Perspective-Aligned AR Mirror with Under-Display Camera

JIAN WANG*, SIZHUO MA, KARL BAYER, YI ZHANG, PEIHAO WANG, and BING ZHOU, Snap Inc., USA
SHREE K. NAYAR, Snap Inc. and Columbia University, USA
GURUNANDAN KRISHNAN, Snap Inc., USA

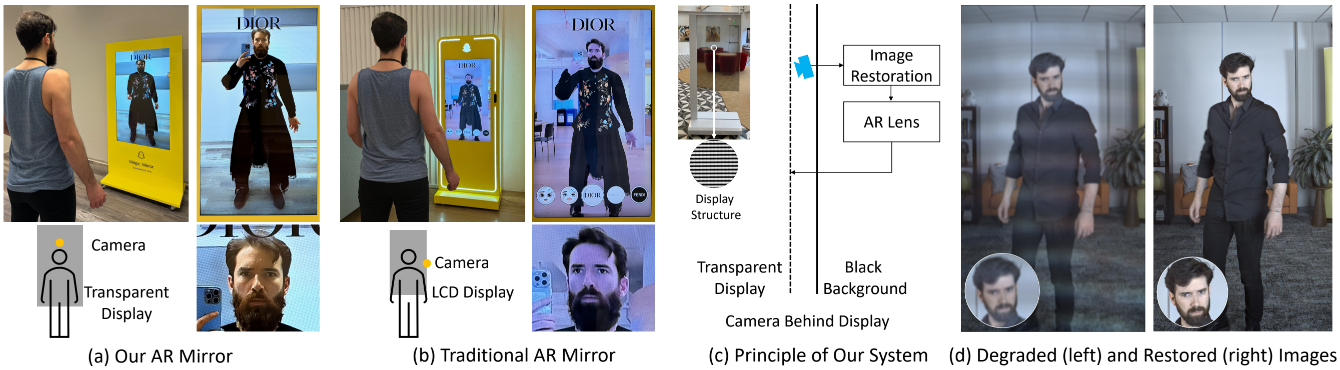

Fig. 1. Our proposed system. (a) Our AR mirror features a camera behind a transparent display, which offers a perspective-aligned experience. (b) A traditional AR mirror places the camera beside the display, resulting in a distorted perspective. (c) Since the camera is placed behind the display, the image suffers from various degradations. We design a processing pipeline which first runs an image restoration algorithm to get an image with improved quality, then passes the result to AR lens rendering to generate the final image for display. (d) Our proposed restoration algorithm effectively improves the visual quality.

Augmented reality (AR) mirrors are novel displays that have great potential for commercial applications such as virtual apparel try-on. Typically the camera is placed beside the display, leading to distorted perspectives during user interaction. In this paper, we present a novel approach to address this problem by placing the camera behind a transparent display, thereby providing users with a perspective-aligned experience. Simply placing the camera behind the display can compromise image quality due to optical effects. We meticulously analyze the image formation process, and present an image restoration algorithm that benefits from physics-based data synthesis and network design. Our method significantly improves image quality and outperforms existing methods especially on the underexplored wire and backscatter artifacts. We then carefully design a full AR mirror system including display and camera selection, real-time processing pipeline, and mechanical design. Our user study demonstrates that the system is exceptionally well-received by users, highlighting its advantages over existing camera configurations not only as an AR mirror, but also for video conferencing. Our work represents a step forward in the development of AR mirrors, with potential applications in retail, cosmetics, fashion, *etc.* The image restoration dataset and code are available at https://perspective-armirror.github.io/.

CCS Concepts: • **Human-centered computing → Displays and imagers**; • **Computing methodologies → Computational photography**.

Additional Key Words and Phrases: Augmented reality mirrors, under display cameras, perspective correction, image quality enhancement, user experience

ACM Reference Format:
Jian Wang, Sizhuo Ma, Karl Bayer, Yi Zhang, Peihao Wang, Bing Zhou, Shree K. Nayar, and Gurunandan Krishnan. 2024. Perspective-Aligned AR Mirror with Under-Display Camera. *ACM Trans. Graph.* 43, 6, Article 185 (December 2024), 11 pages. https://doi.org/10.1145/3687995

*Shree served as the direction lead, Gurunandan as the project lead, and Jian as the tech lead and IC (individual contributor). Sizhuo and Yi contributed equally overall. Yi and Peihao contributed equally to the image restoration experiments. Jian is the corresponding author.

Authors' addresses: Jian Wang, jwang4@snapchat.com; Sizhuo Ma, sma@snapchat.com; Karl Bayer, karlsbayer@gmail.com; Yi Zhang, zhangyi3.link@gmail.com; Peihao Wang, peihaowang@utexas.edu; Bing Zhou, bzhou@snapchat.com, Snap Inc., 229 W. 43rd St 6th Floor, New York, NY, 10036, USA; Shree K. Nayar, nayar@cs.columbia.edu, Snap Inc. and Columbia University, New York, USA; Gurunandan Krishnan, guru@gurukrishnan.com, Snap Inc., New York, USA.

ACM 0730-0301/2024/12-ART185
https://doi.org/10.1145/3687995

1 INTRODUCTION

Recent advances in consumer electronics and computer graphics algorithms lead to the advent of a new type of display, augmented reality (AR) mirrors, which are display devices that show the scene in front of it just like conventional mirrors, while augmenting the imagery with virtual objects or effects. Usually a camera is used to capture the scene, and a large-format screen (*e.g.*, over 50-inch) is used to show the augmented image to mimic a mirror. AR mirrors have already seen early commercialization, particularly for virtual try-on[1]. These AR mirrors enable customers to see how they look in different clothes without physically wearing them, demonstrating great potential for a growing market.

One key challenge for existing AR mirrors is that the camera has to be placed either above or beside the screen. As users tend to center themselves, the camera necessarily captures the users from a slanted view angle, causing an uncomfortable experience.

[1]https://zero10.ar/, https://carrier.huawei.com/en/success-stories/Industries-5G/5G-AR

This effect is demonstrated in Fig. 1(b): It is impossible to see a perspective-aligned view of yourself as in a true mirror.

