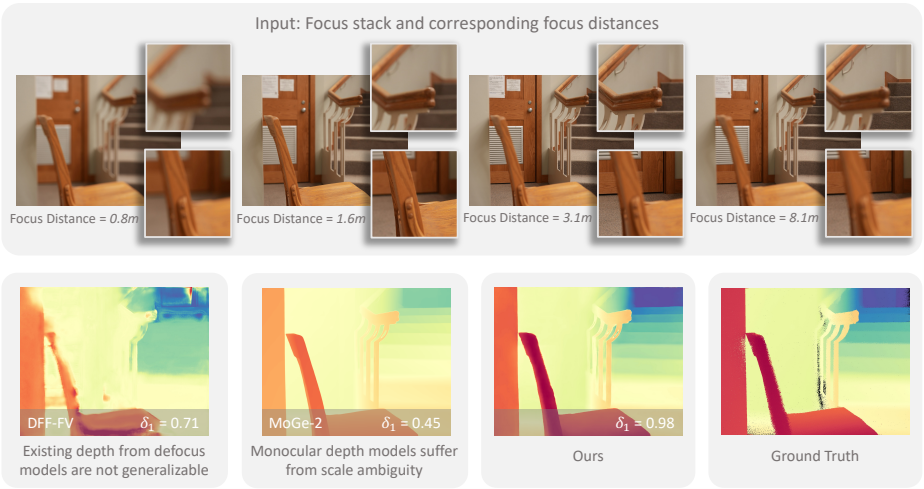
Zero-Shot Depth from Defocus

Yiming Zuo*, Hongyu Wen*, Venkat Subramanian*, Patrick Chen,
Karhan Kayan, Mario Bijelic, Felix Heide, and Jia Deng

Department of Computer Science, Princeton University
{zuoym,hongyu.wen,venkat.subra,patrickchen,
karhan,mario.bijelic,fheide,jiadeng}@princeton.edu

Abstract. Depth from Defocus (DfD) is the task of estimating a dense metric depth map from a focus stack. Unlike previous works overfitting to a certain dataset, this paper focuses on the challenging and practical setting of zero-shot generalization. We first propose a new real-world DfD benchmark ZEDD, which contains 8.3× more scenes and significantly higher quality images and ground-truth depth maps compared to previous benchmarks. We also design a novel network architecture named FOSSA. FOSSA is a Transformer-based architecture with novel designs tailored to the DfD task. The key contribution is a stack attention layer with a focus distance embedding, allowing efficient information exchange across the focus stack. Finally, we develop a new training data pipeline allowing us to utilize existing large-scale RGBD datasets to generate synthetic focus stacks. Experiment results on ZEDD and other benchmarks show a significant improvement over the baselines, reducing errors by up to 55.7%. The ZEDD benchmark is released at `https://zedd.cs.princeton.edu`. The code and checkpoints are released at `https://github.com/princeton-vl/FOSSA`.



*Equal contribution.

1 Introduction

In modern photography, a technique called focus bracketing is widely adopted. By sweeping the focal plane from near to far, the camera captures a set of images, also known as a *focus stack*. While focus bracketing is traditionally used to create an all-in-focus image, it contains rich information about the scene geometry and depth relationships: as the focal plane sweeps through the scene, objects at different depths sequentially come into focus while others become blurred.

In this paper, we address the task of Depth from Defocus (DfD), which aims at estimating a dense metric depth map from a focus stack. The depth map is useful for various downstream applications, such as post-capture defocus control [22, 25], novel view synthesis [18, 21], relighting [46], etc.

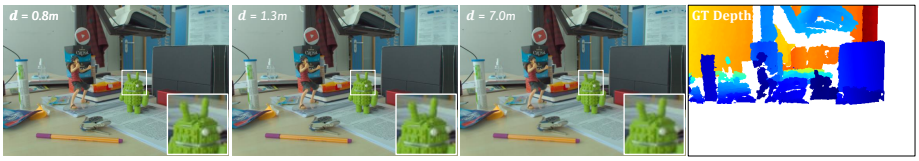
While various previous works have tackled the DfD task over the years [12–14,16,20,37,41,44,45,48], they primarily focus on the in-domain setting. As a result, these models do not generalize well to unseen domains, which hinders their application in real-world scenarios. In contrast, we tackle the challenging problem of *zero-shot DfD*, where the models are evaluated on datasets and domains unseen during training. We address two main problems: *1. Lack of high-quality DfD benchmarks* and *2. DfD models not generalizing well*.

Lack of High Quality Benchmarks. Previous works [12, 34, 44, 45] evaluate their models with synthetically generated focus stacks using a 2D point spread functions (PSF) on RGBD datasets such as NYU [23]. However, this 2D processing method only approximates defocus and cannot fully reproduce real optical blur, which depends on 3D scene geometry. DDFF [14] is the only dataset with real focus stacks and dense depth ground-truth, but has several limitations. Firstly, the focus stack is composed from a light-field image, with the aperture being so small that the defocus is barely noticeable. Secondly, the depth ground-truth is collected from a structured light sensor, which is low-resolution, noisy, and measures only up to 3.5m (Fig. 1).

In this paper, we propose a new benchmark **ZEDD** (**ZE**ro-shot **D**epth from **D**efocus), a high-quality and diverse real-world benchmark for DfD. We collect images with a 4K resolution DSLR camera under multiple apertures, up to F/1.4. We design a pipeline for accurate calibration and control of the focus distances through remote control of the lens motor position. As for the depth ground-truth, we use a high-end Ouster Lidar and accumulate points through time, achieving high accuracy and dense spatial coverage. Finally, our dataset has high diversity: we collect 100 unique scenes, an order of magnitude greater than previous datasets. Detailed comparisons can be found in Tab. 1.

DfD Models Not Generalizing Well. Previous works [12, 44, 45] follow the paradigm of *in-domain* evaluation, *i.e.*, the models are trained and tested on the same datasets, often with limited scales (*e.g.*, both DDFF [14] and NYU [23] have $<$ 1000 samples). Moreover, their network designs lack sufficient capacity and scalability, preventing them from fully leveraging large-scale training.

We propose a modern network architecture **FOSSA** (**FO**cu**S** **S**tack **A**ttention Transformer) for zero-shot DfD. It is based on the powerful Vision Transformer (ViT) [8] but with novel designs tailored to the DfD task. We propose a *stack*



(a) DDFF Dataset

(b) ZEDD (Ours), F/2.8

Fig. 1: Qualitative comparison between ZEDD and DDFF [14]. DDFF uses a very small aperture, so the defocus effect is barely noticeable even when zoomed in. In contrast, our focus stacks exhibit clear, smooth defocus effects. Our depth ground-truth is of higher resolution and denser, and it contains no missing regions due to occlusions.

