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

* Corresponding author
This paper is the short version of ECCV'22 and is NEVER considered an official publication.

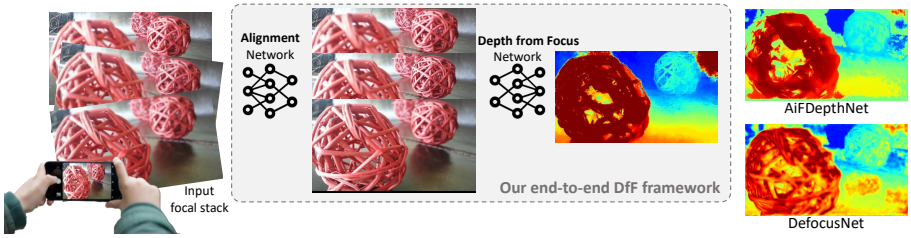

Fig. 1. Results of our true end-to-end DfF framework with comparisons to state-of-the-art methods.

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 and limited DoFs. A straightforward way is to find correspondences between the left and right sub-images [3]. 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 [11]. 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 [11]. 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 [8]. These shortcomings motivate us to examine depth from focus (DfF) as an alternative.

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 recent CNN-based DfD/DfF works [4,10,14] 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. Third, we also propose an efficient down-sampling (EFD) module for the DfF framework. 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 DDFF benchmark [4]. Ablation studies indicate that each of these technical contributions appreciably improves depth prediction accuracy.

2 Methodology

Our network is composed of two major components: One is an image alignment model for sequential defocused images. Another component is a focused feature representation, which encodes the depth information of scenes.

2.1 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 [5], called focal breathing. Because of the focal breathing, an image sharpness cannot be accurately measured on the same pixel coordinates across focal slices. 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. 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. We note that needed values to calculate relative FoVs 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 [5]. 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.2, our alignment network has 3-level encoder-decoder structures, similar to the previous optical flow network [7]. 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 [13]. 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

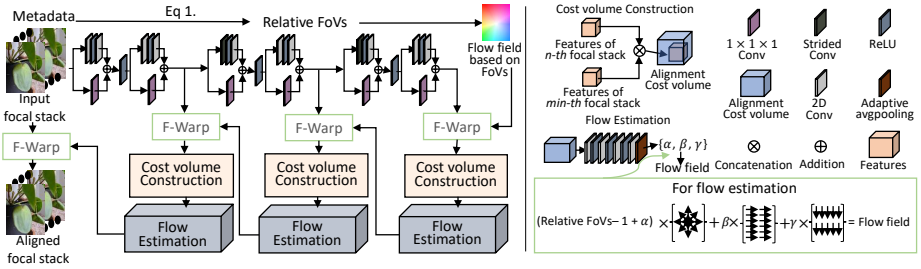

Fig. 2. An illustration of our alignment network. Given initially-aligned images with camera metadata, this network produces an aligned focal stack. In the flow estimation, we use three basis functions to model radial, horizontal and vertical motions of VCMs.

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 [9], as follows:

$$L_{align} = \sum_{n=0}^{N} \rho(I_n(\Gamma + D(\Gamma)) - I_{min}(\Gamma)), \tag{1}$$

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.

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. 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.

2.2 Focal Stack-oriented Feature Extraction

For high-quality depth prediction, we consider two requirements that must be imposed on our network. First: feature downsampling such as a convolution with strides and pooling layers is necessary to reduce the computations in low-level computer vision task. Second: Feature representations for DfF need to identify subtle distinctions in blur magnitudes between input images.

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. In Fig.3 (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

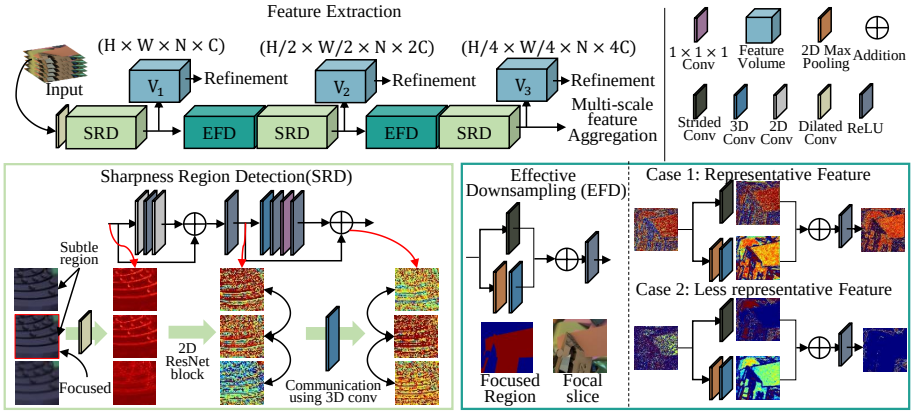

Fig. 3. An architecture of our feature extraction. If feature maps from neighbor focal slices have similar values, our SRD gives an attention score to the sharpest focal slice. Our EFD preserves informative defocus feature representation during downsampling.

Table 1. Quantitative evaluation on DDFF 12-Scene [4]. We directly refer to the results from [14]. Since the result of DefocusNet [10] is not uploaded in the official benchmark, we only bring the MSE value from [10]. **bold**: Best, Underline: Second best. Unit: pixel.

Method	MSE #	RMSE log #	AbsRel #	SqRel #	Bump #	$\delta = 1.25$ "	$\delta = 1.25^2$ "	$\delta = 1.25^3$ "
DDFF [4]	$9.7e^{-4}$	0.32	0.29	**0.01**	0.6	61.95	85.14	92.98
DefocusNet [10]	$9.1e^{-4}$	-	-	-	-	-	-	-
AiFDepthNet [14]	$8.6e^{-4}$	0.29	0.25	**0.01**	0.6	68.33	87.40	93.96
Ours	$\mathbf{5.7e^{-4}}$	**0.21**	**0.17**	**0.01**	0.6	**77.96**	**93.72**	**97.94**

variations in weakly texture-less regions by communicating the features with neighbor focal slices.

