Bridging Unsupervised and Supervised Depth from Focus via All-in-Focus Supervision
[Supplementary Material]

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

Contents

1. Datasets

Totally five datasets are used in the quantitative experiments and visual comparisons. We provide detailed descriptions in the following sections, while the brief descriptions are summarized in the main paper.

1.1. Rendered Synthetic Datasets

(i) *DefocusNet* [6]. They create a synthetic dataset using Blender [2] and release 500 scenes, which consist of 400 training scenes and 100 test scenes. For each scene, they release five RGB images per focal stack, five defocus maps, and one depth image. The AiF images, however, haven't been publicly released yet. Therefore, for this dataset, we can only train our model supervisedly using the ground truth depth.

1.2. Synthetically Blurred Datasets from 4D Light Fields

(i) *DDFF 12-Scene* [3]. They use a light-field camera to capture 4D light-fields and then generate focal stacks with these light-fields. Besides the light-field images, they also provide ground truth depth maps captured from an RGB-D sensor. There are 6 scenes for training and 6 scenes for testing. Generated from 600 light-field images, they provide around 5000 focal stacks for training and 1000 focal stacks for validation with a stack size of 10. Similar to the DefocusNet dataset [6], we cannot apply our unsupervised learning on this dataset as ground truth AiF images are not provided. We use the same train/test split setting as in DDFF 12-Scene and DefocusNet.

(ii) *4D Light Field Dataset* [5]. We use the same data generation method and configuration as described in DDFF 12-Scene [3], and the same train/test split setting for fair comparisons. The focus positions are within $[-2.5\,2.5]$ with a stack size of 10. This dataset includes both AiF and disparity ground truth, which could be used for unsupervised and supervised learning, respectively. For this reason, most of our ablation studies are conducted on this dataset. Note that we predict disparity maps instead of depth maps on this dataset since the ground truth data are disparity maps.

Table 1: **Ablation study on the uniformity of input stack samples.** When the images of the input focal stack are sampled uniformly, the quality of output disparity maps are better than the one with a random sampled stack.

Image index	MAE↓	MSE↓	RMSE↓
1, 3, 5, 7, 9	**0.1663**	**0.0952**	**0.2964**
1, 2, 5, 6, 9	0.1811	0.1235	0.3374

Table 2: **Ablation study on different input stack sizes between training and testing.** The model trained with a fixed stack size of 10 performs poorly with different stack sizes at testing. The model trained with the same input stack size as testing performs the best. However, the model trained with arbitrary input stack sizes performs favorably against the ones trained with the same stack size as testing data. This demonstrates the generalization ability of our method across different input stack sizes.

Test stack size (index)	Training stack size								
	Fixed 10			Arbitrary 2∼10			Same as testing		
	MAE↓	MSE↓	RMSE↓	MAE↓	MSE↓	RMSE↓	MAE↓	MSE↓	RMSE↓
3 (1, 5, 9)	0.7603	0.9566	0.9642	0.2703	0.1482	0.3842	0.2274	0.1551	0.3841
5 (1, 3, 5, 7, 9)	0.2770	0.1545	0.3795	0.1480	0.0681	0.2537	0.1663	0.0952	0.2964
9 (1, 2, ..., 9)	0.0912	0.0493	0.2083	0.1197	0.0616	0.2379	0.0953	0.0591	0.2236

Table 3: **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↓	absRel↓	sqrRel↓	BadPix(0.07)↓	Bumpiness↓	Secs.↓
3D	✓	0.0788	0.0472	0.2014	0.9837	0.4905	0.1894	1.5776	0.0984
		0.0851	0.0461	0.1984	1.0682	0.4709	0.2342	2.3168	0.077
2D	✓	0.1070	0.0577	0.2259	1.2529	0.6129	0.3063	2.1435	0.0488
		0.1179	0.0576	0.2291	1.2493	0.6086	0.4085	2.3168	0.0441

1.3. Synthetically Blurred Real Datasets from RGB Images and Ground Truth Depth

(i) *Middlebury Stereo Datasets* [8]. This dataset includes 15 real-world scenes with left/right image pairs and ground truth disparity maps. We synthesize defocus images with the left RGB images and the disparity maps using the method in [1]. The focus positions are linearly sampled in the disparity range $[10, 60]$ with a stack size of 15. Similar to 4D Light Field Dataset [5], we predict disparity maps instead of depth maps on this dataset. We only use this dataset for testing because the size of this dataset is not sufficiently large for training.

