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

1 Detail of Depth Estimation Network

As shown in Fig.1, we depict our depth estimation network in detail. In our network, output depth map D_i is produced based on probability distributions $P_{i,j}$ of feature maps $K_{i,j}$ from the refinement network where i and j denote an index of the hour-glass level and of focal slices, respectively. In this paper, this is defined as below:

$$D_i = \sum_{j=1}^{N} P_{i,j} * F_j \quad \text{s.t.} \quad P_{i,j} = \frac{\ln(1+\exp(K_i))}{\sum_{j=1}^{N}\left(\ln\left(1+\exp(K_{i,j})\right)\right)} \tag{1}$$

where F_j is a focus distance of the j-th focal slice and N is the number of focal slices of the input focal stack. $*$ means an element-wise multiplication.

2 AIF reconstruction

As discussed in [3], if we use a SoftMax at the last layer of our depth estimation network instead of SoftPlus normalization, all-in-focus images are obtained as below:

$$I_{AiF_i} = \sum_{j=1}^{N} P_{i,j}^{AiF} * I_j \tag{2}$$

$$P_{i,j}^{AiF} = \frac{\exp(K_i)}{\sum_{j}^{N}\exp(K_{i,j})} \tag{3}$$

where $P_{i,j}^{AiF}$ is the probability distribution for all-in-focus image reconstruction. I_{AiF_i} is the AiF image from the $i-th$ hourglass. I_j is the $j-th$ focal slice.

Since the quality of all-in-focus images depends of its depth quality in general, our network shows better performances than that of [3] in Tab.1 whose examples are displayed in Fig.2 and Fig.3.

* Corresponding author

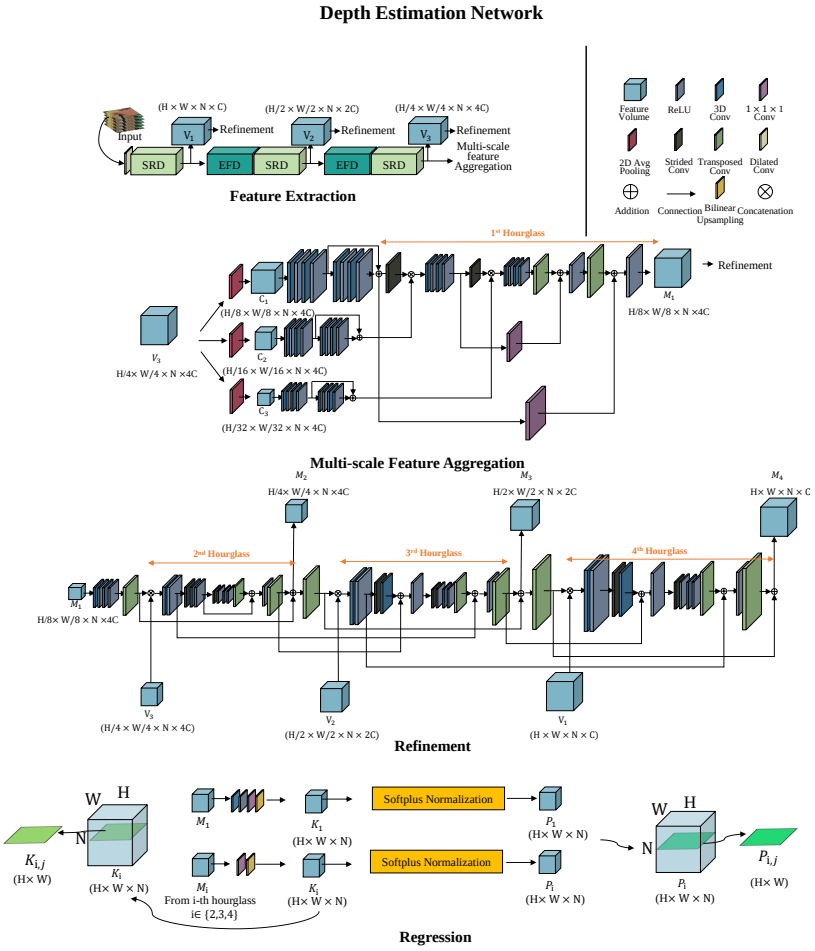

Fig. 1. An illustration of our depth estimation network. C is 8 in our network. Softplus normalization is shown in Eq.(1). From more details, please see the code.