Datasets	# Scenes	Ground Truth	Image Resolution	Defocus Effect	Aperture	Focus Distances	# Images in Focus Stack
MobileDepth [35]	11	No GT	640 × 360	Real	–	–	10
DDFF [14]	12	Structured Light	552 × 383	Synthetic (Light Field)	F15.0*	0.5 - 7.0m	10
ZEDD (Ours)	100	Lidar	4104 × 2736	Real	F1.4/2.0/2.8/ 4.0/5.6/16.0	0.8 - 8.1m	9

Table 1: Compared to existing DfD benchmarks MobileDepth [35] and DDFF [14], our ZEDD has 8.3× more scenes, higher quality ground-truth, higher resolution images, and realistic defocus effect under multiple f-numbers. *: full-frame equivalent aperture computed from specifications of the Lytro camera.

attention layer with a special focus distance embedding, which enables efficient extraction of the defocus cues across the focus stack. We then fuse the individual focus stack features into a global feature map at an early stage, effectively reducing the computation cost of the subsequent layers.

Due to the lack of large-scale focus stack datasets, we train fully on synthetically generated focus stacks, allowing us to utilize any existing RGBD datasets. To achieve better generalization ability, we randomize parameters such as focus distance, aperture size, and the shape of the point spread function (PSF).

We evaluate FOSSA extensively on diverse benchmarks, including ZEDD, DDFF [14], and multiple real-world RGBD datasets. FOSSA significantly outperforms the baselines. It reduces AbsRel by 55.7% compared to the second-best method DepthPro [2] on ZEDD, and MSE by 40.4% on DDFF compared to previous state-of-the-art DualFocus [44]. We also demonstrate that FOSSA is robust to various factors, including the aperture size and the focus stack size.

2 Related Works

2.1 Depth from Defocus Methods

Recent deep-learning-based methods focus on advancing network designs. DefocusNet [20] processes each image with individual branches and fuses information

through repetitive pooling. DFV [45] proposes a network based on differential focus volume, which explicitly compares the features of neighboring frames in a focus stack. HybridDepth [12] proposes a multi-stage framework that first aligns relative monocular depth to metric DfD depth, and then performs refinements. DualFocus [44] introduces a variational approach that predicts gradient constraints and integrates them into depth maps with a least-squares solver. DEReD [34] explores the fully self-supervised setting, trained through a differentiable layer for synthesized defocus effect and photometric reconstruction. Unlike DDFS [10], which carefully models and embeds full camera settings, we only embed focus distances as metric anchors, but maintain robustness to focal length and aperture by training the model on diverse combinations.

Compared to previous works, our method has two unique benefits. *1. Simple and Scalable:* previous works are built heavily on custom network layers and designs, making it hard for them to scale up. Comparatively, FOSSA is simple yet effective: it is based on Vision Transformer (ViT) and is built with standard transformer blocks, which is easy to scale up and benefit from the powerful pretrained models such as DepthAnything [47]. *2. Generalizable:* While previous works achieve good accuracy on dedicated benchmarks, they do not generalize well beyond certain datasets, and require separate data and training for each benchmark. In comparison, FOSSA has a single set of weights that works well across all datasets in a zero-shot manner, thanks to our scalable model design (Sec. 3.2) and data pipeline (Sec. 3.3) aiming for maximum generalization.

2.2 Depth from Defocus Datasets

Defocus effects are common in real photos. Multiple datasets have been collected for depth estimation on those images. iDFD [24], D3Net [4], and Talegaonkar *et al*. [36] collect scenes with paired wide and narrow aperture images, but at the same focus distance, therefore not suitable for the DfD task.

Several real-world datasets have been collected specifically for DfD, but they are all limited in their scale and quality. The MobileDepth dataset [35] has only 11 scenes captured with a mobile phone, with no depth ground-truth available for quantitative evaluation. DDFF [14] is composed of 720 focus stacks from only 12 scenes, therefore has high content overlap among focus stacks. Moreover, the focus stacks are not realistic: the images are not from a regular lens but synthesized from a light-field camera. Finally, the ground-truth is collected from a structured light sensor, which has limited accuracy, FoV, and range coverage. Compared to DDFF [14], our dataset has higher diversity (100 distinct scenes), higher quality defocus images (4K resolution with multiple apertures up to F/1.4), and more accurate ground-truth (from a high-end Ouster Lidar).

Others use graphics engines to render synthetic focus stack images. FoD500 [20] creates toy-like scenes in Blender with randomly scattered objects, and renders defocused images at 256×256 resolution. We propose a new synthetic benchmark (Sec. 5.2) based on Infinigen [28, 29] adopting a similar rendering strategy, but adopt procedural scene generation for systematic layout control and achieve more realistic scene structures, higher photorealism, and higher resolution.

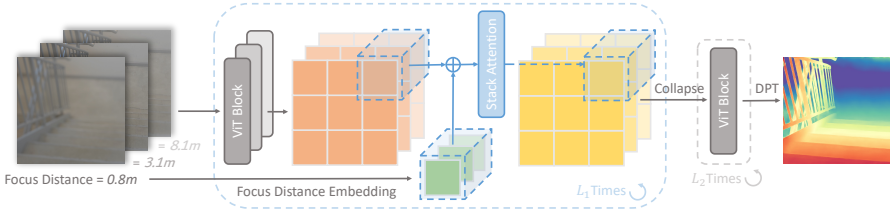


Fig. 2: Overview of the FOSSA pipeline. FOSSA consists of two main stages: focus stack feature extraction (blue box), and feature refinement (gray box). For each focus stack extraction layer, the image features are first processed individually and then pass through a stack attention layer for efficient information exchange across the stack. The features are then collapsed along the stack dimension, and pass through a sequence of refinement ViT blocks and finally a DPT head for dense depth output.

3 Methods

3.1 Problem Definition

The input to the task of Depth from Defocus (DfD) is a focus stack and the corresponding focus metadata. Specifically, the focus stack is a set of M pixel-aligned RGB images $\mathbf{I} = (\mathbf{I}_1, \ldots, \mathbf{I}_M) \in \mathbb{R}^{M\times 3\times H\times W}$, captured from a static scene using a fixed viewpoint and the same aperture, but with different focus distances. The focus metadata is the focus distances $\mathbf{d} = (d_1, \ldots, d_M) \in \mathbb{R}^M$ in meters. The goal is to predict a dense metric depth map $\hat{\mathbf{D}} \in \mathbb{R}^{H\times W}$.

