

# Aberration-Aware Depth-from-Focus

Xinge Yang, *Student Member, IEEE*, and Qiang Fu, and Mohamed Elhoseiny, *Member, IEEE*,  
and Wolfgang Heidrich, *Fellow, IEEE*

**Abstract**—Computer vision methods for depth estimation usually use simple camera models with idealized optics. For modern machine learning approaches, this creates an issue when attempting to train deep networks with simulated data, especially for focus-sensitive tasks like Depth-from-Focus. In this work, we investigate the domain gap caused by off-axis aberrations that will affect the decision of the best-focused frame in a focal stack. We then explore bridging this domain gap through aberration-aware training (AAT). Our approach involves a lightweight network that models lens aberrations at different positions and focus distances, which is then integrated into the conventional network training pipeline. We evaluate the generality of network models on both synthetic and real-world data. The experimental results demonstrate that the proposed AAT scheme can improve depth estimation accuracy without fine-tuning the model for different datasets. The code will be available in [github.com/vccimaging/Aberration-Aware-Depth-from-Focus](https://github.com/vccimaging/Aberration-Aware-Depth-from-Focus).

**Index Terms**—Depth from Focus, Optical Aberration, Ray Tracing, Point Spread Function

## 1 INTRODUCTION

MODERN deep-learning techniques have made significant progress in understanding 3D scenes from 2D RGB images, including depth estimation [1], [2], [3], [4], object detection [5], [6], multiple views [7], [8], and camera tracking and mapping [9], [10], [11]. However, these methods assume that 2D training images are aberration-free, relying on an idealized pinhole camera model that fails to account for out-of-focus effects and optical aberrations present in real camera lenses. This inaccuracy leads to a domain gap between experimental and real-world images, particularly given that modern lenses possess large apertures resulting in shallow depth-of-field (DoF), as well as a large field-of-view resulting in off-axis aberrations. The domain gap undermines the generalizability of trained deep learning models [12], necessitating engineers to fine-tune models with real data for each end device.

Recent studies in depth-from-focus (DfF) [13], [14], [15], [16], [17], [18], [19], [20], [21], [22] have recognized the impact of out-of-focus effects and explored using defocus cues for depth estimation and all-in-focus image estimation. DfF methods estimate the probability that a pixel is the sharpest in a focal stack and then interpolate the input focus distances based on the probability, assuming that the sharpest frame is the best-focused frame. The probability can also be used to interpolate focused images for all-in-focus image synthesis [20], [23], [24], [25]. Although some of these methods [18], [19], [20], [21], [22] have claimed to bridge the domain gap between experimental and real-world images by considering that the sharpest pixel in the focal stack is independent of semantic information, their experiments have relied on a thin lens model that neglects off-axis optical aberrations commonly present in real lenses.

Such aberrations, including field curvature, can cause the focus distance to vary across the image plane or result in asymmetric blurs that make it difficult to measure the most in-focus accurately frame. Moreover, studies [26], [27], [28], [29] have highlighted the impact of optical aberrations in depth estimation. Therefore, another domain gap arises due to the presence of optical aberrations.

To overcome the domain gap resulting from optical aberrations, we introduce aberration-aware training (AAT) that enables the network to learn these optical aberrations during the training. Our AAT method consists of a lightweight point spread function (PSF) network and a re-rendering process to simulate aberrated training images. First, we compute the spatially-varying PSF of an optical lens using ray tracing [30], [31], and train a multilayer perceptron (MLP) to represent it. Once trained, the network can efficiently estimate the PSF for different object positions and focus distances. Next, we render training images to apply off-axis aberrations. Given the depth map for each image, we select a group of focus distances and determine the PSF for each pixel. Then we perform local convolution with all-in-focus images to create a focal stack containing both depth-of-field and off-axis aberrations. During the training of the depth estimation network, the network learns to determine the best-focused frame for each pixel under optical aberrations, thereby enhancing the generalizability. Furthermore, the depth estimation accuracy can be improved as long as the training and testing images are captured/simulated using the same lens.

To assess the effectiveness of our proposed AAT scheme, we conduct experiments to test the generalizability of the pre-trained DfF model on various datasets, including both simulated and real-world datasets. First, we simulate focal stacks using a real lens and a thin lens and train the corresponding DfF models. Then we test two models on focal stacks simulated/captured by the same lens without any fine-tuning. The experimental results demonstrate that the AAT model generalizes better than the non-AAT model in terms of depth estimation accuracy. Specifically, the AAT

- 
- W. Heidrich is with the Department of Computer Science and Electrical and Computer Engineering, King Abdullah University of Science and Technology, Saudi Arabia, 23955.  
E-mail: [wolfgang.heidrich@kaust.edu.sa](mailto:wolfgang.heidrich@kaust.edu.sa)
  - X. Yang, Q. Fu, M. Elhoseiny is with King Abdullah University of Science and Technology.

model successfully resolves small depth differences between adjacent objects, whereas the non-AAT model fails to do so. Moreover, the AAT model proves more successful at suppressing the transfer of color texture detail into the depth geometry. We demonstrate results on two recent state-of-the-art architectures [20], [22], and the results show that our training approach is agnostic to the specific DfF network architecture.

In summary, our contributions are two-fold:

- We propose a lightweight network that can represent the PSF of a real lens at different focus distances and object positions. This PSF network can then simulate aberrated and realistic images for aberration-aware training.
- We reveal the domain gap between real-world and simulated data arising from off-axis aberrations, which impacts the determination of the best-focused frame in a focus stack and reduces the accuracy of depth estimation. We propose an AAT scheme to address the problem, the DfF models trained with the AAT scheme can better generalize to test focal stacks captured or simulated by the same lens.

## 2 RELATED WORKS

### 2.1 Depth from Focus

Depth estimation is a fundamental task in computer vision, and can be tackled using different cues, such as semantic information [2], [4], stereo [32], [33], [34], and defocus [19], [20], [22]. Among these cues, the defocus cue is considered domain invariant, as the defocus pattern is only related to the optical properties of the imaging lens. The depth-from-focus problem learns the depth map from the focus stack, using the idea that each pixel must have one frame that is best in focus.

Conventional approaches solve the depth-from-focus (DfF) problem as an optimization task [13], [35]. Learning-based approaches then use 2D convolutional neural networks (CNNs) to improve the accuracy of feature extraction and depth estimation [15], [16], [18], [19]. To allow for communication between different 2D CNNs, a shared pooling layer is used, and an intermediate defocus map [19] is learned during training to aid the depth estimation. Later, 3D CNNs [20], [21] started to replace 2D architectures, as they perform better in extracting shared information from an image stack. While the idea of the intermediate defocus map is kept, it is replaced by a 3D cost volume [20], [21], [22].

The training and testing datasets are obtained through various methods, including real-world captures and image simulation. For real-world captures, focal swap [13], [15] and light-field camera rendering [16] are two common methods used to obtain focused images, and additional Lidar or ToF cameras are used to get the depth map of the scene. Capturing real-world focal stacks with ground truth depth is quite expensive. Thus, image simulation methods, including PSF convolution [18], stereo image rendering [20], and software rendering [19], are also used to generate simulated focal stacks for training data augmentation. However, existing simulated focal stacks all rely on the idealized thin

lens model and thus lack optical aberrations. Moreover, the real-world captured focal stacks lack focus distance information and lens data. In this work, we simulate aberrated focal stacks with the real lens model and capture real-world focal stacks with the necessary information. We use simulated focal stacks for training, and after training, we test the models on both simulated and real-world focal stacks without fine-tuning.

