Efficient Depth- and Spatially-Varying Image Simulation for Defocus Deblur

Xinge Yang[1*] Chuong Nguyen[2] Wenbin Wang[2] Kaizhang Kang[1]
Wolfgang Heidrich[1] Xiaoxing Li[2]

KAUST[1] Meta Reality Labs[2]

Abstract

Modern cameras with large apertures often suffer from a shallow depth of field, resulting in blurry images of objects outside the focal plane. This limitation is particularly problematic for fixed-focus cameras, such as those used in smart glasses, where adding autofocus mechanisms is challenging due to form factor and power constraints. Due to unmatched optical aberrations and defocus properties unique to each camera system, deep learning models trained on existing open-source datasets often face domain gaps and do not perform well in real-world settings. In this paper, we propose an efficient and scalable dataset synthesis approach that does not rely on fine-tuning with real-world data. Our method simultaneously models depth-dependent defocus and spatially varying optical aberrations, addressing both computational complexity and the scarcity of high-quality RGB-D datasets. Experimental results demonstrate that a network trained on our low resolution synthetic images generalizes effectively to high resolution (12MP) real-world images across diverse scenes.

1. Introduction

The demand for computational photography algorithms to perform well in defocus scenarios is growing rapidly, as modern optical lenses often employ large apertures [3, 4, 35, 52]. Large aperture sizes can reduce noise levels; however, they also result in a shallow depth of field, causing out-of-focus objects to appear blurry [21]. In many cases, the defocus effect degrades the image quality and reduces the amount of information that can be extracted from the physical world, especially for edge-device cameras, for example, those on smart glasses.

Incorporating autofocus modules [13, 17, 27] or novel extended depth-of-field optics [37, 52, 53] can ensure a wide focus range; however, these solutions are often constrained by factors such as device form factor, battery consumption, and immature manufacturing technologies. On the algorithmic side, both classical image deblurring methods [11, 23, 42, 59] and deep learning approaches [1, 25, 28, 30, 32, 33, 49, 55] have been explored for defocus de-

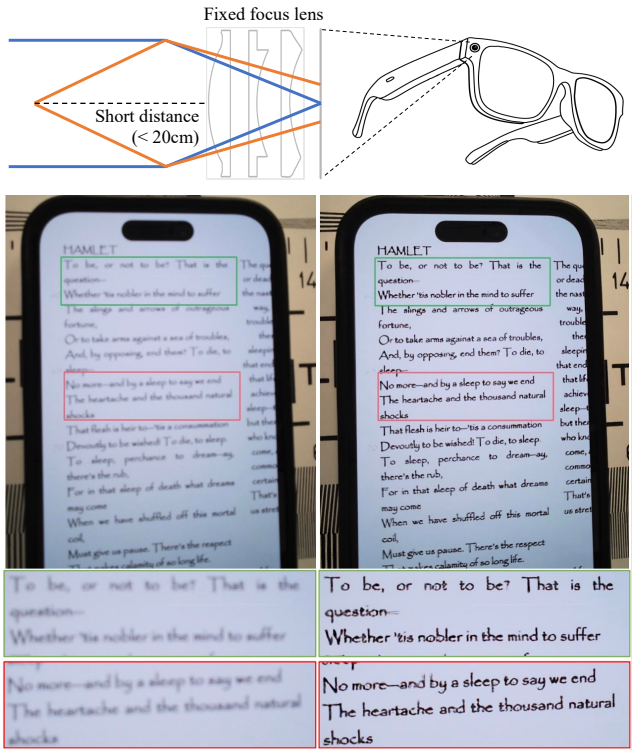

Figure 1. **Motivation and real-world results for our proposed depth-varying dataset synthesis approach.** Large aperture fixed-focus lenses struggle to capture clear images at short distances (typically <20 cm), creating challenges for devices like smart glasses to perceive the physical world. Our efficient depth-varying dataset synthesis approach enhances computational photography algorithms for real-world defocus scenes. Bottom row compares the raw captured image (left) with our restored result (right), demonstrating promising results in defocus deblurring, optical aberration correction, and noise reduction.

*Work partially done during Meta internship. Inquiries can be directed to: xinge.yang@kaust.edu.sa

blurring and restoration of image aberrations. Among these methods, neural networks typically yield higher image quality with fewer restoration artifacts. However, mismatches in optical aberrations, defocus scales, and noise statistics across different camera sensors prevent models trained on existing open-source datasets [1, 22, 25, 31] from being directly applied to customized cameras, resulting in a significant domain gap.

To address the dataset gap, either new real-world datasets can be captured, or synthetic datasets can be employed. Capturing real-world datasets is generally expensive and time-consuming due to the requirement of covering varying spatial positions and depths. Synthetic dataset generation for defocus scenarios necessitates the simultaneous simulation of both depth-dependent defocus (caused by different pixel depths) and spatially-varying optical aberrations (caused by different pixel radial positions). However, existing optical simulation approaches either ignore depth-dependent defocus, considering only the focal plane [8, 41, 43], or overlook spatially-varying optical aberrations [14, 45]. There are two major challenges: First, simultaneously accounting for both optical spatial and depth variances renders the image simulation process computationally expensive [51]. Second, there is a lack of high-quality, high-resolution RGBD datasets suitable for photorealistic dataset synthesis. These two challenges hinder large-scale dataset synthesis for dynamic real-world scenes. In summary, suitable training datasets for defocus deblur networks are crucial yet often lacking, especially when working with specific camera systems.

In this paper, we propose an efficient and scalable dataset synthesis approach for optical systems with depth- and spatially-varying effects. We first demonstrate (Sec.4) that synthetic datasets assuming planar depth perform poorly in real-world scenarios exhibiting significant defocus effects. Subsequently, we use a smart glasses fixed-focus camera as a test case to evaluate our proposed dataset synthesis approach. We comprehensively model depth-dependent defocus, spatially-varying optical aberrations, sensor quantization errors, and sensor noise. To address the limited availability of RGBD datasets, we apply DepthAnythingV2 [50] to high-quality RGB datasets and appropriately scale the estimated depth maps within our pipeline. Recognizing that spatial variance is minimal within small image patches from a 12-megapixel camera sensor, we disregard local spatial variance in low-resolution training batches. Instead, we incorporate positional encoding for each pixel to capture global spatial variance and encode ISO values to represent noise levels. This approach efficiently addresses the computational and dataset challenges inherent in defocus image simulation, enabling large-scale training data generation (Sec. 3).