3.2 FOSSA Network Architecture

As shown in Fig. 2, FOSSA consists of two stages. It first extracts a per-image feature map for all images in the focus stack, then collapses them into a single global representation for depth prediction. In the first stage, the network produces a feature map for each image in the focus stack using a feature extractor containing L_1 number of layers. Each layer contains (i) a weight-shared Vision Transformer (ViT) [8] block applied to each focus image individually and (ii) a novel stack attention layer that enables information exchange across the focus stack dimension. In the second stage, we collapse the per-focus feature maps into one global feature map, which is further refined by additional L_2 ViT blocks and finally fed to a DPT [30] head to regress a dense metric depth map.

ViT Blocks. We adopt standard ViT [8] blocks as the backbone for feature encoding. ViT has become a strong backbone for depth estimation and is widely adopted in prior work [6, 40, 47]. A ViT block contains multi-head self-attention followed by a feed-forward MLP, with residual connections and normalization.

Focus Stack Feature Extraction. For each image $\mathbf{I}_i \in \mathbb{R}^{3\times H\times W}$ in the stack, we partition it into $p \times p$ patches and embed each patch as a token. This yields a patch-level feature map $\mathbf{F}_i \in \mathbb{R}^{C\times(\frac{H}{p}\times\frac{W}{p})}$, where C is the feature dimension.

A naïve way is to process each $\mathbf{F}_i$ individually using a sequence of ViT blocks to obtain an image feature map, as in monocular depth estimation. In the DfD

task setting, however, the key geometric signal is not contained in any single image, but across the focus stack. Consider a certain patch in the focus stack: as the focus distance sweeps from near to far, the patch turns from blurred to sharp when the focus distance approaches its depth, and becomes blurred again as the focus distance further increases. The visual variation, when interpreted together with the focus distances $\mathbf{d}$, provides a strong cue for metric depth. Exploiting this structure requires explicit information exchange across the focus stack.

To this end, we introduce a stack attention layer and insert it between each ViT block. For each image $\mathbf{I}_i$, we embed its focus distance d_i using a two-layer MLP to obtain a C-dimensional vector, and add it to the image tokens. We then concatenate the feature maps across the focus stack to form a 4D feature map $\mathbf{F}' \in \mathbb{R}^{M \times C \times (\frac{H}{p} \times \frac{W}{p})}$ and apply attention along the first (stack) dimension. For each patch location, the module performs self-attention across the M focus settings, enabling the network to aggregate defocus cues across the stack.

This design is computationally efficient: the quadratic attention cost is incurred only over the stack size M, which is much smaller than the number of spatial tokens, providing effective and efficient cross-image interaction.

Feature Collapse and Refinement. After extracting M per-focus feature maps, we collapse them into one global feature map by averaging the features over the focus stack dimension. The global feature has the same shape as a regular image feature ($\mathbb{R}^{C \times (\frac{H}{p} \times \frac{W}{p})}$), and is then refined by subsequent L_2 ViT blocks and a standard DPT decoder to produce final depth prediction $\hat{\mathbf{D}}$.

3.3 Training

The main challenge for training a zero-shot DfD model is that there are no large-scale DfD datasets that can support training. Therefore, we train our model on existing RGBD datasets using focus stacks generated with synthetic blur.

The strength of blur depends on multiple factors, including the focal length, aperture, focus distance, and object distance. This strength is often described using the circle of confusion (CoC). Given an image $\mathbf{I}$ with ground truth depth $\mathbf{D}$, the CoC is calculated as:

$$\text{CoC} = \frac{|\mathbf{D} - d|}{\mathbf{D}} \frac{f^2}{N(d - f)}, \tag{1}$$

where d is the focus distance, f is the focal length, and N is the f-number. Based on CoC, we compute a point spread function (PSF) as a pixel-wise blur kernel, which is applied to $\mathbf{I}$ through convolution. Unlike prior work that assumes a Gaussian PSF [13,34], we randomize the PSF shape to better reflect real imaging physics, which varies between diffraction-limited and defocus-dominated regimes. We also randomize the f-number and the focus distance distribution for better generalization. More details can be found in Appendix Sec. A4.

The loss is a combination of the SiLog loss [9] with $\lambda = 0.5$ to balance metric and relative supervisions, and the Gradient Matching loss [31]:

$$L(\hat{\mathbf{D}}, \mathbf{D}) = \text{SiLog}(\hat{\mathbf{D}}, \mathbf{D}) + 0.1 \cdot \text{GradMatching}(\hat{\mathbf{D}}, \mathbf{D}). \tag{2}$$

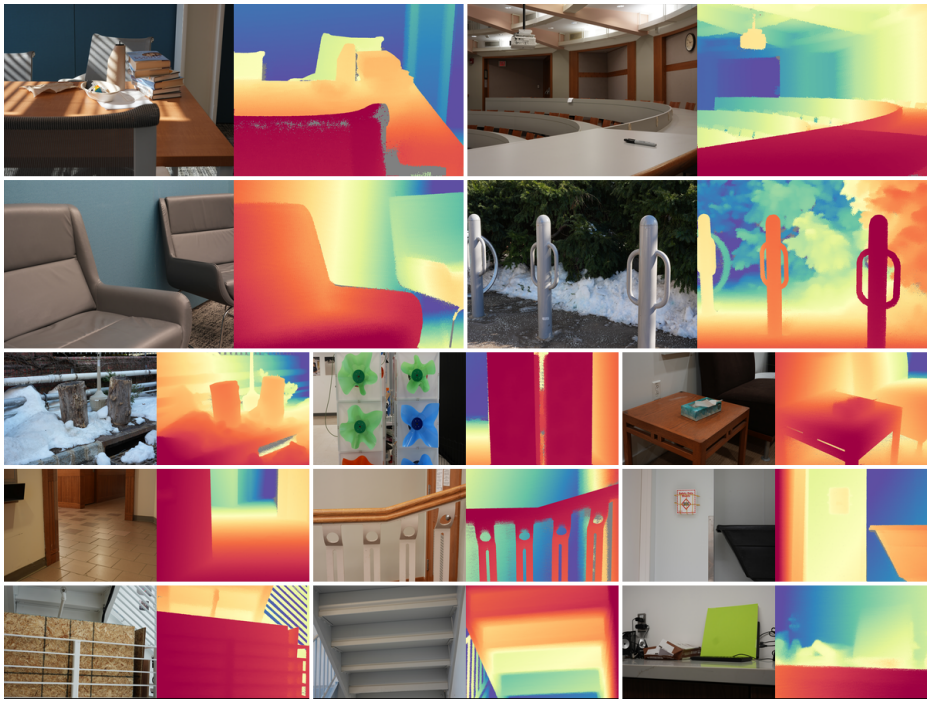


Fig. 3: A gallery of our ZEDD benchmark. ZEDD includes 100 diverse scenes spanning a wide range of indoor and outdoor environments. The dataset features rich geometric structure and provides high-quality ground-truth.

