Bridging Unsupervised and Supervised Depth from Focus
via All-in-Focus Supervision

Ning-Hsu Wang[1,2]   Ren Wang[1]   Yu-Lun Liu[1]   Yu-Hao Huang[1]
Yu-Lin Chang[1]   Chia-Ping Chen[1]   Kevin Jou[1]
[1]MediaTek Inc.   [2]National Tsing Hua University
https://github.com/albert100121/AiFDepthNet

Abstract

Depth estimation is a long-lasting yet important task in computer vision. Most of the previous works try to estimate depth from input images and assume images are all-in-focus (AiF), which is less common in real-world applications. On the other hand, a few works take defocus blur into account and consider it as another cue for depth estimation. In this paper, we propose a method to estimate not only a depth map but an AiF image from a set of images with different focus positions (known as a focal stack). We design a shared architecture to exploit the relationship between depth and AiF estimation. As a result, the proposed method can be trained either supervisedly with ground truth depth, or un-supervisedly with AiF images as supervisory signals. We show in various experiments that our method outperforms the state-of-the-art methods both quantitatively and qual-itatively, and also has higher efficiency in inference time.

1. Introduction

Depth estimation has been one of the most fundamental computer vision problems in decades. Many downstream tasks, such as augmented reality (AR), virtual reality (VR) and autonomous driving, highly rely on this research topic. Recently, it also enabled an increasing number of applications for smartphone photography, such as depth-of-field adjustment, background substitution, and changing focus after the picture is taken.

Consequently, depth sensing has become a fundamental component for capturing devices. Active depth sensing solutions such as Time-of-Flight (ToF) and structured light are often expensive and power-consuming due to the need for specialized hardware. Passive techniques, such as binocular or multi-view stereo, are more cost and power-efficient but prone to errors in textureless regions.

Deep learning based stereo matching methods tackle this problem in a data-driven way by learning depth estimation directly from input images. However, they require a large amount of high-quality paired training data, which are time-consuming and expensive to acquire. They also suffer when the training data are imperfect: synthesized and unrealistic input images, or inaccurately registered depth maps.

Some unsupervised learning approaches [9, 42] were proposed to address this problem. They usually use image reconstruction loss and consistency loss without the need for ground truth depth data. They can also mitigate domain gaps by training directly with real-world stereo images without corresponding registered depth maps.

Another relatively under-explored cue for depth estimation is defocus blur. The task of depth-from-focus (or defocus) aims to estimate the depth of a scene from a focal stack, *i.e.*, images taken at different focal positions by the same camera. This allows consumer auto-focus monocular cameras to estimate depth without additional hardware.

Conventional optimization based depth-from-focus approaches [37, 36, 26] estimate the level of sharpness for each pixel and often suffer from textureless objects or aperture problems. Deep learning techniques [11, 23] help to overcome these issues but need ground truth depth data for supervised learning. It's difficult and time-consuming to retrieve focal stacks with registered depth maps, let alone the imperfect depth data obtained by hardware solutions such as ToF sensors [11]. One could synthesize defocus blur on a synthetic dataset with synthetic depth maps [23]. However, it is still questionable whether the thin lens synthesis model could represent real-world optics well.

In this paper, we propose a novel method to estimate depth and an all-in-focus (AiF) image jointly from an input focal stack. We exploit the relationship between these two tasks and design a shared common network. Moreover, the proposed network can be trained either supervisedly with ground truth depth maps or unsupervisedly with only ground truth AiF images. Compared to high-quality labeled depth, acquiring AiF images is relatively easier because AiF images can be captured with smaller apertures along with longer

exposures. However, collecting the corresponding focal stack of an AiF image might be difficult because of the focus breathing phenomenon, where a camera's field of view changes as the lens moves. To address this problem, we use synthetic data without such effects during training, and apply a calibration process on real data during testing.

Our **contribution** is three-fold:

- When trained supervisedly, our method outperforms the state-of-the-art methods in various comparisons, while our method also runs faster.

- To our knowledge, the proposed method is the first that can learn depth estimation unsupervisedly from only AiF images and performs favorably against the state-of-the-art methods.

- Domain gaps can be mitigated by our method with test-time optimization on real-world data, especially when ground truth depth data are not available.

2. Related Work

2.1. Depth from Focus

Depth estimation is a fundamental computer vision task that aims to use different cues such as color, semantics, stereo, or the difference in image sequences to predict or fuse depth maps [5, 40]. Most of the previous works assume that input images are all-in-focus, whereas in real-world scenarios, images are usually considered to be defocused in the background or with a shallow depth-of-field (DoF). Nevertheless, some approaches elaborate on depth estimation with defocused images. Conventional optimization based approaches [37, 36] proposed to directly estimate depth from focal stacks, and a variant approach [26] tries to generate an index map in which every pixel is assigned to the focus position leading to the maximal sharpness. Chen *et al*. [6] found the relationship between relative blur and disparity, and make use of it to enhance the robustness of matching. Depth from focus sweep video [18] targets estimating depth from images with successive focus positions. Recently, deep learning based approaches [11, 24] could model the blurriness more precisely and achieve much better depth quality. On the other hand, some works [2, 4, 35] use deep learning techniques to remove the defocus blur for single images.

2.2. Multi-focus Image Fusion

Although real-world images usually have defocus blur, most computer vision applications are supposed to cooperate with all-in-focus (AiF) images. Therefore, more and more attention is given on shallow-DoF image deblurring and multi-focus image fusion. A multi-focus image fusion approach [41] applies Laplacian to different scales of images. Zhan *et al*. [43] proposed a guided filter to help edge preserving during multi-focus image fusion. Nejati *et al*. [27] learn

a sparse representation of relative sharpness measurement and produce a pixel-level score map for decision through pooling. An unsupervised learning approach [32] was proposed to fuse either multi-exposure or multi-focus images to generate a high dynamic range (HDR) or an AiF image. Liu *et al*. [22] proposed to learn a focus map and a segmentation map through deep neural networks, and fuse images by integrating these maps.