Our experimental results show that a simple network trained on low-resolution synthetic images can deliver promising results on 12-megapixel full-resolution images across diverse real-world scenes (Sec. 4). The proposed approach is efficient as it supports fast on-the-fly training image synthesis, removes the need for point spread function (PSF) calibration or real-world image fine-tuning. With our proposed approach, a fixed-focus camera can image clearly for close objects, extending the usable scenarios for many applications. Building on this success, we demonstrate downstream applications for daily use cases with smart glasses, including short-distance optical character recognition (OCR), and 3D digital asset generation, where our proposed approach can greatly improve the final quality. In summary, our key contributions are as:

- We propose an efficient and scalable dataset synthesis approach that simultaneously models both spatially varying optical aberrations and depth-dependent defocus.
- We address the lack of high-resolution RGBD datasets by applying pseudo depth maps, generated using state-of-the-art depth estimation models, to augment existing high-quality RGB datasets.
- We establish an end-to-end training pipeline that effectively generalizes from low-resolution synthetic training data to high-resolution real-world images.

2. Related works

2.1. Photorealistic Synthetic Dataset

High-fidelity synthetic datasets with accurate physical modeling are effective to train networks that can generalize well to the real world [2, 5, 8, 47]. Till now, research works have been done for noise models [5, 47], spatially-varying optical aberrations [8, 41, 43], with the promising pipeline unprocessing images from the sRGB space back to the RAW image space to simulate sensor noise [5, 48] and optical aberrations in the RAW domain [8].

However, current research works usually focus only on optical simulation at the focus plane, while ignoring defocus effects [8, 41, 43], or modeling defocus effects while ignoring spatially varying optical aberrations [14, 34, 44, 45]. The challenge arises from the rapidly changing optical aberrations, including both variations across the imaging plane and variations with distance from the camera. Accurately modeling these effects not only requires a large degree of freedom to store the PSFs, but is also computationally expensive for high-resolution dataset generation, as each pixel has independent spatial positions and depths. The limited existing works [24, 50] are not applicable for large-scale training dataset synthesis. In this work, we propose an efficient and accurate image simulation approach that considers both depth-dependent defocus effects and spatially varying aberrations. The proposed method not only greatly reduces the computational time required for synthetic training data

generation but also maintains high fidelity in the simulated images.

2.2. Single Image Defocus Deblur

Image deblurring is a long-standing problem that aims to recover sharp and clear images from various types of blur, such as motion blur [20, 26, 39, 40, 57], defocus blur [22, 25, 28–31, 33, 34], and environment blur [19, 36, 38]. Both classical methods [11, 23, 42, 59] and deep learning approaches [1, 25, 28, 30, 32, 33, 49] have been explored and have shown promising results. Typically, large amounts of high-quality data are required to train deep networks effectively. The existing defocus deblurring dataset [1, 22, 25, 31] is quite limited. Additionally, considering that the defocus characteristics and optical aberrations vary across different lenses, a network model trained on the open-source dataset often cannot be directly applied to images captured with another lens. Capturing enough high-quality training data for each lens is impractical and time-consuming. In this work, we propose a synthetic dataset generation approach that allows machine learning engineers to train image deblurring networks directly on synthetic data, with promising performance on real captured images.

3. Methods

We first discuss our efficient depth-varying defocus and spatially-varying aberrated data generation pipeline in Sec. 3.1, as well as illustrated in Fig. 2. In Sec. 3.2, we discuss the dataset preparation and more implementation details

3.1. Efficient Defocus and Aberration Simulation

To generate realistic synthetic images, we unprocess RGB images to RAW signal space by inverting the image signal processing (ISP) pipeline. This is based on two main considerations: first, the PSF of the lens is typically defined in the radiance space, while the camera ISP will apply nonlinear processing to the RAW signals [8]. Second, the sensor noise, which greatly affects the image reconstruction results, also suffers from postprocessing [5]. The unprocessing and PSF convolution pipeline can be expressed as

$$\mathbf{I}^{'} = \mathbf{P} * \mathcal{F}^{-1}(\mathbf{I}), \tag{1}$$

where $\mathbf{I}$ is the input image, $\mathcal{F}^{-1}$ represents the "unprocess" as described in [5], $\mathbf{P}$ represents the PSF function and $*$ denotes the convolution operation. $\mathbf{I}^{'}$ is the blurred signal in the RAW signal space. Note that the PSF of the camera changes across both the image plane and different depths, which means each image pixel has independent imaging characteristics.

For an accurate image simulation, ideally, we have to perform convolution between each pixel with spatially-varying depth-dependent PSF $\mathbf{P}(u, v, z_{[u,v]})$, which can be expressed as

$$\mathbf{I}^{'}_{[u,v]} = \mathbf{P}(u, v, z_{[u,v]}) * \mathcal{F}^{-1}(\mathbf{I})_{[u,v]}, \tag{2}$$

where u and v denote the normalized spatial position of the pixel on the image plane, and z denotes the corresponding depth value. However, both the storage of per-pixel PSF and the per-pixel convolution are computationally expensive. Although some recent works successfully speed up the PSF representation problem [10, 51, 54], the per-pixel convolution computation is still a time-consuming problem, particularly for large-scale dataset generation.