Fig. 2. Visual comparison for AiF image reconstruction on Middlebury Dataset.

Fig. 3. Visual comparison for AiF image reconstruction on 4D Light Field Dataset.

Table 1. Quantitative comparison for AiF image reconstruction on Middlebury dataset and 4D Light field Dataset. **Bold: best**

	Middlebury [2]		4D Light Field [1]	
Method	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑
AiFDepthNet [3]	0.9508	31.5780	0.9645	**35.5918**
Ours	**0.9511**	**31.8223**	**0.9652**	34.9326

Table 2. Quantitative comparison for depth estimation on 4D Light field Dataset. All methods are unsupervised DfF using AiF images. **Bold: best**

Method	MAE ↓	MSE ↓	RMSE ↓	Bump ↓
AiFDepthNet [3]	0.1671	0.0746	0.2698	**2.58**
Ours	**0.1533**	**0.0648**	**0.2447**	3.075

3 Unsupervised DfF using AiF images

Similar with [3], since our network has no learnable parameters after K_i, our network can be trained in an unsupervised manner by leveraging all-in-focus images.

We first reconstruct all-in-focus images via Eq.(2). We use a training loss function ($L_{unsupervised}$) proposed in [3] as below:

$$L_{unsupervised} = \sum_{i=1}^{4} w_i * (L_{aif_i} + 0.002 * L_{sub_i}) \quad \text{s.t.} \tag{4}$$

$$L_{aif_i} = ||I_{AiF_i} - I_{AiF_{gt}}||_1 \quad \text{and} \tag{5}$$

$$L_{sub_i} = \exp(-a * \sum_{c=1}^{3}(|\frac{\partial I_{AiF_i}}{\partial x}|)) * |\frac{\partial D_i}{\partial x}| + \exp(-a * \sum_{c=1}^{3}(|\frac{\partial I_{AiF_i}}{\partial y}|)) * |\frac{\partial D_i}{\partial y}| \tag{6}$$

where $||\cdot||_1$ means a l_1 loss and $I_{AiF_{gt}}$ indicates a ground truth all-in-focus image. I_{AiF_i} is the reconstructed all-in-focus image. c is the channel dimension. x and y are axis of image coordinates. $i \in \{1, 2, 3, 4\}$ means a 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. a is empirically set to 150.

The quantitative results are reported in Tab.2. The qualitative results are shown in Fig.4. Even though our network is not designed in the unsupervised manner, our network achieves competitive results with AiFDepthNet [3].

4 Simulator

Previous synthetic datasets do not account for the changes of FoVs or intrinsic parameters during focal sweeping. We thus propose a simulator which considers the changes. We determine error distributions in the parameters through an experiment, which is described in section 3.1 of our main manuscript. In this section, we describe the error distibutions according to the camera models. We test four smartphone models and capture 50 images per each model.