To solve this perspective mismatch, can we place the camera in the center, but behind the display? Fortunately, this is actually feasible thanks to recently developed transparent displays. Such displays are made of minimally-sized electronics with transparent space between rows of pixels so that people can see through the open slits and observe what lies behind. If we place a camera only millimeters behind the display, these micro-scale pixel structures cause various degradations on the image.

In this paper, we focus on the problem of enabling perspective-aligned AR mirrors with an under-display camera (UDC). We systematically analyze the interaction between the display and the camera and rigorously derive an image formation model. We would like to highlight that, while previous work [Qin et al. 2016] has analyzed the image formation model for UDC on smartphones, this work aims to optimize the user experience on a large-format display, which turns out to involve a different set of underexplored challenges. Specifically, we assume (1) that all the pixels are on during the camera exposure to avoid unpleasant flickering, (2) that the camera can be tilted downwards to enable capturing the full body of the user without sacrificing image resolution, and (3) that the pixel pitch is larger than on phones. We then analyze various degradations in the imaging process, namely the spatially-varying blur due to diffraction, the additive backscatter light caused by the always-on display pixels, and the multiplicative intensity modulation (*wire* artifacts) due to occlusion by the display pixels.

Based on the imaging model, we correct the degradations through careful data synthesis and model design. We calibrate the model parameters and use them to synthesize training image pairs in a physics-based way. We demodulate the wire artifacts as a preprocessing step. Inspired by state-of-the-art image restoration algorithms, we design a network that effectively removes the blur and backscatter. The algorithm is lightweight and runs in real time.

We build an AR mirror with an off-the-shelf transparent OLED display and a machine vision camera, with careful mechanical design and optimized computation pipeline. We quantitatively and qualitatively evaluate the efficacy of the image restoration algorithm on images captured in the wild. Although the system is designed as an AR mirror that displays an image of the users themselves, it can also be used for providing telepresence of the user for video conference applications. We conduct user studies on both applications to demonstrate that the aligned perspective significantly improve user experience, without introducing noticeable visual artifacts.

To summarize, the main contributions of this work are:

- Our key technical novelty lies in a novel image formation model of a camera behind a large-format display, including the backscatter (which has never been discussed in previous work) and wire artifacts (which have only been briefly mentioned in the appendix of [Zhou et al. 2021]). Such degradations are not evident on smartphone UDCs, but can lead to significant drop in image quality on AR mirrors.
- Based on this analysis, we develop an efficient reconstruction algorithm with a focus on the new degradations. Specifically, wire artifacts are removed by a simple division as a preprocessing step based on image formation analysis and physics-based calibration. Backscatter artifacts are removed by careful data simulation and novel network design.
- We mechanically and computationally build a stable, real-time AR mirror system with improved user experience.
- We conduct experiments and user studies to show the efficacy of the image restoration algorithm and the improvement of user experience by the aligned perspective.

2 RELATED WORK

AR mirror. The most common implementation of AR mirrors is to combine the imagery computationally, where virtual objects are digitally inserted into the captured real scene, which is subsequently displayed on a large, often non-transparent, screen. [Blum et al. 2012; Bork et al. 2017, 2019; Meng et al. 2013] use a non-transparent display in conjunction with a Microsoft Kinect to overlay organs onto the real human body for anatomy education. Although this approach has already been implemented in commercial products [ZERO10 2024], the side/top placement of the camera leads to a perspective misalignment. We address this challenge by placing the camera behind the display, enhancing the overall user experience.

An alternative implementation of AR mirrors is to combine a semi-transparent mirror and a non-transparent display, or a semi-transparent display with a non-transparent mirror, such that the real image and the virtual objects are blended *optically* [Anderson et al. 2013; Jacobs et al. 2019; Saakes et al. 2016]. Since the real image and the virtual objects appear at different depths, they cannot be perfectly aligned in 3D. Due to the optical blending nature, only additive AR is enabled [Luo et al. 2021], which limits its application.

Under-display camera. Most works on under-display cameras use the term “display” to refer to a specific area on a smartphone screen directly above the front-facing camera. To get rid of the camera hole/notch, the pixel density in this small area is intentionally reduced, allowing light to enter the camera through the gaps [Motorola 2021; Samsung 2023; Xiaomi 2021; ZTE 2020]. On the other hand, LG released a 55-inch transparent display in 2019 mainly for retail display use (the only one in the market to our knowledge) [LG 2019]. While [Lim et al. 2021] has made a preliminary attempt to put a camera behind the display for video conferencing, they did not analyze the image formation and restoration from the first principles. In all these works, the camera is placed upright (the optical axis is perpendicular to the screen), and the screen is off during capturing (or flickering rapidly during video recording). In contrast, we tilt the camera and keep the screen continuously on for a better experience. Such design choices lead to new imaging challenges, which are the main focus of our restoration algorithm.

Image restoration for UDC. UDC image restoration is an active research area with two open challenges held in the past [Feng et al. 2022; Zhou et al. 2020]. Among these methods, [Sundar et al. 2020][2] proposes a learning-based guided filter. DISCNet [Feng et al. 2021] utilizes HDR data to synthesize the training data and uses dynamic skip connections to remove flares. MIMO-UDC [Zhu et al. 2023][3] uses multi-resolution network design. UDCUNet [Liu et al. 2022][4]

[2] 2nd place in the first UDC challenge.
[3] 1st place in the second UDC challenge.
[4] 2nd place in the second UDC challenge.

uses SFT layer [Wang et al. 2018b] to guide the restoration. [Kwon et al. 2021] considers spatially-varying PSFs. BNUDC [Koh et al. 2022] separates the processing of low-frequency and high-frequency information to better handle saturations. [Feng et al. 2023] proposes to learn the ground truth from a side camera. [Ahn et al. 2023] recently proposes a new real UDC dataset. Despite their progress in tackling degradations like blur and saturation, none of them is designed to deal with the wire and backscatter artifacts.