Based on two observations that (1) neural networks process high-resolution images by smaller patches, and (2) in a small image patch of a high resolution sensor, in-plane spatial positions (u, v) between different pixels do not vary too much, we believe **it is reasonable to ignore the in-plane spatial variance during the network training stage**. This simplification allows us to only focus on the depth variance of pixels, which significatnly improve the image simulation efficiency. Consequently, we simplify Eq.(2) to

$$\mathbf{I}^{'}_{[u,v]} = \mathbf{P}(z_{[u,v]}) * \mathcal{F}^{-1}(\mathbf{I})_{[u,v]}. \tag{3}$$

For a PSF at an unknown depth, it can be linearly interpolated by its neighbour as long as the sampling is dense enough. Also, the convolution operation is a linear operation, we can further rewrite Eq.(3) as

$$\mathbf{I}^{'}_{[u,v]} \approx (\sum_{z \in \mathbf{Z}} \alpha_{[u,v]} \mathbf{P}_z) * \mathcal{F}^{-1}(\mathbf{I})_{[u,v]} \tag{4}$$

$$\approx \sum_{z \in \mathbf{Z}} \alpha_{[u,v]} (\mathbf{P}_z * \mathcal{F}^{-1}(\mathbf{I}))_{[u,v]}, \tag{5}$$

where $\mathbf{Z}$ is a discrete set of predefined depths, and α denotes the weight of interpolation. Eq.(4) denotes the interpolation in the PSF space. However, a costly per-pixel convolution is still required. To further simplify computation, Eq.(5) converts the problem to first compute convolution between base PSF functions at different depth $z \in \mathbf{Z}$ and input images, then interpolate in the image space. This greatly reduces the computation resources as matrix production and image convolution can be sped up with modern computation algorithms and hardware. Fig. 3 compares synthetic images with and without depth-dependent PSF.

We follow the read and shot noise model as described in [5] to simulate the sensor raw. Here, $\lambda_{read}, \lambda_{shot}$ are the read and shot noise variance that depends on both analogue and digital gain. These two gain levels are set as a direct function of the ISO light sensitivity level, chosen manually by the user or automatically by the camera. The ISO can be read from the metadata.

$$\mathbf{I}^{''} \sim \mathcal{N}(\mu = \mathbf{I}^{'}, \sigma^2 = \lambda_{read} + \lambda_{shot}\mathbf{I}^{'}) \tag{6}$$

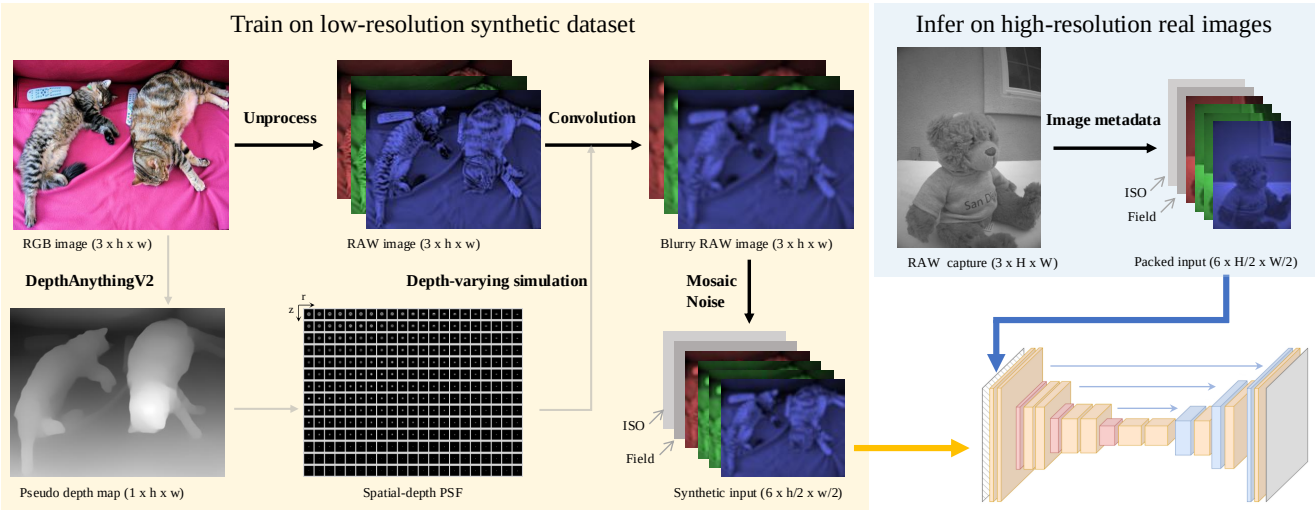

Figure 2. **Training and inference pipeline of the proposed approach.** Left: Images are unprocessed from RGB space to RAW space to simulate defocus blur, optical aberrations, sensor quantization, and noise. A pseudo depth map is predicted using the pretrained DepthAnythingV2 [50] model, then randomly scaled and utilized in the depth-varying defocus and spatially-vary aberration simulation. Noise signal at a random ISO level is added to the blurry RAW image. The image data, ISO channel, and radial field map are then packaged as network inputs. Top Right: During the inference stage on real-world images, the ISO value is read from photograph metadata, and the field map is computed on full-resolution images. Bottom Right: Instead of relying on complicated network architectures, a simple network (NAFNet [7]) is adopted for image reconstruction.

Finally, consider a b-bit sensor (typically $b = 10$), the b-bit $\mathbf{I}^{'''}$ can then be computed from $\mathbf{I}^{''}$ as follows:

$$\mathbf{I}^{'''} = \lfloor f_b(\mathbf{I}^{''}) \rceil \tag{7}$$

where $\lfloor . \rceil$ is *rounding* to the nearest integer operator and $f_b(\mathbf{y}) = \min(\mathbf{y}, 2^b - 1)$ is a b-bit clipping function.

The network $\mathcal{U}$ is trained to reconstruct clear RAW images $\mathcal{F}^{-1}(\mathbf{I})$ from blurry and noisy RAW inputs $\mathbf{I}^{'''}$. The loss function is defined in the RGB space to prioritize the quality of the final RGB images. Specifically, the loss function is formulated as

$$\mathcal{L} = \mathcal{L}\left(\mathcal{F}'\left(\mathcal{F}^{-1}(\mathbf{I})\right), \mathcal{F}'(\mathcal{U}(\mathbf{I}^{'''})\right), \tag{8}$$

where $\mathcal{F}'$ denotes the ISP, which can differ from $\mathcal{F}$ used in the unprocessing stage. $\mathcal{U}(\mathbf{I}^{'''})$ is the network output, given the noisy RAW $\mathbf{I}^{'''}$ as input. Notably, we set the gamma parameter in $\mathcal{F}'$ to 2.0 to emphasize dark regions in the reconstruction results. The loss function $\mathcal{L}$ comprises both pixel loss (L_1) and perceptual loss (LPIPS [58]).

3.2. Scalable Dataset Preparation

For synthetic training data generation, we use RGB images from the Adobe5k dataset [6] with the unprocessing manner to obtain simulated RAW captures [5]. This RGB to RAW space unprocessing is of great importance especially when we want to generate dataset in specific scenarios, for example close distance optical character recognition (OCR).