4 ZEDD: Real-World DfD Benchmark

4.1 Scene Coverage and Statistics

We captured 100 focus stacks in 100 unique scenes, covering various indoor and outdoor locations, such as classrooms, hallways, robotics labs, offices, kitchens, and gardens, providing a diverse scene coverage.

For each focus stack, we capture images at 9 focus distances, ranging from 0.82 to 8.10m. We capture at 5 larger apertures (F1.4/2.0/2.8/4.0/5.6), and a small aperture (F16) for all-in-focus images, resulting in $6\times9 = 54$ images in total for each scene. This rich combination of focus distances and apertures allows us to study the sensitivity of the models' performance to each factor (Sec. 5.4).

We provide a dense ground-truth depth map for each scene under the resolution of 1824×1216, captured with a high-accuracy Lidar.

4.2 Focus Stack Capture Procedure

Hardware. We capture images using a DSLR camera (Sony α1-II) together with a prime lens with a large and variable aperture (Sony FE 50mm F1.4 GM).

This combination allows us to capture high-resolution 4K images with multiple defocus strengths.

Intrinsics Calibration. Our goal is to get accurate camera intrinsics. However, the FoV changes as we adjust the focus distance, a phenomenon known as lens breathing. To resolve this, we define a canonical image space, where we set the lens focus distance to be 3.08m. The intrinsics for the canonical space is calibrated using Kalibr [11], and images captured under all other focus distances are warped into the canonical space using an open-source focus stacking toolbox [1].

Focus Distance Calibration. The autofocus lens we use allows us to read out the focus motor positions as a 4-digit hexcode. The hexcode is a 1-to-1 mapping to the actual focus distance, but this mapping is lens-specific and unknown. Therefore, we recover this mapping as a sparse but accurate lookup table through calibration. Specifically, we place a calibration board parallel to the camera, manually adjust the motor until the board is sharply in focus, and capture an image (I_0). We then fix the motor and record the hexcode, and move the board around to capture multiple images ($I_{1...N}$). We finally co-calibrate the intrinsics and extrinsics for all images ($I_{0...N}$) using OpenCV [3]. The z component of the extrinsics of I_0 is used as the ground-truth focus distance.

Software Control. We capture the focus stack through software control of the camera, providing two main benefits: 1. The software allows us to set the focus motor precisely at the hexcode we have calibrated, making the image-to-focus-distance mapping accurately reproducible. 2. This eliminates possible mechanical vibrations during the capture, so the images in the stack are perfectly aligned.

4.3 Depth Ground-Truth Acquisition

Hardware. We use the Ouster OS0-128 Lidar, which provides a range of up to 100m with a sub-centimeter range accuracy. We select this Lidar for its longer ranges and higher accuracy compared to other commonly used depth sensors such as Intel RealSense.

Point Accumulation. OS0-128 has a fixed sparse scanning pattern with 128 distributed layers. To achieve a dense depth coverage, we handhold the lidar to move it and accumulate the points for about 600 frames. The frames are registered using an ICP algorithm [33] with Open3D [50] with noise reduction. The accumulated point cloud is denoted as $\mathbf{P}_{world}$.

Registration. Once a dense ground truth per scene $\mathbf{P}_{world}$ was captured, we aim to transfer it to the camera frame. To achieve this, we need to estimate the projection matrix ${}^{cam}\mathbf{T}_{world}$ from the world coordinate to the camera coordinate. The camera and the Lidar are rigidly mounted on a tripod, related by a transform ${}^{cam}\mathbf{T}_{lidar} \in \mathrm{SE}(3)$, which is calibrated separately offline. Therefore, it suffices to localize the Lidar on the tripod in the world coordinate frame. To this end, we capture at every camera location a single Lidar point cloud $\mathbf{P}_{lidar}$ and solve for ${}^{lidar}\mathbf{T}_{world}$ using ICP between $\mathbf{P}_{lidar}$ and $\mathbf{P}_{world}$. Finally, the camera-space point cloud is computed as $\mathbf{P}_{cam} = {}^{cam}\mathbf{T}_{lidar}\,{}^{lidar}\mathbf{T}_{world}\mathbf{P}_{world}$.

Projection. We project $\mathbf{P}_{cam}$ into a depth map with z-buffering [5] using the calibrated intrinsics. Note, the projected pointcloud $\mathbf{P}_{cam}$ is dense (up to 1080p

resolution) and does not have missing regions due to the parallax effect problem as in many other RGBD datasets [14, 27]. We finally do manual quality checks and depth map clean-ups, as detailed in the Appendix Sec. A1.2.

5 Experiments

5.1 Implementation Details

We implement FOSSA with 2 different sizes (ViT-S, ViT-B). We use $L_1 = 4$ focus stack feature extraction layers and $L_2 = 8$ refinement ViT layers. We initialize the ViT blocks from the pretrained DepthAnythingv2 [47] indoor metric checkpoint, and zero-initialize the MLP in the stack attention layers. We use the CSTM-label [2,15,49] normalization using ground-truth focal length and predict in the normalized depth space. We provide the ground-truth focal length to both our method and baselines [2, 26, 42] for fair comparison.

We train FOSSA on a combined dataset of Hypersim [32] (66k samples) for indoor coverage and TartanAir [43] (307k samples) for outdoor coverage, using synthetically generated focus stacks. We train the model for 40 epochs using an exponential learning rate scheduler, with a batch size of 8. Training takes 2 days on 4 $\times$ L40 GPUs (48GB memory).

We train with a stack size of 5 images. We randomize the near and far bounds of the focus distances, as well as how the other focus distances are interpolated in between. We also randomize the blur kernel by altering the PSF shape and the f-number. See more details in Appendix Sec. A4.

5.2 Datasets and Evaluation Metrics

ZEDD. Our real-world dataset is the main evaluation benchmark. We randomly divide it into a validation and a test split, with 50 samples in each split. Out of the 9 focus distances (FD), we subsample with stride 2 and create a focus stack of 5 images, with FD=$\{0.8, 1.7, 3.1, 4.9, 8.1\}m$, and aperture=F/2.8.

