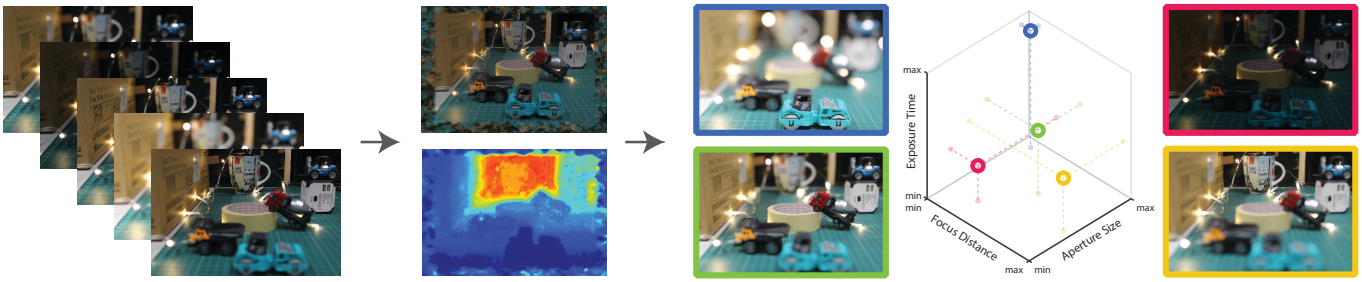
An Implicit Neural Representation for the Image Stack: Depth, All in Focus, and High Dynamic Range

CHAO WANG, Max-Planck-Institut für Informatik, Germany
ANA SERRANO, Universidad de Zaragoza, I3A, Spain
XINGANG PAN, Max-Planck-Institut für Informatik, Germany & Nanyang Technological University, Singapore
KRZYSZTOF WOLSKI, Max-Planck-Institut für Informatik, Germany
BIN CHEN, Max-Planck-Institut für Informatik, Germany
KAROL MYSZKOWSKI, Max-Planck-Institut für Informatik, Germany
HANS-PETER SEIDEL, Max-Planck-Institut für Informatik, Germany
CHRISTIAN THEOBALT, Max-Planck-Institut für Informatik, Germany
THOMAS LEIMKÜHLER, Max-Planck-Institut für Informatik, Germany



Sparse Input TAF Stack

Learned Canonical HDR AiF Image and Depth Map

Continuous 3D Domain

Fig. 1. Our approach is capable of accurately reconstructing a canonical all-in-focus (AiF), high-dynamic-range (HDR) radiance map alongside with depth information, using a sparse Time-Aperture-Focus (TAF) stack as input. This gives full post-processing control over focus, aperture, and exposure conditions.

In everyday photography, physical limitations of camera sensors and lenses frequently lead to a variety of degradations in captured images such as saturation or defocus blur. A common approach to overcome these limitations is to resort to image stack fusion, which involves capturing multiple images with different focal distances or exposures. For instance, to obtain an all-in-focus image, a set of multi-focus images is captured. Similarly, capturing multiple exposures allows for the reconstruction of high dynamic range. In this paper, we present a novel approach that combines neural fields with an expressive camera model to achieve a unified reconstruction of an all-in-focus high-dynamic-range image from an image stack. Our approach is composed of a set of specialized implicit neural representations tailored to address specific sub-problems along our pipeline: We use neural implicits to predict flow to overcome misalignments arising from lens breathing, depth, and all-in-focus images to account for depth of field, as well as tonemapping to deal with sensor responses and saturation – all trained using a physically inspired supervision structure with a differentiable thin lens model at its core. An important benefit of our approach is its ability to handle these tasks simultaneously or independently, providing flexible post-editing capabilities such as refocusing and exposure adjustment. By sampling the three primary factors in photography within our framework (focal distance, aperture, and exposure time), we conduct a thorough exploration to gain valuable insights into their significance and impact on overall reconstruction quality. Through extensive validation, we demonstrate that our method outperforms existing approaches in both depth-from-defocus and all-in-focus image reconstruction tasks. Moreover, our approach exhibits promising results in each of these three dimensions, showcasing its potential to enhance captured image quality and provide greater control in post-processing.

CCS Concepts: • **Computing methodologies** → **Computational photography**; **Image representations**.

Additional Key Words and Phrases: High Dynamic Range Imaging, Depth from Defocus, Neural Fields

ACM Reference Format:
Chao Wang, Ana Serrano, Xingang Pan, Krzysztof Wolski, Bin Chen, Karol Myszkowski, Hans-Peter Seidel, Christian Theobalt, and Thomas Leimkühler. 2023. An Implicit Neural Representation for the Image Stack: Depth, All

Authors' addresses: Chao Wang, chaowang@mpi-inf.mpg.de, Max-Planck-Institut für Informatik, Saarbrücken, Germany; Ana Serrano, anase@unizar.es, Universidad de Zaragoza, I3A, Zaragoza, Spain; Xingang Pan, xingang.pan@ntu.edu.sg, Max-Planck-Institut für Informatik, Saarbrücken, Germany & and Nanyang Technological University, Singapore; Krzysztof Wolski, kwolski@mpi-inf.mpg.de, Max-Planck-Institut für Informatik, Saarbrücken, Germany; Bin Chen, binchen@mpi-inf.mpg.de, Max-Planck-Institut für Informatik, Saarbrücken, Germany; Karol Myszkowski, karol@mpi-inf.mpg.de, Max-Planck-Institut für Informatik, Saarbrücken, Germany; Hans-Peter Seidel, hpseidel@mpi-sb.mpg.de, Max-Planck-Institut für Informatik, Saarbrücken, Germany; Christian Theobalt, theobalt@mpi-inf.mpg.de, Max-Planck-Institut für Informatik, Saarbrücken, Germany; Thomas Leimkühler, thomas.leimkuehler@mpi-inf.mpg.de, Max-Planck-Institut für Informatik, Saarbrücken, Germany.

0730-0301/2023/12-ART221
https://doi.org/10.1145/3618367

in Focus, and High Dynamic Range. *ACM Trans. Graph.* 42, 6, Article 221 (December 2023), 11 pages. https://doi.org/10.1145/3618367

1 INTRODUCTION

Capturing and reproducing the real world is a fundamental goal of image processing. However, due to imaging device limitations, capturing all relevant information in a single shot is often challenging. Image fusion techniques improve image quality by combining multiple images. Traditionally, all-in-focus (AiF) image reconstruction [Bouzos et al. 2019; Li et al. 2013; Liu et al. 2017; Si et al. 2023; Suwajanakorn et al. 2015; Zhang and Levine 2016], depth estimation [Favaro 2010; Hazirbas et al. 2019; Lin et al. 2013; Maximov et al. 2020; Moeller et al. 2015; Si et al. 2023; Suwajanakorn et al. 2015; Won and Jeon 2022; Yang et al. 2022], and high-dynamic-range (HDR) reconstruction [Debevec and Malik 1997; Mertens et al. 2007] are treated separately, requiring multiple images with varying focus and exposure settings. This process is time-consuming and leads to redundant image capture.

In this work, we address these challenges by focusing on the reconstruction of AiF HDR radiance maps from a sparse image burst, where each image may contain entangled information from both defocus blur and different exposure conditions. Our approach simplifies image capture, improves fusion efficiency, and enables post-editing capabilities such as refocusing, bokeh reproduction, and exposure adjustment. We combine an implicit neural representation with an expressive camera model to reconstruct both the AiF HDR image and the corresponding depth from the sparse image burst without the need for ground-truth supervision.