To address the lack of large-scale high-resolution RGBD datasets for close-up scenes, we employ the DepthAnythingV2 [50] model for depth estimation from RGB images before the training stage. Since the state-of-the-art depth estimation models only give relative depth maps, we scale them to absolute depths within our target depth range with multiple random scaling strategies, including linear, quadratic, and exponential functions.

Besides the augmentation coming from the random depth scaling, data augmentation is also applied at multiple stages, including geometric and pixel augmentation of RGBD images, unprocessing from RGB to RAW images, PSF augmentation, and post-processing from RAW to RGB images. Particularly, for PSF augmentation, we randomly apply a Gaussian blur with a small standard deviation to the PSFs to simulate the lens manufacturing and assembly errors in the real world.

Experiments are conducted using Meta Ray-Ban smart glass camera with a fixed focus distance set to infinity. The camera lens has a large aperture size (f-number 2.2), causing objects at short distances to appear significantly blurry due to defocus effects. PSFs at different spatial positions are computed using ZEMAX [56] with internal lens data, encompassing 20 depth stops from 10 cm to infinity and 20 radial stops from the sensor origin to the maximum field-of-view (FoV). Noise statistics for the camera sensor are calibrated as discussed in existing literature [5, 15] and supplementary materials.

Table 1. **Quantitative evaluation on synthetic datasets across different synthetic dataset generation approaches.** The best performance for PSNR, SSIM, and LPIPS is highlighted. For each metric, ↑/↓ indicates that higher/lower values are better, respectively. The results demonstrate that both depth-varying simulation and the incorporation of auxiliary channels in the network input enhance image reconstruction performance.

Method	Training Dataset	Defocus	Input Data	PSNR ↑	SSIM ↑	LPIPS ↓
PolyBlur [9]	✗	✗	RGB	26.18	0.7205	0.3235
LaDKNet [33]	Captured	✓(Unmatched)	RGB	25.90	0.7165	0.4441
Chen et al. [7]	Synthetic	✗	RAW-ISO-Field	27.14	0.8196	0.2469
Ablation #1	Synthetic	✓	RAW	27.60	0.8223	0.2233
Ablation #2	Synthetic	✓	RAW-ISO	28.73	0.8589	0.1960
Ablation #3	Synthetic	✓	RAW-Field	28.14	0.8317	0.2013
Ours	Synthetic	✓	RAW-ISO-Field	**30.03**	**0.8808**	**0.1553**

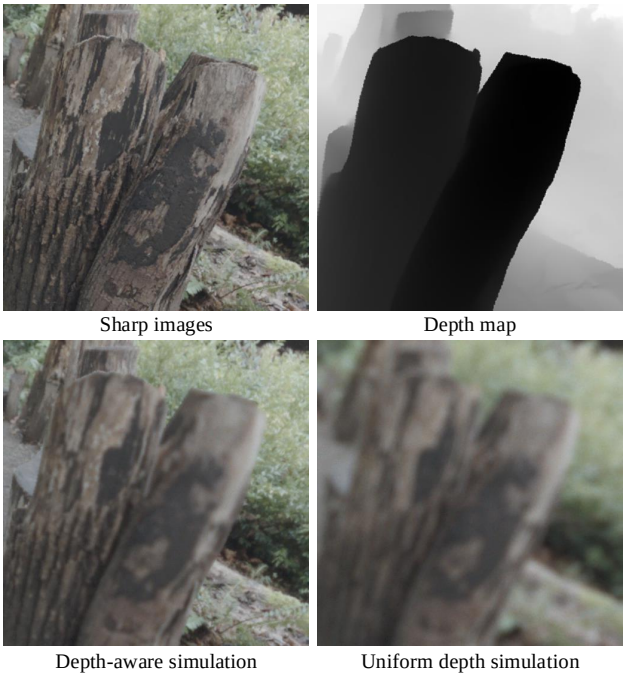

Figure 3. **Comparison of synthetic training image with and without depth-varying simulation.** Incorporating depth-varying defocus allows for more realistic simulations, reflecting real-world scenarios where objects at varying distances from the camera exhibit different levels of defocus.

3.3. Tunable Auxiliary Channels

We employee auxiliary channels to tackle the blind deblur problem, including a single-channel ISO and a single-channel normalized radial position map stacked with the input image (Fig. 2). For the ISO channel, we employ a uniform map with a constant ISO value, which is scaled by 0.001 to match the range of the input data. The use of auxiliary channels serves two primary purposes. First, image noise statistics are highly dependent on the ISO value, and blur profiles within an image patch are significantly affected by radial position. Without explicit ISO information during training and inference, the network must independently infer noise levels, increasing its complexity and potentially leading to averaged outputs. Including the radial position channel enables the network to perform patch-specific deblurring for different regions, thereby simplifying its task. Second, during inference, the auxiliary channels can be adjusted to achieve the desired subjective image quality. For instance, doubling the ISO value enhances denoising to produce smoother images, while halving it reduces denoising to better preserve texture details.

We select the NAFNet [7] for image reconstruction due to its efficiency and low latency. Training was performed using 256×256 Bayer RGGB images comprising 6 input channels (4 RGGB Bayer patterns and 2 auxiliary channels). The network was trained on 8 H100 GPUs with a batch size of 64 using the AdamW optimizer, initialized with a learning rate of 10^{-4}. Given the variability in spatial locations, gains, and depth scaling strategies, the training process was designed to cover a wide range of scenarios, thereby necessitating an extended training duration. In our experiments, training the network for 500 epochs takes approximately two days.

4. Results

We assess the effectiveness of our dataset synthesis approach by comparing it with various alternative methods. Specifically, we evaluate different dataset synthesis techniques, as well as state-of-the-art deep learning models trained on open-source defocus deblurring datasets.

4.1. Inference on Real-World 12-megapixel Images

After training network with our proposed synthetic dataset, we directly use the network for inference on 12-megapixel real-world images captured across various scenes. The RAW captured data with the camera metadata are used as the network input, and the output RAW data undergoes a minimum post-processing for visualization.