1.4. Real Datasets

(i) *Mobile Depth* [9]. The Mobile Depth dataset [9] captures multiple real-world focal stacks with a hand-held mobile phone (Samsung Galaxy S3) and a hand-held camera (Nikon D80). Focal stacks from the mobile phone are captured during auto focus processes with arbitrary stack sizes. The Nikon D80 is used due to the fact that the mobile phone does not allow manual control on focus positions. This dataset respectively provides focal stacks before and after calibration alignment, as well as AiF images stitched by a multi-labeled MRF. This real-world dataset is used for our qualitative evaluations.

2. Additional Evaluations

In this section, we show additional quantitative results for comparisons to the state-of-the-art method DefocusNet [6]. We also provide additional ablation studies and qualitative results for more information.

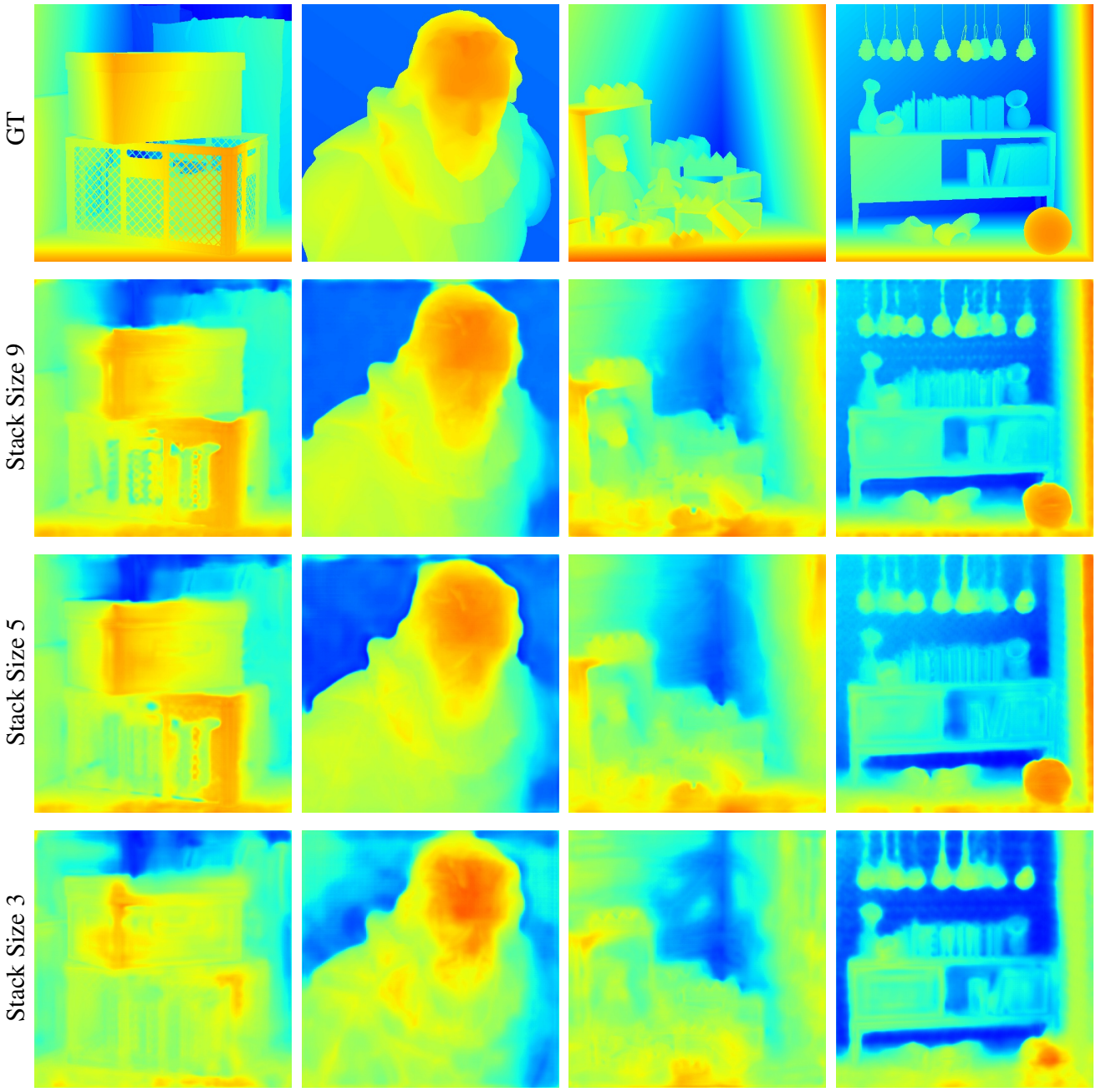

Figure 1: **Visual comparison on different focal stack sizes.** The quality of disparity map improves as the stack size increases.