The problem we address is notably challenging and requires a comprehensive approach due to the interplay of three crucial elements in the image capture process: *focal distance*, *aperture*, and *exposure time*, which we model in our approach to simulate the imaging pipeline (Sec. 3). *Focal distance* determines which parts of the scene will appear sharp and in focus. During focal distance adjustment in a focal stack, lens breathing usually appears, leading to pixel misalignment and ghosting artifacts in captured images. The *aperture* controls lens opening, affecting both the amount of light reaching the sensor and depth of field. Our framework incorporates a differentiable thin lens formulation to model both focal distance and aperture. Additionally, we introduce a novel approach to mitigate lens breathing using an implicit flow, which enables warping each image in the focal stack to a canonical view, eliminating the need for a reference image and preserving all pixel information. While few existing learning-based methods have attempted to address lens breathing, they focus on a single specific problem such as depth from defocus (DfD) [Moeller et al. 2015; Suwajanakorn et al. 2015; Won and Jeon 2022] and often result in a loss of pixel information. Our framework also tackles the problem of DfD by leveraging a thin-lens formulation as a physical constraint to model shallow depth of field in the input images, outperforming the state of the art in both depth estimation and AiF image reconstruction (Sec. 4.1). Finally, *exposure time* controls the light exposure of the image sensor. Longer exposures capture more light, resulting in a brighter image. However, extended exposures can also saturate the sensor and lead to clipping. To model this process, we introduce an implicit tone mapper that supports the reconstruction of the HDR radiance map from the different exposures in the low-dynamic-range (LDR) input images.

Relying on these components, our framework achieves high flexibility. In contrast to previous approaches [Debevec and Malik 1997; Hasinoff and Kutulakos 2007; Mertens et al. 2007; Si et al. 2023; Suwajanakorn et al. 2015; Won and Jeon 2022; Yang et al. 2022], it supports as input any set of images with different unstructured combinations of focal distance, aperture, and exposure to learn an AiF HDR image and its corresponding depth in an unsupervised manner. We term this stack of unstructured images *Time-Aperture-Focus stack* (TAF). Our approach can handle an arbitrary number of input images and produces outstanding results with only five images. To achieve this, the estimated depth serves as an input to our lens model, which produces a differentiable disk kernel. This kernel is applied to blur the AiF image, effectively modeling the combined effects of focus and aperture. In a final step, the blurred image is tone mapped to produce a defocused LDR image. Then, a reconstruction loss is computed by comparing this output image and the input image from the TAF stack, guiding the iterative refinement of the predicted AiF image and depth map. After optimization, our framework further enables post-editing capabilities such as focus, aperture, and exposure editing (Fig. 1). More results can be found in the *supplementary material* and *video*.

Further, we extensively investigate the impact of different combinations of focal distance, aperture, and exposure on the overall reconstruction performance of our method (Sec. 4.2), revealing two effective strategies for improving AiF HDR image reconstruction and depth estimation. On one hand, utilizing a moderate aperture size and multiple exposure times with focal sweeping yields better AiF HDR image reconstruction. On the other hand, maintaining a fixed exposure time while varying the aperture size during focal sweeping leads to improved depth estimation.

In summary, we present the following contributions:

- We introduce a unified implicit neural representation that takes as input any set of images with different unstructured combinations of focal distance, aperture, and exposure (TAF stack), guided by a physically-inspired supervision structure centered around a differentiable thin lens model. This representation enables the reconstruction of an AiF HDR image and depth, and supports flexible post-editing, such as refocusing, as well as aperture and exposure adjustment in a well-disentangled manner.
- Our proposed method achieves outstanding results in DfD and AiF image reconstruction tasks, outperforming the state of the art.
- To support our analysis and validation, we introduce a new dataset of focal stacks comprising 10 synthetic scenes simulating lens-breathing effects (including five HDR scenes), and 25 captured scenes using various cameras and lenses.

Our code and datasets are available under https://taf.mpi-inf.mpg.de/.

2 RELATED WORK

We first discuss previous work addressing the problem of obtaining depth and AiF from the focal stack, which is one of the main

applications of our framework. Then, we focus on implicit neural representations and their recent use in applications close to ours.

2.1 Depth and AiF from Focal Stacks

Traditional depth estimation from a focal stack can be categorized into depth from focus (DfF) and depth from defocus (DfD) methods. DfF approaches face challenges in determining a suitable focus criterion that accurately captures the focus information across various scenes and conditions, and effectively detecting the highest focus values and accurately localizing the corresponding depths. DfD approaches, on the other hand, face challenges in accurately capturing depth information from defocused images and balancing depth estimation accuracy with image resolution. Despite notable advancements [Favaro 2010; Lin et al. 2013; Moeller et al. 2015; Suwajanakorn et al. 2015], the quality of depth estimation in traditional methods still falls short of achieving satisfactory results.

With the advent of deep learning, the differences between DfF and DfD have diminished, as both approaches usually utilize focal stacks as input and ground-truth depth as supervision to train end-to-end models [Hazirbas et al. 2019; Maximov et al. 2020; Won and Jeon 2022; Yang et al. 2022]. Wang et al. [2021] presented a novel approach for jointly estimating depth and an AiF image from an input focal stack by designing a shared common network that can be trained either supervised with ground-truth depth maps or unsupervised with only ground-truth AiF images. Recently, in concurrent work, Si et al. [2023] proposed a self-supervised framework for DfD which also considers a thin lens model for predicting both depth and AiF images from focal stacks while being supervised through input reconstruction. However, their end-to-end network is trained on a synthetic focal stack dataset, which may not generalize well to real-world scenarios. Additionally, they do not consider lens breathing and struggle with a small number of input images. We show and discuss the limitations of their approach through quantitative and qualitative comparisons.

Despite significant advancements in learning-based approaches, multiple problems remain unsolved. For instance, obtaining real focal stacks with ground-truth depth for training is challenging due to factors like lens breathing. Although synthetic datasets can circumvent this issue, they often introduce domain gaps [Si et al. 2023; Yang et al. 2022]. Moreover, while many learning-based methods treat DfD and DfF as a regression problem, defocus blur can be more effectively modeled as a physical phenomenon that holds implicit cues for direct depth inference. Our approach does not rely on ground-truth data for supervision during training and, unlike previous works, is highly flexible as it can handle diverse combinations of focus, aperture and exposure in the TAF stack as input to reconstruct the AiF HDR image and corresponding depth. Further, our approach outperforms state-of-the-art approaches in reconstructing depth from the focal stack.

2.2 Implicit Neural Representations and Their Applications

Recent literature has demonstrated the potential of fully connected networks for memory-efficient and continuous implicit representations, known as implicit neural representations. Sitzmann et al. [2020] demonstrated how to use these representations to effectively model various signals, including images and 3D shapes, in particular using periodic activation functions and Multi-Layer Perceptrons (MLPs). Neural Radiance Fields (NeRF) [Mildenhall et al. 2021] employ MLPs to parametrize 5D radiance fields, enabling high-fidelity novel-view synthesis and 3D reconstruction. Building upon NeRF, Dark-NeRF [Mildenhall et al. 2022] uses raw images to train the model and maximize the available information for tasks like denoising, HDR reconstruction, and refocusing. HDR-NeRF [Huang et al. 2022] directly learns HDR from LDR inputs and uses a tone mapper for re-projecting HDR content to different exposure intervals. Recently, Jun-Seong et al. [2022] explored HDR radiance fields using a plenoptic function as a scene representation, eliminating the need for exposure information during training. Wu et al. [2022] identified that the quality of NeRF decreases when the input images have shallow depth of field (DoF) and introduced a differentiable circle of confusion (CoC) representation to simulate radiance scattering, allowing the synthesis of AiF images and DoF rendering. For recovering clear scene representations from blurred images, Ma et al. [2022] learned a deformable kernel as a degradation model. Other implicit neural representation applications include image alignment and layering [Nam et al. 2022], video fitting with latent codes [Feng et al. 2022], video editing using layered 2D atlases parameterized by MLPs [Kasten et al. 2021], and improving mixed reality rendering coherence through the learned camera characteristics [Mandl et al. 2021]. Our framework draws inspiration from these works and leverages implicit neural representations.

3 METHOD

Our goal is to design an implicit neural representation that, by learning from an image stack, models the AiF HDR image and depth map of the scene. The pipeline of our approach is shown in Fig. 2. We first introduce how we represent the depth and the AiF HDR image (Sec. 3.1). Importantly, we present an implicit flow model to compensate for the lens breathing effect in the image stack. Next, we describe our differentiable thin lens model that renders a defocused HDR image from the depth and AiF image (Sec. 3.2). We then introduce an implicit tone mapper that maps the defocused HDR image to an LDR image conditioned on the level of exposure (Sec. 3.3). Finally, we discuss the loss functions used to train the implicit neural representation (Sec. 3.4). After training, the representation can be used to synthesize images under new focal distances, apertures, and exposures.

