

Learning Depth from Focus in the Wild

Changyeon Won[Ⓛ] and Hae-Gon Jeon^{Ⓛ*}

Gwangju Institute of Science and Technology
cywon1997@gm.gist.ac.kr and haegonj@gist.ac.kr

Abstract. For better photography, most recent commercial cameras including smartphones have either adopted large-aperture lens to collect more light or used a burst mode to take multiple images within short times. These interesting features lead us to examine depth from focus/defocus. In this work, we present a convolutional neural network-based depth estimation from single focal stacks. Our method differs from relevant state-of-the-art works with three unique features. First, our method allows depth maps to be inferred in an end-to-end manner even with image alignment. Second, we propose a sharp region detection module to reduce blur ambiguities in subtle focus changes and weakly texture-less regions. Third, we design an effective downsampling module to ease flows of focal information in feature extractions. In addition, for the generalization of the proposed network, we develop a simulator to realistically reproduce the features of commercial cameras, such as changes in field of view, focal length and principal points. By effectively incorporating these three unique features, our network achieves the top rank in the DDFF 12-Scene benchmark on most metrics. We also demonstrate the effectiveness of the proposed method on various quantitative evaluations and real-world images taken from various off-the-shelf cameras compared with state-of-the-art methods. Our source code is publicly available at <https://github.com/wcy199705/DfFintheWild>.

Keywords: depth from focus, image alignment, sharp region detection and simulated focal stack dataset.

1 Introduction

As commercial demand for high-quality photographic applications increases, images have been increasingly utilized in scene depth computation. Most commercial cameras, including smartphone and DSLR cameras have two interesting configurations: large-aperture lens and a dual-pixel (DP) sensor. Both are reasonable choices to collect more light and to quickly sweep the focus through multiple depths. Because of this, images appear to have a shallow depth of field (DoF) and are formed as focal stacks with corresponding meta-data such as focal length and principal points. One method to accomplish this is to use single dual-pixel (DP) images which have left and right sub-images with narrow baselines

* Corresponding author

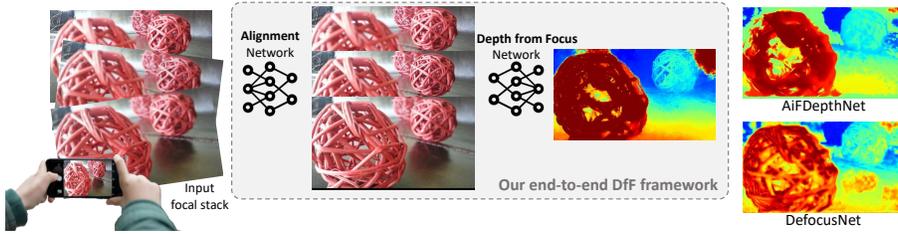


Fig. 1. Results of our true end-to-end DfF framework with comparisons to state-of-the-art methods.

and limited DoFs. A straightforward way is to find correspondences between the left and right sub-images [36,7,41]. Despite an abundance of research, such methods are heavily dependent on the accurate retrieval of correspondences due to the inherent characteristics of DP images. Pixel disparities between the two sub-images result in blurred regions, and the amount of spatial shifts is proportional to the degree of blurrings. Another group of approaches solves this problem using different angles. The out-of-focus regions make it possible to use depth-from-defocus (DfD) techniques to estimate scene depths [2,24,40]. Since there is a strong physical relationship between scene depths and the amount of defocus blurs, the DfD methods account for it in data-driven manners by learning to directly regress depth values. However, there is a potential limitation to these works [2,24,40]. A classic issue, an aperture effect, makes an analysis of defocus blur in a local window difficult. In addition, some of them recover deblurred images from input, but image deblurring also belongs to a class of ill-posed inverse problems for which the uniqueness of the solution cannot be established [19].

These shortcomings motivate us to examine depth from focus (DfF) as an alternative. DfF takes in a focal stack to a depth map during focus sweeping, which is available in most off-the-shelf cameras, and determines the focus in the input focal stack. In particular, the inherent operations of convolutional neural networks (CNNs), convolution and maxpooling, are suitable for measuring the values obtained from derivatives of the image/feature map based on the assumption that focused images contain sharper edges [12,21,37]. Nevertheless, there is still room for improvements with respect to model generalization, due to the domain gap between public datasets and real-world focal stack images, and an alignment issue that we will discuss.

In this work, we achieve a high-quality and well-generalized depth prediction from single focal stacks. Our contributions are threefold (see Fig.1): First, we compensate the change in image appearance due to magnification during the focus change, and the slight translations from principal point changes. Compared to most CNN-based DfD/DfF works [12,21,37] which either assume that input sequential images are perfectly aligned or use hand-crafted feature-based alignment techniques, we design a learnable context-based image alignment, which works well in defocusing blurred images. Second, the proposed sharp region detection (SRD) module addresses blur ambiguities resulting from subtle defocus changes

in weakly-textured regions. SRD consists of convolution layers and a residual block, and allows the extraction of more powerful feature representations for image sharpness. Third, we also propose an efficient downsampling (EFD) module for the DfF framework. The proposed EFD combines output feature maps from upper scales using a stride convolution and a 3D convolution with maxpooling and incorporates them to both keep the feature representation of the original input and to ease the flow of informative features for focused regions. To optimize and generalize our network, we develop a high performance simulator to produce photo-realistic focal stack images with corresponding meta-data such as camera intrinsic parameters.

With this depth from focus network, we achieve state-of-the-art results over various public datasets as well as the top rank in the DDFB benchmark [12]. Ablation studies indicate that each of these technical contributions appreciably improves depth prediction accuracy.

2 Related Work

The mainstream approaches for depth prediction such as monocular depth estimation [6,8,9], stereo matching [3,32] and multiview stereo [10,16] use all-in-focus images. As mentioned above, they overlook the functional properties of off-the-shelf cameras and are out-of-scope for this work. In this section, we review depth from defocus blur images, which are closely related to our work.