### 2.2 PSF Estimation

Multiple approaches have been introduced to calculate the PSF of an optical lens/system. The commonly used idealized optical model [18], [36] calculates the PSF under paraxial principles, resulting in a truncated Gaussian function called the ‘‘Circle-of-Confusion’’ (CoC). The diameter of the CoC is computed by

$$CoC = \frac{f}{N} \frac{|z - f_d|}{z} \frac{f}{f_d - f}, \quad (1)$$

where  $f_d$  is the focal distance,  $z$  is the distance between an object to the lens,  $f$  is the focal length, and  $N$  is the F-number of the camera lens. However, due to the paraxial approximation, this approach can not represent spatially varying PSF and off-axis aberrations like field curvature, coma, and astigmatism in actual optical lenses.

Ray tracing through optical lenses [30], [37], [38] can provide a more accurate spatially-varying PSF and has been widely utilized in commercial software such as ZEMAX and CodeV. This approach involves shooting rays from each object point source and calculating the distribution of each ray on the sensor plane to obtain the PSF. However, tracing rays through a sequence of optical elements is computationally expensive as it requires finding the intersection points of each ray using iterative algorithms. Recently, researchers have started exploring the use of neural networks to represent the PSF [39], which has shown promising results in both accuracy and speed. However, nobody has explored representing a lens with both varying focus distances and spatial positions. In this work, we extend the network estimation approach to represent the PSF for varying focus distances and spatial positions.

Another widely used method, especially when the design space of optical lenses is unavailable, is PSF calibration [40], [41]. PSF calibration approaches measure the lens response to the given point light source and then estimate the PSF at unmeasured positions [42], [43], [44], [45], [46]. Among all the PSF estimation works, the recently proposed low-rank model [43], [47] is an efficient and accurate way to estimate unmeasured PSF with only a few measurements. However, PSF calibration also comes with its own challenges, such as noise and quantization (especially for long tails), since PSF estimation requires either the direct measurement of (weak) point sources or the solution of a deconvolution problem.

## 3 METHODOLOGY

As shown in Fig. 1(c), a classical DfF network takes the focus stack and corresponding focus distances as input and outputs a depth map. However, in the existing training

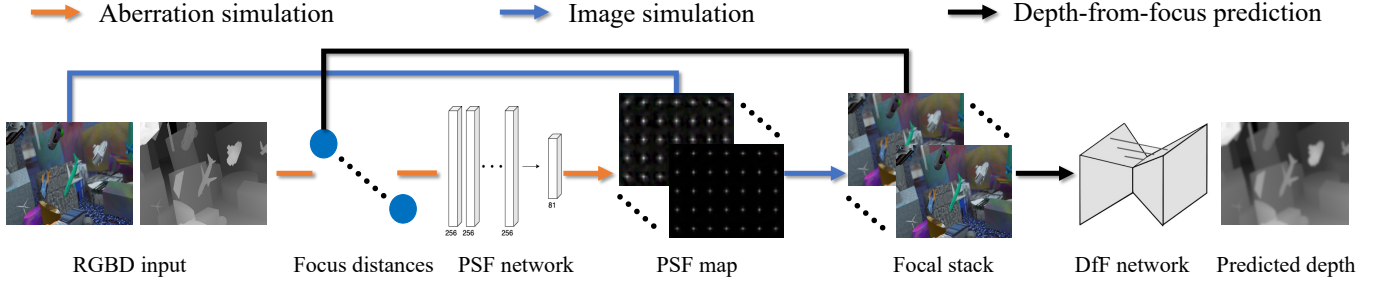


Fig. 1. **Aberration-Aware Training pipeline.** (a) all-in-focus RGB images and corresponding depth maps are given as the input. (b) different focus distances are selected to simulate the focal swap process. The PSF network estimates PSF for different object positions and focus distances (orange path). Then the PSF is convolved with the all-in-focus image to get the focal stack (blue path). (c) the DfF network takes the focal stack and focus distances to estimate the depth map (black path).

pipeline, the focal stacks are pre-given and the network is blind to the optical system during training. Therefore, the pre-trained model is hard to generalize to real-world data, and computer vision engineers are required to fine-tune the model for different end devices. With the idea of embedding optical characteristics into the network training, the AAT scheme integrates the data simulation module into the network training pipeline, as illustrated in Fig. 1(b). All-in-focus images and their corresponding depth maps are used as input during the training process (Fig. 1(a)). A series of focus distances are selected, and the PSF estimation network is applied to estimate the PSF for each pixel, which is then used to render the focal stack. The entire pipeline can be executed end-to-end, and the AAT scheme allows the network to learn the estimation of depth maps in the presence of complex optical aberrations.

### 3.1 Depth-from-Focus Network

We do not modify the architecture of the DfF network but instead use existing architectures such as AiFNet [20] and DFVNet [22]. Here we use AiFNet for illustration. Given a batch of focal stacks with shapes  $(B, S, C, H, W)$ , where  $B$  is the batch size,  $S$  is the stack size,  $C$  is the number of channels, and  $H, W$  are the image height and width, a 3D convolutional encoder is used to extract multilayer image features and create an intermediate attention map. This attention map functions as an in-focus probability map, which is then used to interpolate focus distances for the final depth estimation. The loss function for the DfF task typically consists of a depth estimation loss and a regularization term:

$$L = L_{depth} + \omega L_{reg}, \quad (2)$$

where  $L_{depth}$  denotes the supervision loss function designed directly on the estimated depth map, and  $L_{reg}$  denotes the regularization term (e.g., one that encourages the depth map to be locally smooth using an edge-aware weighting as in [32]).  $\omega$  is the weight coefficient balancing the two parts, and we adopt the hyperparameter settings from the original paper.

### 3.2 Aberration Simulator

Our contribution lies in the accurate simulation of aberrations for the training data to improve on the classical

thin lens model. The PSF characterizes the optical lens response to a point source of light. We can convolve the per-pixel PSF with the object image to simulate the image captured by a camera. The Gaussian PSF (Eq. 1) assumes shift-invariance across the same depth plane. However, this idealized optical model does not accurately account for off-axis optical aberrations in a real camera lens. In Fig. 2(b), we show the field curvature aberration, which is common in real lenses. Due to the field curvature, the pixel at the edge (P1) is blurry, but the pixel on the axis (P2) is sharp, although this image is well-focused to depth  $d_1$ .

Ray tracing through optical lenses is a well-established technique for obtaining a more accurate point spread function (PSF) [30], [37]. This involves tracing a group of rays from a point source through the lens group to the sensor plane, resulting in a spot diagram. We can then convert the spot diagram into sensor pixels and obtain the PSF [36], [48], [49] by:

$$\text{PSF}(\mathbf{o}_p) = \sum_{k=1}^{spp} u_k \cdot \sigma(|(\mathbf{o}_p - \mathbf{o}_k) \cdot \hat{\mathbf{e}}_x|/L) \cdot \sigma(|(\mathbf{o}_p - \mathbf{o}_k) \cdot \hat{\mathbf{e}}_y|/L), \quad (3)$$

where  $\mathbf{o}_p$  denotes the coordinate of the pixel, and  $\mathbf{o}_k$  represents the intersection point of the  $k$ th ray with the sensor plane. The variable  $spp$  stands for “samples per pixel”, corresponding to the number of rays emitted from each point source, which is set to 2048 in our experiments. We assume that the energy of each ray, denoted by  $u_k$ , is equal to 1.  $\hat{\mathbf{e}}_x$  and  $\hat{\mathbf{e}}_y$  are unit vectors in the sensor plane, and  $L$  denotes the physical width of a sensor pixel. The  $\sigma$  function is defined as:

$$\sigma(x) = \begin{cases} 1-x & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}, \quad (4)$$

which assesses a ray’s impact on its surrounding pixels, with a greater impact attributed to rays in closer proximity. The total impact of a ray on the four surrounding pixels sums to one. By leveraging sub-pixel information, the  $\sigma$  function can more accurately represent the actual light distribution using a limited number of samples. Since DfF is based on imaging with large apertures, it is possible to neglect diffraction effects which would dominate in optical systems with a small aperture. Moreover, we assume that chromatic aberrations

are well corrected compared to the other aberrations and out-of-focus effects, which is the case for most commercial-grade lenses. This allows us to simulate the optical system at a single wavelength of 589nm.