2.3. Light Field

A light field camera captures spatially distributed light rays. By re-rendering through digitized multi-view images, variable aperture pictures can be generated after capturing. With this characteristic, a post-capture refocusing can also be accomplished by properly manipulating those light field images [28]. As light field camera provides multi-view information from different poses, it could help many computational photography applications. As the number of views grows, more information could be retrieved.

Light field images often suffer from low spatial resolution as the sampling resources are limited. Cheng *et al*. [7] and Jin *et al*. [17] proposed to apply deep learning on constructing super-resolution images from light field images to cover this situation. Depth estimation could also be achieved by leveraging the abundant information from light field images [20, 12, 13, 8, 38, 21, 45, 29, 31]. Levoy *et al*. [20] proposed to use the epipolar plane images (EPIs), which contain both spatial and angular information, for depth estimation from light field images. Heber *et al*. [12] proposed to learn the end-to-end mapping between the 4D light field and its corresponding 4D depth field accompanied with a high-order regularization refinement. By following [12], they then designed an encoder-decoder architecture to extract the geometric information from light field images [13]. Some works [15, 33] focus on providing light field datasets with ground truth depth or AiF images for further use.

2.4. Realistic Data Synthesis

Deep learning significantly improves the quality of computer vision tasks with a large amount of training data. However, collecting real-world data is often costly and time-consuming. Therefore, many works instead target synthesizing realistic data to provide sufficient data for training. Barron *et al*. [3] proposed to synthesize defocus images by a layer-wised rendering scheme with a z-buffer. Wadhwa *et al*. [39] combine a person segmentation mask with a depth map calculated from a dual-pixel camera to generated shallow depth-of-field images. Gur *et al*. [10] use a differentiable point spread function, namely the PSF convolution layer, to synthesize realistic defocus images and train a depth from defocus network to generate depth maps. Herrmann *et al*. [14] proposed a learning-based approach for autofocus and recently provided a real dataset for training. However, the

dataset was not open to the public at the submission time.

3. Method

We describe the proposed method in this section. Sec. 3.1 gives an overview of our method. Then, a shared network used for both of depth estimation and AiF image reconstruction is illustrated in Sec. 3.2. The attention mechanism that bridges the two tasks is introduced in Sec. 3.3. We then depict the core concept to turn AiF images into supervisory signals for unsupervised depth estimation in Sec. 3.4. Finally, Sec. 3.5 shows the loss functions for training with depth supervision and AiF supervision.

3.1. Overview

Depth from focus aims to recover the depth from a focal stack through defocus cues. As shown in Fig. 1, given a focal stack $\mathbf{S} \in \mathbb{R}^{H \times W \times 3 \times F}$ of F images with gradually varying focus positions $\mathbf{P} \in \mathbb{R}^{H \times W \times 1 \times F}$, our method produces several attention representations through a shared network. Then a depth map $\mathbf{D} \in \mathbb{R}^{H \times W \times 1}$ and an all-in-focus (AiF) image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ of this scene can be generated with these attention maps.

3.2. Network Architecture

DDFF [11] uses 2D ConvNets to address the sharpness measurement problem, which is one of the main challenges in depth from focus. On the other hand, DefocusNet [24] applies a global pooling layer as a communication tool between several weights-sharing 2D ConvNets. Another objective of this architecture is to allow focal stacks with arbitrary sizes. However, some important information across the stack dimension might not be effectively captured due to the limitation of 2D convolution and simple global pooling.

For this reason, we adopt the Inception3D [1] as the backbone of our model. As shown in Fig. 1, our model is an encoder-decoder network consisting of 3D convolutions. With the 3D convolution, defocus cues could be better captured across frames and thus facilitate the tasks of depth estimation and AiF image reconstruction. Moreover, our model can also handle focal stacks with arbitrary sizes attributed to the nature of 3D convolution.

3.3. Attention Mechanism

The output of our network is an intermediate attention $\mathbf{M} \in \mathbb{R}^{H \times W \times 1 \times F}$. The underlying expectation of the intermediate attention $\mathbf{M}$ is that it should reflect the probability of each focus position leading to the maximal sharpness. Then it can benefit both of depth estimation and AiF image reconstruction.

For depth estimation, we propose to normalize the intermediate attention $\mathbf{M}$ into a depth attention $\mathbf{M}^{\text{depth}}$ via a softplus normalization:

$$\mathbf{M}^{\text{depth}} = \varsigma(\mathbf{M}), \tag{1}$$

where

$$\mathbf{M}_{i,j,1,t}^{\text{depth}} = \frac{\ln\left(1 + \exp\left(\mathbf{M}_{i,j,1,t}\right)\right)}{\sum_{n=1}^{F} \ln\left(1 + \exp\left(\mathbf{M}_{i,j,1,n}\right)\right)}. \tag{2}$$

The softplus function is a smooth version of ReLU.

The depth attention $\mathbf{M}^{\text{depth}}$ can also be interpreted as the probability distribution because the softplus function ensures non-negativeness and normalizes $\mathbf{M}^{\text{depth}}$ into a valid probability distribution. Then the expected depth value of each pixel can be derived via:

$$\mathbf{D}_{i,j,1} = \Sigma_{t=1}^{F}(\mathbf{M}^{\text{depth}} \cdot \mathbf{P})_{i,j,1,t}. \tag{3}$$

For AiF image reconstruction, we perform similar procedures except that the normalization function changes from softplus to softmax. That is, the AiF attention $\mathbf{M}^{\text{AiF}}$ is obtained as:

$$\mathbf{M}^{\text{AiF}} = \sigma(\mathbf{M}), \tag{4}$$

where

$$\mathbf{M}_{i,j,1,t}^{\text{AiF}} = \frac{\exp\left(\mathbf{M}_{i,j,1,t}\right)}{\sum_{n=1}^{F} \exp\left(\mathbf{M}_{i,j,1,n}\right)}. \tag{5}$$

The AiF attention $\mathbf{M}^{\text{AiF}}$ can then be used for AiF image reconstruction. The expected AiF image of each pixel is expressed as:

$$\mathbf{I}_{i,j,k} = \Sigma_{t=1}^{F}(\mathbf{M}^{\text{AiF}} \cdot \mathbf{S})_{i,j,k,t}. \tag{6}$$

The reason why we adopt the softmax and softplus normalizations separately for depth and AiF image estimation is to tackle the problem of *sparse* focal stacks. By sparse, we mean that the stack size is small, and the focus positions inside the focal stack are not dense.