Depth from Defocus. Some unsupervised monocular depth estimation [33,11] approaches utilize a defocus blur cue as a supervisory signal. A work in [33] proposes differentiable aperture rendering functions to train a depth prediction network which generates defocused images from input all-in-focus images. The network is trained by minimizing distances between ground truth defocused images and output defocused images based on an estimated depth map. Inspired by [33], a work in [11] introduces a fast differentiable aperture rendering layer from hypothesis of defocus blur. In spite of depth-guided defocus blur, both these works need all-in-focus images as input during an inference time. Anwar *et al.* [2] formulate a reblur loss based on circular blur kernels to regularize depth estimation, and design a CNN architecture to minimize input blurry images and reblurred images from output deblurring images as well. Zhang and Sun [40] propose a regularization term to impose a consistency between depth and defocus maps from single out-of-focus images.

Depth from DP images.

Starting with the use of traditional stereo matching, CNN-based approaches have been adopted for depth from DP images [36]. A work in [7] introduces that an affine ambiguity exists between a scene depth and its disparity from DP data, and then alleviates it using both novel 3D assisted loss and folded loss. In [41], a dual-camera with DP sensors is proposed to take advantage of both stereo matching and depth from DP images. In [27], unsupervised depth estimation by modeling a point spread function of DP cameras. The work in [24] proposes an end-to-end CNN for depth from single DP images using both defocus blur and

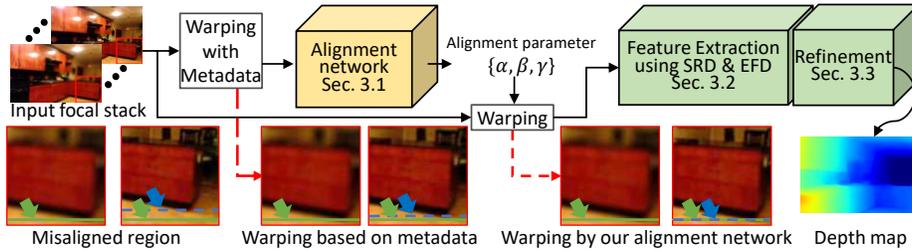


Fig. 2. An overview of the proposed network.

correspondence cues. In addition, they provide a simulator that makes a synthetic DP dataset from all-in-focus images and the corresponding depth map. In [39], single DP images are represented via multi-plane images [35] with a calibrated point spread function for a certain DP camera model. The representation is used for both unsupervised defocus map and all-in-focus image generation.

Depth from Focus. DfF accounts for changes in blur sizes in the focal stack and determines scene depths according to regions adjacent to the focus [26,19,21]. In particular, conventional DfF methods infer depth values from a focal stack by comparing the sharpness of a local window at each pixel [17,28,34]. The research in [12] introduces a CNN-based DfF by leveraging focal stack datasets made with light field and RGB-D cameras. In [21], domain invariant defocus blur is used for dealing with the domain gap. The defocus blur is supervised to train data-driven models for DfF as an intermediate step, and is then utilized for permutation-invariant networks to achieve a better generalization from synthetic datasets to real photos. In addition, the work uses a recurrent auto-encoder to handle scene movements which occur during focal sweeps¹. In [37], a CNN learns to make an intermediate attention map which is shared to predict scene depth prediction and all-in-focus images reconstruction from focal stack images.

3 Methodology

Our network is composed of two major components: One is an image alignment model for sequential defocused images. It is a prerequisite that we should first address the non-alignment issue on images captured with smartphones whose focus is relayed to focus motors adjusting locations of camera lenses. Another component is a focused feature representation, which encodes the depth information of scenes. To be sensitive to subtle focus changes, it requires two consecutive feature maps of the corresponding modules from our sharp region detector (SRD) and an effective downsampling module for defocused images (EFD). The overall procedure is depicted in Fig.2.

¹ Unfortunately, both the source codes for training/test and its pre-trained weight are not available in public.

3.1 A Network for Defocus Image Alignment

Since camera field of views (FoVs) vary according to the focus distance, a zoom-like effect is induced during a focal sweep [13], called focal breathing. Because of the focal breathing, an image sharpness cannot be accurately measured on the same pixel coordinates across focal slices. As a result, traditional DfF methods perform feature-based defocus image alignment to compensate this, prior to depth computations. However, recent CNN-based approaches disregard the focal breathing because either all public synthetic datasets for DfF/DfD, whose scale is enough to generalize CNNs well, provide well-aligned focal stacks, or are generated by single RGB-D images. Because of this gap between real-world imagery and easy to use datasets, their generality is limited. Therefore, as a first step to implementing a comprehensive, all-in-one solution to DfF, we introduce a defocus image alignment network.

Field of view. Scene FoVs are calculated by work distances, focus distances, and the focal length of cameras in Eq.(1). Since the work distances are fixed during a focal sweep, relative values of FoVs (Relative FoVs) are the same as the inverse distance between sensor and lens. We thus perform an initial alignment of a focal stack using these relative FoVs as follows:

$$FoV_n = W \times \frac{A}{s_n} \tag{1}$$

$$Relative\ FoV_n = \frac{FoV_n}{FoV_{min}} = \frac{s_{min}}{s_n} \quad (s_n = \frac{F_n \times f}{F_n - f}),$$

where s_n is the distance between the lens and the sensor in an n -th focal slice. A is the sensor size, and W is the working distance. f and F_n are the focal length of the lens and a focus distance, respectively. min denotes an index of a focal slice whose FoV is the smallest among focal slices. In this paper, we call this focal slice with the min index as the target focal slice. We note that the values are available by accessing the metadata information in cameras without any user calibration.

Nevertheless, the alignment step is not perfectly appropriate for focal stack images due to hardware limitations, as described in [13]. Most smartphone cameras control their focus distances by spring-installed voice coil motors (VCMs). The VCMs adjust the positions of the camera lens by applying voltages to a nearby electromagnet which induces spring movements. Since the elasticity of the spring can be changed by temperature and usage, there will be an error between real focus distances and values in the metadata. In addition, the principal point of cameras also changes during a focal sweep because the camera lens is not perfectly parallel to the image sensor, due to some manufacturing imperfections. Therefore, we propose an alignment network to adjust this mis-alignment and a useful simulator to ensure realistic focal stack acquisition.

