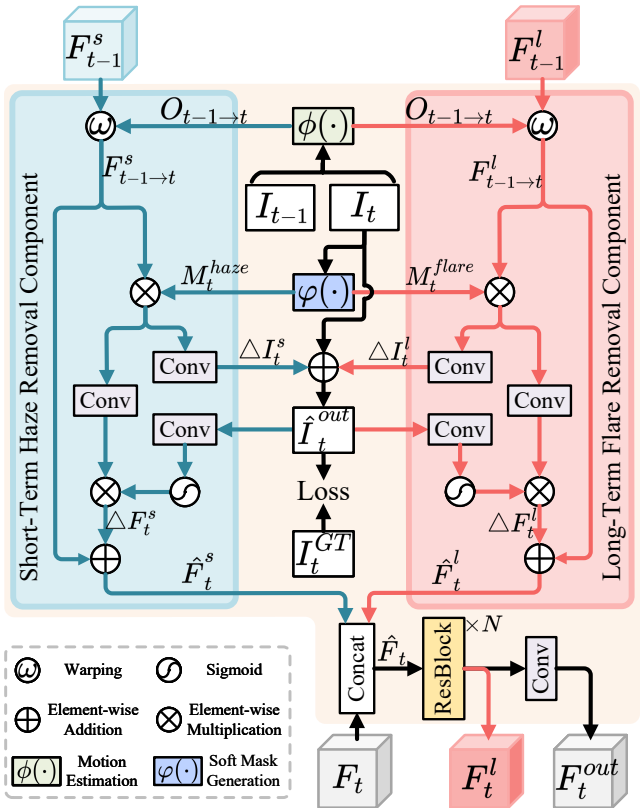

Figure 4: Structure of the Decoupling Attention Module (DAM). From top to bottom, it mainly consists of a soft mask generation function $\varphi(\cdot)$ for decoupling the flare M_t^{flare} and haze M_t^{haze}, and flare and haze removal components handle the respective degradations using long- and short-term features, respectively.

Formally, we first align the given long-term feature F_{t-1}^l and short-term feature F_{t-1}^s to the current frame with motion estimation network $\phi(\cdot)$ and warping operation $\mathcal{W}(\cdot)$, formulated as:

$$\begin{aligned} F_{t-1\to t}^l &= \mathcal{W}(F_{t-1}^l, O_{t-1\to t}), \\ F_{t-1\to t}^s &= \mathcal{W}(F_{t-1}^s, O_{t-1\to t}), \\ O_{t-1\to t} &= \phi(I_{t-1}, I_t), \end{aligned} \tag{6}$$

where $F_{t-1\to t}^l$ and $F_{t-1\to t}^s$ are the output aligned features. $O_{t-1\to t}$ represents the optical flow. Flares are usually caused by strong glare signals and occur with channel saturation. Therefore, inspired by the separation of overexposed regions in HDR images (Cao et al. 2023; Eilertsen et al. 2017; Liu et al. 2023b), an essential soft mask generation function $\varphi(\cdot)$ is proposed to generate the corresponding map of flare M_t^{flare} and haze M_t^{haze}, formulated as:

$$\begin{aligned} M_t^{flare} &= \varphi(I_t), \quad M_t^{haze} = 1 - \varphi(I_t), \\ \varphi(I_t^{\{r,g,b\}}) &= \frac{max(0, max_c(I_t^r, I_t^g, I_t^b) - \tau)}{1 - \tau}, \end{aligned} \tag{7}$$

where $max_c(\cdot)$ denotes taking the maximum value in the channel dimension. τ is an empirical parameter used to partition the flare and haze maps. M_t^{flare} measures the reliability of the flare and locates the region where the flare occurs. The value in M_t^{flare} is a linear ramp starting from pixel values at a threshold τ, and ending at the maximum pixel value. M_t^{haze} opposite to it.

Further, benefiting from the progress of the supervised attention mechanism (Cho et al. 2021; Zamir et al. 2021) in image restoration tasks, we generate intermediate results for supervising the training in each stage. Specifically, with the guidance of M_t^{flare} and M_t^{haze}, the generated intermediate result $\hat{I}_t^{out}$ can be formulated as:

$$\begin{aligned} \hat{I}_t^{out} &= I_t + \triangle I_t^l + \triangle I_t^s, \\ \triangle I_t^l &= Conv(M_t^{flare} \otimes F_{t-1\to t}^l), \\ \triangle I_t^s &= Conv(M_t^{haze} \otimes F_{t-1\to t}^s), \end{aligned} \tag{8}$$

where $Conv(\cdot)$ denotes the convolutional layer. $\otimes$ denotes the element-wise multiplication. Compared to existing restoration methods (Suin, Purohit, and Rajagopalan 2020; Zhang et al. 2019) that directly predict images at each stage and input them to subsequent stages, the introduction of supervision between the intermediate results $\hat{I}_t^{out}$ and the corresponding ground truth in each stage contributes to feature learning and performance gains.