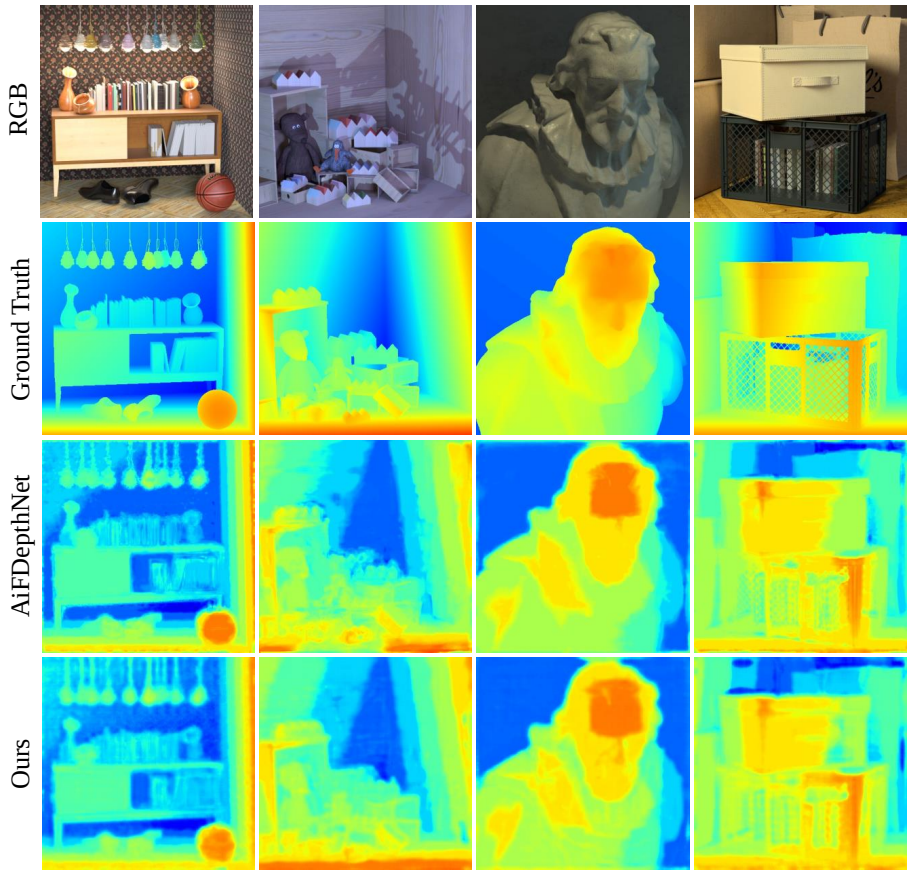

Fig. 4. Visual comparison for depth estimation on 4D Light Field Dataset. All methods are unsupervised DfF using AiF images.

As shown in Fig.5, errors in the FoVs decrease in proportion to 1/focus distance. And, there is no distinct relation between principal points errors and focus distances. We thus assume the error distribution of principal points as the normal distribution with mean and variance of the observed errors in principal points. We incorporate the error distribution and metadata of the smartphones in our simulator.

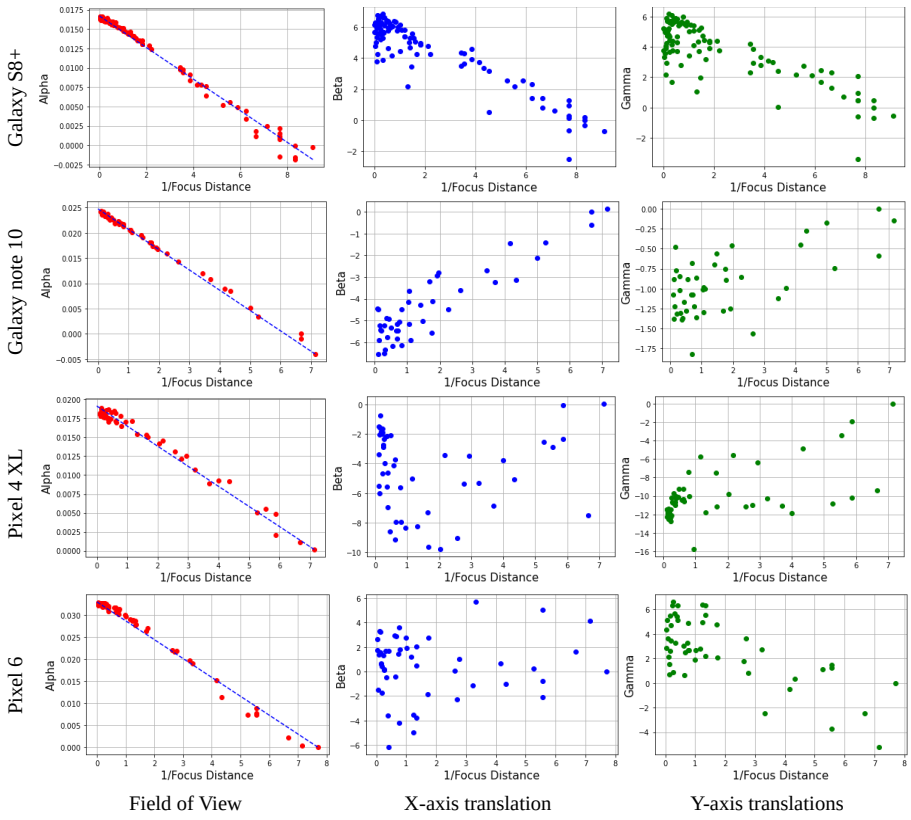

Fig. 5. Error distribution of intrinsic parameters in our experiment.

References

1. 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)
2. 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)
3. 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)