Alignment network. As shown in Fig.3, our alignment network has 3-level encoder-decoder structures, similar to the previous optical flow network [15].

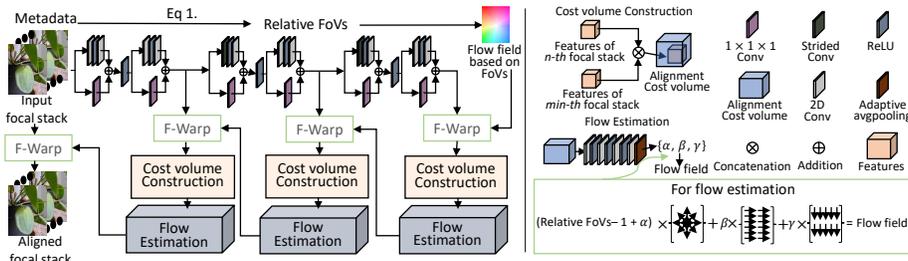


Fig. 3. An illustration of our alignment network. Given initially-aligned images with camera metadata, this network produces an aligned focal stack. In the flow estimation, we use three basis functions to model radial, horizontal and vertical motions of VCMs.

The encoder extracts multi-scale features, and multi-scale optical flow volumes are constructed by concatenating the features of a reference and a target focal slice. The decoder refines the multi-scale optical flow volumes in a coarse-to-fine manner using feature warping (F-warp). However, we cannot directly use the existing optical flow framework for alignment because defocus blur breaks the brightness constancy assumption [34].

To address this issue, we constrain the flow using three basis vectors with corresponding coefficients (α, β, γ) for each scene motion. To compute the coefficients instead of the direct estimation of the flow field, we add an adaptive average pooling layer to each layer of the decoder. The first basis vector accounts for an image crop which reduces errors in the FoVs. We elaborate the image crop as a flow that spreads out from the center. The remaining two vectors represent x - and y -axis translations, which compensate for errors in the principal point of the cameras. These parametric constraints of flow induce the network to train geometric features which are not damaged by defocus blur. We optimize this alignment network using a robust loss function L_{align} , proposed in [20], as follows:

$$L_{align} = \sum_{n=0}^N \rho(I_n(\Gamma + D(\Gamma)) - I_{min}(\Gamma)), \quad (2)$$

where $\rho(\cdot) = (|\cdot| + \varepsilon)^q$. q and ε are set to 0.4 and 0.01, respectively. I_n is a focal slice of a reference image, and I_{min} is the target focal slice. $D(\Gamma)$ is an output flow of the alignment network at a pixel position, Γ .

We note that the first basis might be insufficient to describe the zooming effects with spatially-varying motions. However, our design for the image crop shows consistently promising results for the alignment, thanks to the combination of the three basis functions that compensate for a variety of motions in real-world.

Simulator. Because public datasets do not describe changes in FoVs or hardware limitations in off-the-shelf cameras, we propose a useful simulator to render realistic sequential defocus images for training our alignment network. Here, the most important part is to determine the error ranges of the intrinsic camera pa-

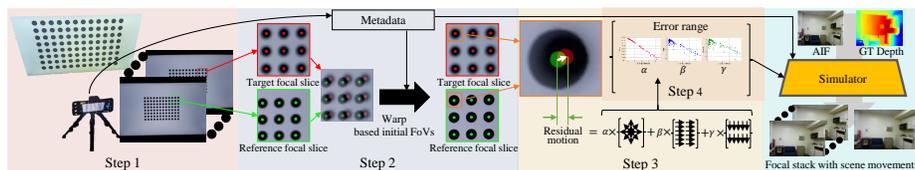


Fig. 4. A pipeline of our simulator. Red and green dots mean a center of a circle. The misalignment error occurs due to inaccurate intrinsic parameters. Our simulator produces misaligned focal stack images because of the hardware limitations for autofocus.

rameters, such as principal points and focal distances. We estimate them as the following process in Fig.4: (1) We capture circle patterns on a flat surface using various smartphone models by changing focus distances. (2) We initially align focal stacks with the recorded focus distances. (3) After the initial alignment, we decompose the residual motions of the captured circles using 3 basis vectors, image crop and x - and y -axis translations. (4) We statistically calculate the error ranges of the principal points and focus distances from the three parameters of the basis vectors. Given metadata of cameras used, our simulator renders focal stacks induced from blur scales based on the focus distance and the error ranges of the basis vector.

3.2 Focal Stack-oriented Feature Extraction

For high-quality depth prediction, we consider three requirements that must be imposed on our network. First, to robustly measure focus in the feature space, it is effective to place a gap space in the convolution operations, as proved in [17]. In addition, even though feature downsampling such as a convolution with strides and pooling layers is necessary to reduce the computations in low-level computer vision tasks like stereo matching [22], such downsampling operations can make a defocused image and its feature map sharper. This fails to accurately determine the focus within the DfF framework. Lastly, feature representations for DfF need to identify subtle distinctions in blur magnitudes between input images.

Initial feature extraction. In an initial feature extraction step, we utilize a dilated convolution to extract focus features. After the dilated convolution, we extract feature pyramids to refine the focal volumes in the refinement step. Given an input focal stack $S \in R^{H*W*N*3}$ where H , W and N denote the height, width and the number of focal slices respectively, we extract three pyramidal feature volumes whose size is $H/2^L \times W/2^L \times N \times C * 2^L$ where $L \in \{0, 1, 2\}$ and C is the number of channels in the focal volume. This pyramidal feature extraction consists of three structures in which SRD and EFD are iteratively performed, as described in Fig.5. Each pyramidal feature volume is then used as the input to the next EFD module. The last one is utilized as an input of the multi-scale feature aggregation step in Sec.3.3.