3.1 Depth and Image Representation with Implicit Flow

Input to our system is an image stack $\{\boldsymbol{I}_i\}_{i=1}^{n}$. For each image $\boldsymbol{I}_i$, we also have its corresponding camera metadata including focal distance, aperture, and exposure, which differ between images in the stack. We aim to learn a shared depth map $\boldsymbol{D}$ and AiF HDR image $\boldsymbol{I}_a$ from the image stack.

In our pipeline, we represent the depth map $\boldsymbol{D}$ and AiF HDR image $\boldsymbol{I}_a$ using coordinate-based MLPs. A simple design is to use two MLPs that take pixel coordinates (x, y) as input and output the depth $S_2 \in \mathbb{R}^+$ and RGB color $\boldsymbol{C} \in \mathbb{R}^3$ respectively. However, this design does not take into account lens breathing, the phenomenon where changing the focus distance causes a slight change in the field of view

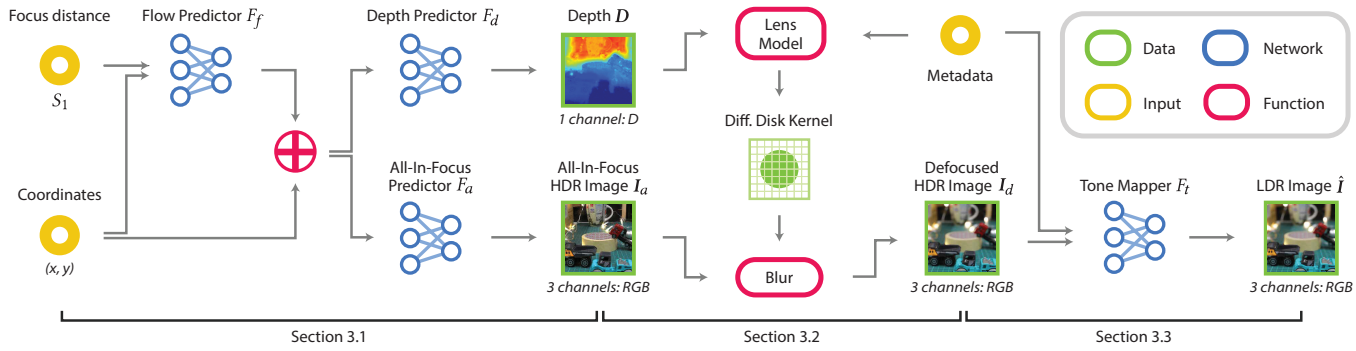


Fig. 2. Pipeline of our method. We represent a depth map and AiF HDR image via coordinate-based MLPs. An implicit flow predictor is adopted to model the lens breathing effect. Given the depth map, AiF HDR image, and the metadata of the camera, we calculate the defocused HDR image via a novel thin-lens model that is differentiable w.r.t. the depth map. Finally, the defocused HDR image is projected to LDR via an implicit tone mapper, again modeled by an MLP.

(FoV) [Gross 2005]. This effect occurs due to the movement of the lens elements inside the lens barrel to adjust the focus distance, and can be observed even when the camera is stationary. Consequently, it introduces misalignments between the same pixel coordinates of images with different focus distances.

Previous techniques for addressing lens breathing typically use the smallest FoV image in a focal stack as reference and align all other images in the stack to it by cropping and resizing using feature extraction and matching [Suwajanakorn et al. 2015; Won and Jeon 2022]. However, this approach may cause a loss of information in the focal stack and can only generate a single FoV AiF image. In this paper, we propose a novel approach to address lens breathing by learning a canonical view from the focal stack that encompasses all pixel information. This approach employs an implicit flow to project the canonical view onto each FoV in the stack, instead of aligning to a specific reference image. To account for lens breathing, we embed the focus distance S_1, which is the key factor in lens breathing, into the coordinate input, and use an MLP F_f as the flow predictor to predict an offset in the x and y coordinates. This continuous offset compensates for the lens breathing effect and warps pixel locations to canonical coordinates:

$$(\triangle x, \triangle y) = F_f\,(x, y, S_1)\,, \tag{1}$$

$$(x', y') = (x + \triangle x, y + \triangle y)\,. \tag{2}$$

This way, pixels corresponding to different focus distances are aligned in the canonical space. An example of the learned flow is shown in Fig. 10.

The canonical coordinates are then used to query the depth and HDR color with two MLPs: $F_d : (x', y') \mapsto S_2$ and $F_a : (x', y') \mapsto \boldsymbol{C}$. Thus, the full depth map $\boldsymbol{D}$ and AiF image $\boldsymbol{I}_a$ can be obtained by querying the MLPs using the coordinates grid of all pixels, as illustrated in the left half of Fig. 2.

3.2 Differentiable Thin Lens Model

Given the AiF HDR image and depth map, our next step is to estimate the defocused HDR image. Defocus blur is a prevalent phenomenon in everyday photography that arises when a camera lens fails to focus light rays onto a single point on the image sensor. The thin-lens model [Potmesil and Chakravarty 1981] can be used to explain defocus blur in an image, where the camera lens is parameterized by a focal length f, focus distance S_1, and aperture N (*i.e.*, F-number). When an object is in focus, the distance between the lens and the image sensor is adjusted to ensure that the object's image is formed on the image sensor. However, if the object is not at the focal distance, the lens will produce a disk on the sensor rather than a single point. This disk is referred to as the circle of confusion (CoC), and its diameter d can be computed as

$$d = \frac{|S_2 - S_1|}{S_2} \frac{f^2}{N\,(S_1 - f)}. \tag{3}$$

Once we have obtained d, we can simulate the defocus blur via a spatially-varying convolution. Let $\boldsymbol{I}_a(x, y)$ be the pixel value of $\boldsymbol{I}_a$ at (x, y) and $W(u, v, x, y)$ be the weight of a spatially-varying kernel, where x and y represent the current position of the convolution in image space, u and v is the location within the kernel. The defocused image $\boldsymbol{I}_d$ of the spatially-varying convolution is given by

$$\boldsymbol{I}_d(x, y) = \iint \boldsymbol{I}_a(x - u, y - v) W(u, v, x, y)\,\mathrm{d}u\,\mathrm{d}v. \tag{4}$$

Following the discussion above, we model W using a disk of unit energy with diameter d. Notice that a disk kernel is a closer approximation to the defocus operation than a commonly used Gaussian kernel [Potmesil and Chakravarty 1981]. However, a conventional disk kernel is discontinuous and therefore not differentiable with respect to its diameter d [Bangaru et al. 2021]. Consequently, the gradients cannot be back-propagated to the depth S_2 and thus the depth map cannot be optimized. In this paper, we propose an approach inspired by Gwosdek et al. [2012], to address this issue by introducing a differentiable disk kernel. This kernel enables smooth optimization while accurately representing the defocus blur effect. To connect the continuous CoC diameter d to the discrete kernel grid, we introduce a soft boundary of the kernel as follows:

$$\hat{W}(u, v, x, y) = \begin{cases} 1, & m \le \frac{d-1}{2} \\ \frac{d+1}{2} - m, & \frac{d-1}{2} < m \le \frac{d+1}{2} \\ 0, & \frac{d+1}{2} < m \end{cases} \tag{5}$$

$$W(u, v, x, y) = \frac{\hat{W}(u, v, x, y)}{\sum_{u,v} \hat{W}(u, v, x, y)} \tag{6}$$

where m represents the distance from (u, v) to the center of the disk, and (u, v) are evaluated on the discrete pixel grid. Notice how this is a simple and efficient approximation of anti-aliasing, which does not require integrating over pixel area. This allows the gradients to be back-propagated through d, ensuring smooth optimization of the depth map. Fig. 3 illustrates our kernels.

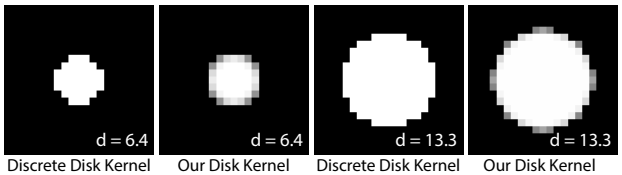


Fig. 3. Discrete disk kernels vs. our differentiable versions.