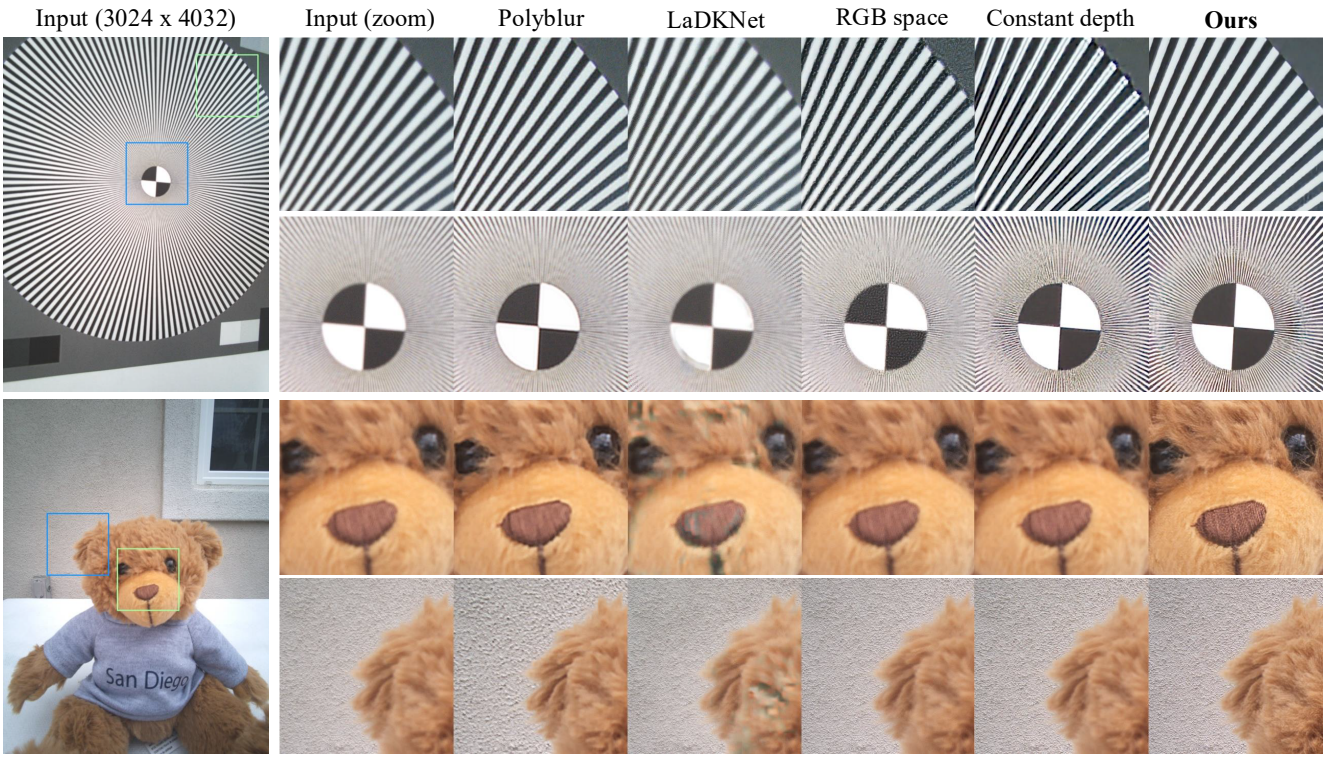

Figure 4. **Qualitative evaluation on 12MP real-world images with different defocus deblur methods and synthetic dataset generation.** From left to right: the classical deblurring algorithm ("Polyblur")[9] exhibits limited capability, failing to produce high-quality images. The network ("LaDKNet")[33], trained on an open dataset [1], cannot be directly applied to our camera captures and produces artifacts due to inconsistent optics and sensor noise. Simulation in RGB image space results in images with structural artifacts due to inaccurate noise modeling. A network trained on synthetic datasets without depth-varying image simulation fails to deblur effectively, and also struggles with scenes that have varying depth ranges. For example, in the second case, the network is confused by the sharp wall and therefore fails to recover the teddy bear. Our proposed synthetic dataset generation helps the network implicitly learn to detect sharp and blurry regions, successfully recovering clear details such as the face of the teddy bear.

In our experiments, we compare the defocus deblurring results obtained using a classical method ("Polyblur") [9], the state-of-the-art deep learning method ("LaDKNet") [33] with pretrained weights on the DPDD dataset [1], and a NAFNet network [7] trained with different synthetic dataset generation approaches. Several synthetic dataset generation approaches are chosen for comparison. In existing works, the most commonly used method is to directly apply PSF convolution to RGB images. Recent work by Chen et al. [8] demonstrates promising results in removing spatially-varying optical aberrations using only synthetic datasets; however, they consider only the spatial variance of optical aberrations within the focus plane and ignore the defocus effect. To conduct a fair comparison with their method, we randomly assign constant depth maps to the input RGB images during image simulation, with a higher probability assigned to closer depths. By comparing with them, we want to prove the importance of depth-varying optical simulation.

In Fig. 4, we show two example images reconstructed by different approaches. From the results, we observe that classical deblurring methods can slightly sharpen the images, but the improvement is quite marginal, and artifacts such as halos are introduced. Pretrained networks on open datasets cannot be generalized to our camera because the optical properties and noise statistics are different. Synthesizing data in the RGB image space leads to reconstructions with artifact noise patterns, especially in low-frequency regions such as the resolution chart. This issue arises because noise signals are not accurately modeled when simulating in the RGB image space, causing the network to fail to recover clear latent signals. For synthetic datasets without depth-varying optical simulation, the chart image is effectively deblurred, but the bear image remains blurry. This likely occurs because the network, trained with a constant defocus level across all pixels, is confused by sharp regions. For instance, the sharp background wall in the second example might lead the network to assume the entire image is in focus, preventing it from deblurring the bear's face. In short, training with constant depth maps makes the network

unable to distinguish spatially varying defocus (see Supplement for more examples).

In contrast, the network trained with our proposed synthetic dataset generation approach performs much better. In the resolution chart image, both high-frequency and low-frequency regions are well reconstructed from blurry and noisy inputs without producing reconstruction artifacts. In the toy bear example, the face of the toy is successfully recovered, along with the text on the label in another spatial location. We believe this is because **with depth-varying optical simulation, the network model implicitly learns defocus detection from the spatially varying blurry inputs**, even though no explicit depth information is given, which explains the success and generality in the real world. Additional comparison examples on real-world captured images can be found in the Supplement.