Relation to coded-aperture cameras. Under-display cameras can also be seen as a type of coded-aperture cameras where the aperture is coded by the display pixel pattern. [Yang and Sankaranarayanan 2021] discussed how to optimize the layout of display pixels such that the blur kernel can be robustly invertible in the presence of noise. [Yang et al. 2023] proposes using phase masks with an existing display panel to avoid challenges in fabricating specific pixel layouts. While we focus on building a system using off-the-shelf optics in this paper, it is possible to integrate the proposed coded aperture approaches in the future to further boost the image quality.

Novel view synthesis. It is possible to correct the perspective computationally, a problem known as novel view synthesis. When the scene is planar, this can be easily achieved by a homography [Hartley and Zisserman 2003]. The problem becomes more challenging considering the great depth variance between the user and the background, and the 34cm baseline between a side camera and the center of a 55-inch display. While NVIDIA Broadcast [NVIDIA 2020] can edit the eyes to enable eye contact in real time, it cannot synthesize a completely new view for a mirror-like experience. Although real-time rendering of NeRF [Mildenhall et al. 2021; Müller et al. 2022] and 3D Gaussian splatting [Kerbl et al. 2023] have achieved great success, real-time *scene-level* reconstruction [Luiten et al. 2023; Zhang et al. 2022] is still challenging and cannot reach photo-level quality, a harder problem than UDC image reconstruction.

3 IMAGE FORMATION

Placing the camera behind the transparent display introduces a series of degradation that affect the quality of captured images. Specifically, the transparent display comprises of rows of OLED pixels with gaps in-between, allowing the camera to capture external scenes through these gaps, as depicted in Fig. 2(a). Moreover, the OLED pixels are enclosed within two layers of glass, adding further complexity to the imaging system. Our derived image formation model can be described as follows,

$$y = [(x * k) + b] \odot w + n, \tag{1}$$

where y is the captured image, x is the undegraded image we aim to recover, $*$ indicates a convolution modified to use a spatially-varying kernel and $\odot$ denotes Hadamard product. The image is subject to four kinds of degradation: k is the point spread function (PSF), a large-support, spatially-varying blur kernel caused by diffraction. b is the backscatter caused by the glass layers. w is a spatial modulation of intensity (*wire* pattern) caused by OLED pixel occlusion. n is the image noise. While the image formation model for under-display cameras has been analyzed before, our model differs from those from prior work because of three important assumptions:

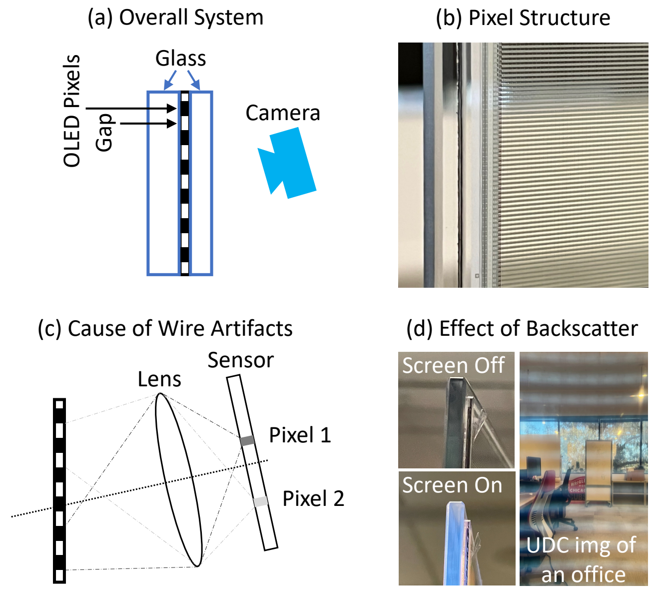

Fig. 2. Imaging system. (a) Overall system: a camera is placed behind a screen, which is composed of rows of OLED pixels encapsulated by two layers of glass. (b) Gaps between rows of pixels. (c) The wire artifact is a spatial beating pattern formed by different amounts of light blocked at different camera pixels. (d) Multi-bounce reflections and scattering between the glass layers cause the additive backscatter.

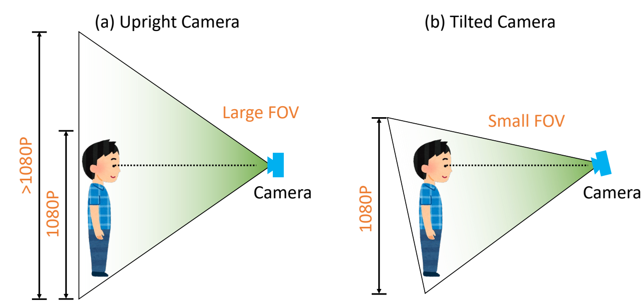

Fig. 3. Why tilting the camera? (a) When an upright camera is used, a large FOV lens (short focal length) and a high resolution sensor are needed. (b) A smaller FOV and lower resolution are sufficient for a tilted camera setup, which allows for more flexible design choices with off-the-shelf components.

- We assume the display pixels are always on during the capture. Traditionally, the screen is turned off during the capture to avoid backscatter, which is feasible for UDC on phones as high-end phones have refresh rates of 120Hz or 144Hz. However, the only commercially-available transparent display refreshes at 60Hz. Switching between on and off will drop the frame rate to 30Hz, causing noticeable flickering [Davis et al. 2015] and is not acceptable for a large-format screen aiming for interactive applications.
- We assume the camera can be tilted downwards instead of perpendicular to the screen. This is because if we want to keep