3.3 Implicit Tone Mapper

In order to infer HDR information from an LDR image stack with different levels of exposure, we require a tone mapping operation that projects HDR to LDR. Typically, the following function is used for tone mapping [Liu et al. 2020; Wang et al. 2022]:

$$\hat{\boldsymbol{I}}(x, y, EV) = \gamma\left(\text{clip}\left(\boldsymbol{I}_d(x, y) \cdot 2^{EV}\right)\right), \tag{7}$$

where $\hat{\boldsymbol{I}}$ is an LDR image, and EV quantifies the level of exposure, *i.e.*, the amount of light that reaches the sensor, depending on aperture and exposure time. The clip operation is employed to constrain the range of values within the interval $[0, 1]$, and $\gamma(\cdot)$ is the gamma correction. However, there is still a gap between this approximation and real camera response curves, and we found that directly incorporating this explicit model into our framework produces artifacts (Sec. 4.3).

To tackle this challenge, we use an implicit tone mapper F_t, again represented by an MLP:

$$\hat{\boldsymbol{I}}_j(x, y, EV) = F_t\left(\boldsymbol{I}_{d,j}(x, y) \cdot 2^{EV}\right), \tag{8}$$

where subscripts j denote color channels, *i.e.*, each channel is processed independently using the same network. To ensure an LDR output, we employ a hyperbolic tangent as the last layer of the MLP and linearly remap its outputs to $[0, 1]$.

3.4 Loss

We jointly optimize all four MLPs using the loss function

$$\mathcal{L} = \sum_{i=1}^{n}\left(\mathcal{L}_{\text{rec}}(\hat{\boldsymbol{I}}_i, \boldsymbol{I}_i) + \alpha\mathcal{L}_{\text{VGG}}(\hat{\boldsymbol{I}}_i, \boldsymbol{I}_i) + \beta\mathcal{L}_{\text{reg}} + \gamma\mathcal{L}_{\text{TV}}\right), \tag{9}$$

where $\mathcal{L}_{\text{rec}}$ is the MSE loss, $\mathcal{L}_{\text{VGG}}$ is a perceptual loss [Johnson et al. 2016], $\mathcal{L}_{\text{reg}}$ is a regularization term for flow, and $\mathcal{L}_{\text{TV}}$ is a total variation regularization penalty for the depth map. α, β and γ are weights to balance the components, which are set to $\alpha = 0.01$, $\beta = 0.5$, and $\gamma = 0.05$ in all our experiments. The combination of MSE and perceptual loss is a common practice [Feng et al. 2022; Liu et al. 2020]. For the perceptual loss $\mathcal{L}_{\text{VGG}}$, we compute squared differences of intermediate feature maps of a pre-trained VGG16 [Simonyan and Zisserman 2014] network (layers 5, 10, and 16). To enforce smoothness in our implicit flow, we include the regularization term

$$\mathcal{L}_{\text{reg}} = \sum \left\| J_{F_f}(x, y, f_d) \right\|_1, \tag{10}$$

where J_{F_f} is a Jacobian matrix assembled from the gradients of F_f with respect to all input coordinates. Finally, to encourage smoothness in the estimated depth values, we apply the total variation loss $\mathcal{L}_{\text{TV}}$ on the depth map $\boldsymbol{D}$, penalizing depth differences of neighboring pixels in both the horizontal and vertical direction.

After training, we can synthesize new images by feeding any focus distance, aperture, and exposure into the pipeline. Results are presented in Fig. 4. Please refer to the *supplementary material* for additional implementation details, including network architectures and training details.

4 EVALUATION

In this section we demonstrate the advantages of our method in a series of experiments. In Sec. 4.1 we evaluate depth-from-defocus (DfD) and all-in-focus (AiF) reconstruction. In Sec. 4.2 we extensively investigate the three factors focal distance, aperture, and exposure time, which our approach for the first time allows to consider in a unified framework. We conduct ablation studies in Sec. 4.3, and give an overview of runtime performance in Sec. 4.4.

4.1 Depth-from-Defocus and All-in-Focus Reconstruction

Here, we consider the one-dimensional problem of DfD reconstruction by comparing to the state-of-the-art methods DFFMobile [Suwajanakorn et al. 2015], DFFWild [Won and Jeon 2022], DEReD [Si et al. 2023], and DFFV [Yang et al. 2022]. Among these methods, DFFMobile and DEReD are also able to reconstruct AiF images, so we consider them as baselines for this task as well. As there is no variation in exposure for both tasks, we do not use our tone mapper in this set of experiments.

To facilitate this comparison, we consider three datasets: *(a)* A novel rendered dataset, consisting of 10 scenes, generated using path tracing, providing ground-truth data for quantitative evaluation; *(b)* a synthetic dataset generated from the RGBD NYUv2 corpus [Silberman et al. 2012] in conjunction with the camera simulator from DFFWild [Won and Jeon 2022], consisting of 100 scenes; *(c)* a novel real dataset for qualitative analysis, captured using a Canon RP camera with EF 50mm/F1.8, RF 24-105mm/F4.0 and RF 85mm/F2.0 lenses, as well as a Canon 6D2 camera with an EF 24-105mm F/3.5-5.6 lens, consisting of 25 scenes. All datasets contain lens breathing. For each experiment, we select five images as input to the methods, each with a different focus distance. In the *supplementary material* we provide further details about the datasets.

Most of our competitors take into account lens breathing by choosing the input image with the smallest effective FoV as reference and performing alignment with respect to it, cropping away pixels close to the image boundaries in all other images. Our approach is markedly different in this regard and does not require any cropping. However, to facilitate meaningful comparisons, in the following we consider only pixels that are visible for all methods.

We use the mean absolute error (MAE), the mean squared error (MSE) and the absolute relative distance (Abs-Rel) as the metrics

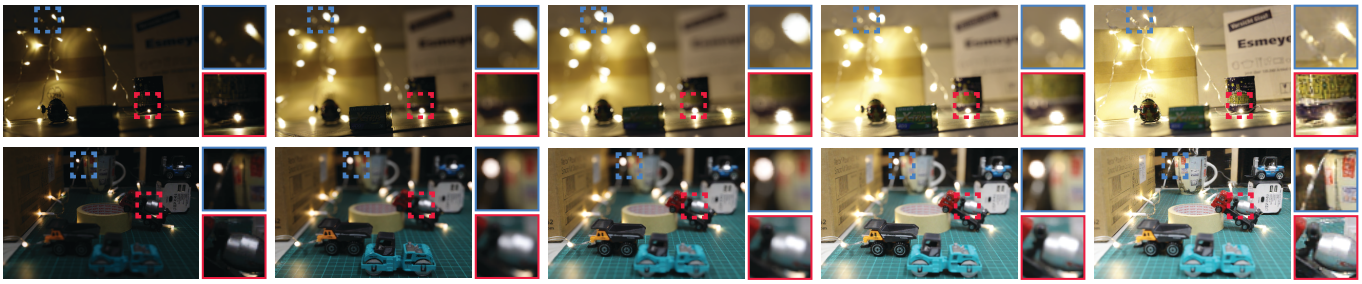


Fig. 4. Novel views with varying exposure time, aperture and focus distance synthesized by our method.

Table 1. Evaluation of DfD on the rendered dataset.

Method	MAE (↓)	MSE (↓)	Abs-Rel (↓)
DFFMobile	0.31 ± 0.14	0.13 ± 0.08	2.05 ± 2.74
DFFWild	0.30 ± 0.14	0.15 ± 0.11	0.39 ± 0.23
DEReD	0.35 ± 0.24	0.21 ± 0.23	0.80 ± 0.29
DFFV	0.37 ± 0.18	0.22 ± 0.14	0.51 ± 0.26
Ours	**0.22 ± 0.11**	**0.09 ± 0.06**	**0.33 ± 0.25**

Table 2. Evaluation of DfD on the NYU dataset.