4.2. Quantitative Evaluation on Synthetic Dataset

To quantitatively assess the effectiveness of our proposed method, we established a synthetic validation dataset that fully simulates spatially varying and depth-dependent optical aberrations, defocus, sensor noise at various ISO levels, and sensor quantitative errors. Specifically, we selected 2,000 RGB images from the EBB! [16] dataset, each containing both close-up subjects and background scenes. The images were center-cropped and downsampled to a 512×512 resolution. Starting with all-in-focus images (with both subject and background in sharp focus), we applied depth estimation, random depth scaling, unprocessing, pixel-varying PSF convolution, and noise injection. For pixel-wise PSF calculation, the given PSF and the estimated depth map are used to interpolate the PSF for each image pixel [54]. For pixel-wise PSF convolution, we adopt the folding calculation method [50]. Notably, the validation data incorporates spatial variance within a local image patch, which better reflects the real world and enables a more realistic evaluation of different methods.

Presented in the Table 1, we compare defocus deblur results between a classical method (PolyBlur [9, 12]), the state-of-the-art defocus deblur model (LaDKNet [33]) trained on open source datasets (DPDD [1]), networks trained without considering varying depth maps [8], and networks trained with our depth-varying synthetic dataset. Experimental results demonstrate that all comparison methods cannot achieve satisfying results. Specifically, classical deblur methods can slightly sharpen the images, while the overall image quality is not satisfying. For LaDKNet, the domain gap between the open source training dataset and the target camera characteristics prevents it from functioning well on a new camera system. Synthetic datasets without depth-varying defocus modeling perform much worse, which we believe is because the network model fails to learn defocusing deblur capabilities from constant depth and invariant defocus maps. In contrast, with our proposed depth-varying dataset synthesis, we observe significant improvements in image restoration quality, demonstrating the effectiveness of our proposed approach.

Table 2. **Efficiency and performance comparison with full optical simulation**. Our proposed synthetic data generation approach significantly reduces rendering time and peak GPU memory consumption, enabling more efficient large-scale data generation while maintaining comparable or better performance.

	Full simulation [51]	Ours
Rendering Time (ms)	1306.9	**16.3**
Peak GPU Memory (GB)	10.6	**1.1**
PSNR (dB)	**30.09**	30.03
SSIM	0.8633	**0.8808**
LPIPS	0.1849	**0.1553**

We perform ablation studies on the same network (NAFNet) trained and evaluated using various auxiliary input channels (spatial position and camera ISO), as shown in Table 1. With camera RAW capture inputs and no auxiliary channels, the reconstructed image quality surpasses previous results but falls short compared to using either the ISO or spatial position ("Field") channels. This likely occurs because the model lacks information about the image patch location and noise level, leading it to generalize across all possible scenarios. When both ISO and spatial position channels are provided, the image quality improves significantly.

We further evaluated the efficiency and performance of our approach against full optical simulation (Table 2). Our comparison encompassed per-pixel PSF convolution for spatially-varying optical aberrations following [51], depth-varying defocus, and sensor noise. The full optical simulation proves substantially more resource-intensive, requiring ∼80 times longer rendering time per training batch and consuming ∼9.6 times more peak GPU memory. These measurements were taken using (1, 3, 512, 512) image batches on a single A100 GPU. Despite these computational differences, our efficient training approach achieves comparable image quality as measured by PSNR, SSIM, and LPIPS metrics (Table 2). These results confirm that **neglecting spatial variance within low-resolution training images for high-resolution camera sensors is a reasonable simplification while maintaining promising final performance**. Our efficient simulation approach significantly increases maximum batch size and reduces training time, accelerating algorithm development and offering broader benefits for computational photography algorithms.

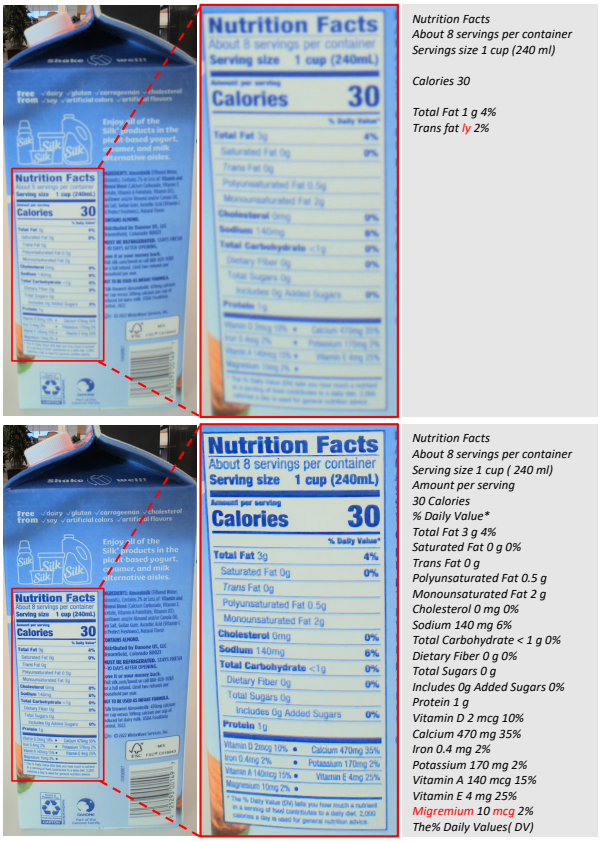

Figure 5. **Performance improvement in OCR for scene understanding.** Given an input image with texts captured at close distance (left), our depth of field extension successfully recovers details (right). Our result significantly improves OCR performance for both accuracy and detection rate. OCR results are generated with online program [46], with errors marked in red.

5. Applications

Based on the success of the proposed approach, we explored two downstream applications: short-distance OCR and small object 3D reconstruction. Due to the small physical scale and fine details of text characters and small objects, these applications present significant challenges in balancing image quality and details for existing fixed-focus cameras in smart glasses, primarily because of their shallow depth of field.

5.1. OCR

Smart glass utilizes OCR technology to enhance user experiences by providing real-time text recognition and interaction capabilities. The OCR technology for smart glasses focuses on text captured from the user's point of view while wearing the glasses. This approach allows for seamless interaction with text in various environments, such as translating menus, adding business card contacts, or creating shopping lists. Image blurriness due to fixed focus can be challenging to text recognition models, particularly when capturing at a close distance. Fig. 5 shows the OCR accuracy improvement using our approach.