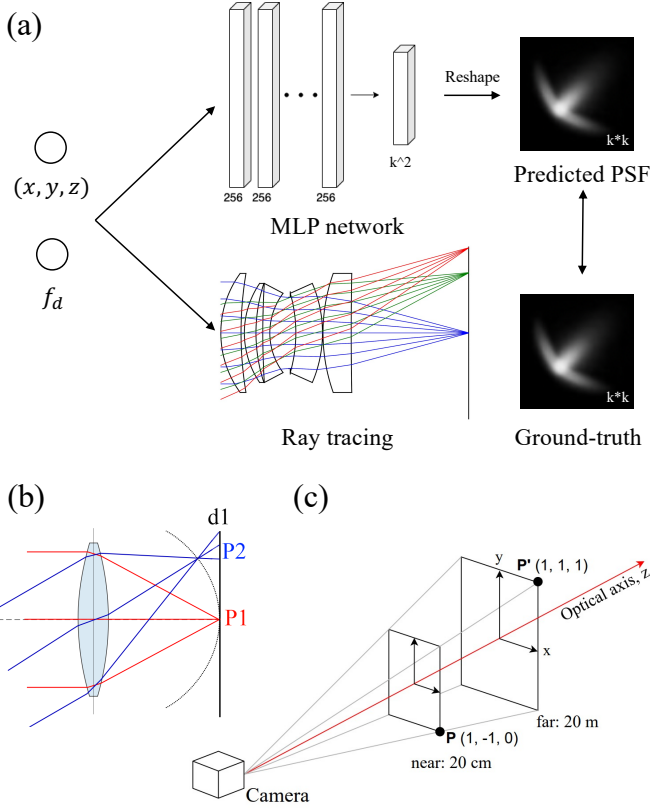


Fig. 2. (a) a MLP network is trained to represent the PSF for different positions and focus distances. We use ray tracing to calculate accurate PSF as the ground truth. The network takes as input the object positions  $(x, y, z)$  and focus distance  $f_d$ , and produces a 2D matrix as output. (b) off-axis optical aberrations blur the pixels at the edge (P2) when the pixel on the axis (P1) is focused. (c) the valid imaging area is a frustum, and we normalize the  $x, y$  coordinates to  $[-1, 1]$ ,  $z$  and  $f_d$  to  $[0, 1]$ .

However, ray tracing is computationally expensive, particularly as objects may appear at different positions and the focus distance may vary. Inspired by [39], we train a network  $H$  with parameters  $\theta$  to represent the PSF as

$$\theta = \arg \min_{\theta} \|\text{PSF}(x, y, z; f_d) - H_{\theta}(x, y, z; f_d)\|_2^2, \quad (5)$$

where  $(x, y, z)$  represent the normalized coordinates of an object point, and  $f_d$  denotes the focus distance. As depicted in Fig. 2(a), we train the network  $H$  to fit an imaging lens by minimizing the difference between the estimated PSF and the ray-traced PSF. In object space, the valid imaging region of a lens is a frustum (as shown in Fig. 2(c)) that is defined by the field of view (FoV), sensor size, minimum depth, and maximum depth. We then normalize the focus distance  $f_d$  and depth  $z$  to  $[0, 1]$ , and normalize  $(x, y)$  to  $[-1, 1]$ .

Once the optical structure and aperture size of a lens is fixed, the PSF is solely determined by the object position and focus distance. To map the four-parameter input into a 2D PSF kernel, we adopt a simple MLP network. The PSF estimation network consists of one input layer, five

hidden layers with 256 neurons each, and an output layer with  $k^2$  neurons, where  $k$  is the width of PSF. We use ReLU activation functions after each input and hidden layer and a Sigmoid activation function after the output layer. The  $k^2$ -channel output is then reshaped into a  $k \times k$  2D tensor. After training, we fix the parameters of the PSF network and use it to estimate the PSF for various object positions and focus distances. Then, we can use the per-pixel PSF to render aberrated images for the subsequent depth estimation task.

## 4 PSF ESTIMATION RESULTS

### 4.1 Implementation Details

We train the PSF network for 400,000 iterations to overfit the lens. In each iteration, we randomly focus the lens to a distance of  $f_d$ , and uniformly select 256 points in object space for training. The ground-truth PSFs are computed by tracing 1024 rays from each object point, and  $(x, y, z, f_d)$  coordinates are provided to the network as input. We use a wavelength of 589 nm and set the PSF size to  $11 \times 11$  sensor pixels. AdamW optimizer [50] and CosineAnnealing learning rate scheduler [51] with default parameters are used to train the PSF network. The initial learning rate is set to  $1 \times 10^{-3}$ . The lens has a minimum imaging depth of 20cm and a maximum depth of 20m, beyond which no relevant depth information can be recovered due to the small baseline of the DfF approach. We use the same depth range for the focus distance. We employ the DeepLens framework [30], [31] for ray tracing computation. The training process is run on a single A100 80G GPU and completed in approximately 6 hours. The network consists of 0.28 million parameters, with a storage size of approximately 1.5 MB.

### 4.2 Evaluation

We use the Canon EOS RF50mm F/1.8 lens [52] for evaluation. The image sensor is simulated with physical dimensions of  $24\text{mm} \times 32\text{mm}$  and a resolution of  $640 \times 960$  to match the resolution of the RGBD data used for training the DfF network. While this image resolution is much lower than that of modern cameras, our experiments show that off-axis aberrations are still visible in the out-of-focus regime.

After training the PSF network, we focus the lens to a distance of 1.5m and evaluate the PSF at depths of 1.2m, 1.5m, and 2m, and three different view angles, as shown in Fig. 3. At the focused depth (1.5m), the PSF is small, but when the lens is out of focus (1.2m and 2m), the PSF becomes larger. At  $0^\circ$ , both the network and the Gaussian model accurately predict the ground truth PSF. However, as the viewing angle increases, the Gaussian PSF becomes increasingly inaccurate due to off-axis aberrations, and a clear difference can be observed between the Gaussian PSF and the ground truth at  $23.5^\circ$ . In contrast, the PSF network estimated accurate results at all depths and view angles. Furthermore, at  $23.5^\circ$ , the ground truth PSF at a depth of 2m is the smallest among the three depths, caused by the field curvature aberration (also illustrated in Fig. 2(b)). However, the Gaussian model can not capture this phenomenon. Comparing the network estimation and the ray tracing results, we find that the network-estimated PSF exhibits less noise