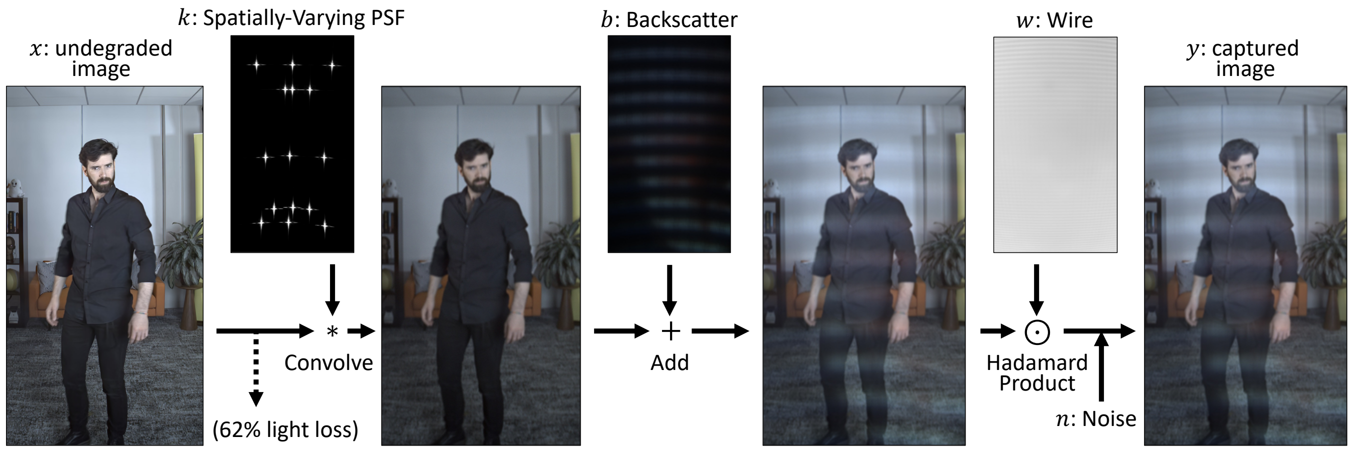

Fig. 4. Image formation model. The scene undergoes 62% light loss and blur through convolution with a spatially-varying PSF, resulting in a darkened and blurred image. Additive backscatter is introduced, followed by pixelwise multiplication with the wire pattern. Noise is introduced during image sensor capture.

eye contact when the user is looking straight forward just like a true mirror, the camera has to be placed at the same height as the eyes. Allowing the camera to tilt gives more freedom in choosing the right camera parameters to ensure best framing, as shown in Fig. 3. Notice that the perspective can be corrected by virtually rotating the camera via a homography as post-processing. **See Sec. 2 in the supplementary material.**

- Large displays have larger pixels, which make the wire artifacts due to display pixel occlusion more evident.

A detailed explanation for each kind of degradation is given below.

Point spread function. The minuscule pixel gaps (Fig. 2(b)) introduces diffraction artifacts which exhibits as image blurring and flare [Feng et al. 2021; Qin et al. 2016]. This effect can be characterized by a Point Spread Function (PSF), which varies spatially with the incidence angle of light. In our setup, the PSF exhibits further spatial variability because a wide-angle lens is needed for the large format. Fig. 4 k clearly shows the spatially-varying nature of the PSFs captured across the field of view (FOV), with the PSFs in the lower portion of the image showing notable curvature.

Backscatter. Since the pixels are always on, the light emitted from the OLEDs undergoes complex interactions with the encapsulating glass layers, involving multiple reflections and scattering. These interactions yield a complex light field when observed by the camera, resulting in an additive layer of image, a phenomenon we term *backscatter* (Fig. 2(d), Fig. 4 b). Apparently, the backscatter depends on the display content, the relative camera pose, and the properties of the display such as glass thickness. Applying an anti-reflection film to the front glass significantly reduces the reflected light, which mitigates but cannot fully remove the backscatter.

Wire pattern. As shown in Fig. 2(c), different camera pixels have different fraction of light blocked by the OLED pixels, thus causing a multiplicative intensity modulation we term "wire pattern" due to its appearance (Fig. 4 w). This wire pattern can also be regarded as the defocused image of the OLED pixels: When a small aperture is used, the wire pattern becomes sharp, black lines. This effect has been noted in previous work [Zhou et al. 2021] which is attributed to the "imperfect adhesion of the display to the camera lens". Notice that perfect adhesion is hard to achieve in practice and impossible when the camera is tilted. As opposed to [Zhou et al. 2021] which blindly learns to remove it via a network, we calibrate and invert the multiplicative wire pattern directly during pre-processing.

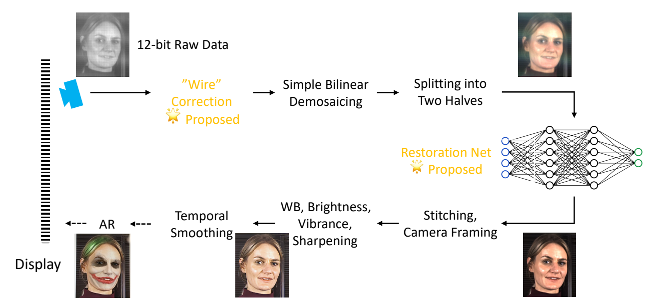

Fig. 5. Image processing pipeline: from RAW data to reconstructed image, to AR lens-applied image. We propose "wire" correction and a network to remove backscatter, blur and noise artifacts. The pipeline runs at 30fps.

Noise. The transparent display reduces 62% of light entering the camera, making shot noise more noticeable.

Reflection of the lens. In addition to backscatter, the glass also creates specular reflections, notably the camera lens. We build a protective enclosure around the camera, painted in black to minimize them. **See Sec 3.2 in the supplementary material.**

4 IMAGE RESTORATION

Informed by the image formation model, we first remove the wire effect through a pixel-wise multiplication as pre-processing, while other artifacts, especially the backscatter, are removed by a carefully designed neural network. Notice that both realistic training data and network design are essential for image restoration in the wild.

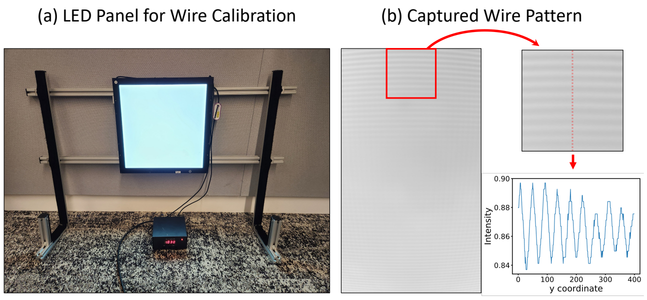

Fig. 6. Wire calibration. (a) We place a LED panel in front of the display to calibrate the wire pattern. (b) The captured wire pattern locally follows a sinusoidal pattern with a small amplitude.