Method	MAE (↓)	MSE (↓)	Abs_rel (↓)
DFFMobile	0.24 ± 0.13	0.10 ± 0.09	0.96 ± 0.01
DFFWild	0.24 ± 0.09	0.09 ± 0.07	0.62 ± 0.7
DEReD	0.21 ± 0.10	0.08 ± 0.07	0.83 ± 0.06
DFFV	0.33 ± 0.18	0.17 ± 0.16	0.65 ± 0.67
Ours	**0.14± 0.06**	**0.03 ± 0.02**	**0.09 ± 0.06**

Table 3. Evaluation of AiF image reconstruction on the rendered and NYU dataset.

Method	Rendered		NYU	
	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)
DFFMobile	16.8 ± 6.5	0.61 ± 0.23	31.9 ± 5.3	0.95 ± 0.07
DEReD	17.7 ± 6.4	0.63 ± 0.25	20.3 ± 3.2	0.71 ± 0.15
Ours	**21.7 ± 7.9**	**0.76 ± 0.27**	**32.6 ± 4.7**	**0.96 ± 0.04**

for DfD evaluations, which are commonly used in this field [Eigen et al. 2014; Laina et al. 2016; Li et al. 2017]. For the evaluation of AiF results, we utilize PSNR and SSIM [Wang et al. 2004] metrics.

Based on the results presented in Tables 1 and 2, our method demonstrates superior performance compared to other approaches on the DfD task. Similarly, from the results shown in Table 3, our method outperforms the baselines in the AiF task.

In Figures 8 and 9 we showcase qualitative results for DfD and AiF, confirming our numerical evaluation. We observe that DFFWild and DFFV, which require full supervision, suffer from a domain gap: While they can generate visually plausible results on the rendered dataset, their performance deteriorates significantly on other datasets. On the other hand, we see that label-free methods like DFFMobile and DEReD generally struggle to produce high-quality results across the spectrum. Regarding the AiF task, our results appear sharper and more natural compared to other methods. DFFMobile's cropping-based approach to account for lens breathing produces undesirable black boundaries, while DEReD, which does not account for lens breathing, tends to produce noticeable ghosting artifacts. For more qualitative results on both AiF and DfD tasks please refer to the *supplementary material*.

We are further interested in the performance of all methods when data is provided that does not exhibit lens breathing. To this end, we re-ran all evaluations on a variant of the NYU dataset that does not contain this effect. On this simpler dataset, our competitors significantly increase their performance – in particular DFFMobile and DFFWild. The performance of our method increases only marginally, but still outperforms all competitors, indicating robustness. We provide all details in the *supplementary material*.

4.2 Exploration of TAF Sampling Strategies

Here, we investigate different sparse sampling strategies of the TAF cube. We are interested in devising strategies that result in highest-quality reconstruction and disentanglement, providing guidelines for capturing TAF stacks.

The final photometric exposure H of an image is determined by both aperture and exposure time:

$$H \propto \frac{t}{N^2}. \tag{11}$$

Our TAF formulation allows to disentangle these factors [Jacobson et al. 2000], while previous methods often struggle with this decomposition [Huang et al. 2022; Wang et al. 2022]. In our exploration setup, we simplify Eq. 11 by switching to the log domain, resulting in the linear relationship (Fig. 5a)

$$T = 2 \cdot log_2(N) + EV, \tag{12}$$

where $EV = log_2(H)$, and $T = log_2(t)$. Based on this parameterization, we conduct a comprehensive set of experiments. We base our evaluation on five path-traced scenes. For each scene, we render $5 \times 5 \times 5 = 125$ samples on a parallelepiped in the linearized TAF space (Fig. 5b). Please refer to the *supplementary material* for details of the dataset creation.

In each of our experiments, we sample five images according to a variety of strategies and evaluate reconstruction quality. We consider 1D, 2D, and 3D sampling strategies.

As 1D strategies, we consider the three colored lines in Fig. 5b, which correspond to specific tasks previously considered in the

literature: DfD, *i.e.*, varying focal distance while fixing aperture and exposure (yellow), HDR fusion [Debevec and Malik 1997], *i.e.*, varying exposure time while fixing focal distance and aperture (blue), and varying aperture photography [Hasinoff and Kutulakos 2007], *i.e.*, varying aperture and fixing exposure time and focal distance (red).

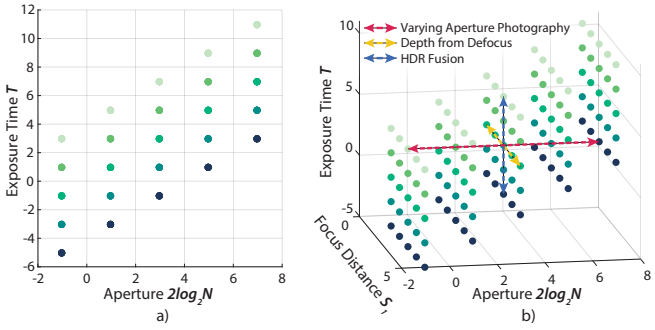


Fig. 5. a) The parallelogram visually depicts the relationship between F-number, exposure time, and photometric exposure (EV) in the logarithmic domain. Dots with the same color signify an equal exposure intensity. b) Focus distance is incorporated, resulting in the construction of a 3D parallelepiped. Previous approaches only deal with a single dimension (arrows), while we consider the entire 3D space.

In the 2D case, we manually select four distinct cross-sections. Three of these cross-sections represent scenarios where the aperture is fixed at small, medium, and large values (denoted as F_S, F_M, and F_L), while varying the exposure time and focus distance. The fourth cross-section represents a scenario where exposure time is fixed, while aperture and focus distance vary (denoted as F_V). Within each cross-section, we design four sampling patterns: two diagonals, a vertical cross, and a horizontal cross. Please refer to the *supplementary material* for more details. In the 3D case, we produce random instances of n-rooks sampling for all three dimensions.

Our main focus lies in exploring the strategies in 2D and 3D, which are in principle capable of reconstructing and disentangling all information, while the 1D results are naturally limited. We evaluate the results in terms quality for AiF HDR as well as depth map reconstruction. We use HDR-VDP-3.0.6 [Mantiuk et al. 2023] and PU21-SSIM [Azimi et al. 2021] to measure AiF HDR results. For depth map evaluation, we again employ the Abs-Rel metric. Table 4 provides the results of this analyis, while more detailed results can be found in the *supplementary material*. We observe that our method demonstrates robustness and achieves good performance under various sampling strategies. Results of our 1D evaluation are found in the *supplementary material*.

In the 2D evaluation, F_S exhibits the lowest performance in AiF HDR reconstruction, as expected due to significant defocus blur. Conversely, F_L demonstrates lower performance in depth estimation, primarily attributed to a wider depth-of-field and limited blur cues available. F_M offers a trade-off, showcasing satisfactory results in both AiF HDR reconstruction and depth estimation.

Remarkably, within the fixed exposure time cross section, F_V demonstrated competitive performance in AiF HDR reconstruction and improved performance in depth estimation. This can be attributed to the fact that, in this cross section, the variation in exposure primarily arises from changes in the aperture size. Consequently, this variation introduces both exposure variation and defocus blur, resulting in a wider range of defocus blur types and richer cues for depth estimation. Both focus distance and aperture size contribute to the strength of the defocus blur, explaining the observed performance variations.

We observe similar trends in the 3D n-rooks sampling scenario, although the performance is marginally inferior to F_V, with a notable increase in standard deviation indicating higher instability. The good quality of depth estimation in n-rooks sampling further verifies our conjecture that jointly changing the focus distance and aperture size is beneficial for accurate depth estimation.

In conclusion, for capturing we recommend using either a moderate aperture size with multiple exposure times or fixing the exposure time while changing the aperture size, along with focal sweeping. These combinations provide a good balance between capturing high-quality images and maintaining flexibility in adjusting the depth and HDR information. We also verify these conclusions on real datasets in the *supplementary material*.

While the results in this section indicate that for TAF stack *capture* it suffices to consider a 2D subspace, we want to emphasize that TAF space *synthesis* is fully 3D, and that our method for the first time enables this kind of systematic analyis.

Table 4. Explorations of different combinations.