Then, with the help of supervised $\hat{I}_t^{out}$, we generate attention maps that allow us to preserve the useful features to refine the long-term features F_{t-1}^l and short-term features F_{t-1}^l, formulated as:

$$\begin{aligned} \hat{F}_t^l &= F_{t-1\to t}^l \oplus \triangle F_t^l, \quad \hat{F}_t^s = F_{t-1\to t}^s \oplus \triangle F_t^s, \\ \triangle F_t^l &= (S(Conv(\hat{I}_t^{out}))) \otimes Conv(M_t^{flare} \otimes F_{t-1\to t}^l), \\ \triangle F_t^s &= (S(Conv(\hat{I}_t^{out}))) \otimes Conv(M_t^{haze} \otimes F_{t-1\to t}^s), \end{aligned} \tag{9}$$

where $\hat{F}_t^l$ and $\hat{F}_t^s$ are the output refined features. $S(\cdot)$ is a sigmoid function for generating attention maps. $\oplus$ denotes the element-wise addition.

Finally, the refined features $\hat{F}_t^l, \hat{F}_t^s$ and the current features F_t are concatenated to update the long-term features while outputting features for reconstruction, formulated as:

$$\begin{aligned} F_t^{out} &= Conv(F_t^l), \\ F_t^l &= RBs(C(F_t, \hat{F}_t^l, \hat{F}_t^s)), \end{aligned} \tag{10}$$

where F_t^{out} is the output of DAM for the final reconstruction. F_t^l is the updated long-term feature used for the next frame inference. $RBs(\cdot)$ denotes the N stacked residual blocks. $C(\cdot)$ is feature concatenation along the channel.

Experiments

Dataset and Metrics

Since no other datasets are available to study this problem, we compare our D^{2}RNet with other SOTA methods on the proposed VidUDC33K dataset. We keep the same evaluation metrics: 1) peak signal-to-noise ratio (PSNR), 2) structural similarity index (SSIM) (Wang et al. 2004) and 3) learned perceptual image patch similarity (LPIPS) (Zhang et al. 2018) as previous works (Liang et al. 2022a,b).

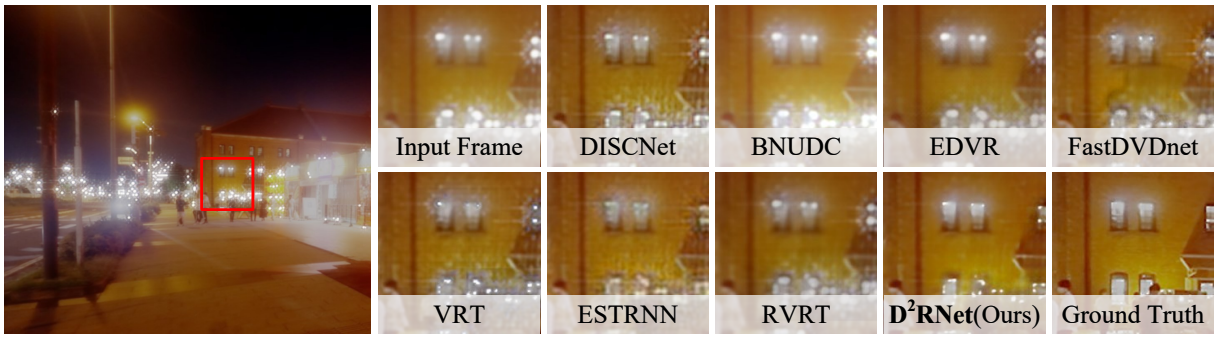

Figure 5: Visual results on proposed VidUDC33K. The method is shown on the bottom. Zoom in to see better visualization.