Infinigen Defocus. We create a synthetic benchmark based on Infinigen Indoors [29], and we render the focus stacks with defocus effect created using the ray tracer (Cycles) in Blender [7]. This benchmark is a digital counterpart of the real dataset, which is less photorealistic but has perfect ground-truth. It contains 200 scenes in total, with FD=$\{0.8, 1.7, 3.0, 4.7, 8.0\}m$, and aperture=F/1.4.

Real-World RGBD Datasets. We further evaluate on 3 real-world RGBD datasets for broader scene coverage of indoors (iBims [19] and DIODE [38]), outdoors (DIODE [38]), and transparent objects (HAMMER [17]). We simulate the focus stack from the ground-truth depth map as described in Sec. 3.3. We sample the focus distances depending on the scene-specific depth statistics, to mimic the real-world scenario where the photographer chooses areas to focus on. We set the aperture to F/1.4. See more details in Appendix Sec. A5.

DDFF [14]. Although the DDFF [14] dataset has several limitations, we still report the numbers for an easy and thorough comparison with DfD baselines. The data split and evaluation metrics follow previous works [12, 20, 44, 45].

Methods	ZEDD (Test Split)						Infinigen Defocus		
	$\delta_{1.05}\uparrow$	$\delta_{1.25}\uparrow$	AbsRel $\downarrow$	MAE $\downarrow$	MSE $\downarrow$	RMSE $\downarrow$	$\delta_{1.05}\uparrow$	$\delta_{1.25}\uparrow$	AbsRel $\downarrow$
Monocular Depth									
DAv2 [47]	0.037	0.261	0.337	1.211	4.295	1.353	0.026	0.133	0.429
DepthPro [2]	0.182	0.665	0.201	0.684	2.020	0.812	0.174	0.692	0.176
UniDepthv2 [26]	0.106	0.605	0.253	0.866	2.439	0.976	0.103	0.464	0.370
MoGe-2 [42]	0.149	0.580	0.278	0.775	1.768	0.908	0.175	0.621	0.233
Depth from Defocus									
DFF-FV [45]	0.153	0.576	0.369	0.868	2.898	1.270	0.045	0.172	0.629
DFF-DFV [45]	0.152	0.546	0.573	1.122	4.515	1.619	0.053	0.205	0.568
DEReD [34]	0.128	0.505	0.347	1.056	4.077	1.368	0.108	0.446	0.355
HybridDepth [12]	0.075	0.347	0.754	1.582	6.024	2.060	0.037	0.157	1.232
Ours (ViT-S)	0.451	0.884	0.098	**0.370**	**0.661**	**0.534**	**0.520**	**0.940**	**0.085**
Ours (ViT-B)	**0.505**	**0.918**	**0.089**	0.371	0.988	0.543	0.420	0.936	0.091

Table 2: Experiment results on ZEDD and Infinigen Defocus datasets. All methods are tested zero-shot on these benchmarks. Our method FOSSA outperforms all monocular depth baselines and DfD baselines by a large margin.

Evaluation Metrics. We use standard metrics in monocular depth estimation: absolute relative error (AbsRel) and the inlier percentage δ_k, defined as follows:

$$\text{AbsRel} := \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{d}_i - d_i|}{d_i}, \quad \delta_k := \frac{1}{N}\sum_{i=1}^{N}\left[\max(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}) < k\right]$$

where N is the number of pixels, d_i is the ground-truth, and $\hat{d}_i$ is the prediction. Note $\delta_{1.25}$ is also commonly denoted as δ_1 in the literature. We also report standard depth metrics MAE, MSE, and RMSE on ZEDD. See previous literature [51] for detailed metrics definition.

5.3 Comparisons with State-of-the-Art Methods

Baselines. We compare against state-of-the-art monocular depth baselines [2, 26, 42, 47] and DfD baselines [12, 34, 45]. Monocular methods take a single all-in-focus image as input (an F/16 image for ZEDD). Note the comparison to monocular methods is not intended as a direct head-to-head comparison, nor do we claim that our method replaces monocular depth. Instead, we aim to position our method relative to widely used depth-estimation baselines. For the DfD baselines providing multiple checkpoints, we report the best-performing one.
Results on ZEDD. Quantitative results are presented in Tab. 2. Our model outperforms all monocular and DfD baselines by a large margin, proving the strong effectiveness of our approach. Compared to the best-performing monocular depth baseline DepthPro [2], the AbsRel is reduced by 55.7%, from 0.201 to 0.089. Note that none of the DfD baselines perform well on the ZEDD benchmark, suggesting their limited generalization ability. The strongest DfD baseline, DFF-FV, is even slightly behind the monocular baselines ($\delta_{1.25} = 0.576$ compared to $\delta_{1.25} = 0.580$ for MoGe-2 [42]).

Qualitative results are shown in Fig. 4. Our results are both sharp and metrically correct. The MoGe-2 [42] results are sharp but suffer from scale ambiguity, whereas HybridDepth [12] fails catastrophically due to the large domain gap.