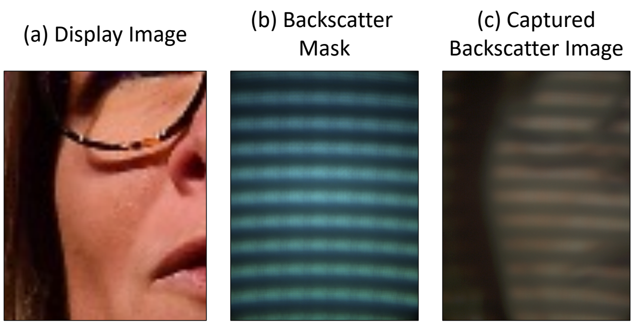

Fig. 7. Backscatter calibration and data collection. (a) Image shown on the display. (2) The backscatter mask: backscatter corresponding to a white image. (3) The captured backscatter image can be seen as the display image undergoing a low-pass filter and modulation by the backscatter mask.

Since ground truth images are hard to capture (especially with our setup), we adopt physics-based data synthesis with carefully calibrated parameters to enable realistic data synthesis. We then design a network architecture specifically focusing on removing the dominant backscatter artifacts by explicitly injecting knowledge about the backscatter.

Fig. 5 shows the image processing pipeline. To enable real-time computation at FHD resolution, we implement a highly optimized pipeline on two GPUs (with another optional GPU for AR effects), which runs at 30Hz with a latency of 24ms. **See Sec. 3.1 in the supplementary material.**

4.1 Physics-Based Calibration and Data Simulation

We start from the clean HDR and LDR images (as HDR data is scarce and there is no HDR video dataset) and add the degradations step by step following Eq. (1). We linearize the LDR images assuming a gamma randomly chosen between 1.8 and 2.2 before synthesizing the degradations. Below we discuss how to calibrate each degradation **(See Sec. 1 in the supplementary material for details).**

Wire effect. To calibrate the wire effect, we place an LED panel in front of the display to capture an all-white scene, as shown in Fig. 6(a). Fig. 6(b) shows a captured wire image. We also plot the intensity change along a specific column. The wire pattern locally follows a sinusoidal pattern with a small amplitude. At inference time, we divide the captured image by this captured wire image as pre-processing. Notice that to strictly invert the model proposed in Eq. (1), denoising must happen before wire removal. In practice we found that removing the wires first does not lead to visible artifacts.

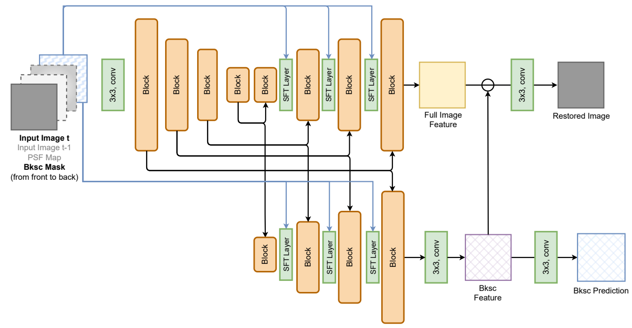

Fig. 8. Our proposed network: using backscatter mask as guidance and backscatter supervision for backscatter removal.

Backscatter. To calibrate the backscatter, we cover the camera and the display with a piece of black cloth such that the captured images only contain the backscatter. Fig. 7 shows an example of captured backscatter images. It appears that the backscatter image can be represented as a function of the display image and a *backscatter mask*, which we define as the backscatter created by displaying an all-white image). Instead of fitting this complicated function, we capture a large number of backscatter images, add them to the synthetic images, and let the restoration network learn to remove backscatter without knowing what content is being shown. We notice the backscatter can be quite strong when the pixels in front of the camera are displaying high intensities. To enhance our model's capabilities of removing strong backscatter, we *balance* the dataset by intentionally increasing the intensity of 1/3 of the images.

Point spread functions. We place a small white LED at 2 meters away where the image of the LED is close to one pixel wide and capture the 3-channel (RGB) HDR point spread function by gradually increasing the camera exposure time [Feng et al. 2021]. Since a Bayer pattern is used and applying 4/8-neighbor demosaicing methods can lead to inaccurate results, we slightly shift the LED's position and use all the data to fit a smooth PSF. We calibrate the PSFs at sampled locations and then interpolate them over the entire image.

Noise. We follow [Wei et al. 2021] to calibrate the dark noise, photon noise, and Gaussian noise.

4.2 Network Design

We observe that the backscatter is the dominant image degradation, as illustrated in Fig. 4. Existing convolution-based architectures cannot handle challenging backscatter due to their spatial-invariant nature. This motivates us to integrate the position and strength information of the backscatter into the proposed architecture.

We introduce the backscatter mask as an additional input to our network. Our framework (Fig. 8) comprises two key branches: the

Fig. 9. Mechanical design of our AR mirror device.

image restoration branch and the backscatter prediction branch. The image restoration branch focuses on extracting deep features and removing common degradation types such as blur [Kwon et al. 2021] and noise, yet it retains the backscatter artifact. Concurrently, the backscatter prediction branch estimates the distribution of backscatter and effectively subtracts it from the pre-restored image.

Both branches are constructed upon the encoder-decoder framework, incorporating skip connections to facilitate efficient information flow. To optimize computational performance and promote information sharing, the encoder and residual connections are shared by both branches. Supervision for these branches comes from clean images and corresponding backscatter ground truth, respectively.

Specifically, the encoder contains four convolutional NAFNet blocks [Chen et al. 2022], each of which has five convolution layers with the SimpleGate activation [Chen et al. 2022]. The main architecture for the two branches is kept the same for simplification. The decoder is composed of four NAFNet blocks, with SFTLayer [Wang et al. 2018b] appended to the output of each block. We use the convolution layer with stride=2 for downsampling the resolution and PixelShuffle for upsampling. The network is compact for real-time performance, which has 8.63M parameters and 225.5 GFLOPs when the input is 1152×1152. The backscatter mask serves as conditional information for each SFTLayer, which allows the model to adapt to spatial variations and effectively remove the backscatter.