Sharp Region Detector. The initial feature of each focal slice is needed to communicate with other neighboring focal slices, to measure the focus of the pixel of interest. A work in [21] extracts focus features using a global pooling

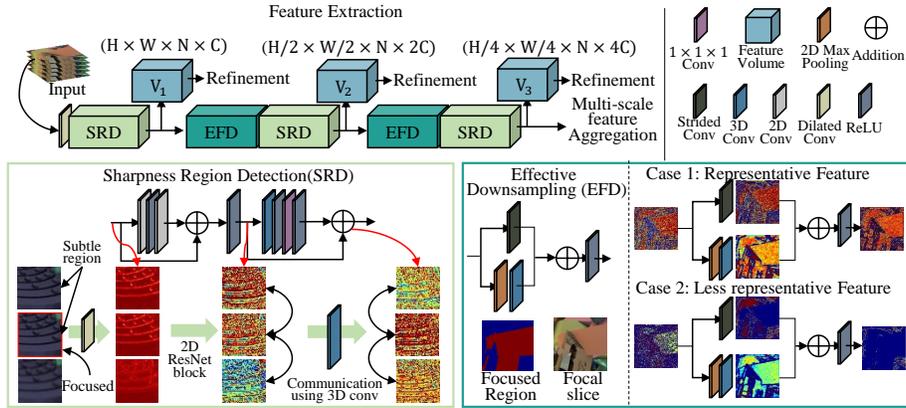


Fig. 5. An architecture of our feature extraction. If feature maps from neighbor focal slices have similar values, our SRD gives an attention score to the sharpest focal slice. Our EFD preserves informative defocus feature representation during downsampling.

layer as a communication tool across a stack dimension. However, we observe that the global pooling layer causes a loss of important information due to its inherent limitation that all values across focal slices become single values.

Using our SRD module consisting of both a 2D convolution and a 3D convolution, we overcome the limitation. In Fig.5 (left), we extract features using a 2D ResNet block and add an attention score which is computed from them by 3D convolutions and a ReLU activation. The 3D convolution enables the detection of subtle defocus variations in weakly texture-less regions by communicating the features with neighbor focal slices. With this module, our network encodes more informative features for the regions than previous works [21,37]. **Effective Downsampling.** Unlike stereo matching networks that use convolutions with strides for downsampling features [3,32], the stride of a convolution causes a loss in spatial information because most of the focused regions may not be selected. As a solution to this issue, one previous DfF work [21] uses a combination of maxpooling and average pooling with the feature extraction step.

Inspired by [21], we propose a EFD module leveraging a well-known fact that a feature has higher activation in a focused region than weakly textured regions in Fig.5 (right). The EFD module employs a 2D max-pooling as a downsampling operation and applies a 3D convolution to its output. Through our EFD module, our network can both take representative values of focused regions in a local window and communicate the focal feature with neighbor focal slices.

3.3 Aggregation and Refinement

Our network produces a final depth map after multi-scale feature aggregation and refinement steps.

Multi-scale feature aggregation. The receptive field of our feature extraction module might be too small to learn non-local features. Therefore, we propose

a multi-scale feature aggregation module using one hour-glass module to expand the receptive field, which is similar to the stereo matching network in [32]. At an initial step, we use three different sizes of kernels (2×2 , 4×4 , 8×8) in the average pooling layer. Unlike [32], the reason for using average pooling is to avoid a memory consumption issue because DfF requires more input images. We then apply a ResBlock on each output of average pooling in order to extract multi-scale features. These features are embedded into the encoder and aggregated by the decoder of the hour-glass module. The aggregated feature volume is utilized as an input in the refinement step.

Refinement and Regression. The refinement module has three hour-glass modules with skip-connections like [3]. Here, we add transposed convolutions to resize the output of each hourglass whose size is the same as each level of a pyramidal feature volume from the feature extraction module. We construct an input focal volume of each hourglass by concatenating pyramidal feature volumes of the feature extraction module with the output focal volume of the previous hourglass. As each hourglass handles increasingly higher resolutions with pyramidal feature volumes, the focal volumes are refined in a coarse to fine manner. To obtain a depth map from the output focal volumes, we multiply a focus distance value and the probability of each focus distance leading to maximal sharpness. The probability is computed by applying a normalized soft-plus in the output focal volumes in a manner similar to [37]. The whole depth prediction network is optimized using a weighted loss function L_{depth} from scratch as follows:

$$L_{depth} = \sum_{i=1}^4 w_i * \|D_i - D_{gt}\|_2 \quad (3)$$

where $\|\cdot\|_2$ means a l_2 loss and D_{gt} indicates a ground truth depth map. $i \in \{1, 2, 3, 4\}$ means the scale level of the hour-glass module. In our implementation, we set w_i to 0.3, 0.5, 0.7 and 1.0, respectively.

Implementation details. We train our network using the following strategy: (1) We first train the alignment network in Sec.3.1 during 100 epochs using the alignment loss in Eq.(2). (2) We freeze the alignment network and merge it with the depth prediction network. (3) We train the merged network during 1500 epochs with the depth loss in Eq.(3). (4) In an inference step, we can estimate the depth map from the misaligned focal stack in an end-to-end manner. We note that our network is able to use an arbitrary number of images as input, like the previous CNN-based DfF/DfD [21,37]. The number of parameters of our alignment network and feature extraction module is 0.195M and 0.067M, respectively. And, the multi-scale feature aggregation module and the refinement module have 2.883M and 1.067M learnable parameters, respectively. That’s, the total parameters of our network is 4.212M. We implement our network using a public PyTorch framework [25], and optimize it using Adam optimizer [18] ($\beta_1 = 0.9$, $\beta_2 = 0.99$) with a learning rate 10^{-3} . Our model is trained on a single NVIDIA RTX 2080Ti GPU with 4 mini-batches, which usually takes three days. For data augmentation, we apply random spatial transforms (rotation, flipping and cropping) and color jittering (brightness, contrast and gamma correction).