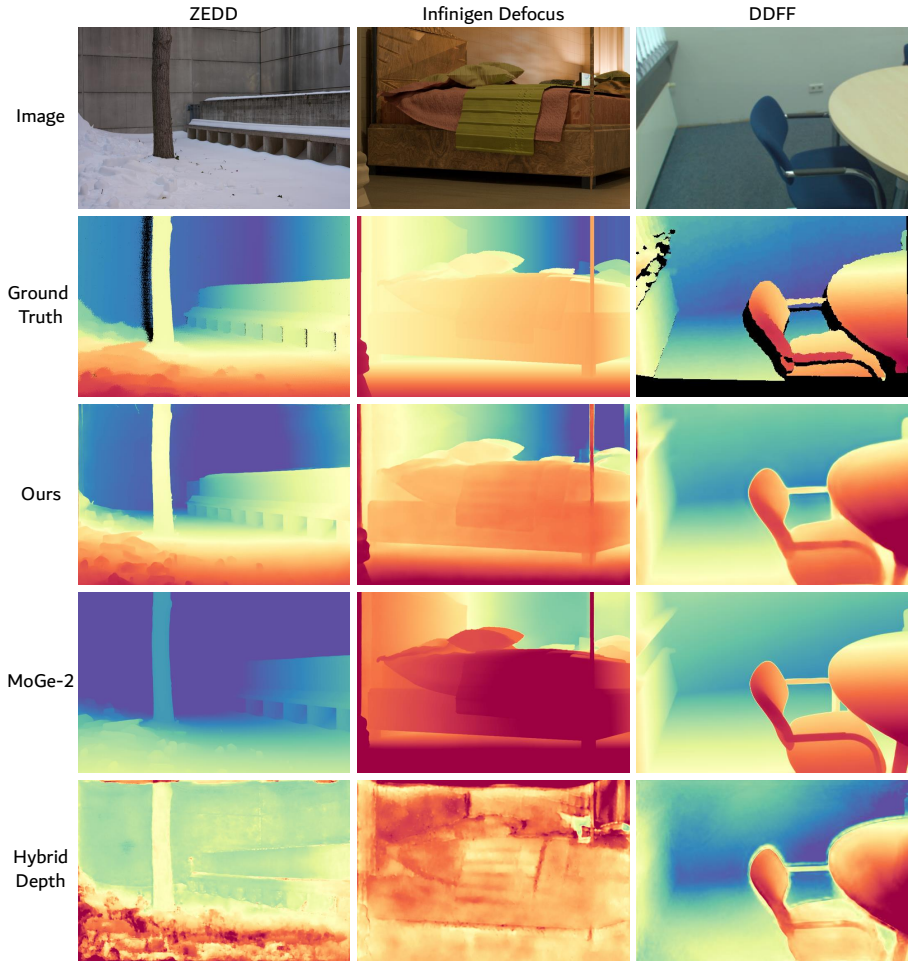


Fig. 4: Qualitative comparisons with baselines on 3 benchmarks. Our results are both sharp and metrically accurate. While monocular depth methods such as MoGe-2 [42] also produce sharp results, the scales are not metrically correct due to the scale ambiguity. DfD methods such as HybridDepth [12] fail on ZEDD and Infinigen Defocus. Even on DDFF where HybridDepth is trained on, their results are not as sharp as ours.

Infinigen Defocus. Results are presented in Tab. 2. Similar to the ZEDD results, our model outperforms all baselines by a large margin. The AbsRel is reduced by 51.7% compared to the second-best DepthPro [2], from 0.176 to 0.085. Qualitative results are shown in Fig. 4. Our model accurately captures thin structures, such as poles, while maintaining the correct global scale.

Real-World RGBD Datasets. We evaluate our model against state-of-the-art methods on iBims [19], DIODE [38], and HAMMER [17] with synthetic focus stacks. These datasets contain diverse indoor and outdoor scenes, where

Methods	iBims [19]		DIODE [38]		HAMMER [17]	
	$\delta_{1.25}$ ↑	AbsRel ↓	$\delta_{1.25}$ ↑	AbsRel ↓	$\delta_{1.25}$ ↑	AbsRel ↓
Monocular Depth						
DAv2 [47]	0.886	0.127	0.440	0.312	0.320	0.541
DepthPro [2]	0.878	0.120	0.462	0.289	0.719	0.272
UniDepthV2 [26]	0.941	0.078	0.680	0.368	0.482	0.371
MoGe-2 [42]	0.886	0.107	0.757	0.341	0.755	0.266
Depth from Defocus						
DFF-FV [45]	0.773	0.166	0.636	0.729	0.832	0.116
DFF-DFV [45]	0.786	0.160	0.671	0.315	0.931	0.100
DEReD [34]	0.263	0.741	0.106	0.886	0.000	5.430
HybridDepth [12]	0.872	0.133	0.580	4.261	0.956	0.089
DFF-DFV†	0.925	0.096	0.751	1.331	0.999	0.044
Ours (ViT-S)	0.954	0.075	0.766	0.178	0.999	0.044
Ours (ViT-B)	**0.963**	**0.070**	**0.779**	**0.160**	**0.999**	**0.017**

Table 3: Results on 3 real-world RGBD datasets, covering various indoor and outdoor scenes. Focus stacks are synthesized from the ground-truth depth maps using PSF. All methods are zero-shot. † means we re-train the model using the exact same data and loss as ours. Our method FOSSA outperforms all baselines.

outdoor scenes are particularly challenging as the defocus effects become weaker as depth values increase. Nevertheless, our method outperforms all baselines across all metrics, as shown in Tab. 3. Our method reduces the AbsRel by 44.6% compared to DepthPro [2] on DIODE [38], from 0.289 to 0.160.

We also fine-tune a DfD baseline DFF-DFV using the same training data as ours. This yields a substantial gain, showing an 11.92% improvement on $\delta_{1.25}$ on DIODE, highlighting the effectiveness of our synthetic training data generation pipeline. On the other hand, the finetuned DFF-DFV still consistently performs worse than our model. Overall, these results suggest that our improvements arise from both the proposed architecture and the improved data pipeline.

DDFF [14]. Results are shown in Tab. 4. For the finetuned experiments, we finetune using the DDFF training set (400 samples, same as baselines) for another 150 epochs using the same loss as in Eq. (2). Thanks to our powerful pretrained weights, the finetuned model significantly outperforms the baselines: the MSE is reduced by 40.4% compared to the previous state-of-the-art DualFocus [44].

5.4 Robustness

On ZEDD validation split, we evaluate robustness to three factors: aperture size, the distribution of focus distances, and the number of images in the focus stack. Results are shown in Fig. 5.

Aperture size. Smaller apertures (larger f-numbers) produce weaker defocus blur and are more challenging. As the aperture changes from F/1.4 to F/5.6, the $\delta_{1.25}$ of our method decreases by only 5%, indicating strong robustness.

Focus distance distribution. Varying the sampling distribution has a limited impact. Even under an extreme setting diverging largely from the training distribution (the 3rd group, using only the smallest three and largest two focus

Methods	DDFF Dataset [14]						
	MSE ↓ ($\times 10^{-4}$)	RMSE ↓	AbsRel ↓	SqRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Zero-Shot							
Ours (ViT-S, pretrained)	15.1	0.0352	0.27	0.0119	0.346	0.812	0.954
Ours (ViT-B, pretrained)	12.9	0.0324	0.21	0.0107	0.608	0.921	0.968
Trained on DDFF							
DDFF [14]	8.9	0.0276	0.24	0.0095	0.613	0.887	0.965
DefocusNet [20]	8.6	0.0255	0.17	0.0060	0.726	0.942	0.979
DFF-DFV [45]	5.7	0.0213	0.17	0.0063	0.767	0.942	0.981
HybridDepth [12]	5.1	0.0200	0.17	0.0060	0.789	0.947	0.981
DualFocus [44]	4.7	0.0194	0.16	0.0056	0.800	0.954	0.982
Ours (ViT-S, finetuned)	4.2	0.0183	**0.11**	0.0045	**0.936**	0.983	0.991
Ours (ViT-B, finetuned)	**2.8**	**0.0148**	**0.11**	**0.0025**	0.932	**0.987**	**0.994**

Table 4: Results on the DDFF [14] dataset. Numbers are adopted from the DualFocus [44] paper. We test under two settings: zero-shot and finetuned on DDFF. Our finetuned checkpoint outperforms all previous DfD baselines by a large margin.