5 EXPERIMENTS

Mechanical Design. We designed and assembled a frame and facade to securely mount the camera (Basler acA2040-120uc with an Edmund Optics 6mm/F1.85 lens) and the display (LG-55EW5G-V), as shown in Fig. 9. The system is designed to maintain the rigidity between the camera and screen to prevent drifting from calibration, while simultaneously providing flexibility to adjust the camera's position and orientation for fast prototyping. A black background and casing is used to hide the camera when observed from the front. **See Sec. 3.2 in the supplementary material.**

Network training details. We trained the proposed model with the carefully synthesized data which includes backscatter, blur, noise, saturation artifacts. Wire artifacts were not included as they were handled in a preprocessing step unlike traditional methods [Koh et al. 2022; Zhou et al. 2021]. We used the HDR dataset [Feng et al. 2021] and LDR dataset Inter4K [Stergiou and Poppe 2022] to synthesize the data. We generated 3000 pairs of data in total with 1080P resolution. During training, the images were randomly cropped into 512×512 patches (better than the normal 256×256 choice). We implemented our model with the PyTorch [Paszke et al. 2019] and BasicSR [Wang et al. 2018a] and use BasicSR's default training parameters. The model was trained on 8 NVIDIA V100 GPUs for two days.

Evaluation on real data. We captured two real datasets for evaluation: (1) Human-TV and (2) Human-Real. We placed an additional camera of the same model beside the display. We then used a 70" TV to display human action videos which were recorded by both the UDC camera and the side camera; after homography transformation and color normalization, the side camera provides the ground truth for the UDC restoration. We call this dataset Human-TV. We also used the stereo to capture real human actions, where the side camera's data cannot be the ground truth due to the stereo disparity but can act as subjective reference. This dataset is called Human-Real.

We compared our method to all existing UDC restoration methods with code available, which are retrained using our dataset for fair comparison. This includes deep atrous guided filter [Sundar et al. 2020], DISCNet [Feng et al. 2021], MIMO-UDC [Zhu et al. 2023], UDCUNet [Liu et al. 2022], and BNUDC [Koh et al. 2022]. Backscatter is similar to the haze which lowers the contrast of the images. Therefore, we included two dehazing methods, dark channel prior (Dehaze-DCR) [He et al. 2010] and the recent learning-based Dehaze-RIDCP [Wu et al. 2023]. Lastly, we also included Restormer [Zamir et al. 2022] which is for general image restoration.

The comparison results are shown in Fig. 10 on Human-TV dataset, Fig. 11 on Human-Real dataset, and Table 1 for metrics on Human-TV dataset as ground truth is available. All the qualitative and quantative results show that our method can remove distortions effectively, especially the backscatter and wire artifacts, and outperforms all other methods significantly.

Ablation study. Our model's strong restoration capability can be attributed to balancing backscatter augmentation (Sec. 4.1) and adding the backscatter knowledge through the backscatter branch. Table 1 shows that the performance drops significantly once either of the two designs is removed. We also propose another variant network that stacks two consecutive frames as 6-channel input ("Ours_2f"), which further improves the results. One explanation is that the optical flows of the human and the backscatter are independent, which provides a cue for the network to separate the two. See Fig. 12 for visual results.

6 USER STUDY

We conduct a comprehensive user study to gauge the user perception and experience. We directly compare our UDC design with a conventional Side Camera Display (SCD) design, which involves the same display setup with a side camera (same model), under identical lighting conditions. The sole variable is the camera's location. The UDC video frames undergo processing through our UDC pipeline, while the SCD frames undergo conventional processing

Fig. 10. Comparisons on real data Human-TV. Our method can remove backscatter and wire effectively and outperforms other methods significantly. Here Ours_1f is our method with only current frame as the input to ensure fair comparisons, and GT_bksc is ground truth of the backscatter.

Fig. 11. Comparisons on real data Human-Real. Our method can remove backscatter and wire effectively while others cannot. Here, Ours_1f and Ours_2f refer to our methods using only the current frame and the current+previous frames, respectively.

Fig. 12. Ablation study. The visual results show that a good training dataset with balanced backscatter simulations, the backscatter guidance in the network design, and previous frame as the input help improve the performance.

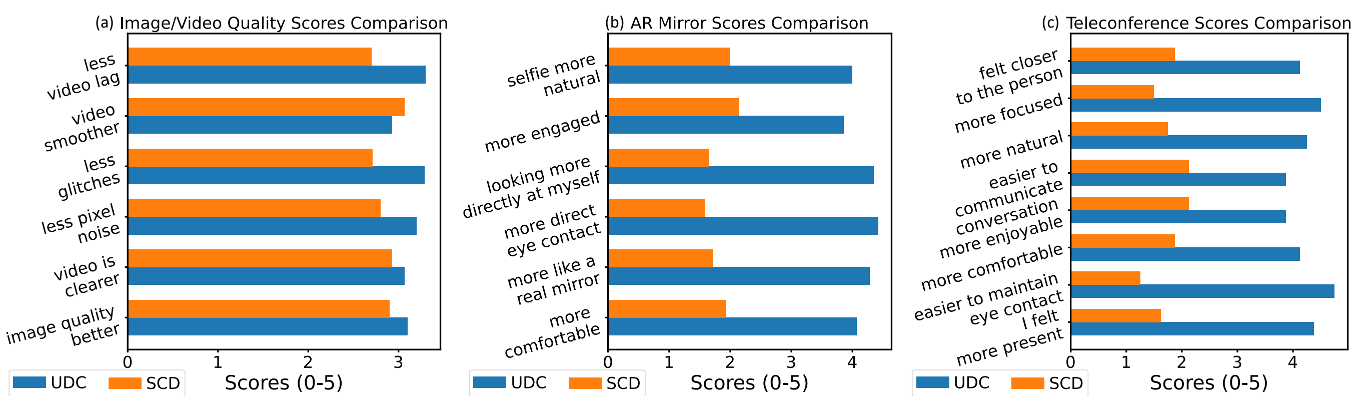

Fig. 13. User study results. We compare our UDC with a conventional Side Camera Display (SCD). Higher score = agree with the statement more.

to ensure optimal image quality. Following the design of similar studies [Lawrence et al. 2021], we asked each participant to interact with both systems, and then provide feedback through Likert-style and open-response questions, addressing aspects such as image quality, comfort level, realism, and user experience.