Table 1. Quantitative evaluation on DDFFF 12-Scene [12]. We directly refer to the results from [37]. Since the result of DefocusNet [21] is not uploaded in the official benchmark, we only bring the MAE value from [21]. **bold**: Best, Underline: Second best. Unit: pixel.

Method	MAE ↓	RMSE log ↓	AbsRel ↓	SqRel ↓	Bump ↓	$\delta = 1.25 \uparrow$	$\delta = 1.25^2 \uparrow$	$\delta = 1.25^3 \uparrow$
Lytro	$2.1e^{-3}$	0.31	0.26	0.01	1.0	55.65	82.00	93.09
VDFF [23]	$7.3e^{-3}$	1.39	0.62	0.05	0.8	8.42	19.95	32.68
PSP-LF [42]	$2.7e^{-3}$	0.45	0.46	<u>0.03</u>	0.5	39.70	65.56	82.46
PSPNet [42]	$9.4e^{-4}$	<u>0.29</u>	0.27	0.01	<u>0.6</u>	62.66	85.90	<u>94.42</u>
DFLF [12]	$4.8e^{-3}$	0.59	0.72	0.07	0.7	28.64	53.55	71.61
DDFF [12]	$9.7e^{-4}$	0.32	0.29	0.01	<u>0.6</u>	61.95	85.14	92.98
DefocusNet [21]	$9.1e^{-4}$	-	-	-	-	-	-	-
AiFDDepthNet [37]	<u>$8.6e^{-4}$</u>	<u>0.29</u>	<u>0.25</u>	0.01	<u>0.6</u>	<u>68.33</u>	<u>87.40</u>	93.96
Ours	$5.6e^{-4}$	0.21	0.17	0.01	<u>0.6</u>	77.96	93.72	97.94

4 Evaluation

We compare the proposed network with state-of-the-art methods related to DfD, DfF and depth from light field images. We also conduct extensive ablation studies to demonstrate the effectiveness of each component of the proposed network. For quantitative evaluation, we use standard metrics as follows: mean absolute error (MAE), mean squared error (MSE), absolute relative error (AbsRel), square relative error (SqRel), root mean square error (RMSE), log root-mean-squared error (RMSE log), bumpiness (Bump), inference time (Secs) and accuracy metric $\delta_i = 1.25^i$ for $i \in \{1, 2, 3\}$. Following [37], we exclude pixels whose depth ranges are out of focus distance at test time.

4.1 Comparisons to State-of-the-art Methods

We validate the robustness of the proposed network by showing experimental results on various public datasets: DDFFF 12-Scene [12], DefocusNet Dataset [21], 4D Light Field Dataset [14], Smartphone [13] as well as focal stack images generated from our simulator. The datasets provide pre-aligned defocused images. We use the training split of each dataset to build our depth estimation network in both Sec.3.2 and Sec.3.3 from scratch, and validate it on the test split.

DDFF 12-Scene [12]. DDFFF 12-Scene dataset provides focal stack images and its ground truth depth maps captured by a light-field camera and a RGB-D sensor, respectively. The images have shallow DoFs and show texture-less regions. Our method shows the better performance than those of recent published works in Tab.1 and achieves the top rank in almost evaluation metrics of the benchmark site².

DefocusNet Dataset [21]. This dataset is rendered in a virtual space and generated using Blender Cycles renderer [4]. Focal stack images consist of only five defocused images whose focus distances are randomly sampled in an inverse depth space. The quantitative results are shown in Tab.2. As shown in Fig.6, our method successfully reconstructs the smooth surface and the sharp depth discontinuity rather than previous methods.

² <https://competitions.codalab.org/competitions/17807#results>

Table 2. Quantitative evaluation on DefocusNet dataset [21] (unit: meter), 4D Light Field dataset [14] (unit: pixel) and Smartphone dataset [13] (unit: meter). For DefocusNet dataset and 4D Light Field dataset, we directly refer to the results from [37]. For Smartphone dataset [13], we multiply confidence scores on metrics ('MAE' and 'MSE') which are respectively denoted as 'MAE*' and 'MSE*'. **bold:** Best.

Method	DefocusNet Dataset [21]			4D Light Field [14]			Smartphone [13]		
	MAE ↓	MSE ↓	AbsRel ↓	MSE ↓	RMSE ↓	Bump ↓	MAE* ↓	MSE* ↓	Secs ↓
DefocusNet [21]	0.0637	0.0175	0.1386	0.0593	0.2355	2.69	0.1650	0.0800	0.1598
AiFDepthNet [37]	0.0549	0.0127	0.1115	0.0472	0.2014	1.58	0.1568	0.0764	0.1387
Ours	0.0403	0.0087	0.0809	0.0230	0.1288	1.29	0.1394	0.0723	0.1269

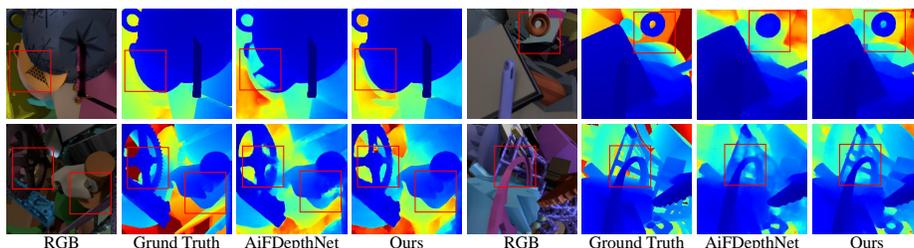


Fig. 6. Examples of depth prediction from AiFDepthNet and ours on DefocusNet dataset.

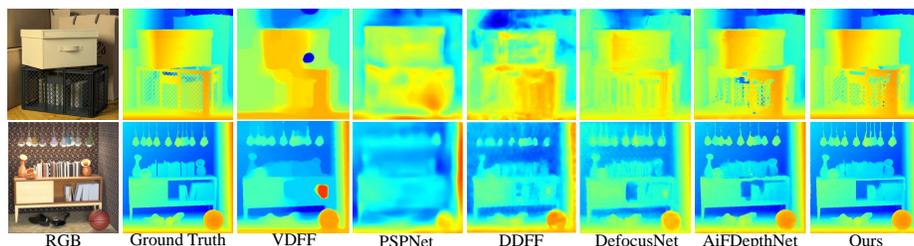


Fig. 7. Qualitative comparison on 4D Light Field dataset.