2.1. Ablation Studies

Focal stack size. Table 2 shows the detailed ablation results on the effect of different input stack sizes between training and testing. Fig. 1 shows the corresponding qualitative results.

Network Architecture. Table 3 shows the full metric version of Table 2 in the main paper.

Unsupervised learning using all-in-focus images. Table 4 is the full metric version of Table 3 in the main paper.

Uniformity of input stack samples. Table 1 ablates the uniformity of input focal stacks. When the images of the stack are sampled uniformly, it retains the stack information and performs better than the one with a non-uniformly sampled stack.

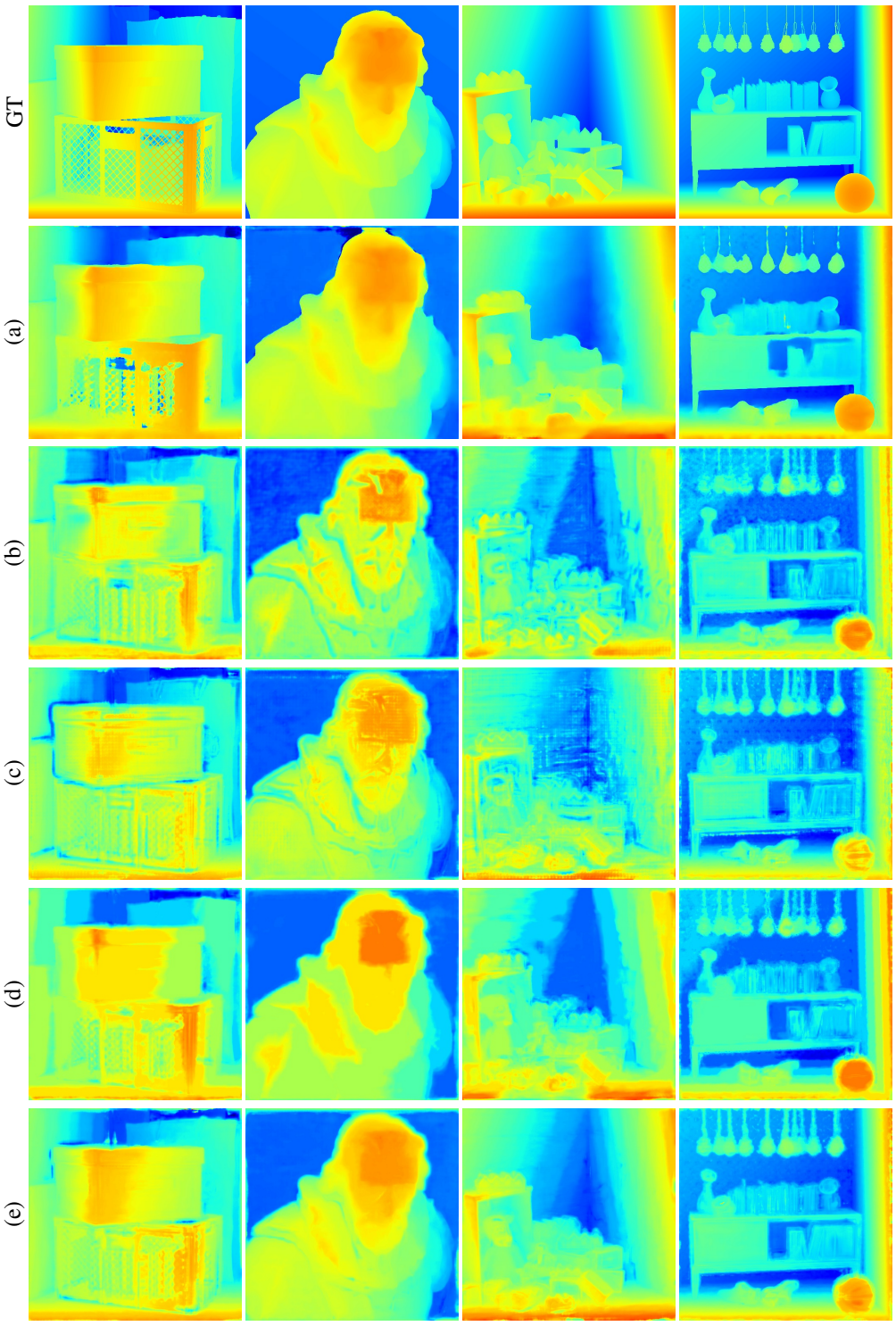

Figure 2: **Visual comparison on different supervision settings.** (a) Supervised learning. (b) Unsupervised learning (AiF supervision) with a stack size = 10. (c) Unsupervised learning with a stack size = 37. (d) Unsupervised learning with a stack size = 10 and the smoothness loss. (e) Unsupervised learning with a stack size = 37 and the smoothness loss. The results of supervised learning with disparity maps as supervision are better than the ones from unsupervised learning. The disparity maps from unsupervised learning always suffer from the quantization effect. By adding the smoothness loss, the disparity results of unsupervised learning become locally smooth and perform favorably against the ones from supervised learning qualitatively.