Method	RT(s)	#P(M)	PSNR↑	SSIM↑	LPIPS↓
DISCNet	0.73	3.80	28.89	0.8405	0.2432
BNUDC	0.09	4.60	28.59	0.8398	0.2728
UDC-UNet	0.37	5.70	28.37	0.8361	0.2561
Alignformer	-	-	28.96	0.8610	0.2200
EDVR	1.17	23.6	28.71	0.8531	0.2416
FastDVDnet	0.45	2.48	28.95	0.8638	0.2203
ESTRNN	0.20	2.47	29.54	0.8744	0.2170
VRT	2.18	17.5	30.61	0.9235	0.1397
RVRT	1.68	13.9	30.89	0.9261	0.1314
D^{2}RNet	0.44	5.76	**31.91**	**0.9313**	**0.1306**

Table 1: Quantitative comparison (PSNR(dB)↑, SSIM↑, and LPIPS↓) on the VidUDC33K dataset. RT and #P indicate the runtimes and parameters, respectively.

Base	IS	SHR	LFR	SMG	PSNR	SSIM	LPIPS
✓					30.49	0.9220	0.1394
✓	✓				31.01	0.9277	0.1352
✓	✓	✓			31.25	0.9301	0.1344
✓	✓		✓		31.37	0.9308	0.1340
✓	✓	✓	✓		31.74	0.9312	0.1314
✓	✓	✓	✓	✓	**31.91**	**0.9313**	**0.1306**

Table 2: Ablation study of each components on the proposed VidUDC33K dataset.

Training Details

For fair comparisons, we follow existing works (Liang et al. 2022b) to use the pre-trained SPyNet (Ranjan and Black 2017) for motion estimation. In multi-scale architecture, s_1, s_2, and s_3 correspond to $2\times$, $4\times$, and $8\times$ down-sampling, where the channels of features are 48, 60, and 72, respectively. In DAM, the threshold τ in the SMG is set to 0.9, and the number of ResBlocks N is set to 5. During training, we use Cosine Annealing scheme and Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rates of the motion estimation and other parts are set as 1.25×10^{-5} and 1×10^{-4}, respectively. We set the batch size as 8 and the input patch size as 256×256. To keep fair comparisons, we augment the data with random horizontal flips, vertical flips, and 90° rotations. Besides, to enable long-range sequence capability, we use sequences with a length of 30 as inputs. The Charbonnier penalty loss, defined as $\mathcal{L}(x, y) = \sqrt{\|x - y\|^2 + \varepsilon^2}$ where $\epsilon = 10^{-3}$, is applied not only to the whole frames between the restored frame I^{out} and ground truth, but also to the whole frames between the intermediate result $\hat{I}^{out}$ and the ground truth. To stabilize the training, we fix the motion estimation network in the first 5K iterations, and make it trainable later. The total number of iterations is 400K.

Comparisons with State-of-the-art Methods

We compare our D^{2}RNet with four UDC image restoration models (Feng et al. 2021; Koh, Lee, and Yoon 2022; Feng et al. 2023; Liu et al. 2022c) and five video restoration models (Liang et al. 2022a,b; Tassano, Delon, and Veit 2020; Wang et al. 2019; Zhong et al. 2023). For fair comparisons, we reproduce results with recommended configurations by the authors' officially released codes.

Quantitative comparison. The performance comparisons on our proposed VidUDC33K dataset are shown in Tab. 1. The image-based UDC restoration method (*e.g.*, Alignformer (Feng et al. 2023)) cannot exploit temporal information, resulting in poor performance, despite having fewer parameters and runtimes. Methods dedicated to video denoise (*i.e.*, FastDVDnet (Tassano, Delon, and Veit 2020)) and video blurring (*i.e.*, ESTRNN (Zhong et al. 2023)) do not yield the ideal performance due to the lack of design to handle diffraction. Moreover, compared to the latest video restoration algorithm (*e.g.*, RVRT (Liang et al. 2022b)), which treat all degradations as equivalent, our method outperforms the latest methods in both objective evaluation metrics PSNR, SSIM and perceptual metrics LPIPS with less runtime and parameters. In particular, our method exceeds the RVRT (Liang et al. 2022b) by **1.02 dB** in PSNR, benefiting from our multi-scale bi-directional recurrent architecture and the design of the decoupled degradation. This large margin demonstrates the power of D^{2}RNet.

Qualitative comparison. To further compare the visual qualities of different algorithms, we show visual results restored by our D^{2}RNet and other SOTA methods in Fig. 5. It can be observed that compared to other algorithms, D^{2}RNet can simultaneously remove flare at brighter windows and haze elsewhere. It verify that D^{2}RNet has a stronger UDC

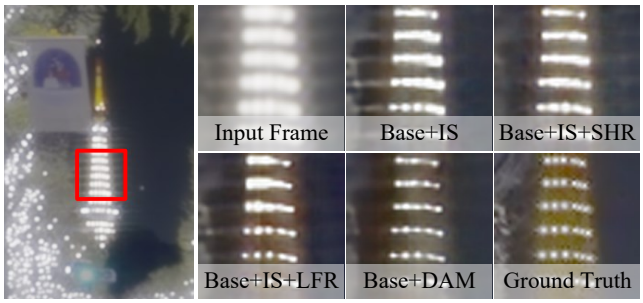

Figure 6: Visual comparison of different components used.