For AiF image reconstruction, it is best to select the sharpest pixel along the stack dimension. Blending multiple pixels inside a sparse stack usually does not help. Therefore, we leverage the softmax function to pursue the peaking phenomenon to extract the clearest pixels.

For depth estimation, the softmax normalization results in severe quantization for sparse focal stacks because it simply selects the nearest focal position of maximal sharpness. On the other hand, the softplus normalization leads to a flatter distribution so that more accurate depth can be predicted by interpolating among sparse focal positions. Fig. 2 shows the effects of this design choice.
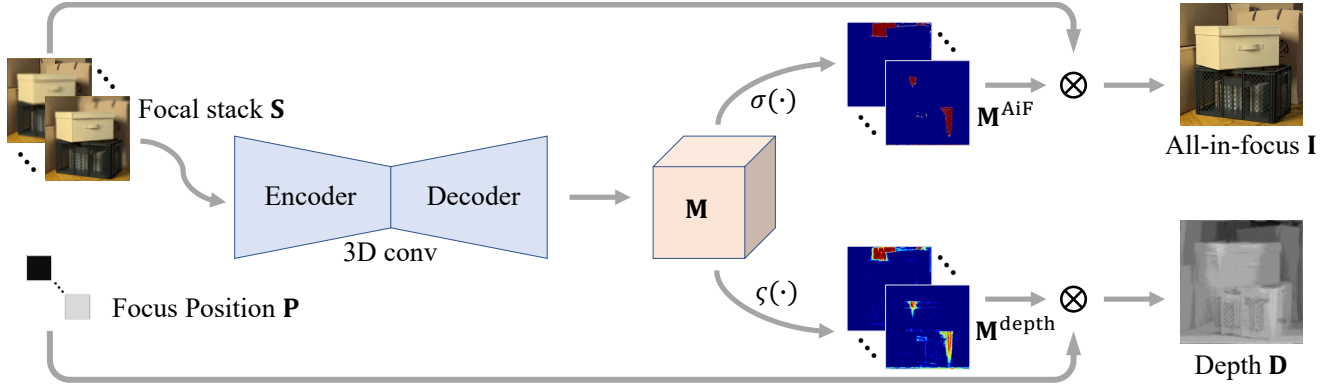
Figure 1: **An overview of the proposed method.** Given a stack of images with varying focus positions, *i.e.*, focal stack, our model first produces an intermediate attention map **M**. The intermediate attention map can be shared between depth estimation and all-in-focus (AiF) image reconstruction. With different normalization functions, the attention map can be further manipulated to generate either depth or AiF results.
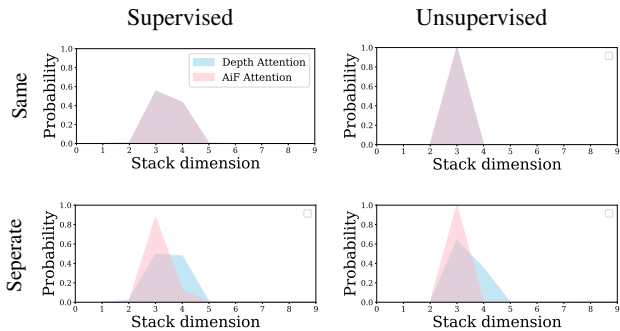

Figure 2: **The effect of normalization on attention maps.** The first row shows that if we adopt the same softmax normalization for depth and AiF attention, both of them tend to be a flatter distribution in supervised learning and face a peaking phenomenon in unsupervised learning (AiF supervision). The second row shows that with separate normalization as described in Sec. 3.3, each of depth and AiF attention becomes a proper distribution in both supervised and unsupervised learning.

3.4. Towards Unsupervised Depth Estimation

Similar to our method, DefocusNet [24] also combines the tasks of depth estimation and AiF image reconstruction via *intermediate defocus maps*. Their intermediate defocus maps necessitate supervisory signals, which can be derived by calculating the circle of confusions from the ground truth depth. Furthermore, they also propose to use a DepthNet following the intermediate defocus maps to predict the output depth. Therefore, their method can only be trained supervisedly when ground truth depth data are available.

Instead, we propose to use the intermediate attention **M** to bridge the two tasks, which does not necessitate intermediate supervisory signals from ground truth depth data. Moreover, there are no learnable parameters after **M**. We only use fixed normalization functions to generate the output depth and AiF images. Therefore, even when ground truth depth data are not available, we can still train the shared network to generate the intermediate attention **M** only via supervisory signals from ground truth AiF images. That is, our method can be trained supervisedly or unsupervisedly with or without ground truth depth data.

3.5. Training Loss

For supervised depth estimation , our model is trained with a simple L_1 loss:

$$L_{\text{supervised}} = L_{\text{depth}} = \mathbb{E}[\|\mathbf{D} - \mathbf{D}_{\text{gt}}\|_1], \qquad (7)$$

where $\mathbf{D}_{\text{gt}}$ stands for the ground truth depth.