4D Light Field Dataset [14]. This synthetic dataset has 10 focal slices with shallow DoFs for each focal stack. The number of focal stacks in training and test split is 20 and 4, respectively. For fair comparison on this dataset, we follow the evaluation protocol in the relevant work [37]. In qualitative comparisons Fig.7, our SRD and EFD enable to capture sharp object boundaries like the box and fine details like lines hanging from the ceiling. In quantitative evaluation of Tab.2 the MSE and RMSE are half of them from the comparison methods [1,38].

Smartphone [13]. This dataset shows real-world scenes captured from Pixel 3 smartphones. Unlike previous datasets, ground truth depth maps are obtained by multiview stereo [30,31] and its depth holes are not considered in the evaluation. As expected, our network achieves the promising performance over the state-of-the-art methods, whose results are reported in Tab.2 and Fig.8. We note that our method consistently yields the best quality depth maps from focal stack images regardless of dataset, thanks to our powerful defocused feature representations using both SRD and EFD.

Generalization across different datasets. Like [37], we demonstrate the generality of the proposed network. For this, we train our network on Flyingth-

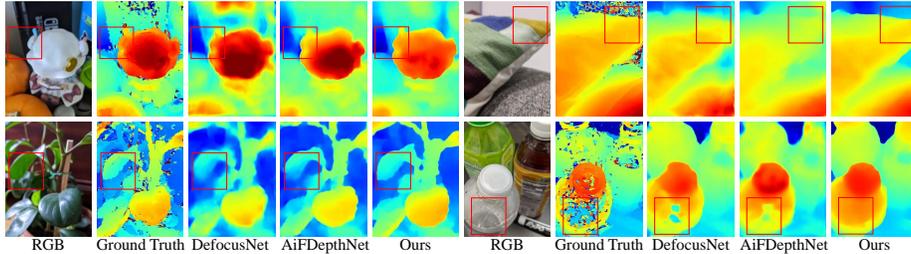


Fig. 8. Qualitative results on Smartphone dataset.

Table 3. Quantitative result across different datasets for generalization of the state-of-the-art methods and ours. We train our depth prediction model on FlyingThings3D and test them on Middlebury stereo (unit: pixel) and DefocusNet dataset (unit: meter). For fair comparison, we directly refer the results of the works [37,21] from [37].

Method	Train Dataset	Test Dataset	MAE ↓	MSE ↓	RMSE ↓	AbsRel ↓	SqRel ↓
DefocusNet [21]			7.408	157.440	9.079	0.231	4.245
AiFDepthNet [37]	FlyingThings3d	Middlebury	3.825	58.570	5.936	0.165	3.039
Ours			1.645	9.178	2.930	0.068	0.376
DefocusNet [21]			0.320	0.148	0.372	1.383	0.700
AiFDepthNet [37]	FlyingThings3d	DefocusNet	0.183	0.080	0.261	0.725	0.404
Ours			0.163	0.076	0.259	0.590	0.360

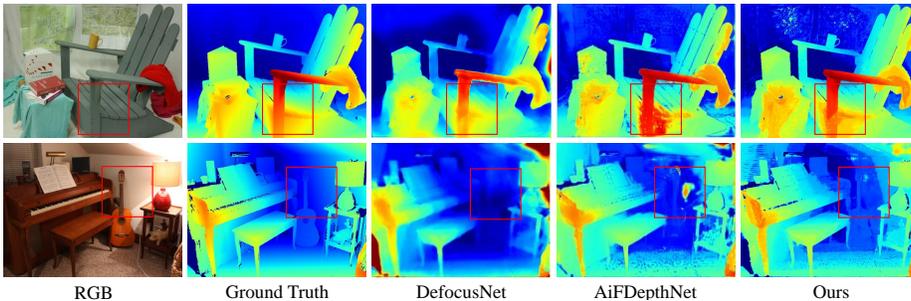


Fig. 9. Qualitative results on Middlebury dataset.

ings3D [22] which is a large-scale synthetic dataset, and test it on two datasets including Middlebury Stereo [29] and DefocusNet dataset [21]. As shown in Tab.3 and Fig.9, our network still shows impressive results on both datasets.

4.2 Ablation studies

We carry out extensive ablation studies to demonstrate the effectiveness of each module of the proposed network.

Alignment network. We first evaluate our alignment network. To do this, we render focal stacks using our simulator which generates defocused images based on a camera metadata. We test our alignment network in consideration of four cases: 1) without any warping, 2) with only initial FoVs in (1), 3) a classical homography method [5], 4) our alignment network with initial FoVs. The quantitative results are reported in Tab.4, whose example is displayed in Fig.10. The results demonstrate that our alignment network achieves much faster and competent performance with the classic homography-based method.

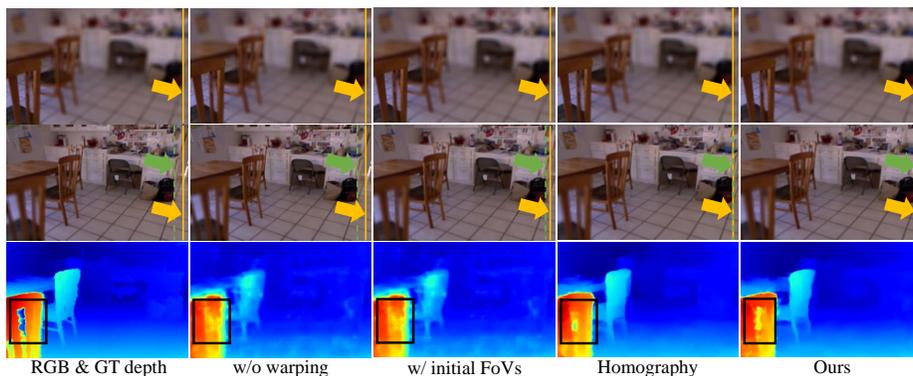


Fig. 10. Ablation study on our alignment network. The first and second row refer a target and reference focal slice whose FoVs have the smallest and the biggest values, respectively. The third row shows depth estimation results in accordance to the alignment methods.