Method	Dim.	HDR-VDP3 (↑)	PU21-SSIM (↑)	Abs-Rel (↓)
F_S	2D	9.503 ± 0.229	0.937 ± 0.031	0.405 ± 0.239
F_M	2D	9.827 ± 0.031	0.964 ± 0.007	0.335 ± 0.159
F_L	2D	9.863 ± 0.042	0.966 ± 0.011	0.555 ± 0.145
F_V	2D	9.807 ± 0.031	0.962 ± 0.007	0.294 ± 0.128
n-rooks	3D	9.805 ± 0.049	0.960 ± 0.009	0.301 ± 0.143

Table 5. Ablation Study

Method	PSNR (↑)	SSIM (↑)
Ours	**33.6 ± 3.0**	**0.95 ± 0.03**
w/o Flow	28.5 ± 3.1	0.89 ± 0.07
Polynomial Flow (k=1)	32.3 ± 3.8	0.93 ± 0.07
Polynomial Flow (k=3)	32.8 ± 3.8	0.94 ± 0.06
Gaussian Kernel	32.0 ± 2.9	0.93 ± 0.05
Explicit Tone Mapper	27.8 ± 2.0	0.89 ± 0.09
Tone Mapper w/o Weight Sharing	31.7 ± 4.1	0.92 ± 0.07
w/o VGG loss	31.5 ± 3.6	0.90 ± 0.11
w/o Jacobian loss	30.2 ± 2.5	0.91 ± 0.05
w/o TV loss	33.2 ± 2.3	0.94 ± 0.05

4.3 Ablations

In this section, we ablate selected components of our pipeline. Numerical results for all ablations are provided in Table 5, based on our real dataset. We observe that removing any component decreases the performance of our method. We elaborate on the components in the following paragraphs.

Implicit Flow Predictor. We choose to model lens breathing using a neural network. We study alternative models in Tab. 5 and Fig. 16. First, we consider removing flow prediction from our pipeline and observe severe ghosting artifacts due to the misalignment between the input images. Second, we consider a polynomial model inspired by established practices in lens distortion correction [Weng et al. 1992]. Specifically, we model radial distortion in polar coordinates, *i.e.*, as a function of radius $r = \sqrt{x^2 + y^2}$ via

$$r'(S_1) = \sum_{i=1}^{k} \alpha_i(S_1) r^i, \tag{13}$$

where r' is the remapped radius, and the coefficients α_i are 3rd-order polynomials of the focal distance S_1. During training, we optimize for the coefficients of the latter polynomials. We study the cases $k = 1$, *i.e.*, simple linear magnification, and $k = 3$, which includes non-linear distortions. We observe that neither model reaches the quality of our neural flow predictor.

Disk vs. Gaussian Kernel. Previous DfD methods [Favaro 2010; Si et al. 2023] commonly employ a Gaussian kernel to simulate defocus blur. Fig. 11 demonstrates that a disk kernel produces more realistic bokeh.

Implicit Tone Mapper. We compare our implicit tone mapper to two alternatives. First, we consider an explicit tone mapping function as per Eq. 7 in Fig. 7, which struggles to reproduce highlights. Second, we consider an alternative implicit tone mapper without weight sharing between the color channels in Fig. 12, which leeds to color shifts. Table 5 reveals that our choice of implicit tone mapping delivers the highest-quality results.

In all our experiments, we have used JPEG images as input. Alternatively, linear/RAW images could be considered, for which a way simpler tone mapper would suffice, at the expense of additional processing steps [Mildenhall et al. 2022]. We emphasize that linear images still are not HDR, therefore requiring at least a clipping operation for "tone mapping".

Loss Terms. Here we investigate the effect of our loss terms. Fig. 13 reveals that the VGG-loss is important to reproduce high-frequency details, avoiding blurry reconstructions. The Jacobian regularization helps to smooth the implicit flow and thus prevents deformation artifacts, as shown in Fig. 14. The TV loss enhances smoothness and continuity of the depth map by reducing noise and abrupt pixel transitions, illustrated in Fig. 15. Again, corresponding numerical results are provided in Table 5.

Number of input images. We investigate depth reconstruction quality as a function of the number of input images to our system, evaluated on the NYU dataset. In Fig. 6 we plot errors relative to our solution which uses five images. We see that less images give inferior results, while more images tend to only marginally improve the reconstruction. Even though these results tend to vary across scenes, they indicate that five input images are a reasonable default choice.

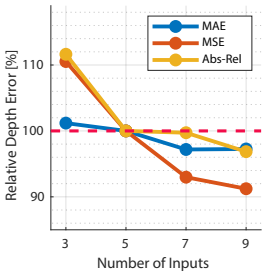


Fig. 6. Relative depth reconstruction error for different stack sizes.

4.4 Runtime

Our method runs at interactive rates, but the inference time highly depends on the aperture size during post editing, as it determines the blur kernel size. When a small aperture is used, one frame takes as little as 22 ms per frame for a resolution of 384×256 pixels, and 85 ms for a resolution of 768×512 pixels on an Nvidia RTX 3090. In the worst-case scenario, when the aperture is big and larger kernel size has to be used, our method provides results within 62 ms and 788 ms per frame for the lower and higher resolution, respectively. These results can be further improved by replacing the differentiable blur kernel with a faster approximation at inference time. Training time is substantial. It takes around 20 minutes for low resolution and 4 hours for the higher resolution on a GPU RTX 8000.

5 CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

In this work, we have proposed a novel approach utilizing implicit neural fields and a differentiable thin-lens model to represent the Time-Aperture-Focus (TAF) stack. This representation allows us to extract complete information from the stack, including depth, all-in-focus, and high-dynamic-range data. Our method achieves state-of-the-art performance in Depth-from-Defocus tasks, and allows a faithful exploration of the three-dimensional space of imaging dimensions. One notable feature of our method is the ability to perform flexible post-editing. After fitting the model, users can adjust parameters such as focus distance, aperture, and exposure, enabling effective disentanglement. Additional results can be found in the *supplementary material*.

Our method does not account for the sensitivity of the sensor (ISO), which affects exposure and noise levels. Additionally, we primarily focus on static scenes and do not address motion blur caused during image capture. Our implementation is not performance optimized, and we expect that leveraging recent advances in neural field training and inference [Müller et al. 2022] will significantly boost the performance. As our approach provides just one depth value per pixel, we are not currently able to handle transparent objects. We hope our method inspires future work on full-dimensional image recovery from sparse measurements.

ACKNOWLEDGMENTS

Ana Serrano received funding from the Spanish Agencia Estatal de Investigación (project PID2022-141539NB-I00). This work was supported by an academic gift from Meta.

REFERENCES

Maryam Azimi et al. 2021. PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR. In *2021 Picture Coding Symposium (PCS)*. IEEE, 1–5.

Sai Bangaru, Jesse Michel, Kevin Mu, Gilbert Bernstein, Tzu-Mao Li, and Jonathan Ragan-Kelley. 2021. Systematically Differentiating Parametric Discontinuities. *ACM Trans. Graph.* 40, 107 (2021), 107:1–107:17.

Odysseas Bouzos, Ioannis Andreadis, and Nikolaos Mitianoudis. 2019. Conditional random field model for robust multi-focus image fusion. *IEEE Transactions on Image Processing* 28, 11 (2019), 5636–5648.

Paul E. Debevec and Jitendra Malik. 1997. Recovering High Dynamic Range Radiance Maps from Photographs. In *Proceedings of the 24th Annual Conference on Computer*

Graphics and Interactive Techniques (SIGGRAPH '97). ACM Press/Addison-Wesley Publishing Co., USA, 369–378. https://doi.org/10.1145/258734.258884

David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems* 27 (2014).

Paolo Favaro. 2010. Recovering thin structures via nonlocal-means regularization with application to depth from defocus. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1133–1140.

Brandon Yushan Feng, Susmija Jabbireddy, and Amitabh Varshney. 2022. Viinter: View interpolation with implicit neural representations of images. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.

Herbert Gross. 2005. Handbook of Optical Systems. (2005).

Pascal Gwosdek, Sven Grewenig, Andrés Bruhn, and Joachim Weickert. 2012. Theoretical foundations of gaussian convolution by extended box filtering. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*. Springer, 447–458.

Samuel W Hasinoff and Kiriakos N Kutulakos. 2007. A layer-based restoration framework for variable-aperture photography. In *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 1–8.