For unsupervised depth estimation, our model can also be trained by a L_1 loss with AiF supervison:

$$L_{\text{AiF}} = \mathbb{E}[\|\mathbf{I} - \mathbf{I}_{\text{gt}}\|_1], \qquad (8)$$

where $\mathbf{I}_{\text{gt}}$ denotes the ground truth AiF image. Furthermore, we also encourage our depth map to be locally smooth using an edge-aware weighting as in [9]. The smoothness loss is defined as:

$$L_{\text{smooth}} = \mathbb{E}[W_x \left| \frac{\partial \mathbf{D}_{i,j,1}}{\partial x} \right| + W_y \left| \frac{\partial \mathbf{D}_{i,j,1}}{\partial y} \right|], \qquad (9)$$

where

$$W_x = \exp\left(-\frac{\lambda}{3} \sum_k \left| \frac{\partial \mathbf{I}_{i,j,k}}{\partial x} \right|\right)$$
$$W_y = \exp\left(-\frac{\lambda}{3} \sum_k \left| \frac{\partial \mathbf{I}_{i,j,k}}{\partial y} \right|\right), \qquad (10)$$

and λ is a hyper-parameter for the edge weighting based on the ground truth AiF image. The total loss of our unsupervised depth estimation is then:

$$L_{\text{unsupervised}} = L_{\text{AiF}} + \alpha L_{\text{smooth}}, \qquad (11)$$

where α indicates the importance of the smoothness loss.

4. Evaluation

In this section, we describe the datasets used in the experiments, and report ablation studies and comparisons with the state-of-the-art methods on depth from focus. In the supplementary material, we present additional results for AiF image reconstruction, and visual comparisons on DDFF 12-Scene [11], Middlebury Stereo Datasets [34], and the DefocusNet dataset [24].

4.1. Datasets

Totally five datasets are used in the quantitative experiments and visual comparisons. Their descriptions and statistics are summarized in Table 1. For more details, please see the supplementary material.

4.2. Evaluation Metrics

In this paper, we evaluate quantitative results with the following metrics: mean-absolute error (MAE), mean-squared error (MSE), root-mean-squared error (RMSE), log root-mean-squared error (logRMS), relative-absolute error (Abs. rel.), relative-squared error (Sqr. rel.), bumpiness (Bump), accuracy with $\delta = 1.25$, per-image inference time (Secs.).

4.3. Implementation Details

We implement our method in PyTorch [30]. Our model is trained from scratch using the Adam optimizer [19] ($\beta_1 = 0.9$, $\beta_2 = 0.999$), with a learning rate of 10^{-4}. For DDFF 12-Scene [11], 4D Light Field Dataset [15], FlyingThings3D [25], and the DefocusNet dataset [24], we use 10, 10, 15, and 5 as the input stack sizes, respectively. In unsupervised learning, the weight of smoothness loss α is set to 0.002 in all experiments. We apply random spatial transformations (flipping, cropping, rotation) and random color jittering (brightness, contrast, and gamma) for data augmentation during training. The training patch size is 256×256 after random cropping except for DDFF 12-Scene [11]. Because the image size of DDFF 12-Scene is 224×224, which is larger than 256×256, we do not apply random cropping on this dataset. All the experiments are conducted with a single NVIDIA GTX 1080 GPU.

4.4. Ablation Studies

To understand the performance of each proposed components in our method, we conduct various ablation studies
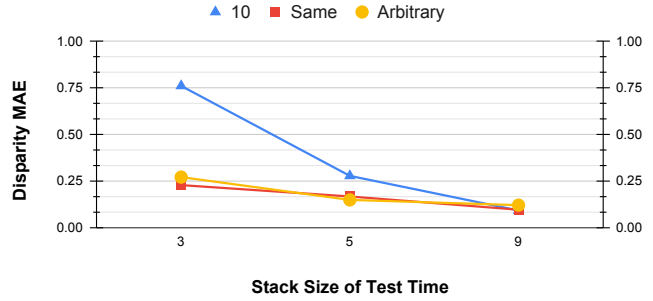

Figure 3: **The effect of focal stack size.** Each line indicates a stack size setting during training. Specifically, *10* and *Arbitrary* respectively represent training with a fixed size of 10 and training with arbitrary sizes, and *Same* refers to using the same setting as test stacks.

on 4D Light Field Dataset [16]. All models are supervised trained with disparity maps if not specified.

Architecture. Table 2 shows the comparison between 2D and 3D convolutions and with/without the proposed attention mechanism. The 3D convolution performs better than 2D convolution as the 3D convolution is able to capture and aggregate features across both spatial and stack dimensions. The design of attention bridges the tasks of depth estimation and AiF image reconstruction, and thus enables the possibility of unsupervised depth estimation as described in Sec. 3.4 while maintaining roughly the same prediction quality.

Focal stack size. Our network can handle arbitrary input stack sizes due to the nature of 3D convolution. Fig. 3 shows a comparison on different input stack sizes. We train the models in three different ways: 1) *10*: training with a *fixed* input stack size of 10. 2) *Arbitrary*: training with *arbitrary* input stack sizes randomly sampled from 2 to 10. 3) *Same*: training with the *same* size as test stacks.

The model trained with a fixed stack size performs poorly while testing with different stack sizes. As expected, the model trained with the same stack size as test data performs the best as the input setting for training and testing are consistent. However, the model trained with arbitrary input stack sizes performs favorably against the ones trained with the same stack size as test data. This demonstrates the robustness of our method across different input stack sizes.

Fig. 4 shows the visual comparison of different input stack sizes, generated with the model trained with arbitrary input stack sizes. It is obvious that the quality of the estimated disparity map improves as the stack size increases.

Unsupervised learning using all-in-focus images. On the basis of the attention mechanism described in Sec. 3.3, our method is able to be trained with either ground truth disparity maps (supervised) or ground truth AiF images (unsupervised). Table 3 and Fig. 5 respectively show the quantitative and qualitative results for both of our supervised and unsu-

Table 1: **Summary of the evaluation datasets.**

Dataset	Image source	Cause of defocus	Disparity/Depth GT	AiF GT
DefocusNet [24]	Synthetic	Blender rendering	Depth	
DDFF 12-Scene [11]	Real	Light-field composition	Depth	
4D Light Field Dataset [16]	Synthetic	Light-field composition	Disparity	✓
Middlebury Stereo Datasets [34]	Real	Disparity rendering	Disparity	✓
Mobile Depth [37]	Real	Real		Stitching by MRF

Table 2: **Ablation study on network architecture.** The 3D convolution leads to better performance on disparity estimation because of its ability to capture features for both spatial and stack dimensions. Although the design of attention does not increase the performance explicitly, it bridges the tasks of depth estimation and AiF image reconstruction and thus enables the unsupervised learning for depth estimation.