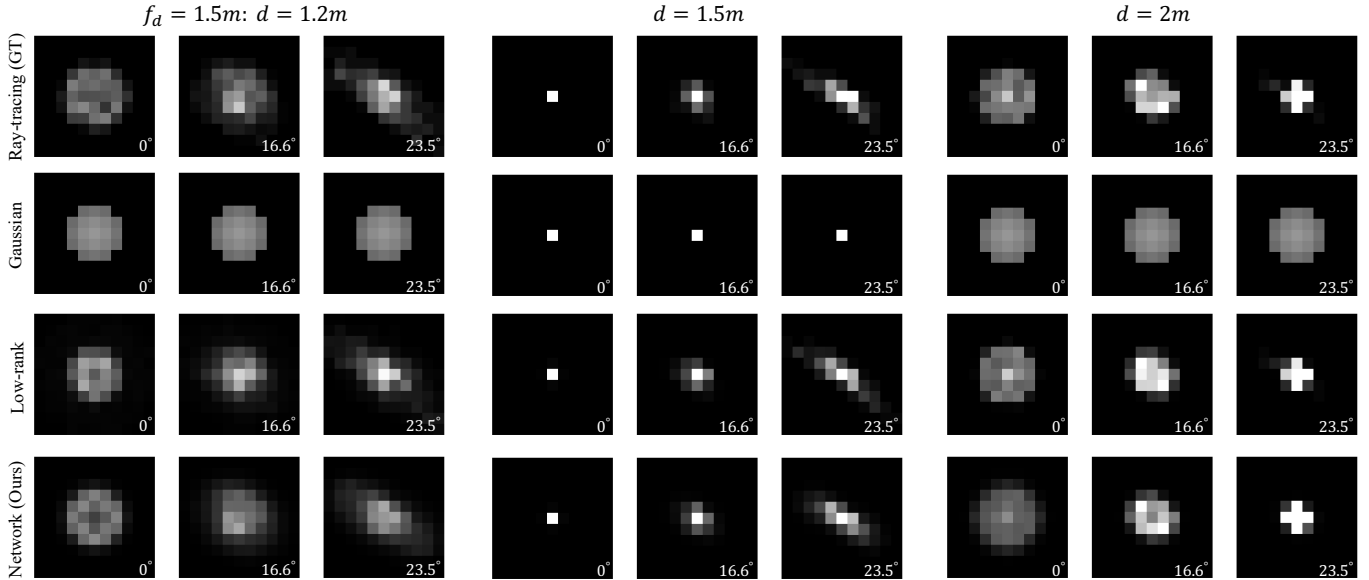


Fig. 3. **PSF estimation results.** We focus the lens to a distance of 1.5m and evaluate the PSF of the four methods at three depths and three view angles. Both our PSF network and low-rank model produce a PSF that is close to the ground truth (ray tracing). In contrast, the Gaussian PSF exhibits significant differences, particularly at large view angles. The PSF of a real lens varies with different view angles due to the presence of off-axis optical aberrations, while the Gaussian PSF model neglects this aberration.

and holds better symmetry, which is more in line with the physical situation.

TABLE 1  
Quantitative comparison of difference PSF estimation methods.

Method	$\ell_1$ error	$\ell_2$ error	time (min)
Ours	$4.68e^{-3}$	$8.23e^{-5}$	2.5
Tseng, et. al. [39]	$7.35e^{-3}$	$2.94e^{-4}$	1.5
Kyrollos, et. al. [47]	$7.33e^{-3}$	$1.43e^{-3}$	87

For comparison purposes, we also tested the low-rank PSF estimation model described in [43], [47]. In this model, we use ray tracing to calculate the PSF of the surrounding 8 positions and employ trilinear interpolation to obtain the center PSF. We divide the object space into 20 depths, with 64 grids in each depth plane. The PSFs of these positions are calculated and used for querying. As depicted in Fig. 3, the low-rank PSF model can estimate PSFs similar to the ground truth. Since the PSF is slowly varying in the object space, using enough sampled PSFs for querying can yield promising results.

For a more detailed evaluation, we selected 20 focus distances, 40 depths, and 8 (height)  $\times$  10 (width) positions per depth as the testing dataset to quantify PSF estimation accuracy. Additionally, we compare our proposed network model with the model proposed by Tseng et al. [39]. The evaluation metrics used are the  $\ell_1$  and  $\ell_2$  errors. As shown in Table 1, our MLP network achieves the most accurate PSF estimation among the three methods. Furthermore, our PSF network can estimate the PSF given discrete spatial points and focus distances without requiring the querying of external data or processing the network output. This

advantage makes our network model suitable for aberrated and focused image simulation.

## 5 ABERRATION-AWARE TRAINING RESULTS

After evaluating the fitting accuracy of the PSF network, we now turn to evaluate the impact of using this network for aberration-aware training of the DfF network. To this end, we conduct experiments on both simulated and real-world data. In the first experiment, we train and test the model on simulated focal stacks. In the second experiment, we train the model on simulated focal stacks and test it on real-world focal stacks. Following the AAT scheme, we simulate the training focal stacks with the same lens that is used to simulate/capture the test data.

For comparison, we simulate training focal stacks using the Gaussian PSF calculated by the thin lens model (“non-AAT”). This serves as the baseline, as used in the existing works. To ensure a fair comparison, the thin lens is set to have the same focal length, F-number, and sensor size as the objective real lens, with the only variable being the off-axis optical aberrations. In the experiment, we select two DfF networks: AiFNet [20] and DFVNet [22], and train them with the same settings as described in the original papers. For each network, we train two models, one with the AAT scheme and one without it, and evaluate their performance on the testing focal stacks without fine-tuning.

### 5.1 Implementation Details

In the first experiment, we use a 50mm F/2.8 lens<sup>1</sup> (see Fig. 2(a)), which has significant off-axis aberrations. The image sensor has a physical size of 24mm  $\times$  32mm and a

1. The lens data comes from [lensnet.com](https://lensnet.com) [53], the actual F-number we calculate is F/1.86

TABLE 2

**Quantitative depth estimation results on simulated focal stacks.** Two DfF networks, AiFNet [20] and DFVNet [22], are chosen for depth estimation. The testing focal stacks are rendered by real lens PSF, while we use Gaussian PSF (“Baseline”) and real lens PSF (“Ours”) to render training focal stacks. Models trained with our AAT scheme deliver better depth estimation accuracy.

Method	MAE ↓	MSE ↓	RMSE ↓	Abs. rel. ↓	Sqr. rel. ↓	$\delta = 1.25 \uparrow$	$\delta = 1.25^2 \uparrow$	$\delta = 1.25^3 \uparrow$
Baseline (AiFNet)	0.4706	0.4805	0.6229	0.2759	0.4906	0.8098	0.9587	0.9713
Ours (AiFNet)	<b>0.2095</b>	<b>0.1536</b>	<b>0.3475</b>	<b>0.1613</b>	<b>0.2982</b>	<b>0.9683</b>	<b>0.9852</b>	<b>0.9895</b>
Baseline (DFVNet)	0.5900	0.8452	0.8025	0.2885	0.5065	0.7390	0.8903	0.9277
Ours (DFVNet)	<b>0.1977</b>	<b>0.1582</b>	<b>0.3446</b>	<b>0.1563</b>	<b>0.2992</b>	<b>0.9669</b>	<b>0.9830</b>	<b>0.9876</b>