Caner Hazirbas, Sebastian Georg Soyer, Maximilian Christian Staab, Laura Leal-Taixé, and Daniel Cremers. 2019. Deep depth from focus. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, 525–541.

Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. 2022. Hdr-nerf: High dynamic range neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18398–18408.

Ralph Jacobson, Sidney Ray, Geoffrey G Attridge, and Norman Axford. 2000. *Manual of Photography*. Taylor & Francis.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 694–711.

Kim Jun-Seong, Kim Yu-Ji, Moon Ye-Bin, and Tae-Hyun Oh. 2022. HDR-Plenoxels: Self-Calibrating High Dynamic Range Radiance Fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*. Springer, 384–401.

Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. 2021. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–12.

Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. 2016. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 239–248.

Jun Li, Reinhard Klein, and Angela Yao. 2017. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*. 3372–3380.

Shutao Li, Xudong Kang, and Jianwen Hu. 2013. Image fusion with guided filtering. *IEEE Transactions on Image processing* 22, 7 (2013), 2864–2875.

Xing Lin, Jinli Suo, Xun Cao, and Qionghai Dai. 2013. Iterative feedback estimation of depth and radiance from defocused images. In *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part IV 11*. Springer, 95–109.

Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. 2017. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion* 36 (2017), 191–207.

Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. 2020. Single-image HDR reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1651–1660.

Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V Sander. 2022. Deblur-nerf: Neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12861–12870.

David Mandl, Peter M Roth, Tobias Langlotz, Christoph Ebner, Shohei Mori, Stefanie Zollmann, Peter Mohr, and Denis Kalkofen. 2021. Neural cameras: Learning camera characteristics for coherent mixed reality rendering. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 508–516.

Rafal K Mantiuk, Dounia Hammou, and Param Hanji. 2023. HDR-VDP-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content. *arXiv preprint arXiv:2304.13625* (2023).

Maxim Maximov, Kevin Galim, and Laura Leal-Taixé. 2020. Focus on defocus: bridging the synthetic to real domain gap for depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1071–1080.

Tom Mertens, Jan Kautz, and Frank Van Reeth. 2007. Exposure fusion. In *15th Pacific Conference on Computer Graphics and Applications (PG'07)*. IEEE, 382–390.

Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. 2022. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16190–16199.

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

Michael Moeller, Martin Benning, Carola Schönlieb, and Daniel Cremers. 2015. Variational depth from focus reconstruction. *IEEE Transactions on Image Processing* 24, 12 (2015), 5369–5378.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.* 41, 4, Article 102 (July 2022), 15 pages.

Seonghyeon Nam, Marcus A Brubaker, and Michael S Brown. 2022. Neural image representations for multi-image fusion and layer separation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*. Springer, 216–232.

Michael Potmesil and Indranil Chakravarty. 1981. A lens and aperture camera model for synthetic image generation. *ACM SIGGRAPH Computer Graphics* 15, 3 (1981), 297–305.

Haozhe Si, Bin Zhao, Dong Wang, Yupeng Gao, Mulin Chen, Zhigang Wang, and Xuelong Li. 2023. Fully Self-Supervised Depth Estimation from Defocus Clue. *arXiv preprint arXiv:2303.10752* (2023).

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. *ECCV (5)* 7576 (2012), 746–760.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. 2020. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems* 33 (2020), 7462–7473.

Supasorn Suwajanakorn, Carlos Hernandez, and Steven M Seitz. 2015. Depth from focus with your mobile phone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3497–3506.

Chao Wang, Ana Serrano, Xingang Pan, Bin Chen, Hans-Peter Seidel, Christian Theobalt, Karol Myszkowski, and Thomas Leimkuehler. 2022. GlowGAN: Unsupervised Learning of HDR Images from LDR Images in the Wild. *arXiv preprint arXiv:2211.12352* (2022).

Ning-Hsu Wang, Ren Wang, Yu-Lun Liu, Yu-Hao Huang, Yu-Lin Chang, Chia-Ping Chen, and Kevin Jou. 2021. Bridging unsupervised and supervised depth from focus via all-in-focus supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12621–12631.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

Juyang Weng, Paul Cohen, Marc Herniou, et al. 1992. Camera calibration with distortion models and accuracy evaluation. *IEEE Transactions on pattern analysis and machine intelligence* 14, 10 (1992), 965–980.

Changyeon Won and Hae-Gon Jeon. 2022. Learning Depth from Focus in the Wild. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*. Springer, 1–18.

Zijin Wu, Xingyi Li, Juewen Peng, Hao Lu, Zhiguo Cao, and Weicai Zhong. 2022. DoF-NeRF: Depth-of-Field Meets Neural Radiance Fields. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1718–1729.

Fengting Yang, Xiaolei Huang, and Zihan Zhou. 2022. Deep Depth from Focus with Differential Focus Volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12642–12651.

Qiang Zhang and Martin D Levine. 2016. Robust multi-focus image fusion using multitask sparse representation and spatial context. *IEEE Transactions on Image Processing* 25, 5 (2016), 2045–2058.

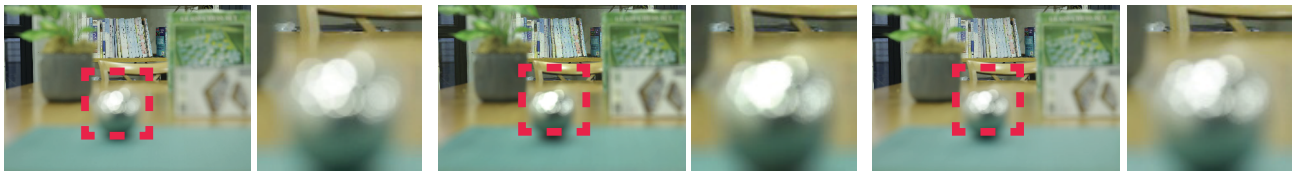


Fig. 7. Comparing LDR outputs of an explicit tone mapper to our implicit approach. The explicit solution struggles with the accurate reproduction of out-of-focus highlights.

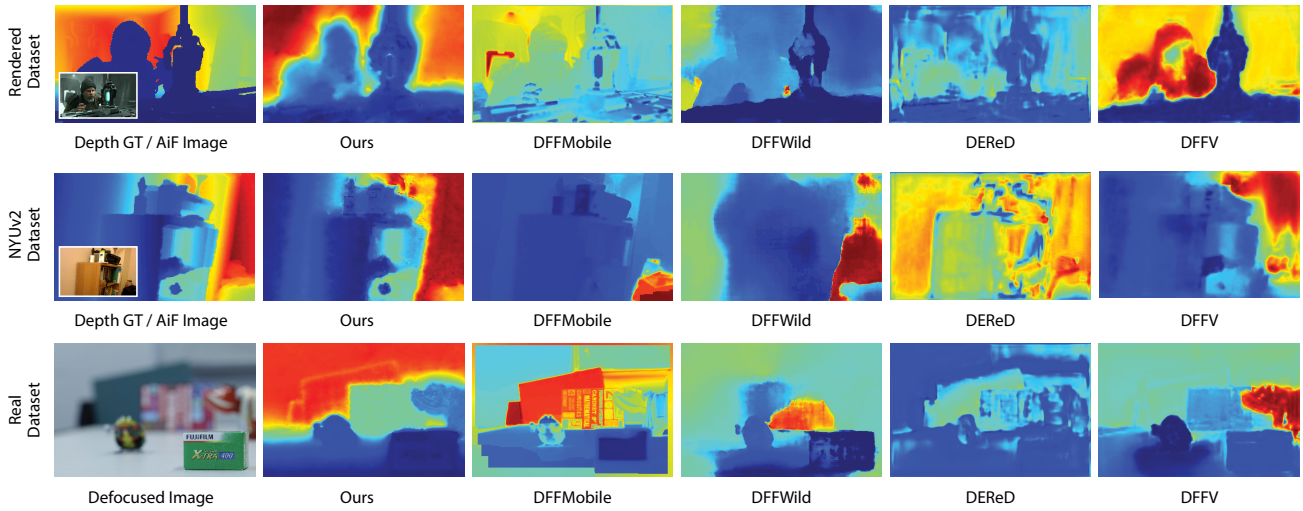


Fig. 8. Depth reconstruction results for the three datasets used in our evaluation. More results can be found in the *supplementary material*.