Architecture	Attention	MAE↓	MSE↓	RMSE↓	Bump.↓
3D	✓	0.0788	0.0472	0.2014	1.5776
		0.0851	0.0461	0.1984	2.3168
2D	✓	0.1070	0.0577	0.2259	2.1435
		0.1179	0.0576	0.2291	2.3168

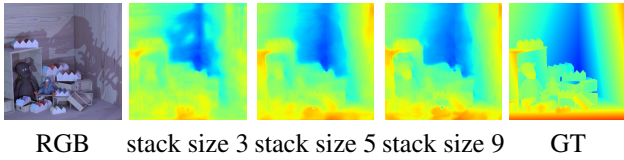RGB stack size 3 stack size 5 stack size 9 GT

Figure 4: **Visual comparison on different focal stack sizes.** The quality of the estimated disparity map improves as the stack size increases.

pervised models. As shown in Table 3, with a large input stack size of 37 and the smoothness loss, the performance of unsupervised learning is able to approximate that of supervised learning with a smaller stack size of 10 .

However, the results of unsupervised learning with a smaller stack size suffer from the quantization problem as described in Sec. 3.4. As shown in Fig. 5, after adding the smoothness loss, the output disparity map becomes locally smooth and perform better qualitatively.

4.5. Comparisons to the State-of-the-art Methods

After ablation on different components of the proposed method, we conduct comparisons with the state-of-the-art methods on various datasets in this section.

DDFF 12-Scene. Table 4 shows the quantitative comparison on DDFF 12-Scene [11]. All methods are supervisedly trained on this dataset with ground truth depth. The proposed method performs favorably against the state-of-the-art

Table 3: **Ablation study on supervision.** The results of supervised learning perform better than the ones from unsupervised learning (AiF supervision). Unsupervised learning often generates disparity maps that suffer from the quantization effect and lead to poor results. After adding the smoothness loss, the output disparity maps become locally smooth and perform better quantitatively.

Supervised	Stack size	L_{smooth}	MAE↓	MSE↓	RMSE↓
Yes	10		0.0788	0.0472	0.2014
No	10		0.2425	0.1174	0.3401
No	37		0.2099	0.1039	0.3202
No	10	✓	0.1671	0.0746	0.2698
No	37	✓	0.1116	0.0584	0.2311

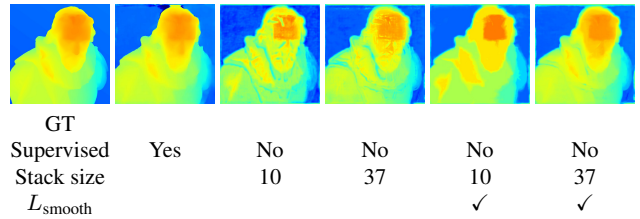|					
GT					
Supervised	Yes	No	No	No	No
Stack size		10	37	10	37
L_{smooth}				✓	✓

Figure 5: **Visual comparison on different supervision settings.** The results of supervised learning are better than the ones from unsupervised learning (AiF supervision). The output disparity maps from unsupervised learning suffer from the quantization effect. By adding the smoothness loss, the disparity results become locally smooth and perform favorably against the ones from supervised learning qualitatively.

methods under all metrics. Furthermore, our method runs faster than another deep learning based method DDFF [11] .

4D Light Field Dataset. Compared to DDFF 12-Scene [11], 4D Light Field Dataset [16] provides the ability to simulate shallower DoF images as its larger baseline leads to a larger synthetic aperture. Table 5 shows the quantitative comparison with state-of-the-art methods. All of them are trained with this dataset. Our supervised model achieves the best MSE (0.0472), which outperforms the state-of-the-art method DefocusNet [24] (0.0593). Furthermore, our unsupervised model even performs better than most of the other

Table 4: **Quantitative comparison on DDFF 12-Scene.** Note that RMSE is not presented due to the mismatched results from the benchmark website and the paper of DDFF [11]. For DefocusNet [24], we only show the original metric reported in their paper, which is MSE. Red text indicates the best, and blue text indicates the second-best performing method.

Method	MSE $\downarrow$	log RMS $\downarrow$	Abs. rel. $\downarrow$	Sqr. rel. $\downarrow$	Bump. $\downarrow$	$\delta = 1.25 \uparrow$	$\delta = 1.25^2 \uparrow$	$\delta = 1.25^3 \uparrow$
Ours	**$8.6e^{-4}$**	**0.29**	**0.25**	**0.01**	0.63	**68.33**	**87.40**	<u>93.96</u>
DefocusNet [24]	<u>$9.1e^{-4}$</u>	-	-	-	-	-	-	-
DDFF [11]	$9.7e^{-4}$	0.32	0.29	**0.01**	0.59	61.95	85.14	92.98
PSPNet [44]	$9.4e^{-4}$	**0.29**	0.27	**0.01**	<u>0.55</u>	<u>62.66</u>	<u>85.90</u>	**94.42**
Lytro	$2.1e^{-3}$	<u>0.31</u>	<u>0.26</u>	**0.01**	1.02	55.65	82.00	93.09
PSP-LF [44]	$2.7e^{-3}$	0.45	0.46	<u>0.03</u>	**0.54**	39.70	65.56	82.46
DFLF [11]	$4.8e^{-3}$	0.59	0.72	0.07	0.65	28.64	53.55	71.61
VDFF [26]	$7.3e^{-3}$	1.39	0.62	0.05	0.79	8.42	19.95	32.68