Table 4. Ablation study for alignment network. Unit: meter

Module	MAE ↓	MSE ↓	RMSE log ↓	AbsRel ↓	SqRel ↓	$\delta = 1.25 \uparrow$	$\delta = 1.25^2 \uparrow$	$\delta = 1.25^3 \uparrow$	Secs ↓	GPU
w/o alignment	0.0247	0.0014	0.0915	0.0067	0.0034	0.9707	0.9970	0.9995	0.0107	2080Ti
w/ initial FoVs	0.0165	0.0009	0.0636	0.0400	0.0019	0.9867	0.9976	0.9994	0.0358	2080Ti
Homography-based	0.0151	0.0007	0.0570	0.0369	0.0015	0.9907	0.9986	0.9997	0.8708	R3600
Ours	0.0151	0.0007	0.0578	0.0365	0.0016	0.9898	0.9984	0.9996	0.0923	2080Ti

SRD and EFD. We compare our modules with other feature extraction modules depicted in Fig.11. We conduct this ablation study on DefocusNet dataset [21] because it has more diverse DoF values than other datasets. The quantitative result is reported in Tab.5.

When we replace our SRD module with either 3D ResNet block or 2D ResNet block only, there are performance drops, even with more learnable parameters for the 3D ResNet block. We also compare our EFD module with four replaceable modules: max-pooling+3D Conv, average pooling+3D Conv, Stride convolution and 3D pooling layer. As expected, our EFD module achieves the best performance because it allows better gradient flows preserving defocus property.

Number of focal slices. Like previous DfF networks [37,21], our network can handle an arbitrary number of focal slices by the virtue of 3D convolutions. Following the relevant work [37], we train our network from three different ways, whose result is reported in Fig.12: The '5' means a model trained using five focal slices; The 'Same' denotes that the number of focal slices in training and test phase is same; The 'Random' is a model trained using an arbitrary number of focal slices.

The '5' case performs poorly when the different number of focal slices is used in the test phase, and the 'Same' case shows promising performances. Nevertheless, the 'Random' case consistently achieves good performances regardless of the number of focal slices

5 Conclusion

In this paper, we have presented a novel and true end-to-end DfF architecture. To do this, we first propose a trainable alignment network for sequential de-

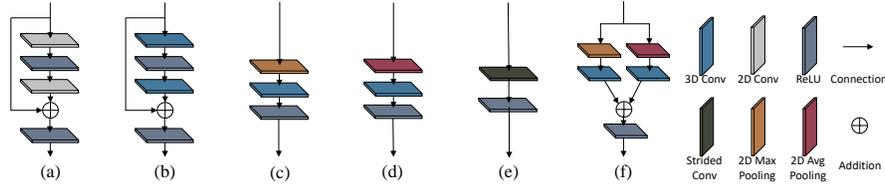


Fig. 11. Candidate modules of our SRD and EFD. (a) 2D ResNet block, (b) 3D ResNet block, (c) Max pooling + 3D Conv, (d) Average pooling + 3D Conv, (e) Strided Conv and (f) 3D pooling layer.

Table 5. Ablation studies for SRD and EFD. Unit: meter

Module	MAE ↓	MSE ↓	RMSE log ↓	AbsRel ↓	SqRel ↓	$\delta = 1.25 \uparrow$	$\delta = 1.25^2 \uparrow$	$\delta = 1.25^3 \uparrow$
SRD → 2D ResNet block	0.0421	0.0095	0.1614	0.0842	0.0142	0.9082	0.9722	0.9873
SRD → 3D ResNet block	0.0409	0.0088	0.1576	0.0818	0.0128	0.9123	0.9725	0.9891
EFD → Maxpooling + 3D Conv	0.0421	0.0094	0.1622	0.0845	0.0143	0.9125	0.9712	0.9849
EFD → Avgpooling + 3D Conv	0.0422	0.0097	0.1628	0.0830	0.0141	0.9126	0.9718	0.9860
EFD → Strided Conv	0.0419	0.0091	0.1630	0.0842	0.0135	0.9144	0.9725	0.9867
EFD → 3D Pooling Layer	0.0414	0.0089	0.1594	0.0843	0.0132	0.9088	0.9747	0.9886
Ours	0.0403	0.0087	0.1534	0.0809	0.0130	0.9137	0.9761	0.9900

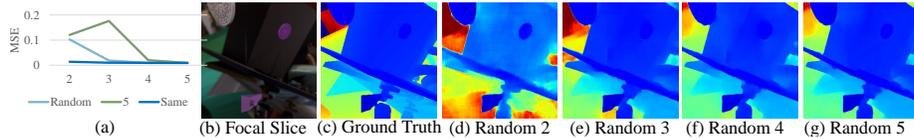


Fig. 12. (a) The performance change according to the number of focal slices in training and test phase. (b) One of focal slices and (c) its ground truth depth map. (d) to (g) output depth maps on the random number of input focal slices in training phase.

focused images. We then introduce a novel feature extraction and an efficient downsampling module for robust DfF tasks. The proposed network achieves the best performance in the public DfF/DfD benchmark and various evaluations.

Limitation. There are still rooms for improvements. A more sophisticated model for flow fields in the alignment network would enhance depth prediction results. More parameters can be useful for extreme rotations. Another direction is to make depth prediction better by employing focal slice selection like defocus channel attention in the aggregation process.