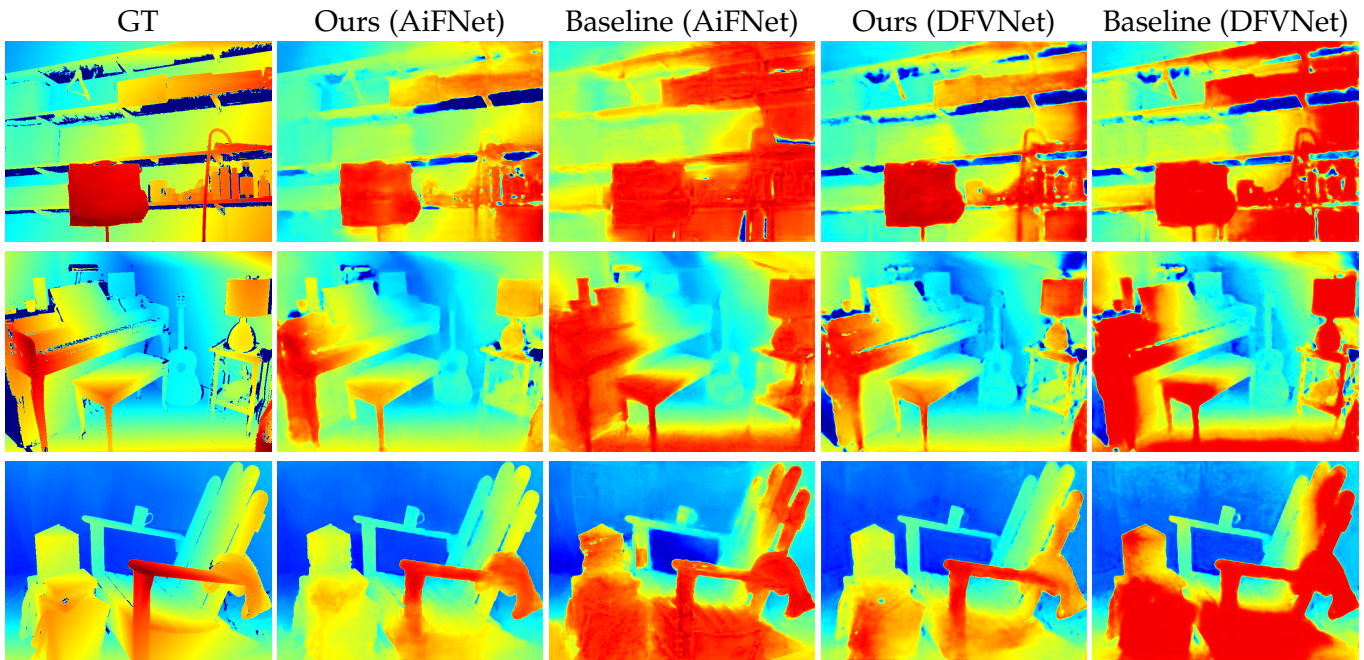


Fig. 4. **Qualitative results on simulated focal stacks.** With the AAT scheme, both network models predict more accurate and finer depth maps, while the non-AAT models fail to distinguish between adjacent objects and also mispredict the depth of some edge objects.

resolution of  $480 \times 640$ , and we train a new PSF network to model this lens. The FlyingThings3D dataset [20], [54] is used as the training dataset, which contains 800 pairs of synthetic RGBD images. The Middlebury2014 dataset [55] is used as the testing dataset, which contains 23 real-world RGBD images. There is a significant domain gap between the synthetic and real-world images, making it suitable for evaluating the generalizability of the DfF models. We simulate focal stacks with the RGBD images for training and testing, using a stack size of 10. The focus distances are chosen linearly from the minimum (20 cm) and maximum (20 m) depth range of each image, with a random perturbation.

We utilize the AdamW optimizer [50] and the CosineAnnealing learning rate scheduler [51] with their default parameters. The training batch size is set to 16, and the initial learning rate is set to  $1e^{-4}$ . Each DfF model is trained for 400 epochs on a single A100 80G GPU which is enough for convergence. After training, we evaluate each model on the testing dataset without fine-tuning.

In the second experiment, we use the Canon EOS R camera and the RF 50mm F/1.8 lens, discussed in Sec-

tion 4. The DfF training procedure remains the same as the previous experiment. We use the Matterport3D dataset preprocessed by Zhang et al. [56], which comprises 117,516 pairs of indoor RGBD images as additional training data and train each model for an additional 20 epochs, followed by an evaluation of real-world captured focal stacks. The real-world captured focal stacks consist of 24 outdoor scenes and 1 indoor scene.

## 5.2 Evaluation on Simulated Focal Stacks

### 5.2.1 Depth estimation

In Fig. 4, we present the qualitative results of the estimated depth maps. Despite the significant domain gap between the original synthetic training data and the real-world test data, both DfF models can estimate promising depth maps with the AAT scheme. This is because we simulate both training and testing focal stacks with the same lens model, and the optical aberrations are independent of semantic information in the images. We observe that small depth differences between adjacent objects are preserved in the estimated depth maps. Furthermore, the absolute depth values of the objects are precisely estimated without significant errors.



Fig. 5. **All-in-focus image synthesis results on simulated focal stacks.** The all-in-focus images synthesized without the AAT scheme exhibit significant estimation errors at the edge of the objects, and the continuous image regions show inconsistent sharpness and blurriness. In contrast, the AAT scheme results in clear and continuous estimated edges.

However, the depth maps estimated by the non-AAT models fail to distinguish between adjacent objects, and also there are errors at the corner of the depth maps. This inaccuracy is caused by the off-axis aberrations, which is the only variable in the experiment. As illustrated in the previous section, the off-axis aberrations, such as field curvature, affects the decision of the best in-focus frame in the focus stack and degrade the performance of pre-trained models. These off-axis aberrations have a dominant effect at smaller out-of-focus depths, leading to the non-AAT model’s inability to discriminate between adjacent objects.

In Table. 2, we evaluate quantitative results with the following metrics: mean-absolute error (MAE), mean-squared error (MSE), root-mean-squared error (RMSE), relative-absolute error (Abs.rel.), relative-squared error (Sqr.rel.), accuracy with  $\delta = 1.25$ . For MAE, MSE, RMSE, Abs.rel. and Sqr.rel., lower values indicate better results, while for the three accuracy metrics, higher values represent better results. As shown in the table, both AiFNet and DFVNet models greatly improve with the AAT compared to the non-AAT results.

### 5.2.2 All-in-Focus Image Synthesis

We also evaluated the results of all-in-focus image synthesis, which utilized the same predicted probability maps but instead interpolated RGB values of focused images for the final output. The synthesized all-in-focus images are presented in Fig. 5, with zoomed image patches in the top-right corner. As observed, the edges of the objects in the synthesized images without the AAT scheme exhibit a fuzzy effect. Furthermore, there are noticeable estimation errors that the continuous image regions exhibit inconsistent sharpness and blurriness. PSNR and SSIM scores of estimated all-in-focus images are presented in Table 3, the AAT scheme leads to higher all-in-focus image synthesis quality compared to the baseline.

TABLE 3  
Quantitative evaluation of AAT scheme on all-in-focus image synthesis.

AiFNet	PSNR	SSIM
Baseline	33.55	0.970
Ours	<b>34.65</b>	<b>0.976</b>

### 5.3 Evaluation on Real-World Focal Stacks

We capture real-world focal stacks using the Canon RF 50 mm F/1.8 lens and test pre-trained DfF models. We select 24 outdoor scenes and one indoor scene for evaluation.

#### 5.3.1 Outdoor Scenes Evaluation

The qualitative results of outdoor scenes are presented in Fig. 6, where we obtain the focus distance from the photo’s EXIF data. We observe that all models estimate good depth maps, but the AAT-trained models provide finer details and smoother results. Moreover, the AAT-trained models give a more hierarchical depth map and can distinguish depth differences between neighboring objects better. In contrast, the non-AAT models estimate less accurate depth maps, and the depth maps tend to confuse objects with little difference in depth, as indicated by the objects marked by boxes in the images. The two models exhibit significant differences at the corner of the estimated depth maps because off-axis lens aberrations are more pronounced at the image edge than at the center.

#### 5.3.2 Indoor Scene Evaluation

In Fig. 7, we present an indoor scene with objects at different positions and depths. We focus the camera on various objects to form the focal stack and measure more accurate focus distances with a ruler. Furthermore, we use the Lidar sensor from the iPhone14 Pro to scan the 3D scene, load the

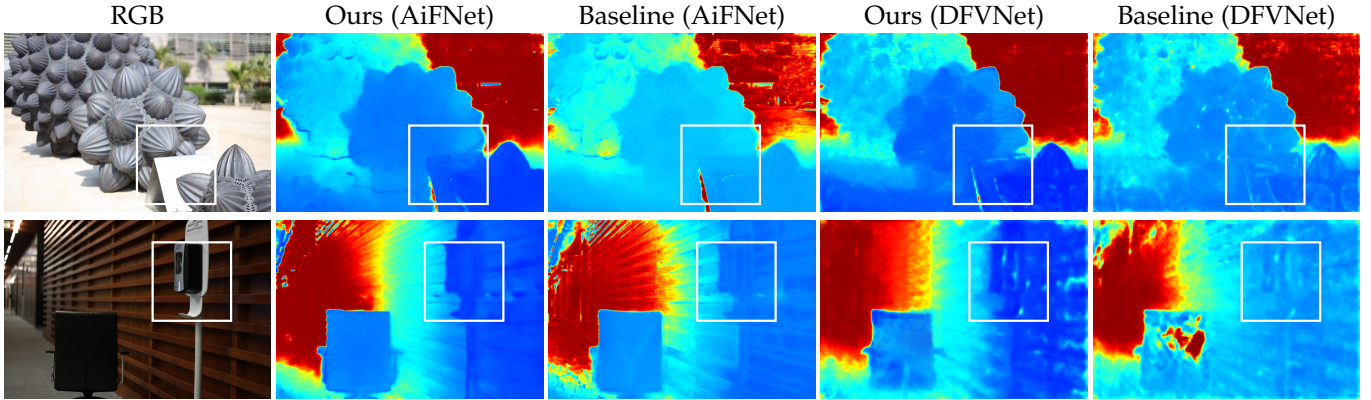


Fig. 6. **Qualitative results on real-world outdoor focal stacks.** We capture real-world focal stacks using the same camera lens that is used for training and test pretrained models on them. The depth map predicted by AAT models is more hierarchical and can better distinguish depth differences between neighboring objects, as labeled by boxes.