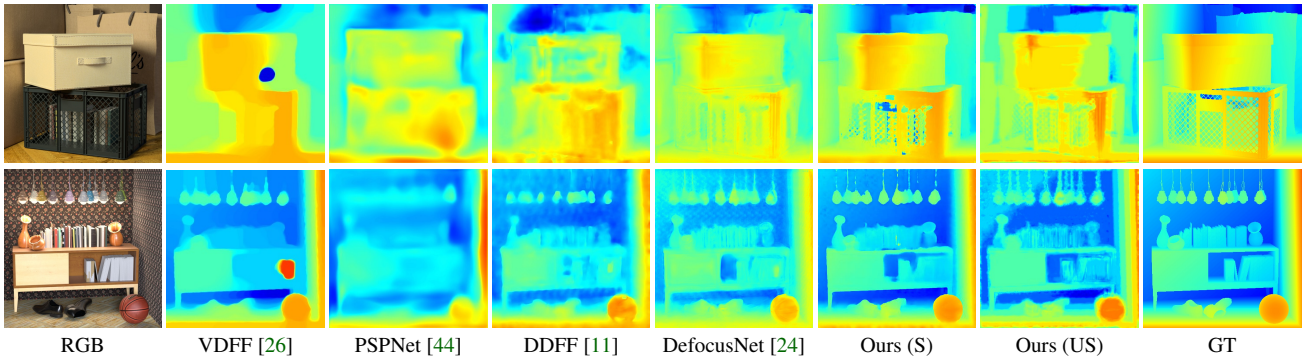| RGB | VDFF [26] | PSPNet [44] | DDFF [11] | DefocusNet [24] | Ours (S) | Ours (US) | GT |

Figure 6: **Visual comparison on 4D Light Field Dataset.** S: supervised. US: unsupervised (AiF-supervision).

Table 5: **Quantitative comparison on 4D Light Field Dataset.** Our supervised model outperforms the state-of-the-art method DefocusNet. Furthermore, our unsupervised model (with AiF supervision) even performs better than most of the other supervised methods. Please see Fig. 6 for the visual comparison. (∗ represents that the model is pre-trained on DDFF 12-Scene.)

Method	Supervised	MSE $\downarrow$	RMSE $\downarrow$	Bump. $\downarrow$
Ours	Yes	**0.0472**	**0.2014**	<u>1.58</u>
DefocusNet [24]	Yes	<u>0.0593</u>	<u>0.2355</u>	2.69
*DDFF [11]	Yes	0.19	0.42	1.92
*PSPNet [44]	Yes	0.37	0.53	**1.21**
VDFF [26]	Yes	1.3	1.15	<u>1.58</u>
Ours	No	0.0746	0.2398	2.58

Table 6: **Quantitative comparison on the DefocusNet dataset.** Our supervised model outperforms the state-of-the-art method DefocusNet. Please refer to the supplementary material for the full metric results.

Method	MAE $\downarrow$	MSE $\downarrow$	RMSE $\downarrow$
Ours	**0.0549**	**0.0127**	**0.1043**
DefocusNet [24]	0.0637	0.0175	0.1207

dataset using physically based rendering (PBR) shaders. Since only a subset of this dataset has been released and there are no AiF images, we can only conduct supervised learning on the provided subset. The quantitative results of our method and DefocusNet [24] are shown in Table 6. Note that the results of DefocusNet are retrained on the provided subset with the released code and setting.

Mobile Depth. The Mobile Depth dataset [37] is a real-world dataset of focal stacks captured with mobile phone and cameras. Due to the lack of training data in this dataset, all compared models are trained on FlyingThings3D [25], and the input focal stacks are synthesized by the rendering technique used in Barron *et al.* [3].

As shown in Fig. 7, the output depth maps of our super-

supervised methods, except for DefocusNet.

Fig. 6 shows the qualitative results of 4D Light Field Dataset [16]. Our supervised model delivers sharper depth boundaries and less noise in textureless regions. Meanwhile, our unsupervised model also achieves visually comparable results with the supervised DefocusNet [24].

DefocusNet. The DefocusNet dataset [24] is a synthetic

Table 7: **Analysis on generalization ability across different datasets.** We train our models and DefocusNet [24] on FlyingThings3D [25] with synthesized focal stacks by the rendering technique used in Barron *et al.* [3], and test these models on Middlebury Stereo Datasets [34] and the DefocusNet dataset. The results show that both of our supervised and unsupervised (with AiF supervision) models have better generalization ability than DefocusNet.

Method	Supervised	Test Dataset	MAE↓	MSE↓	RMSE↓	absRel↓	sqrRel↓	Sec.↓
Ours	Yes	Middlebury	**3.8249**	**58.5698**	**5.9355**	**2.4776**	**45.5886**	0.0287
Ours	No	Middlebury	5.4499	99.6029	8.2606	3.8954	80.7120	**0.0283**
DefocusNet [24]	Yes	Middlebury	7.4084	157.4397	9.0794	3.4698	63.6797	0.3798
Ours	Yes	DefocusNet	0.1827	0.0795	0.2607	72.4664	40.4281	0.021
Ours	No	DefocusNet	**0.1816**	**0.0627**	**0.2380**	**59.4459**	**14.0227**	**0.01959**
DefocusNet [24]	Yes	DefocusNet	0.3200	0.1478	0.3722	138.2917	70.0229	0.0527

vised model are smoother and exhibit less ambiguity than DefocusNet [24]. But the results of our unsupervised model (with AiF supervision) are worse than DefocusNet. Nonetheless, since our method allows training without ground truth depth, we can perform test-time optimization on this dataset with only AiF images. After test-time optimization, our unsupervised model performs better than DefocusNet, and both of our supervised and unsupervised models perform favorably against Mobile Depth. We also show the output AiF images in the supplementary material.