τ	0.98	0.95	0.90	0.85
PSNR	31.30	31.74	**31.91**	31.78
SSIM	0.9299	0.9310	**0.9313**	0.9310

Table 3: Ablation of τ in soft mask generation function.

video restoration capability and has a great improvement in visual quality, especially for flare-rich videos.

Ablation Study

In this section, we first conduct ablation for each component in DAM. After that, we study the effect of the τ in the SMG and the effect of the multi-scale architecture.

Individual components. Based on our proposed model, we directly use ResBlock (He et al. 2016) to replace the decoupling attention module as the "Base" model and progressively add the intermediate supervision (IS), the short-term haze removal component (SHR), the long-term flare removal component (LFR), and the soft mask generation function (SMG) for comparisons. As shown in Tab. 2, the PSNR can be improved from 30.49 dB to 31.74 dB with the addition of IS, SHR, and LFR, verifying the powerful ability of the-supervised attention mechanism and the haze/flare removal component. When SMG is involved, degraded haze and flare are decoupled and learned separately, and the performance is improved to 31.91 dB. These demonstrate the superiority of each part in DAM. We further explore the visual differences as shown in Fig. 6. LFR can eliminate the content loss caused by flare, and SHR can eliminate the content fuzziness caused by haze. Decoupling flare and haze in UDC videos to remove them separately can produce clearer textures.

The effect of τ. To explore the effect of τ used in the soft mask generation function on performance. In Tab. 3, we use different τ to decouple the flare and haze. It can be seen that too large τ does not completely separate the flare region, which affects the recovery of the lost content. On the contrary, too small τ will result in incomplete removal of content fuzziness. It demonstrates the effectiveness of the soft mask generation function in decoupling video degradation. We set τ as 0.9 in the final model.

The effect of the multi-scale architecture. To alleviate the scale-changing problem in sequences, we discuss the effect of multi-scale architecture on performance. In Tab. 4, we use $2\times$, $4\times$, and $8\times$ to denote s_1, s_2, and s_3 as described in Sec. , respectively. The results show that our method can restore clearer content as the scale increases. The performance can improve PSNR from 30.45 dB to 31.91 dB, indicating that multi-scale architecture can adapt to scale-changing problems in sequences. In our model, we use all three scales to achieve the best performance.

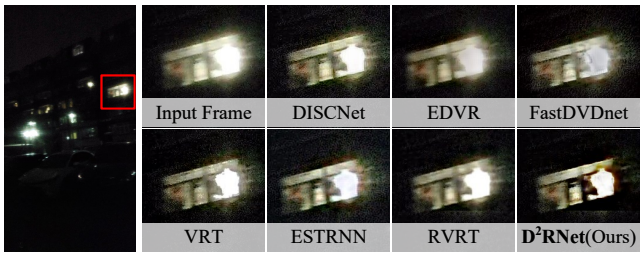

Figure 7: Visual comparison of real UDC video.

Scale	(2×)	(2×,4×)	(2×,4×,8×)
PSNR	30.45	31.36	**31.91**
SSIM	0.9192	0.9308	**0.9313**

Table 4: Ablation of the multi-scale structure.

Evaluation on Real UDC Video

In addition to the simulated dataset described above, we conduct compare on real videos collected by the same device. As shown in Fig. 7, for diffraction-induced flare and haze, our D^{2}RNet can produce clearer textures. This demonstrates the robustness of our D^{2}RNet.

Conclusion

In this paper, we study the effects of the intensity of the incident light and the motion information in UDC video degradation, and introduce a new perspective to handle them by decoupling different types of degradation in advance. In particular, we propose a novel video restoration network to enable effective UDC video representation learning, dubbed D^{2}RNet, in which the core decoupling attention module (DAM) provides an effective and targeted solution for eliminating various degradations. Experimental results show significant performance improvements and clear visual margins between D^{2}RNet and existing SOTA models. To the best of our knowledge, we are the first work to study this problem and propose the first large-scale UDC video benchmark. Our perspective on UDC video has the potential to inspire more diffraction-limited video restoration works. In the future, we will further improve the generality and robustness of our model, and extend it to other low-level vision tasks through more exploration.