Fig. 13(a) shows the aggregated scores on image quality. Our UDC system exhibits comparable or even better performance over SCD in all aspects, demonstrating that, with our processing pipeline, the UDC design does not compromise perceptual image quality. Fig. 13(b) shows that the UDC design is significantly better received

Table 1. Quantitative results on real dataset Human-TV. Our single-frame method (Ours_1f) outperforms existing methods. By stacking two consecutive frames as input (Ours_2f), the results are further improved. Ablation study show that our design choices around backscatter help effectively.

Method	Input	Dehaze1	Dehaze2	D.a.g.f.	DISCNet	MIMO-UDC	UDCUNet	BNUDC	Restormer	Ours_1f	Ours_2f	Ours w/o Bal. Bksc	Ours W/o Bksc Guid.
PSNR ↑	21.50	17.44	14.76	25.00	23.01	23.55	23.79	24.59	25.00	<u>32.32</u>	**33.46**	25.20	29.34
SSIM ↑	.7654	.6117	.4762	.8003	.8255	.8504	.7727	.8432	.8329	<u>.9273</u>	**.9314**	.8362	.8443
LPIPS ↓	.2926	.4268	.5223	.2104	.1246	.1645	.1654	.1576	.1280	<u>.1151</u>	**.1085**	.1592	.1154

(a) How People Interact with Our AR Mirror

(b) Gallery of Selfies Captured through Our AR Mirror

Fig. 14. Our AR mirror in the field.

by the participants, proving the substantial impact of user perspective and eye contact on overall user experience. In addition to functioning as an AR mirror, the two display systems can also be used for teleconferencing, where perspective and eye contact are similarly important. Fig. 13(c) highlights that the UDC system ensure easier, more natural and comfortable communication. **See Sec. 4 in the supplementary material for detailed user study design and participants' responses to open questions.**

7 CONCLUSION AND FUTURE WORK

In conclusion, we propose a novel AR mirror system that allows a seamless, perspective-aligned user experience, which is enabled by placing the camera behind a transparent display. We derive on a rigorous image formation model, which allows us to design a restoration method that deals with the underexplored wire and backscatter artifacts. Our system significantly improves user experience without compromising perceptual image quality.

Further improving the computational efficiency. Although the current AR mirror system can run in real-time, it requires two high-end GPUs, which lifts cost and limits deployability. It is possible to adopt knowledge distillation, neural architecture search, weights quantization or other state-of-the-art approaches to further improve the efficiency of the proposed system.

Under-display camera array for true mirror experience. Although the proposed system provides an aligned perspective when a user of average height stands in the center of the display, the perspective is not perfect when an extremely tall (or short) user stands in the corner of the field-of-view. To realize true-perspective mirror experience, it is possible to build an array of cameras behind the display, such that at any instant the closest camera to the user is activated. Furthermore, it is also possible to employ novel view synthesis to generate interpolated views between cameras [Lim et al. 2021], which may further improve the accuracy of eye contact.

Calibration simplification and motorized camera. While the mechanical design of the system ensures that the camera calibration does not drift over time, the multi-step calibration still need to be performed on each camera which could limit mass production. It is possible to develop auto-calibration techniques to simplify this process. For example, a standard calibration pattern can be shown on the screen, and the difference between the captured backscatter images can be used to estimate the relative camera pose. Auto-calibration also enables a motorized camera that can move vertically, adapting to users of different heights.

Future adoption. In addition to in-lab evaluation and user study, we have also shown our AR mirror to the public at events, which received widespread appreciation for the perspective-aligned experience, as illustrated in Fig. 14. The positive reception from users validates the real-world applicability and user-centric design of our approach. This enthusiastic response signals a promising trajectory for the future adoption and integration of perspective-aligned displays in various interactive environments.

ACKNOWLEDGEMENT

We thank Bobby Murphy, Betsy Kenny Lack, Adrian Bradford, Dhritiman Sagar, Yining Liang, and Eric Hu for their support on the project. We also appreciate those who assisted with our experiments and the user study. Lastly, we are grateful to the anonymous reviewers for their feedback, which improved this paper.

REFERENCES

Kyusu Ahn, Byeonghyun Ko, HyunGyu Lee, Chanwoo Park, and Jaejin Lee. 2023. UDC-SIT: A Real-World Dataset for Under-Display Cameras. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Fraser Anderson, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2013. YouMove: enhancing movement training with an augmented reality mirror. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 311–320.

Tobias Blum, Valerie Kleeberger, Christoph Bichlmeier, and Nassir Navab. 2012. mirracle: An augmented reality magic mirror system for anatomy education. In *2012 IEEE Virtual Reality Workshops (VRW)*. IEEE, 115–116.

Felix Bork, Roghayeh Barmaki, Ulrich Eck, Pascal Fallavolita, Bernhard Fuerst, and Nassir Navab. 2017. Exploring non-reversing magic mirrors for screen-based augmented reality systems. In *2017 IEEE virtual reality (VR)*. IEEE, 373–374.

Felix Bork, Leonard Stratmann, Stefan Enssle, Ulrich Eck, Nassir Navab, Jens Waschke, and Daniela Kugelmann. 2019. The benefits of an augmented reality magic mirror system for integrated radiology teaching in gross anatomy. *Anatomical sciences education* 12, 6 (2019), 585–598.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. 2022. Simple baselines for image restoration. In *European Conference on Computer Vision*. Springer, 17–33.

James Davis, Yi-Hsuan Hsieh, and Hung-Chi Lee. 2015. Humans perceive flicker artifacts at 500 Hz. *Scientific reports* 5, 1 (2015), 7861.

Ruicheng Feng, Chongyi Li, Huaijin Chen, Shuai Li, Jinwei Gu, and Chen Change Loy. 2023. Generating Aligned Pseudo-Supervision from Non-Aligned Data for Image Restoration in Under-Display Camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5013–5022.

Ruicheng Feng, Chongyi Li, Huaijin Chen, Shuai Li, Chen Change Loy, and Jinwei Gu. 2021. 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*. 662–671.