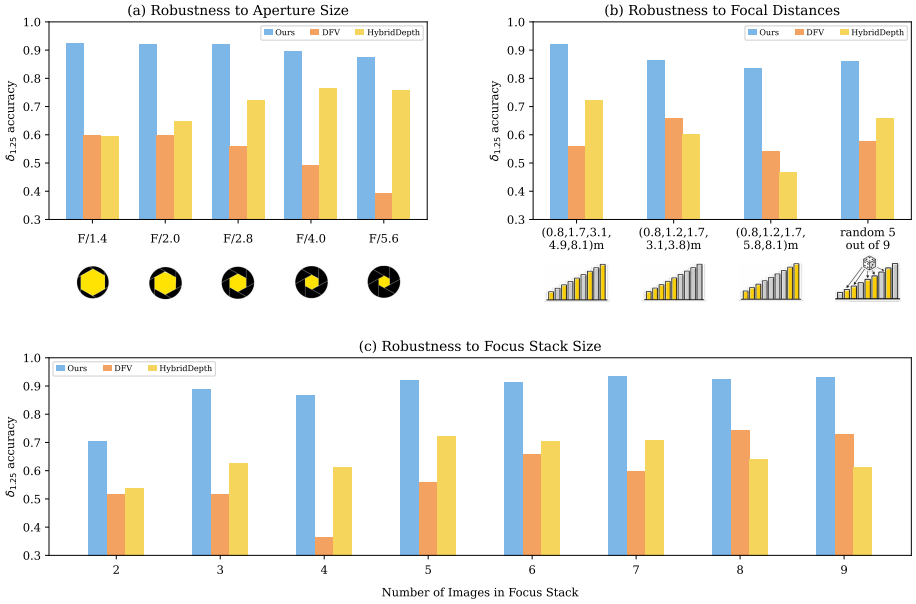


Fig. 5: FOSSA is robust to the aperture sizes, the focus distance distribution, and the focus stack size. The variance is small and we consistently outperform baselines under all configurations. Although always trained with a focus stack size of 5, our method can work with as few as 2 images. All results are on the ZEDD validation split.

distances out of nine), $\delta_{1.25}$ drops by only 9% relative to uniform sampling (the 1st group).

Focus stack size. Although the model is trained with 5 images per stack, it remains effective with fewer input images and performance degrades only gradually as the stack size decreases. Conversely, providing more than 5 images consistently improves performance.

Variants	ZEDD		Infinigen		Params	Mem	Speed	GFLOPS
	$\delta_{1.05}\uparrow$	Rel $\downarrow$	$\delta_{1.05}\uparrow$	Rel $\downarrow$	(M)	(GB)	(ms)	
Stack Attention Layer								
1) No stack attention layer	0.092	0.444	0.135	0.320	24.8	1.42	32.6	206.2
2) No FD embedding	0.090	0.399	0.123	0.365	37.8	1.52	45.7	326.7
Where to Collapse								
3) Collapse at $L_1 = 2$	0.372	0.108	0.332	0.105	33.7	1.49	34.8	241.0
4) Collapse at $L_1 = 6$	0.524	0.085	0.548	0.073	51.4	1.62	69.5	499.8
5) Fuse in DPT [6]	0.061	0.464	0.075	0.376	30.6	1.57	78.1	457.9
Domain Randomization During Training								
6) No random blur kernel	0.351	0.128	0.367	0.118	42.5	1.56	51.9	370.4
7) No random FD distribution	0.268	0.136	0.294	0.171	42.5	1.56	51.9	370.4
Ours	0.445	0.099	0.520	0.085	42.5	1.56	51.9	370.4

Table 5: Ablation studies on ZEDD (Val split) and Infinigen Defocus. All models are ViT-S, under 700×512 resolution on a single L40 GPU. Details: **2)** Using the regular sinusoidal embedding [39] based on frame index. **5)** Processing all images in parallel until the end, and fuse the features using stack attention in DPT (design as in VideoDA [6]). **6)** Training with F/1.4 and GaussPSF [34]. **7)** Training with scene-dependent focus distances, mimicking a photographer-in-the-loop (Appendix Sec. A4.4).

5.5 Ablation Studies

We study the effectiveness of our designs, as presented in Tab. 5. Our designs and hyperparameters are general and not tuned to overfit certain datasets, as the same conclusions can be drawn from ZEDD and Infinigen Defocus.

Stack Attention Layer. Stack attention and focus distance embedding are crucial to achieving good performance. Removing either component leads to a substantial drop, with $\delta_{1.05}$ decreasing from 0.445 to around 0.09.

Where to Collapse. Collapsing at a later stage (*i.e.*, more focus stack feature extraction layers) brings higher capacity and improves performance, but also higher computational cost. We choose $L_1 = 4$ to balance accuracy and efficiency. In contrast, fusion within the DPT head such as VideoDA [6] performs poorly on the DfD task, likely because the DPT head has insufficient capacity to effectively fuse features across the entire focus stack.

Domain Randomization During Training. Both randomizing the blur kernel and the focus distance distribution are important for robustness, as it reduces overfitting to specific defocus patterns. Removing these two sources of randomization leads to the $\delta_{1.25}$ drops of 21.12% and 39.78%, respectively.

6 Conclusion

This paper studies zero-shot depth from defocus (DfD), aiming to estimate dense metric depth from a focus stack. We introduce ZEDD, a real-world benchmark with broad scene coverage and high-quality ground truth, and propose FOSSA, a network built around a stack attention layer. Our method consistently outperforms all DfD and monocular baselines by a substantial margin on various benchmarks. Our paper paves the road for future research on the DfD task.

Acknowledgments

This work was partially supported by the National Science Foundation. Felix Heide is a co-founder of Algolux (now Torc Robotics), Head of AI at Torc Robotics, and a co-founder of Cephia AI.