Acknowledgements

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

References

Babbar, G.; and Bajaj, R. 2022. Homography Theories Used for Image Mapping: A Review. In *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, 1–5. IEEE.

Cao, G.; Zhou, F.; Liu, K.; Wang, A.; and Fan, L. 2023. A decoupled kernel prediction network guided by soft mask for single image HDR reconstruction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2s): 1–23.

Chan, K. C.; Zhou, S.; Xu, X.; and Loy, C. C. 2022. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, 5972–5981.

Cho, S.-J.; Ji, S.-W.; Hong, J.-P.; Jung, S.-W.; and Ko, S.-J. 2021. Rethinking coarse-to-fine approach in single image deblurring. In *ICCV*, 4641–4650.

Dai, Y.; Li, C.; Zhou, S.; Feng, R.; and Loy, C. C. 2022. Flare7k: A phenomenological nighttime flare removal dataset. *NeurIPS*, 35: 3926–3937.

Eilertsen, G.; Kronander, J.; Denes, G.; Mantiuk, R. K.; and Unger, J. 2017. HDR image reconstruction from a single exposure using deep CNNs. *ACM TOG*, 36(6): 1–15.

Feng, R.; Li, C.; Chen, H.; Li, S.; Gu, J.; and Loy, C. C. 2023. Generating Aligned Pseudo-Supervision from Non-Aligned Data for Image Restoration in Under-Display Camera. In *CVPR*, 5013–5022.

Feng, R.; Li, C.; Chen, H.; Li, S.; Loy, C. C.; and Gu, J. 2021. Removing diffraction image artifacts in under-display camera via dynamic skip connection network. In *CVPR*, 662–671.

Feng, R.; Li, C.; Zhou, S.; Sun, W.; Zhu, Q.; Jiang, J.; Yang, Q.; Loy, C. C.; and Gu, J. 2022. Mipi 2022 challenge on under-display camera image restoration: Methods and results. *arXiv preprint arXiv:2209.07052*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Huang, Y.; Wang, W.; and Wang, L. 2015. Bidirectional recurrent convolutional networks for multi-frame super-resolution. *NeurIPS*, 28.

Huang, Y.; Wang, W.; and Wang, L. 2017. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE TPAMI*, 40(4): 1015–1028.

Isobe, T.; Jia, X.; Gu, S.; Li, S.; Wang, S.; and Tian, Q. 2020. Video super-resolution with recurrent structure-detail network. In *ECCV*, 645–660. Springer.

Kim, T. H.; Sajjadi, M. S.; Hirsch, M.; and Scholkopf, B. 2018. Spatio-temporal transformer network for video restoration. In *ECCV*, 106–122.

Koh, J.; Lee, J.; and Yoon, S. 2022. Bnudc: A two-branched deep neural network for restoring images from under-display cameras. In *CVPR*, 1950–1959.

Kwon, K.; Kang, E.; Lee, S.; Lee, S.-J.; Lee, H.-E.; Yoo, B.; and Han, J.-J. 2021. Controllable image restoration for under-display camera in smartphones. In *CVPR*, 2073–2082.

Li, D.; Shi, X.; Zhang, Y.; Cheung, K. C.; See, S.; Wang, X.; Qin, H.; and Li, H. 2023. A Simple Baseline for Video Restoration With Grouped Spatial-Temporal Shift. In *CVPR*, 9822–9832.

Li, D.; Xu, C.; Zhang, K.; Yu, X.; Zhong, Y.; Ren, W.; Suominen, H.; and Li, H. 2021. Arvo: Learning all-range volumetric correspondence for video deblurring. In *CVPR*, 7721–7731.

Liang, J.; Cao, J.; Fan, Y.; Zhang, K.; Ranjan, R.; Li, Y.; Timofte, R.; and Van Gool, L. 2022a. VRT: A video restoration transformer. *arXiv preprint arXiv:2201.12288*.

Liang, J.; Fan, Y.; Xiang, X.; Ranjan, R.; Ilg, E.; Green, S.; Cao, J.; Zhang, K.; Timofte, R.; and Gool, L. V. 2022b. Recurrent video restoration transformer with guided deformable attention. *NeurIPS*, 35: 378–393.

Lin, J.; Cai, Y.; Hu, X.; Wang, H.; Yan, Y.; Zou, X.; Ding, H.; Zhang, Y.; Timofte, R.; and Van Gool, L. 2022. Flow-guided sparse transformer for video deblurring. *arXiv preprint arXiv:2201.01893*.