Generalization ability analysis. Middlebury Stereo Datasets[34] is a real-world dataset of stereo images along with ground truth disparity maps. We use this dataset to analyze the generalization ability of our method. To this end, we firstly let our models and DefocusNet [24] be trained on FlyingThings3D [25] with synthesized focal stacks by the rendering technique used in Barron *et al.* [3], and then test these models on Middlebury Stereo Datasets as well as the DefocusNet dataset. The quantitative results are shown in Table 7. One can see that both of our supervised and unsupervised (with AiF supervision) models achieve better generalization than DefocusNet on these two datasets.

Running time. As indicated in Table 7, our method is faster than the state-of-the-art method DefocusNet [24]. The main reason might be that our model does not have a bottleneck like the global pooling layer in DefocusNet.

5. Conclusion

We have proposed a method to jointly estimate the depth map and the all-in-focus (AiF) image from an input focal stack with a shared network. By the design of the proposed attention mechanism, the shared network can be trained either supervisedly, or unsupervisedly with AiF images. Our method can further mitigate domain gaps even in the absence of ground truth depth. Experimental results show that our method outperforms the state-of-the-art methods while also having higher efficiency in inference time.
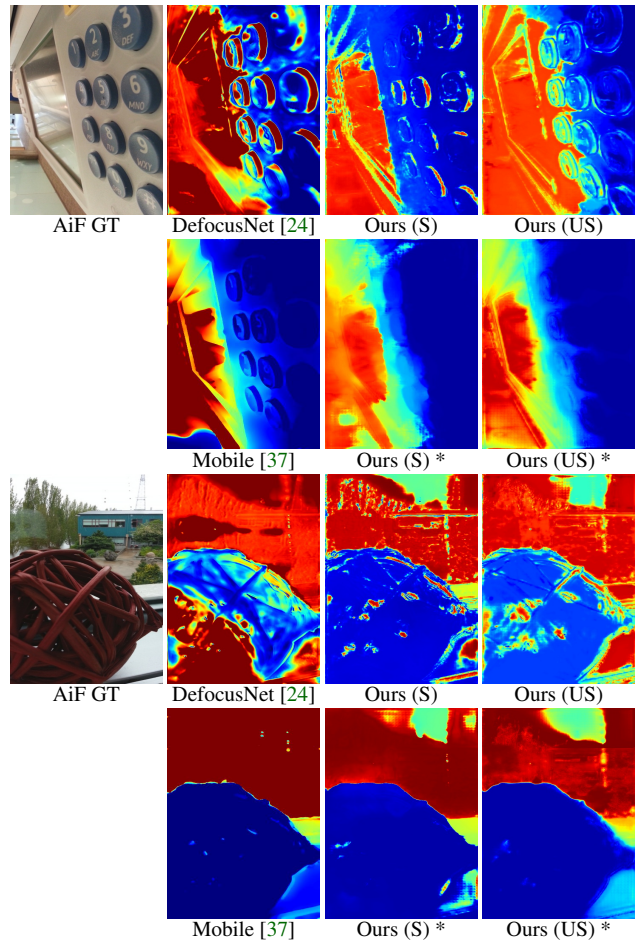

Figure 7: **Visual comparison on the Mobile Depth dataset.** With test-time optimization, our models perform better than DefocusNet [24] and favorably against Mobile Depth [37] qualitatively. (S: supervised. US: unsupervised (AiF supervision). *: test-time optimization.)

References

[1] Jean-Baptiste Alayrac, João Carreira, and Andrew Zisserman. The visual centrifuge: Model-free layered video representations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2457–2466. Computer Vision Foundation / IEEE, 2019. 3

[2] Saeed Anwar, Zeeshan Hayder, and Fatih Porikli. Depth estimation and blur removal from a single out-of-focus image. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. BMVA Press, 2017. 2

[3] Jonathan T. Barron, Andrew Adams, YiChang Shih, and Carlos Hernández. Fast bilateral-space stereo for synthetic defocus. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4466–4474. IEEE Computer Society, 2015. 2, 7, 8

[4] Marcela Carvalho, Bertrand Le Saux, Pauline Trouvé-Peloux, Andrés Almansa, and Frédéric Champagnat. On regression losses for deep depth estimation. In *2018 IEEE International Conference on Image Processing, ICIP 2018, Athens, Greece, October 7-10, 2018*, pages 2915–2919. IEEE, 2018. 2

[5] Yu-Lin Chang, Wei-Yin Chen, Jing-Ying Chang, Yi-Min Tsai, Chia-Lin Lee, and Liang-Gee Chen. Priority depth fusion for the 2d to 3d conversion system. In Brian D. Corner, Masaaki Mochimaru, and Robert Sitnik, editors, *Proceedings of the Conference on Three-Dimensional Image Capture and Applications 2008, San Jose, CA, USA, January 28-29, 2008*, volume 6805 of *SPIE Proceedings*, page 680513. SPIE, 2008. 2

[6] Ching-Hui Chen, Hui Zhou, and Timo Ahonen. Blur-aware disparity estimation from defocus stereo images. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 855–863. IEEE Computer Society, 2015. 2

[7] Zhen Cheng, Zhiwei Xiong, Chang Chen, and Dong Liu. Light field super-resolution: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1804–1813. Computer Vision Foundation / IEEE, 2019. 2

[8] Donald G. Dansereau, Bernd Girod, and Gordon Wetzstein. Liff: Light field features in scale and depth. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8042–8051. Computer Vision Foundation / IEEE, 2019. 2

[9] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6602–6611. IEEE Computer Society, 2017. 1, 4

[10] Shir Gur and Lior Wolf. Single image depth estimation trained via depth from defocus cues. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7683–7692. Computer Vision Foundation / IEEE, 2019. 2

[11] Caner Hazirbas, Sebastian Georg Soyer, Maximilian Christian Staab, Laura Leal-Taixé, and Daniel Cremers. Deep depth from focus. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part III*, volume 11363 of *Lecture Notes in Computer Science*, pages 525–541. Springer, 2018. 1, 2, 3, 5, 6, 7

[12] Stefan Heber and Thomas Pock. Convolutional networks for shape from light field. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3746–3754. IEEE Computer Society, 2016. 2