EFfective Downsampling. Unlike stereo matching networks that use convolutions with strides for downsampling features [12], the stride of a convolution causes a loss in spatial information because most of the focused regions may not be selected. 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 Evaluation

3.1 Comparisons to State-of-the-art Methods

We validate the robustness of the proposed network by showing experimental results on various public pre-aligned datasets.

DDFF 12-Scene [4]. DDFF 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

Table 2. Quantitative evaluation on DefocusNet dataset [10] (unit: meter), 4D Light Field dataset [6]) and Smartphone dataset [5] (unit: meter). For DefocusNet dataset and 4D Light Field dataset.For Smartphone dataset [5], we multiply confidence scores on metrics ('MAE' and 'MSE') which are respectively denoted as 'MAE*' and 'MSE*'.

	DefocusNet Dataset [10]			4D Light Field [6]			Smartphone [5]		
Method	MAE #	MSE #	AbsRel #	MSE #	RMSE #	Bump #	MAE* #	MSE* #	Secs #
DefocusNet [10]	0.0637	0.0175	0.1386	0.0593	0.2355	2.69	0.1650	0.0800	0.1598
AiFDepthNet [14]	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**

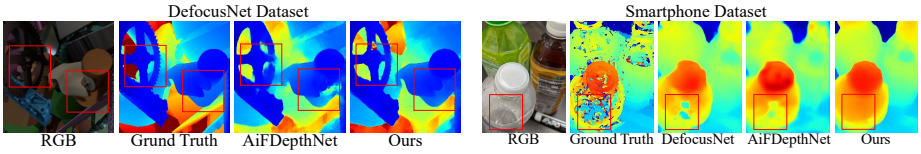

Fig. 4. Qualitative results on DefocusNet dataset and Smartphone dataset.

works in Tab.1 and achieves the top rank in almost evaluation metrics of the benchmark site.

DefocusNet Dataset [10]. This dataset is rendered in a virtual space and generated using Blender Cycles renderer [1]. 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.4, our method successfully reconstructs the smooth surface and the sharp depth discontinuity rather than previous methods.

4D Light Field Dataset [6]. 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.

Smartphone [5]. This dataset shows real-world scenes captured from Pixel 3 smartphones.As expected, our network achieves the promising performance over the state-of-the-art methods, whose results are reported in Tab.2 and Fig.4.

Table 3. Ablation studies for SRD and EFD.

Module	MAE #	MSE #	RMSE log #	AbsRel #	SqRel #	$\delta = 1.25$ "	$\delta = 1.25^2$ "	$\delta = 1.25^3$ "
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 Poolying 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**

3.2 Ablation Study

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

Alignment network. To evaluate our alignment network, we render focal stacks using our simulator which generates defocused images based on a camera

https://competitions.codalab.org/competitions/17807#results

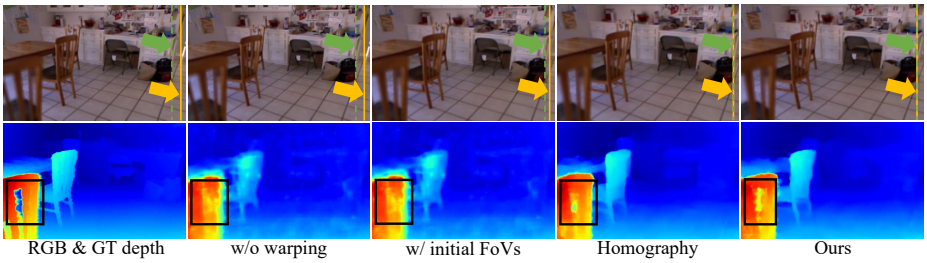

Fig. 5. Ablation study on our alignment network. The first row refer a target and reference focal slice whose FoVs have the smallest and the biggest values, respectively. The second row shows depth estimation results in accordance to the alignment methods. Homography denotes a classical homography method [2].

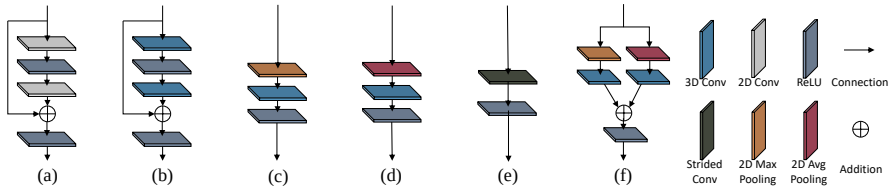

Fig. 6. Candidate modules of our SRD and EFD. (a) 2D ResNet block, (b) 3D ResNet block, (c) Max pooling + 3D Conv, (d) Average pooling + 3D Conv, (e) Strided Conv and (f) 3D pooling layer.

metadata.The qualitative results are reported in Fig.5. By using GPUs, our alignment network achieves much faster and competable performance with the classic homography-based method.

SRD and EFD. We compare our modules with other feature extraction modules depicted in Fig.6. The quantitative result is reported in Tab.3.

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.

4 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 defocused 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.

Remarks. This paper is a summary presentation of the paper which has been published in ECCV2022 by request of the IW-FCV2023 program committee to share the research results.

References

1. Community, B.O.: Blender—a 3d modelling and rendering package. Blender Foundation (2018)
2. 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)
3. 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)
4. 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)
5. 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)
6. 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)
7. 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)
8. 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)
9. 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)
10. 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)
11. 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)
12. 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)
13. 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)
14. 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)