TABLE 4  
Quantitative results on real-world indoor focal stacks.

AiFNet	MAE ↓	MSE ↓	Abs.Rel. ↓	$\delta = 1.25 \uparrow$
Baseline	0.1775	0.0836	0.1736	0.7978
Ours	<b>0.1526</b>	<b>0.0770</b>	<b>0.1486</b>	<b>0.9029</b>

meshes into Blender software, and calibrate the camera’s position and view. After calibration, we render the depth map as the ground truth and calculate the score metrics. In Table. 4, the quantitative scores indicate that the AAT pre-trained model performs better than the baseline (“non-AAT”). We also observe from the zoomed image patches in Fig. 7 that the non-AAT model incorrectly treats texture information in RGB images as depth information and loses edge information. For more results on the evaluation, please refer to the Supplementary.

#### 5.4 Efficiency Evaluation of AAT Scheme

In an optical lens, aberrations typically become more severe as the off-axis angle increases. Therefore, in this section, we evaluate the effectiveness of the AAT scheme for across the FoV. We calculate the relevant error map for the estimated depth maps and average the results, as shown in Fig. 8. In the figure, the red color represents a large estimation error, while the blue color represents a small estimation error. When trained without the AAT scheme, the network model produces depth maps with significant errors in the corners of the image. These errors are caused by optical aberrations, which affect the determination of the best in-focus frame in a focal stack. In contrast, our proposed AAT scheme takes optical aberrations into account during the training of the DfF models. As a result, the final estimated depth map is virtually unaffected by off-axis optical aberrations and is more uniform compared to the baseline results.

#### 5.5 Training Time

We also analyze the extra time introduced by the AAT scheme, which results from the image re-rendering process. We use DFVNet to evaluate the training speed with varying

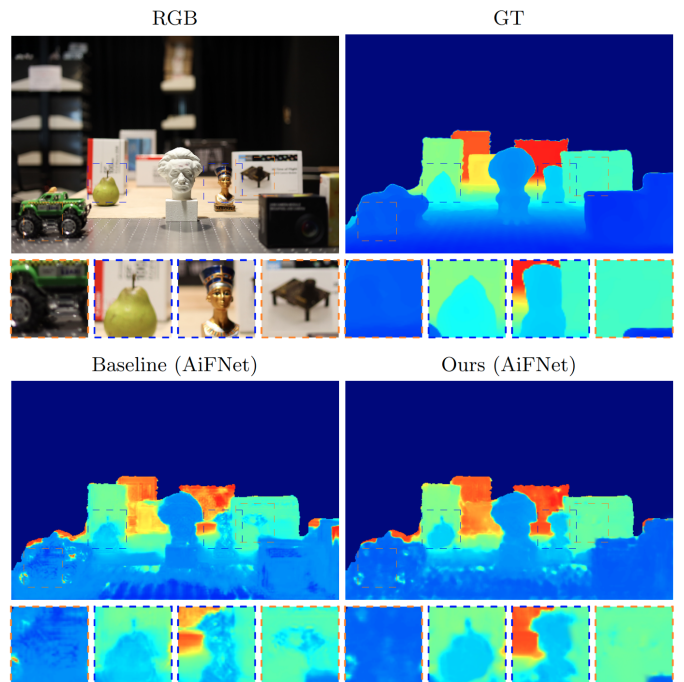


Fig. 7. **Qualitative results on real-world indoor focal stacks.** We set up an indoor scene with objects placed at different positions. The non-AAT model incorrectly treats texture information in RGB images as depth information and also loses edge details in the estimated depth map.

numbers of stacks. In Fig. 9, we report the average number of batches per second during training. The image resolution is  $480 \times 640$ . For AAT, we render the focus stack for each batch during the training. For comparison, we render all focused images before the experiment and only load them to form focus stacks during the training. The average speed is calculated over 200 batches within a training epoch, and we run the model for 5 epochs to reduce the variance. When the stack size is small ( $\leq 4$ ), the AAT training scheme reduces the training speed. However, in DfF applications, we typically use a stack size larger than 5 in our experiments, and the speed difference is negligible. Particularly, when the number of stacks is large ( $\geq 8$ ), AAT training is faster since it reduces the overhead of image loading.



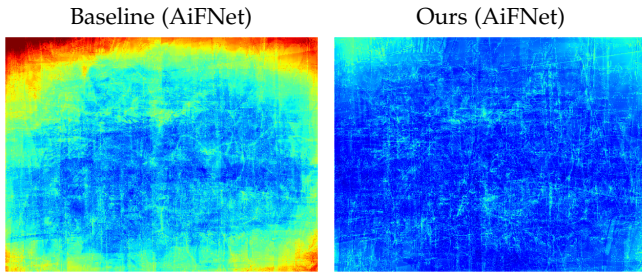


Fig. 8. **Depth estimation error map.** The baseline model incorrectly estimates depth information in the corners of the image. In contrast, the AAT scheme eliminates the effect of optical aberrations in depth estimation and leads to a uniform error map.

When the stack size equals 1, we analyze the additional time cost introduced in monocular image processing. The AAT scheme reduces the training speed by approximately 30%. However, this is also highly dependent on the network architecture. The DFVNet used for evaluation is a small and simple network; thus, the AAT scheme introduces a significant impact. However, we believe that this extra time can be negligible for more complex and larger network structures.

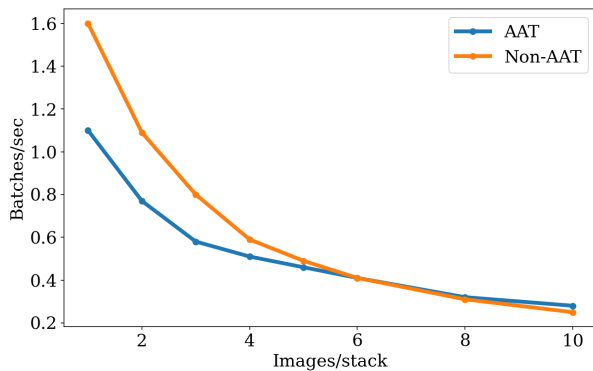


Fig. 9. **Comparison of training speed for different stack sizes.** The AAT scheme has a significantly slower speed when the focal size is small. But when the stack size is large, rendering focal stacks on-the-fly shows no time delay compared to loading focused images from the computer memory.

## 6 DISCUSSION

The experiments conducted in the previous section demonstrate that optical aberrations can degrade the generalizability of pretrained DfF models, but we can improve the results with the proposed AAT scheme. However, there are also some limitations that need to be considered, especially in real-world practice.

Firstly, the impact of optical aberrations is significant only for small defocus distances, as the PSF is dominated by defocus phenomena when object points are far from the focus plane. Therefore, the AAT scheme is more effective in improving depth estimation accuracy for adjacent objects with less significant defocus. However, in sparse scenes, the improvement may not be as significant.