Ruicheng Feng, Chongyi Li, Shangchen Zhou, Wenxiu Sun, Qingpeng Zhu, Jun Jiang, Qingyu Yang, Chen Change Loy, Jinwei Gu, Yurui Zhu, et al. 2022. Mipi 2022 challenge on under-display camera image restoration: Methods and results. In *European Conference on Computer Vision*. Springer, 60–77.

Richard Hartley and Andrew Zisserman. 2003. *Multiple view geometry in computer vision*. Cambridge university press.

Kaiming He, Jian Sun, and Xiaoou Tang. 2010. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence* 33, 12 (2010), 2341–2353.

Rachel Jacobs, Holger Schnädelbach, Nils Jäger, Silvia Leal, Robin Shackford, Steve Benford, and Roma Patel. 2019. The performative mirror space. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics* 42, 4 (2023), 1–14.

Jaihyun Koh, Jangho Lee, and Sungroh Yoon. 2022. BNUDC: a two-branched deep neural network for restoring images from under-display cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1950–1959.

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

Jason Lawrence, Danb Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G Desloge, Tommy Fortes, Eric M Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, et al. 2021. Project starline: A high-fidelity telepresence system. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–16.

LG. 2019. Transparent OLED Displays. https://www.lg.com/us/business/digital-signage/oled-signage/transparent-oled-displays. [Online; accessed 1-Jan.-2024].

Sehoon Lim, Luming Liang, Yatao Zhong, Neil Emerton, Tim Large, and Steven Bathiche. 2021. 18-3: Free Viewpoint Teleconferencing Using Cameras Behind Screen. In *SID Symposium Digest of Technical Papers*, Vol. 52. Wiley Online Library, 218–221.

Xina Liu, Jinfan Hu, Xiangyu Chen, and Chao Dong. 2022. UDC-UNet: Under-Display Camera Image Restoration via U-shape Dynamic Network. In *European Conference on Computer Vision*. Springer, 113–129.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. 2023. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713* (2023).

Katie Luo, Guandao Yang, Wenqi Xian, Harald Haraldsson, Bharath Hariharan, and Serge Belongie. 2021. Stay positive: Non-negative image synthesis for augmented reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10050–10060.

Ma Meng, Pascal Fallavollita, Tobias Blum, Ulrich Eck, Christian Sandor, Simon Weidert, Jens Waschke, and Nassir Navab. 2013. Kinect for interactive AR anatomy learning. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 277–278.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Motorola. 2021. Motorola Edge X30. https://www.gsmarena.com/motorola_edge_x30-11262.php. [Online; accessed 1-Jan.-2024].

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)* 41, 4 (2022), 1–15.

NVIDIA. 2020. NVIDIA Broadcast App. https://www.nvidia.com/en-us/geforce/broadcasting/broadcast-app/. [Online; accessed 17-May.-2024].

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

Zong Qin, Yu-Hsiang Tsai, Yen-Wei Yeh, Yi-Pai Huang, and Han-Ping David Shieh. 2016. See-Through Image Blurring of Transparent Organic Light-Emitting Diodes Display: Calculation Method Based on Diffraction and Analysis of Pixel Structures. *Journal of Display Technology* 12, 11 (Nov. 2016), 1242–1249. https://doi.org/10.1109/JDT.2016.2594815

Daniel Saakes, Hui-Shyong Yeo, Seung-Tak Noh, Gyeol Han, and Woontack Woo. 2016. Mirror mirror: An on-body t-shirt design system. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 6058–6063.

Samsung. 2023. A look at the Under Display Camera (UDC) on the Galaxy Z Fold4 and Fold5. https://www.samsung.com/latin_en/support/mobile-devices/a-look-at-the-under-display-camera-udc-on-the-galaxy-z-fold4-and-fold5/. [Online; accessed 1-Jan.-2024].

Alexandros Stergiou and Ronald Poppe. 2022. Adapool: Exponential adaptive pooling for information-retaining downsampling. *IEEE Transactions on Image Processing* 32 (2022), 251–266.

Varun Sundar, Sumanth Hegde, Divya Kothandaraman, and Kaushik Mitra. 2020. Deep atrous guided filter for image restoration in under display cameras. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 379–397.

Xintao Wang, Ke Yu, Kelvin CK Chan, Chao Dong, and Chen Change Loy. 2018a. BasicSR: Open source image and video restoration toolbox. *GitHub* (2018).

Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2018b. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 606–615.

Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. 2021. Physics-based noise modeling for extreme low-light photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 8520–8537.

Rui-Qi Wu, Zheng-Peng Duan, Chun-Le Guo, Zhi Chai, and Chongyi Li. 2023. RIDCP: Revitalizing Real Image Dehazing via High-Quality Codebook Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22282–22291.

Xiaomi. 2021. Xiaomi MIX 4. https://en.wikipedia.org/wiki/Xiaomi_MIX_4. [Online; accessed 1-Jan.-2024].

Anqi Yang, Eunhee Kang, Hyong-Euk Lee, and Aswin C Sankaranarayanan. 2023. Designing Phase Masks for Under-Display Cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10637–10645.

Anqi Yang and Aswin C Sankaranarayanan. 2021. Designing display pixel layouts for under-panel cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 7 (2021), 2245–2256.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5728–5739.

ZERO10. 2024. ZERO10 is a fashion AR try-on company. https://zero10.ar/. [Online; accessed 24-Jan.-2024].

Xiaoshuai Zhang, Sai Bi, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. 2022. Nerfusion: Fusing radiance fields for large-scale scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5449–5458.

Yuqian Zhou, Michael Kwan, Kyle Tolentino, Neil Emerton, Sehoon Lim, Tim Large, Lijiang Fu, Zhihong Pan, Baopu Li, Qirui Yang, et al. 2020. UDC 2020 challenge on image restoration of under-display camera: Methods and results. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 337–351.

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

Yurui Zhu, Xi Wang, Xueyang Fu, and Xiaowei Hu. 2023. Enhanced Coarse-to-Fine Network for Image Restoration from Under-Display Cameras. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*. Springer, 130–146.

ZTE. 2020. ZTE Axon 20 5G: Path To New Vision The World's First Under-Display Camera Smartphone. https://ztedevices.com/en-gl/zte-axon-20-5g/. [Online; accessed 1-Jan.-2024].