Acknowledgement This work is in part supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) (No.2021-0-02068, Artificial Intelligence Innovation Hub), Vehicles AI Convergence Research & Development Program through the National IT Industry Promotion Agency of Korea (NIPA), ‘Project for Science and Technology Opens the Future of the Region’ program through the INNOPOLIS FOUNDATION (Project Number: 2022-DD-UP-0312) funded by the Ministry of Science and ICT (No.S1602-20-1001), the National Research Foundation of Korea (NRF) (No. 2020R1C1C10 12635) grant funded by the Korea government (MSIT), the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program (P0019797), and the GIST-MIT Collaboration grant funded by the GIST in 2022.

References

1. Abuolaim, A., Brown, M.S.: Defocus deblurring using dual-pixel data. In: Proceedings of European Conference on Computer Vision (ECCV) (2020)
2. Anwar, S., Hayder, Z., Porikli, F.: Depth estimation and blur removal from a single out-of-focus image. In: BMVC (2017)
3. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
4. Community, B.O.: Blender—a 3d modelling and rendering package. Blender Foundation (2018)
5. Evangelidis, G.D., Psarakis, E.Z.: Parametric image alignment using enhanced correlation coefficient maximization. *IEEE transactions on pattern analysis and machine intelligence* **30**(10), 1858–1865 (2008)
6. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
7. Garg, R., Wadhwa, N., Ansari, S., Barron, J.T.: Learning single camera depth estimation using dual-pixels. In: Proceedings of International Conference on Computer Vision (ICCV) (2019)
8. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
9. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of International Conference on Computer Vision (ICCV) (2019)
10. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
11. Gur, S., Wolf, L.: Single image depth estimation trained via depth from defocus cues. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
12. Hazirbas, C., Soyer, S.G., Staab, M.C., Leal-Taixé, L., Cremers, D.: Deep depth from focus. In: Proceedings of Asian Conference on Computer Vision (ACCV) (2018)
13. Herrmann, C., Bowen, R.S., Wadhwa, N., Garg, R., He, Q., Barron, J.T., Zabih, R.: Learning to autofocus. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
14. Honauer, K., Johannsen, O., Kondermann, D., Goldluecke, B.: A dataset and evaluation methodology for depth estimation on 4d light fields. In: Proceedings of Asian Conference on Computer Vision (ACCV) (2016)
15. Hui, T.W., Tang, X., Loy, C.C.: Liteflownet: A lightweight convolutional neural network for optical flow estimation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
16. Im, S., Jeon, H.G., Lin, S., Kweon, I.S.: Dpsnet: End-to-end deep plane sweep stereo. *arXiv preprint arXiv:1905.00538* (2019)
17. Jeon, H.G., Surh, J., Im, S., Kweon, I.S.: Ring difference filter for fast and noise robust depth from focus. *IEEE Transactions on Image Processing* **29**, 1045–1060 (2019)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)

19. Levin, A., Fergus, R., Durand, F., Freeman, W.T.: Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)* **26**(3), 70–es (2007)
20. Liu, P., King, I., Lyu, M.R., Xu, J.: DdfLOW: Learning optical flow with unlabeled data distillation. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (2019)
21. Maximov, M., Galim, K., Leal-Taixé, L.: Focus on defocus: bridging the synthetic to real domain gap for depth estimation. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020)
22. Mayer, N., Ilg, E., Haussler, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016)
23. Moeller, M., Benning, M., Schönlieb, C., Cremers, D.: Variational depth from focus reconstruction. *IEEE Transactions on Image Processing* **24**(12), 5369–5378 (2015)
24. Pan, L., Chowdhury, S., Hartley, R., Liu, M., Zhang, H., Li, H.: Dual pixel exploration: Simultaneous depth estimation and image restoration. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021)
25. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32**, 8026–8037 (2019)
26. Pertuz, S., Puig, D., Garcia, M.A.: Analysis of focus measure operators for shape-from-focus. *Pattern Recognition* **46**(5), 1415–1432 (2013)
27. Punnappurath, A., Abuolaim, A., Affi, M., Brown, M.S.: Modeling defocus-disparity in dual-pixel sensors. In: *2020 IEEE International Conference on Computational Photography (ICCP)*. pp. 1–12. IEEE (2020)
28. Sakurikar, P., Narayanan, P.: Composite focus measure for high quality depth maps. In: *Proceedings of International Conference on Computer Vision (ICCV)* (2017)
29. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: *German conference on pattern recognition*. Springer (2014)
30. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016)
31. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: *Proceedings of European Conference on Computer Vision (ECCV)* (2016)
32. Shen, Z., Dai, Y., Rao, Z.: Cfnet: Cascade and fused cost volume for robust stereo matching. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021)
33. Srinivasan, P.P., Garg, R., Wadhwa, N., Ng, R., Barron, J.T.: Aperture supervision for monocular depth estimation. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018)
34. Suwajanakorn, S., Hernandez, C., Seitz, S.M.: Depth from focus with your mobile phone. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015)
35. Tucker, R., Snavely, N.: Single-view view synthesis with multiplane images. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020)

36. Wadhwa, N., Garg, R., Jacobs, D.E., Feldman, B.E., Kanazawa, N., Carroll, R., Movshovitz-Attias, Y., Barron, J.T., Pritch, Y., Levoy, M.: Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics (ToG)* **37**(4), 1–13 (2018)
37. 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: *Proceedings of International Conference on Computer Vision (ICCV)* (2021)
38. Wanner, S., Goldluecke, B.: Globally consistent depth labeling of 4d light fields. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE (2012)
39. Xin, S., Wadhwa, N., Xue, T., Barron, J.T., Srinivasan, P.P., Chen, J., Gkioulekas, I., Garg, R.: Defocus map estimation and deblurring from a single dual-pixel image. In: *Proceedings of International Conference on Computer Vision (ICCV)* (2021)
40. Zhang, A., Sun, J.: Joint depth and defocus estimation from a single image using physical consistency. *IEEE Transactions on Image Processing* **30**, 3419–3433 (2021)
41. Zhang, Y., Wadhwa, N., Orts-Escolano, S., Häne, C., Fanello, S., Garg, R.: Du 2 net: Learning depth estimation from dual-cameras and dual-pixels. In: *Proceedings of European Conference on Computer Vision (ECCV)* (2020)
42. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017)