[13] Stefan Heber, Wei Yu, and Thomas Pock. Neural epi-volume networks for shape from light field. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2271–2279. IEEE Computer Society, 2017. 2

[14] Charles Herrmann, Richard Strong Bowen, Neal Wadhwa, Rahul Garg, Qiurui He, Jonathan T Barron, and Ramin Zabih. Learning to autofocus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2230–2239, 2020. 2

[15] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato, editors, *Computer Vision - ACCV 2016 - 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III*, volume 10113 of *Lecture Notes in Computer Science*, pages 19–34. Springer, 2016. 2, 5

[16] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision*, pages 19–34. Springer, 2016. 5, 6, 7

[17] Jing Jin, Junhui Hou, Jie Chen, and Sam Kwong. Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 2257–2266. IEEE, 2020. 2

[18] Hyeongwoo Kim, Christian Richardt, and Christian Theobalt. Video depth-from-defocus. In *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, pages 370–379. IEEE Computer Society, 2016. 2

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[20] Marc Levoy and Pat Hanrahan. Light field rendering. In John Fujii, editor, *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, New Orleans, LA, USA, August 4-9, 1996*, pages 31–42. ACM, 1996. 2

[21] Haiting Lin, Can Chen, Sing Bing Kang, and Jingyi Yu. Depth recovery from light field using focal stack symmetry. In *Pro-

ceedings of the IEEE International Conference on Computer Vision, pages 3451–3459, 2015. 2

[22] Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36:191–207, 2017. 2

[23] Maxim Maximov, Kevin Galim, and Laura Leal-Taixé. 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*, pages 1071–1080, 2020. 1

[24] Maxim Maximov, Kevin Galim, and Laura Leal-Taixé. Focus on defocus: Bridging the synthetic to real domain gap for depth estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1068–1077. IEEE, 2020. 2, 3, 4, 5, 6, 7, 8

[25] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134. 5, 7, 8

[26] Michael Möller, Martin Benning, Carola Schönlieb, and Daniel Cremers. Variational depth from focus reconstruction. *IEEE Trans. Image Process.*, 24(12):5369–5378, 2015. 1, 2, 7

[27] M. Nejati, S. Samavi, and S. Shirani. Multi-focus image fusion using dictionary-based sparse representation. *Inf. Fusion*, 25:72–84, 2015. 2

[28] Ren Ng, Marc Levoy, Mathieu Bredif, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light field photography with a hand-held plenoptic camera. In *Stanford University Computer Science Tech Report CSTR 2005-02*. Stanford University, 2005. 2

[29] In Kyu Park, Kyoung Mu Lee, et al. Robust light field depth estimation using occlusion-noise aware data costs. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2484–2497, 2017. 2

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[31] Jiayong Peng, Zhiwei Xiong, Yicheng Wang, Yueyi Zhang, and Dong Liu. Zero-shot depth estimation from light field using a convolutional neural network. *IEEE Transactions on Computational Imaging*, 6:682–696, 2020. 2

[32] K. Ram Prabhakar, V. Sai Srikar, and R. Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy,*

October 22-29, 2017, pages 4724–4732. IEEE Computer Society, 2017. 2

[33] Parikshit Sakurikar, Ishit Mehta, Vineeth N. Balasubramanian, and P. J. Narayanan. Refocusgan: Scene refocusing using a single image. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *European Conference on Computer Vision ECCV 2016: Computer Vision – ECCV 2016*, pages 842–857. Springer, 2016. 2

[34] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nesic, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In Xiaoyi Jiang, Joachim Hornegger, and Reinhard Koch, editors, *Pattern Recognition - 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings*, volume 8753 of *Lecture Notes in Computer Science*, pages 31–42. Springer, 2014. 5, 6, 8

[35] Pratul P. Srinivasan, Rahul Garg, Neal Wadhwa, Ren Ng, and Jonathan T. Barron. Aperture supervision for monocular depth estimation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6393–6401. IEEE Computer Society, 2018. 2

[36] Jaeheung Surh, Hae-Gon Jeon, Yunwon Park, Sunghoon Im, Hyowon Ha, and In So Kweon. Noise robust depth from focus using a ring difference filter. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2444–2453. IEEE Computer Society, 2017. 1, 2

[37] Supasorn Suwajanakorn, Carlos Hernández, and Steven M. Seitz. Depth from focus with your mobile phone. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3497–3506. IEEE Computer Society, 2015. 1, 2, 6, 7, 8

[38] Yu-Ju Tsai, Yu-Lun Liu, Ming Ouhyoung, and Yung-Yu Chuang. Attention-based view selection networks for light-field disparity estimation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 12095–12103. AAAI Press, 2020. 2

[39] Neal Wadhwa, Rahul Garg, David E. Jacobs, Bryan E. Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T. Barron, Yael Pritch, and Marc Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Trans. Graph.*, 37(4):64:1–64:13, 2018. 2

[40] Ning-Hsu Wang, Bolivar Solarte, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. 360sd-net: 360° stereo depth estimation with learnable cost volume. In *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*, pages 582–588. IEEE, 2020. 2

[41] Wencheng Wang and Faliang Chang. A multi-focus image fusion method based on laplacian pyramid. *J. Comput.*, 6(12):2559–2566, 2011. 2

[42] Junyuan Xie, Ross B. Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convo-

lutional neural networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 842–857. Springer, 2016. 1

[43] Kun Zhan, Jicai Teng, Qiaoqiao Li, and Jinhui Shi. A novel explicit multi-focus image fusion method. *J. Inf. Hiding Multim. Signal Process.*, 6(3):600–612, 2015. 2

[44] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6230–6239. IEEE Computer Society, 2017. 7

[45] Wenhui Zhou, Enci Zhou, Yuxiang Yan, Lili Lin, and Andrew Lumsdaine. Learning depth cues from focal stack for light field depth estimation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1074–1078. IEEE, 2019. 2