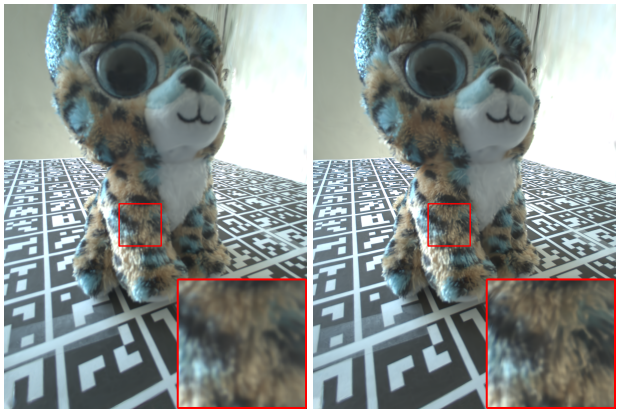

Figure 6. **Improved performance in 3D digital assets reconstruction.** Using Gaussian Splatting [18], we reconstruct a small object with either captured photos (left) or our deblurred results (right) as inputs. A novel view is rendered as above for evaluation.

5.2. 3D digital asset generation

Smart glasses have revolutionized the way we interact with our surroundings, and 3D digital assets are one of the most exciting applications of this technology. By leveraging the camera and sensor capabilities of smart glasses, users can create detailed 3D digital assets of their interest. Reconstructing 3D objects at close distances is, in particular, important, as it can capture detailed textures and features for high-quality reconstruction. Our approach significantly improves the accuracy and performance of 3D reconstruction (Fig. 6 and the supplementary video).

6. Conclusion

In this paper, we introduce an efficient depth-varying dataset synthesis pipeline for defocus deblur and spatially-varying optical aberration correction in computational photography. Our method effectively bridges the gap between synthetic and real-world data by incorporating depth-varying defocus into spatially varying simulations, thereby providing a scalable and robust solution without requiring extensive real-world data collection. Experimental results demonstrate the superiority of our approach over alternative methods, both in terms of image restoration quality and in simulation speed and memory efficiency. This advancement establishes a strong foundation for future research in the field and significantly shortens the development cycle for computational photography algorithms.

References

[1] Abdullah Abuolaim and Michael S. Brown. *Defocus Deblurring Using Dual-Pixel Data*, page 111–126. Springer International Publishing, 2020. 1, 2, 3, 6, 7

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 9205–9214. IEEE, 2021. 2

[3] Vladan Blahnik and Oliver Schindelbeck. Smartphone imaging technology and its applications. *Advanced Optical Technologies*, 10(3):145–232, 2021. 1

[4] David J. Brady and Nathan Hagen. Multiscale lens design. *Optics Express*, 17(13):10659, 2009. 1

[5] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T. Barron. Unprocessing images for learned raw denoising. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019. 2, 3, 4

[6] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 4

[7] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. *Simple Baselines for Image Restoration*, page 17–33. Springer Nature Switzerland, 2022. 4, 5, 6

[8] Shiqi Chen, Huajun Feng, Dexin Pan, Zhihai Xu, Qi Li, and Yueting Chen. Optical aberrations correction in postprocessing using imaging simulation. *ACM Transactions on Graphics*, 40(5):1–15, 2021. 2, 3, 6, 7

[9] Mauricio Delbracio, Ignacio Garcia-Dorado, Sungjoon Choi, Damien Kelly, and Peyman Milanfar. Polyblur: Removing mild blur by polynomial reblurring. *IEEE Transactions on Computational Imaging*, 7:837–848, 2021. 5, 6, 7

[10] Loïc Denis, Eric Thiébaut, Ferréol Soulez, Jean-Marie Becker, and Rahul Mourya. Fast approximations of shift-variant blur. *International Journal of Computer Vision*, 115 (3):253–278, 2015. 3

[11] Laurent D'Andres, Jordi Salvador, Axel Kochale, and Sabine Susstrunk. Non-parametric blur map regression for depth of field extension. *IEEE Transactions on Image Processing*, 25 (4):1660–1673, 2016. 1, 3

[12] Thomas Eboli, Jean-Michel Morel, and Gabriele Facciolo. Breaking down polyblur: Fast blind correction of small anisotropic blurs. *Image Processing On Line*, 12:435–456, 2022. 7

[13] Frans C. A. Groen, Ian T. Young, and Guido Ligthart. A comparison of different focus functions for use in autofocus algorithms. *Cytometry*, 6(2):81–91, 1985. 1

[14] Shir Gur and Lior Wolf. Single image depth estimation trained via depth from defocus cues. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 7675–7684. IEEE, 2019. 2

[15] G.E. Healey and R. Kondepudy. Radiometric ccd camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(3):267–276, 1994. 4

[16] Andrey Ignatov, Jagruti Patel, and Radu Timofte. Rendering natural camera bokeh effect with deep learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 1676–1686. IEEE, 2020. 7

[17] N. Kehtarnavaz and H.-J. Oh. Development and real-time implementation of a rule-based auto-focus algorithm. *Real-Time Imaging*, 9(3):197–203, 2003. 1

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 8

[19] Norman S. Kopeika. lt;titlegt;blur in imaging through the atmosphere: a system engineering approach to imaginglt;/titlegt;. In *Propagation and Imaging through the Atmosphere II*, page 320–331. SPIE, 1998. 3

[20] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019. 3

[21] S Kuthirummal, H Nagahara, Changyin Zhou, and S K Nayar. Flexible depth of field photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):58–71, 2011. 1

[22] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. 2, 3

[23] Shaojun Liu, Qingmin Liao, Jing-Hao Xue, and Fei Zhou. Defocus map estimation from a single image using improved likelihood feature and edge-based basis. *Pattern Recognition*, 107:107485, 2020. 1, 3

[24] Jun Luo, Yunfeng Nie, Wenqi Ren, Xiaochun Cao, and Ming-Hsuan Yang. Correcting optical aberration via depth-aware point spread functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5541–5555, 2024. 2

[25] Haoyu Ma, Shaojun Liu, Qingmin Liao, Juncheng Zhang, and Jing-Hao Xue. Defocus image deblurring network with defocus map estimation as auxiliary task. *IEEE Transactions on Image Processing*, 31:216–226, 2022. 1, 2, 3

[26] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 257–265. IEEE, 2017. 3

[27] Alex Ning. Auto-focus (af) lens and process, 2004. US Patent App. 10/778,785. 1

[28] Yuhui Quan, Zicong Wu, and Hui Ji. Neumann network with recursive kernels for single image defocus deblurring. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 5754–5763. IEEE, 2023. 1, 3

[29] Yuhui Quan, Xin Yao, and Hui Ji. Single image defocus deblurring via implicit neural inverse kernels. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 12566–12576. IEEE, 2023.