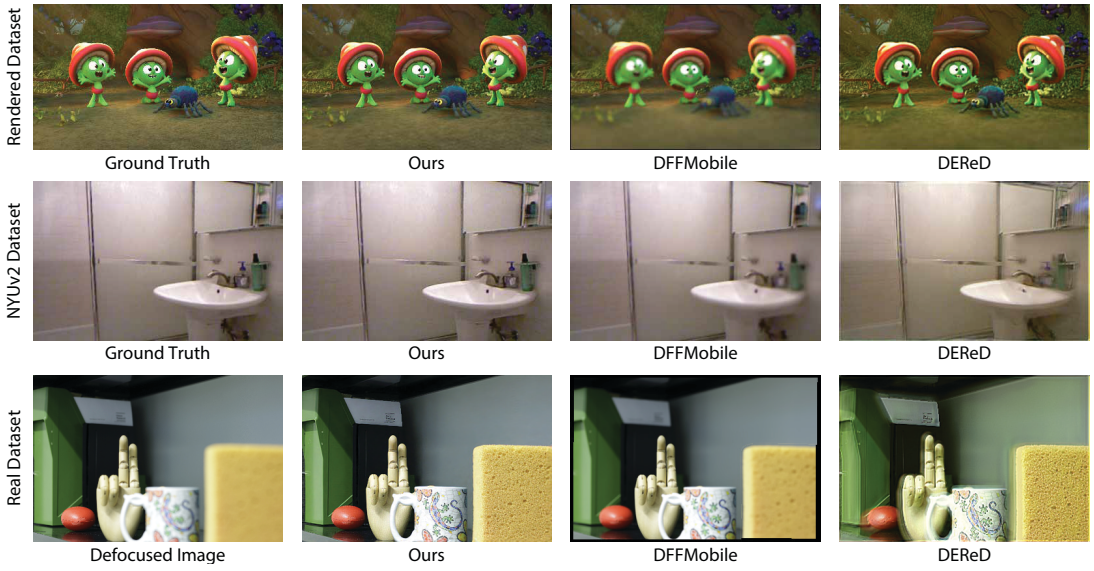


Fig. 9. All-in-focus image reconstruction for the three datasets used in our evaluation. Our method successfully recovers the AiF image, while other methods introduce a variety of artifatcs: DFFMobile struggles with resolving the blur and also produces undesirable black boundaries on the Real Dataset. DEReD produces results with visible hue shifts and ghosting artifacts. More results can be found in the *supplementary material*.

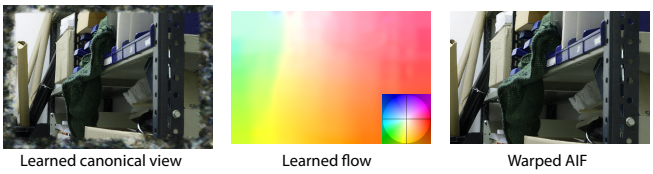


Fig. 10. Visualization of our learned flow. We warp the learned canonical view to images corresponding to different focal distances, simulating lens breathing.

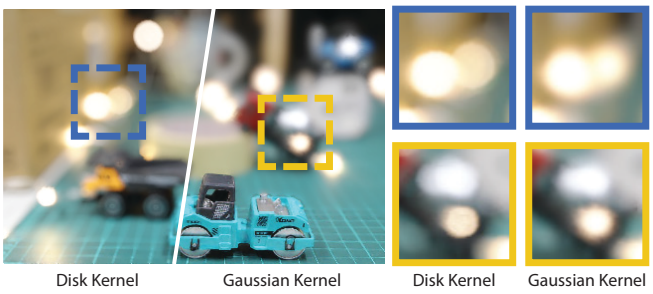


Fig. 11. Comparisons of bokeh generation between a disk and a Gaussian kernel for focus editing. The disk kernel demonstrates a more natural bokeh.

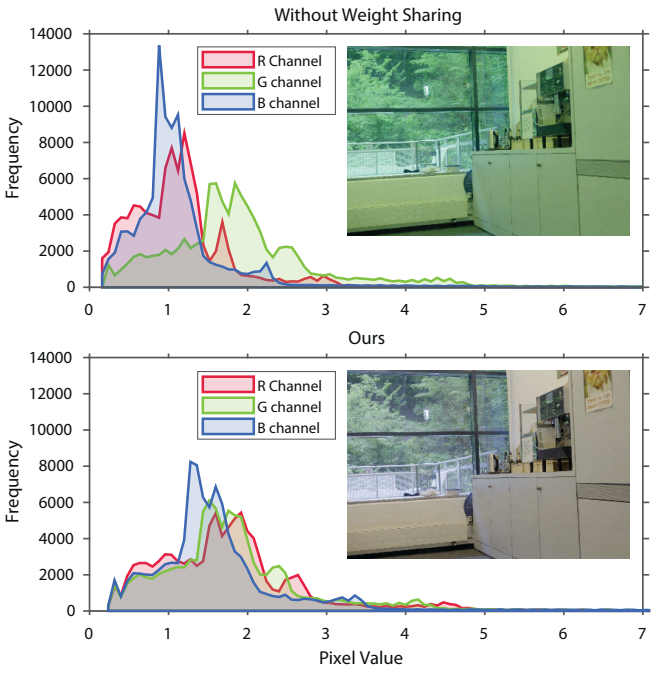


Fig. 12. Comparisons of reconstructed HDR using a tone mapper without and with (Ours) weight sharing, revealing that weight sharing is essential in preventing hue shifts.

Fig. 13. Comparisons of the reconstructed HDR without and with the VGG loss (Ours). Our method synthesizes sharper details.

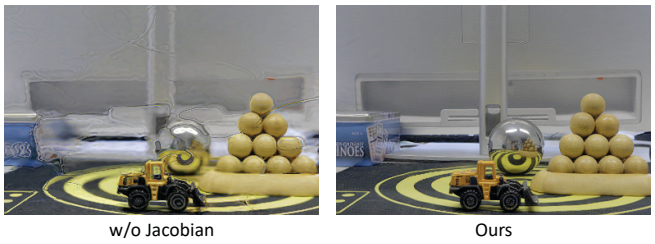


Fig. 14. Comparisons of the reconstructed HDR without and with the Jacobian regularization (Ours). The introduction of Jacobian regularization allows the method to recover artifact-free images.

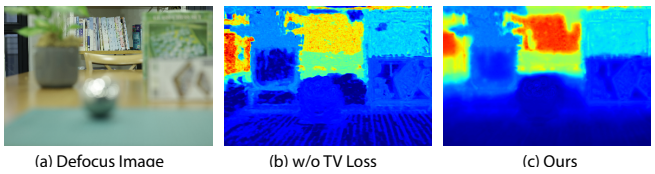


Fig. 15. Comparisons of estimated depth map with (Ours) and without the TV loss. This loss allows the method to reconstruct smoother depth maps and reduces local fluctuations caused by textures of the captured objects.

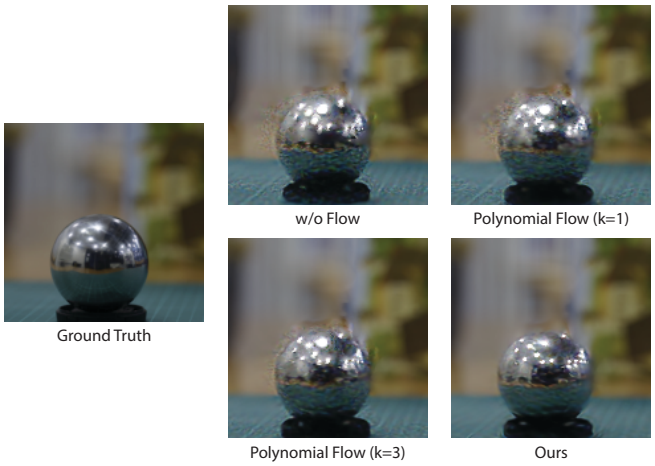


Fig. 16. Comparison of novel-view image quality for different flow predictors. Only our implicit flow-based approach gets rid of ghosting artifacts arising from misalignment caused by lens breathing.