Liu, C.; Wang, X.; Li, S.; Wang, Y.; and Qian, X. 2023a. FSI: Frequency and Spatial Interactive Learning for Image Restoration in Under-Display Cameras. In *ICCV*, 12537–12546.

Liu, C.; Yang, H.; Fu, J.; and Qian, X. 2022a. Learning trajectory-aware transformer for video super-resolution. In *CVPR*, 5687–5696.

Liu, C.; Yang, H.; Fu, J.; and Qian, X. 2023b. 4D LUT: learnable context-aware 4d lookup table for image enhancement. *IEEE TIP*, 32: 4742–4756.

Liu, C.; Yang, H.; Fu, J.; and Qian, X. 2023c. TTVFI: Learning trajectory-aware transformer for video frame interpolation. *IEEE TIP*.

Liu, S.; Lu, Y.; Jiang, H.; Ye, N.; Wang, C.; and Zeng, B. 2022b. Unsupervised Global and Local Homography Estimation With Motion Basis Learning. *IEEE TPAMI*.

Liu, X.; Hu, J.; Chen, X.; and Dong, C. 2022c. UDC-UNet: Under-Display Camera Image Restoration via U-shape Dynamic Network. In *ECCV*, 113–129. Springer.

Liu, Y.; Yan, Z.; Chen, S.; Ye, T.; Ren, W.; and Chen, E. 2023d. Nighthazeformer: Single nighttime haze removal using prior query transformer. In *ACM MM*, 4119–4128.

Panikkasseril Sethumadhavan, H.; Puthussery, D.; Kuriakose, M.; and Charangatt Victor, J. 2020. Transform domain pyramidal dilated convolution networks for restoration of under display camera images. In *ECCVW*, 364–378. Springer.

Qin, Z.; Qiu, R.; Li, M.; Yu, X.; and Yang, B.-R. 2021. P-78: Simulator-Based Efficient Panel Design and Image Retrieval for Under-Display Cameras. In *SID Symposium Digest of Technical Papers*, volume 52, 1372–1375. Wiley Online Library.

Ranjan, A.; and Black, M. J. 2017. Optical flow estimation using a spatial pyramid network. In *CVPR*, 4161–4170.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Sajjadi, M. S.; Vemulapalli, R.; and Brown, M. 2018. Frame-recurrent video super-resolution. In *CVPR*, 6626–6634.

Suin, M.; Purohit, K.; and Rajagopalan, A. 2020. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *CVPR*, 3606–3615.

Tao, X.; Gao, H.; Shen, X.; Wang, J.; and Jia, J. 2018. Scale-recurrent network for deep image deblurring. In *CVPR*, 8174–8182.

Tassano, M.; Delon, J.; and Veit, T. 2020. FastDVDnet: Towards real-time deep video denoising without flow estimation. In *CVPR*, 1354–1363.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. EDVR: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 0–0.

Wang, Y.; Lu, Y.; Gao, Y.; Wang, L.; Zhong, Z.; Zheng, Y.; and Yamashita, A. 2022. Efficient video deblurring guided by motion magnitude. In *ECCV*, 413–429. Springer.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4): 600–612.

Ye, N.; Wang, C.; Fan, H.; and Liu, S. 2021. Motion basis learning for unsupervised deep homography estimation with subspace projection. In *ICCV*, 13117–13125.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *CVPR*, 14821–14831.

Zhang, H.; Dai, Y.; Li, H.; and Koniusz, P. 2019. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*, 5978–5986.

Zhang, H.; Xie, H.; and Yao, H. 2022. Spatio-temporal deformable attention network for video deblurring. In *ECCV*, 581–596. Springer.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 586–595.

Zhong, Z.; Gao, Y.; Zheng, Y.; Zheng, B.; and Sato, I. 2023. Real-world video deblurring: A benchmark dataset and an efficient recurrent neural network. *IJCV*, 131(1): 284–301.

Zhou, Y.; Kwan, M.; Tolentino, K.; Emerton, N.; Lim, S.; Large, T.; Fu, L.; Pan, Z.; Li, B.; Yang, Q.; et al. 2020. UDC 2020 challenge on image restoration of under-display camera: Methods and results. In *ECCVW*, 337–351. Springer.

Zhou, Y.; Ren, D.; Emerton, N.; Lim, S.; and Large, T. 2021. Image restoration for under-display camera. In *CVPR*, 9179–9188.