Secondly, the AAT scheme can provide more significant improvements for lenses with larger aberrations. Current

DfF datasets contain fairly low-resolution images, in large part, because the depth sensors used to generate ground truth data have a limited pixel count. At these low image resolutions, most commercial-grade lenses only exhibit very modest aberrations. However, we believe that the AAT scheme will be vital for extending the DfF approach to the modern image sensor resolutions, as well as in the context of end-to-end learned compact computational imaging lenses [38], [49], [57].

Thirdly, the focus distance information obtained from the EXIF data of photography cameras is often inaccurate, leading to reduced depth estimation performance. However, modern imaging and display systems typically include Lidar or depth cameras that can provide more accurate focus distance information. By incorporating such information, we believe the depth estimation performance can be further improved, especially in scenarios where multiple cameras are employed.

## 7 CONCLUSION

In this work, we address the domain gap caused by off-axis optical aberrations, which has been overlooked by most existing works. To this end, we propose an AAT scheme to bridge this gap. Specifically, we develop a network to estimate the PSF of a real camera lens for different positions and focus distances. We then use the estimated PSF to simulate aberrated training images, enabling the network to learn to extract more accurate image features in the presence of optical aberrations. We evaluate the AAT scheme on two DfF networks and demonstrate its generalizability through both simulated and real-world experiments. The experimental results indicate that the AAT scheme improves the generalizability of the DfF models. The DfF models trained with AAT can estimate more accurate depth maps without fine-tuning compared to the baseline. Furthermore, we believe that the AAT scheme is not limited to the DfF task and can be applied to improve the generalizability of other computer vision tasks.

## ACKNOWLEDGMENTS

This work was supported by the King Abdullah University of Science and Technology (KAUST) individual baseline funding.

## REFERENCES

- [1] S. Niklaus, L. Mai, J. Yang, and F. Liu, "3D Ken Burns effect from a single image," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 1–15, 2019. 1
- [2] S. F. Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 4009–4018. 1, 2
- [3] Z. Xie, Z. Geng, J. Hu, Z. Zhang, H. Hu, and Y. Cao, "Revealing the dark secrets of masked image modeling," *arXiv preprint arXiv:2205.13543*, 2022. 1
- [4] V. Patil, C. Sakaridis, A. Liniger, and L. Van Gool, "P3Depth: Monocular depth estimation with a piecewise planarity prior," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 1610–1621. 1, 2
- [5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 779–788. 1

- [6] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H.-Y. Shum, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *The Eleventh International Conference on Learning Representations*, 2022. **1**
- [7] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "DeepMVS: Learning multi-view stereopsis," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 2821–2830. **1**
- [8] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 5038–5047. **1**
- [9] M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. J. Davison, "CodeSLAM — learning a compact, optimisable representation for dense visual SLAM," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 2560–2568. **1**
- [10] C. Tang and P. Tan, "BA-Net: Dense bundle adjustment network," *arXiv preprint arXiv:1806.04807*, 2018. **1**
- [11] K. Tateno, F. Tombari, I. Laina, and N. Navab, "CNN-SLAM: Real-time dense monocular slam with learned depth prediction," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 6243–6252. **1**
- [12] J. M. Facil, B. Ummenhofer, H. Zhou, L. Montesano, T. Brox, and J. Civera, "CAM-ConvS: Camera-aware multi-scale convolutions for single-view depth," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 11 826–11 835. **1**
- [13] S. Suwajanakorn, C. Hernandez, and S. M. Seitz, "Depth from focus with your mobile phone," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3497–3506. **1, 2**
- [14] S. Anwar, Z. Hayder, and F. Porikli, "Depth estimation and blur removal from a single out-of-focus image." in *BMVC*, vol. 1, 2017, p. 2. **1**
- [15] M. Carvalho, B. Le Saux, P. Trounev-Peloux, A. Almansa, and F. Champagnat, "Deep depth from defocus: how can defocus blur improve 3D estimation using dense neural networks?" in *Eur. Conf. Comput. Vis.*, 2018, pp. 0–0. **1, 2**
- [16] C. Hazirbas, S. G. Soyer, M. C. Staab, L. Leal-Taixé, and D. Cremers, "Deep depth from focus," in *Asian conference on computer vision*. Springer, 2018, pp. 525–541. **1, 2**
- [17] P. P. Srinivasan, R. Garg, N. Wadhwa, R. Ng, and J. T. Barron, "Aperture supervision for monocular depth estimation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 6393–6401. **1**
- [18] S. Gur and L. Wolf, "Single image depth estimation trained via depth from defocus cues," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 7683–7692. **1, 2**
- [19] M. Maximov, K. Galim, and L. Leal-Taixé, "Focus on defocus: bridging the synthetic to real domain gap for depth estimation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 1071–1080. **1, 2**
- [20] N.-H. Wang, R. Wang, Y.-L. Liu, Y.-H. Huang, Y.-L. Chang, C.-P. Chen, and K. Jou, "Bridging unsupervised and supervised depth from focus via all-in-focus supervision," in *Int. Conf. Comput. Vis.*, 2021, pp. 12 621–12 631. **1, 2, 3, 5, 6**
- [21] C. Won and H.-G. Jeon, "Learning depth from focus in the wild," in *European Conference on Computer Vision*. Springer, 2022, pp. 1–18. **1, 2**
- [22] F. Yang, X. Huang, and Z. Zhou, "Deep depth from focus with differential focus volume," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 12 642–12 651. **1, 2, 3, 5, 6**
- [23] X. Zhang, "Deep learning-based multi-focus image fusion: A survey and a comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4819–4838, 2021. **1**
- [24] T. Broad and M. Grierson, "Light field completion using focal stack propagation," in *ACM SIGGRAPH 2016 Posters*, 2016, pp. 1–2. **1**
- [25] L. Ruan, B. Chen, J. Li, and M.-L. Lam, "AIFNet: All-in-focus image restoration network using a light field-based dataset," *IEEE Trans. Comput. Imaging*, vol. 7, pp. 675–688, 2021. **1**
- [26] P. Trounev, F. Champagnat, G. Le Besnerais, J. Sabater, T. Avignon, and J. Idier, "Passive depth estimation using chromatic aberration and a depth from defocus approach," *Appl. Opt.*, vol. 52, no. 29, pp. 7152–7164, 2013. **1**
- [27] J. Chang and G. Wetzstein, "Deep optics for monocular depth estimation and 3D object detection," in *Int. Conf. Comput. Vis.*, 2019, pp. 10 193–10 202. **1**
- [28] S.-H. Baek, H. Ikoma, D. S. Jeon, Y. Li, W. Heidrich, G. Wetzstein, and M. H. Kim, "Single-shot hyperspectral-depth imaging with learned diffractive optics," in *Int. Conf. Comput. Vis.*, 2021, pp. 2651–2660. **1**
- [29] M. Marquez, P. Meza, F. Rojas, H. Arguello, and E. Vera, "Snapshot compressive spectral depth imaging from coded aberrations," *Opt. Express*, vol. 29, no. 6, pp. 8142–8159, 2021. **1**
- [30] C. Wang, N. Chen, and W. Heidrich, "dO: A differentiable engine for deep lens design of computational imaging systems," *IEEE Trans. Comput. Imaging*, vol. 8, pp. 905–916, 2022. **1, 2, 3, 4**
- [31] X. Yang, Q. Fu, and W. Heidrich, "Curriculum learning for ab initio deep learned refractive optics," *arXiv preprint arXiv:2302.01089*, 2023. **1, 4**
- [32] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 270–279. **2, 3**
- [33] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 5410–5418. **2**
- [34] Z. Shen, Y. Dai, and Z. Rao, "CFNet: Cascade and fused cost volume for robust stereo matching," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 13 906–13 915. **2**
- [35] J. Surh, H.-G. Jeon, Y. Park, S. Im, H. Ha, and I. So Kweon, "Noise robust depth from focus using a ring difference filter," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 6328–6337. **2**
- [36] M. H. Freeman, *Optics*. Butterworth-Heinemann, 1990. **2, 3**
- [37] C. Kolb, D. Mitchell, and P. Hanrahan, "A realistic camera model for computer graphics," in *Proceedings of the 22nd annual conference on computer graphics and interactive techniques*, 1995, pp. 317–324. **2, 3**
- [38] Q. Sun, C. Wang, F. Qiang, D. Xiong, and H. Wolfgang, "End-to-end complex lens design with differentiable ray tracing," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–13, 2021. **2, 9**
- [39] E. Tseng, A. Mosleh, F. Mannan, K. St-Arnaud, A. Sharma, Y. Peng, A. Braun, D. Nowrouzezahrai, J.-F. Lalonde, and F. Heide, "Differentiable compound optics and processing pipeline optimization for end-to-end camera design," *ACM Trans. Graph.*, vol. 40, no. 2, pp. 1–19, 2021. **2, 4, 5**
- [40] Y. Li, Y.-L. Wu, P. Hoess, M. Mund, and J. Ries, "Depth-dependent PSF calibration and aberration correction for 3D single-molecule localization," *Biomed. Opt. Express*, vol. 10, no. 6, pp. 2708–2718, 2019. **2**
- [41] F. Chen, J. Y. Cheng, V. Taviani, V. R. Sheth, R. L. Brunsing, J. M. Pauly, and S. S. Vasanaawala, "Data-driven self-calibration and reconstruction for non-cartesian wave-encoded single-shot fast spin echo using deep learning," *J. Magn. Reson. Imaging*, vol. 51, no. 3, pp. 841–853, 2020. **2**
- [42] J. D. Rego, K. Kulkarni, and S. Jayasuriya, "Robust lensless image reconstruction via PSF estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 403–412. **2**
- [43] K. Yanny, N. Antipa, W. Liberti, S. Dehaeck, K. Monakhova, F. L. Liu, K. Shen, R. Ng, and L. Waller, "Miniscope3D: optimized single-shot miniature 3D fluorescence microscopy," *Light Sci. Appl.*, vol. 9, no. 1, p. 171, 2020. **2, 5**
- [44] L. Denis, E. Thiébaud, F. Soulez, J.-M. Becker, and R. Mourya, "Fast approximations of shift-variant blur," *Int. J. Comput. Vis.*, vol. 115, pp. 253–278, 2015. **2**
- [45] L. Denis, E. Thiébaud, and F. Soulez, "Fast model of space-variant blurring and its application to deconvolution in astronomy," in *IEEE Int. Conf. Image Process.* IEEE, 2011, pp. 2817–2820. **2**
- [46] F. Sroubek, J. Kamenicky, and Y. M. Lu, "Decomposition of space-variant blur in image deconvolution," *IEEE Signal Process. Lett.*, vol. 23, no. 3, pp. 346–350, 2016. **2**
- [47] K. Yanny, K. Monakhova, R. W. Shuai, and L. Waller, "Deep learning for fast spatially varying deconvolution," *Optica*, vol. 9, no. 1, pp. 96–99, 2022. **2, 5**
- [48] F. A. Jenkins and H. E. White, "Fundamentals of optics," *Indian J. Phys.*, vol. 25, pp. 265–266, 1957. **3**
- [49] Z. Li, Q. Hou, Z. Wang, F. Tan, J. Liu, and W. Zhang, "End-to-end learned single lens design using fast differentiable ray tracing," *Opt. Lett.*, vol. 46, no. 21, pp. 5453–5456, 2021. **3, 9**
- [50] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017. **4, 6**
- [51] —, "SGDR: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016. **4, 6**
- [52] J. Ichimura, "Optical system and image pickup apparatus having the same," U.S. Patent US20210263286A1, 8 26, 2021, uS Patent. [Online]. Available: <https://patents.google.com/patent/US20210263286A1/en> **4**