References

1. Aimonen, P.: Focus stacking toolbox. `https://github.com/PetteriAimonen/focus-stack/`
2. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. In: ICLR (2025)
3. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
4. Carvalho, M., Le Saux, B., Trouvé-Peloux, P., Almansa, A., Champagnat, F.: On regression losses for deep depth estimation. ICIP (2018)
5. Catmull, E.E.: A subdivision algorithm for computer display of curved surfaces. The University of Utah (1974)
6. Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., Kang, B.: Video depth anything: Consistent depth estimation for super-long videos. In: CVPR (2025)
7. Community, B.O.: Blender - a 3D modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Amsterdam (2018), `http://www.blender.org`
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
9. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NeurIPS (2014)
10. Fujimura, Y., Iiyama, M., Funatomi, T., Mukaigawa, Y.: Deep depth from focal stack with defocus model for camera-setting invariance. IJCV **132**(6), 1970–1985 (2024)
11. Furgale, P., Rehder, J., Siegwart, R.: Unified temporal and spatial calibration for multi-sensor systems. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1280–1286. IEEE (2013)
12. Ganj, A., Su, H., Guo, T.: Hybriddepth: Robust metric depth fusion by leveraging depth from focus and single-image priors. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE (2025)
13. Gur, S., Wolf, L.: Single image depth estimation trained via depth from defocus cues. In: CVPR. pp. 7683–7692 (2019)
14. Hazirbas, C., Soyer, S.G., Staab, M.C., Leal-Taixé, L., Cremers, D.: Deep depth from focus. In: ACCV. pp. 525–541. Springer (2018)
15. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. IEEE TPAMI **46**(12), 10579–10596 (2024)
16. Ikoma, H., Nguyen, C.M., Metzler, C.A., Peng, Y., Wetzstein, G.: Depth from defocus with learned optics for imaging and occlusion-aware depth estimation. In: 2021 IEEE International Conference on Computational Photography (ICCP). IEEE (2021)

17. Jung, H., Ruhkamp, P., Zhai, G., Brasch, N., Li, Y., Verdie, Y., Song, J., Zhou, Y., Armagan, A., Ilic, S., et al.: Is my depth ground-truth good enough? hammer–highly accurate multi-modal dataset for dense 3d scene regression. arXiv preprint arXiv:2205.04565 (2022)
18. Khan, N., Xiao, L., Lanman, D.: Tiled multiplane images for practical 3d photography. In: ICCV (2023)
19. Koch, T., Liebel, L., Fraundorfer, F., Korner, M.: Evaluation of cnn-based single-image depth estimation methods. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018)
20. Maximov, M., Galim, K., Leal-Taixé, L.: Focus on defocus: bridging the synthetic to real domain gap for depth estimation. In: CVPR (2020)
21. Mescheder, L., Dong, W., Li, S., Bai, X., Santos, M., Hu, P., Lecouat, B., Zhen, M., Delaunoy, A., Fang, T., Tsin, Y., Richter, S.R., Koltun, V.: Sharp monocular view synthesis in less than a second. arXiv preprint arXiv:2512.10685 (2025)
22. Mu, C.W.T., Huang, J.B., Liu, Y.L.: Generative refocusing: Flexible defocus control from a single image. arXiv preprint arXiv:2512.16923 (2025)
23. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
24. Nazir, S., Qiu, Z., Coltuc, D., Mart'ınez-S'anchez, J., Arias, P.: idfd: A dataset annotated for depth and defocus. In: Image Analysis: 23rd Scandinavian Conference, (SCIA). Springer (2023)
25. Peng, J., Cao, Z., Luo, X., Lu, H., Xian, K., Zhang, J.: Bokehme: When neural rendering meets classical rendering. In: CVPR (2022)
26. Piccinelli, L., Sakaridis, C., Yang, Y.H., Segu, M., Li, S., Abbeloos, W., Van Gool, L.: Unidepthv2: Universal monocular metric depth estimation made simpler. IEEE TPAMI (2025)
27. Qiu, J., Cui, Z., Zhang, Y., Zhang, X., Liu, S., Zeng, B., Pollefeys, M.: Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In: CVPR (2019)
28. Raistrick, A., Lipson, L., Ma, Z., Mei, L., Wang, M., Zuo, Y., Kayan, K., Wen, H., Han, B., Wang, Y., Newell, A., Law, H., Goyal, A., Yang, K., Deng, J.: Infinite photorealistic worlds using procedural generation. In: CVPR (2023)
29. Raistrick, A., Mei, L., Kayan, K., Yan, D., Zuo, Y., Han, B., Wen, H., Parakh, M., Alexandropoulos, S., Lipson, L., Ma, Z., Deng, J.: Infinigen indoors: Photorealistic indoor scenes using procedural generation. In: CVPR (2024)
30. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021)
31. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE TPAMI **44**(3), 1623–1637 (2020)
32. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: ICCV (2021)
33. Rusinkiewicz, S., Levoy, M.: Efficient variants of the icp algorithm. In: Proceedings third international conference on 3-D digital imaging and modeling. pp. 145–152. IEEE (2001)
34. Si, H., Zhao, B., Wang, D., Gao, Y., Chen, M., Wang, Z., Li, X.: Fully self-supervised depth estimation from defocus clue. In: CVPR (2023)
35. Suwajanakorn, S., Hernandez, C., Seitz, S.M.: Depth from focus with your mobile phone. In: CVPR (2015)

36. Talegaonkar, C., Suresh, N.G., Novack, Z., Belhe, Y., Nagasamudra, P., Antipa, N.: Repurposing marigold for zero-shot metric depth estimation via defocus blur cues. arXiv preprint arXiv:2505.17358 (2025)
37. Tang, H., Cohen, S., Price, B., Schiller, S., Kutulakos, K.N.: Depth from defocus in the wild. In: CVPR (2017)
38. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., et al.: Diode: A dense indoor and outdoor depth dataset. arXiv preprint arXiv:1908.00463 (2019)
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS (2017)
40. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR (2025)
41. Wang, N.H., Wang, R., Liu, Y.L., Huang, Y.H., Chang, Y.L., Chen, C.P., Jou, K.: Bridging unsupervised and supervised depth from focus via all-in-focus supervision. In: CVPR (2021)
42. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. arXiv preprint arXiv:2507.02546 (2025)
43. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4909–4916. IEEE (2020)
44. Woo, S., Lee, S.: Dualfocus: Depth from focus with spatio-focal dual variational constraints. arXiv preprint arXiv:2509.21992 (2025)
45. Yang, F., Huang, X., Zhou, Z.: Deep depth from focus with differential focus volume. In: CVPR (2022)
46. Yang, H.H., Chen, W.T., Luo, H.L., Kuo, S.Y.: Multi-modal bifurcated network for depth guided image relighting. In: CVPR (2021)
47. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. NeurIPS (2024)
48. Yang, X., Fu, Q., Elhoseiny, M., Heidrich, W.: Aberration-aware depth-from-focus. IEEE TPAMI (2023)
49. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from a single image. In: ICCV (2023)
50. Zhou, Q.Y., Park, J., Koltun, V.: Open3d: A modern library for 3d data processing. arXiv preprint arXiv:1801.09847 (2018)
51. Zuo, Y., Yang, W., Ma, Z., Deng, J.: Omni-dc: Highly robust depth completion with multiresolution depth integration. In: ICCV (2025)