[30] Yuhui Quan, Zicong Wu, Ruotao Xu, and Hui Ji. Deep single image defocus deblurring via gaussian kernel mixture learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11361–11377, 2024. 1, 3

[31] Lingyan Ruan, Bin Chen, Jizhou Li, and Miu-Ling Lam. Aifnet: All-in-focus image restoration network using a light field-based dataset. *IEEE Transactions on Computational Imaging*, 7:675–688, 2021. 2, 3

[32] Lingyan Ruan, Bin Chen, Jizhou Li, and Miuling Lam. Learning to deblur using light field generated and real defocus images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 16283–16292. IEEE, 2022. 1, 3

[33] Lingyan Ruan, Mojtaba Bemana, Hans-peter Seidel, Karol Myszkowski, and Bin Chen. Revisiting image deblurring with an efficient convnet. *arXiv preprint arXiv:2302.02234*, 2023. 1, 3, 5, 6, 7

[34] Lingyan Ruan, Martin Bálint, Mojtaba Bemana, Krzysztof Wolski, Hans-Peter Seidel, Karol Myszkowski, and Bin Chen. Self-supervised video defocus deblurring with atlas learning. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24*, page 1–11. ACM, 2024. 2, 3

[35] Conor J. Sheil and Alexander V. Goncharov. Large aperture camera lens with minimalistic refocus for smartphone portraiture photography. *Optics Communications*, 440:207–213, 2019. 1

[36] Zheng Shi, Yuval Bahat, Seung-Hwan Baek, Qiang Fu, Hadi Amata, Xiao Li, Praneeth Chakravarthula, Wolfgang Heidrich, and Felix Heide. Seeing through obstructions with diffractive cloaking. *ACM Transactions on Graphics*, 41(4): 1–15, 2022. 3

[37] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics*, 37(4): 1–13, 2018. 1

[38] Binbin Song, Xiangyu Chen, Shuning Xu, and Jiantao Zhou. Under-display camera image restoration with scattering effect. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 12546–12555. IEEE, 2023. 3

[39] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 3

[40] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 769–777. IEEE, 2015. 3

[41] Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1383–1393. IEEE, 2020. 2

[42] P. Trouve, F. Champagnat, G. Le Besnerais, and J. Idier. Single image local blur identification. In *2011 18th IEEE International Conference on Image Processing*, page 613–616. IEEE, 2011. 1, 3

[43] Ethan Tseng, Ali Mosleh, Fahim Mannan, Karl St-Arnaud, Avinash Sharma, Yifan Peng, Alexander Braun, Derek Nowrouzezahrai, Jean-François Lalonde, and Felix Heide. Differentiable compound optics and processing pipeline optimization for end-to-end camera design. *ACM Transactions on Graphics*, 40(2):1–19, 2021. 2

[44] Chao Wang, Ana Serrano, Xingang Pan, Krzysztof Wolski, Bin Chen, Karol Myszkowski, Hans-Peter Seidel, Christian Theobalt, and Thomas Leimkühler. An implicit neural representation for the image stack: Depth, all in focus, and high dynamic range. *ACM Transactions on Graphics*, 42(6):1–11, 2023. 2

[45] Ning-Hsu Wang, Ren Wang, Yu-Lun Liu, Yu-Hao Huang, Yu-Lin Chang, Chia-Ping Chen, and Kevin Jou. Bridging unsupervised and supervised depth from focus via all-in-focus supervision. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 12601–12611. IEEE, 2021. 2

[46] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024. 8

[47] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2755–2764. IEEE, 2020. 2

[48] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8520–8537, 2021. 2

[49] Shumian Xin, Neal Wadhwa, Tianfan Xue, Jonathan T. Barron, Pratul P. Srinivasan, Jiawen Chen, Ioannis Gkioulekas, and Rahul Garg. Defocus map estimation and deblurring from a single dual-pixel image. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021. 1, 3

[50] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 2, 4, 7

[51] Xinge Yang, Qiang Fu, Mohamed Elhoseiny, and Wolfgang Heidrich. Aberration-aware depth-from-focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–11, 2024. 2, 3, 7

[52] Xinge Yang, Qiang Fu, and Wolfgang Heidrich. Curriculum learning for ab initio deep learned refractive optics. *Nature Communications*, 15(1), 2024. 1

[53] Xinge Yang, Matheus Souza, Kunyi Wang, Praneeth Chakravarthula, Qiang Fu, and Wolfgang Heidrich. End-to-end hybrid refractive-diffractive lens design with differentiable ray-wave model. In *SIGGRAPH Asia 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 1

[54] Kyrollos Yanny, Nick Antipa, William Liberti, Sam Dehaeck, Kristina Monakhova, Fanglin Linda Liu, Konlin Shen, Ren Ng, and Laura Waller. Miniscope3d: optimized single-shot miniature 3d fluorescence microscopy. *Light: Science & Applications*, 9(1), 2020. 3, 7

[55] Kyrollos Yanny, Kristina Monakhova, Richard W. Shuai, and Laura Waller. Deep learning for fast spatially varying deconvolution. *Optica*, 9(1):96, 2022. 1

[56] Zemax, Inc. *Zemax OpticStudio® - Design and Analysis Software for Optical Systems*. Zemax, Inc., Fremont, CA, USA, version 23.0 edition, 2023. 4

[57] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 5971–5979. IEEE, 2019. 3

[58] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018. 4

[59] Shaojie Zhuo and Terence Sim. Defocus map estimation from a single image. *Pattern Recognition*, 44(9):1852–1858, 2011. 1, 3