- [53] G. Côté, J.-F. Lalonde, and S. Thibault, "Deep learning-enabled framework for automatic lens design starting point generation," *Opt. Express*, vol. 29, no. 3, pp. 3841–3854, 2021. [5](#)
- [54] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 4040–4048. [6](#)
- [55] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German conference on pattern recognition*. Springer, 2014, pp. 31–42. [6](#)
- [56] Y. Zhang and T. Funkhouser, "Deep depth completion of a single RGB-D image," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 175–185. [6](#)
- [57] E. Tseng, S. Colburn, J. Whitehead, L. Huang, S.-H. Baek, A. Majumdar, and F. Heide, "Neural nano-optics for high-quality thin lens imaging," *Nat. Commun.*, vol. 12, no. 1, p. 6493, 2021. [9](#)



**Xinge Yang** received the B.S. degree in physics (major) and computer science (minor) from University of Science and Technology of China in 2020 and the M.Sc. in computer science from King Abdullah University of Science and Technology in 2022. He is currently a Ph.D. student at King Abdullah University of Science and Technology. His research interests include computational imaging, computational lens design, and differentiable optical simulation.



**Qiang Fu** received the B.S. degree in Mechanical Engineering from University of Science and Technology of China, Hefei, China in 2007 and the Ph.D. degree in Optical Engineering from University of Chinese Academy of Sciences, Beijing, China in 2012. He is currently a Research Scientist at the Visual Computing Center, KAUST. Previously, he was a Research Associate Professor at ShanghaiTech University (2016-2017), Postdoctoral Research Fellow at KAUST (2014-2016), Assistant Research Fellow

at Academy of Opto-Electronics, Chinese Academy of Sciences (2012-2014). His research interests lie in the multidisciplinary research areas of computational imaging, including optical system design, diffractive optics, micro/nano-fabrication, hyperspectral imaging, polarization imaging, high dynamic range imaging, wavefront sensing etc.



**Mohamed Elhoseiny** is an assistant professor of Computer Science at KAUST. Researcher at Facebook Research. Previously, he was a visiting Faculty at Stanford Computer Science department (2019-2020), Visiting Faculty at Baidu Research Silicon Valley Lab (2019), and Postdoc researcher at Facebook AI Research (2016-2019). Dr. Elhoseiny did his Ph.D. in 2016 at Rutgers University during which he spent time at Adobe Research (2015-2016) for more than a year and at SRI International in 2014 (a best

intern award). Dr. Elhoseiny received an NSF Fellowship in 2014 for the Write-a-Classifier project (ICCV13) and the Doctoral Consortium award at CVPR 2016. Dr. Elhoseiny was also a Stanford Igniter as a venture idea generator; Stanford Ignite is an entrepreneurship program at Stanford Graduate School of Business. His primary research interest is in computer vision and especially in efficient multimodal learning with limited data in areas like zero/few-shot learning, Vision & Language, language guided visual perception. He is also interested in Affective AI and especially to produce novel art and fashion with AI. His creative AI work was featured in MIT Tech Review, New Scientist Magazine, and HBO Silicon Valley.



**Wolfgang Heidrich** (Fellow, IEEE) is a Professor of Computer Science and Electrical and Computer Engineering in the KAUST Visual Computing Center, for which he also served as director from 2014 to 2021. Heidrich joined KAUST in 2014, after 13 years as a faculty member at the University of British Columbia. He received his PhD in from the University of Erlangen in 1999, and then worked as a Research Associate in the Computer Graphics Group of the Max-Planck-Institute for Computer Science

in Saarbrücken, Germany, before joining UBC in 2000. Heidrich's research interests lie at the intersection of imaging, optics, computer vision, computer graphics, and inverse problems. His more recent interest is in computational imaging, focusing on hardware-software co-design of the next generation of imaging systems, with applications such as High-Dynamic Range imaging, compact computational cameras, hyperspectral cameras, to name just a few. His work on HDR displays serves as the basis for many modern-day HDR consumer devices today. Heidrich is a Fellow of the IEEE, AAIA, and Eurographics, and the recipient of a Humboldt Research Award as well as the ACM SIGGRAPH Computer Graphics Achievement Award.