Table 4: **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 and qualitatively. See Fig. 2 for the visual comparison.

Supervised	Stack size	L_{smooth}	MSE↓	MAE↓	RMSE↓	absRel↓	sqrRel↓	BadPix↓	Bump.↓	Secs.↓
Yes	10		0.0472	0.0788	0.2014	0.9837	0.4905	0.1894	1.5776	0.0673
No	10		0.1174	0.2425	0.3401	1.9612	0.7882	0.7879	4.1254	0.0657
No	37		0.1039	0.2099	0.3202	1.7174	0.6648	0.6508	4.1827	0.0886
No	10	✓	0.0746	0.1671	0.2698	1.5957	0.6045	0.6610	2.5836	0.0509
No	37	✓	0.0584	0.1116	0.2311	1.0637	0.4676	0.2826	2.6989	0.0775

Table 5: **Quantitative comparison on the DefocusNet [6] dataset.** Please refer to Fig. 4 for the visual comparison.

Method	MAE↓	MSE↓	RMSE↓	AbsRel↓	SqrRel↓	Sec↓
Ours	**0.0549**	**0.0127**	**0.1043**	**11.1528**	**1.8668**	**0.018**
DefocusNet [6]	0.0637	0.0175	0.1207	13.8610	3.3104	0.039

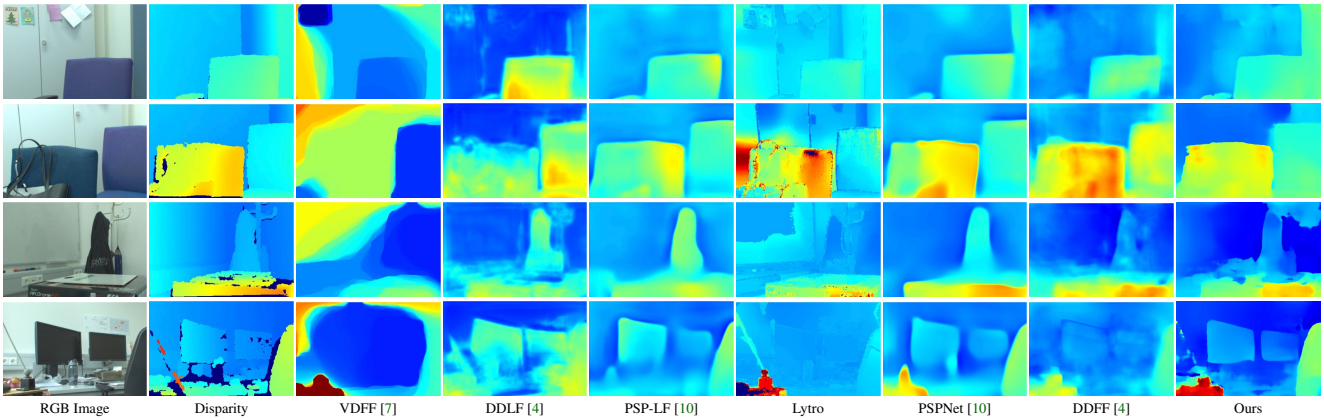

Figure 3: **Visual comparison on DDFF 12-Scene [3].** The other state-of-the-art methods often generate blurry disparity boundaries. In contrast, our method restores fine details at disparity discontinuities.

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

Depth Results. We show detailed quantitative and qualitative comparisons on the DefocusNet [6] dataset with our supervised model in Table 5 and Fig. 4, respectively. Fig. 3 is the corresponding qualitative results of Table 4 in the main paper. For the Mobile Depth [9] dataset, we show extra depth qualitative results in Fig. 5.

AiF Results. Our model is able to output AiF images in both supervised and unsupervised settings. We show the qualitative results of predicted AiF images as well as the error maps on 4D Light Field Dataset [5], Middlebury Stereo Datasets [8], and the Mobile Depth [9] dataset since they all contain ground truth AiF images. As shown in Figs. 7–9, although the model trained on ground truth depth provides visually pleasing AiF results, the model trained on ground truth AiF images outperforms the aforementioned results.

3. Limitations

Our method assumes that the images of the focal stack are perfectly aligned to form a reasonable 4D attention map. Thus, not considered are those factors that would bring about misalignment, *e.g.*, focus breathing, camera motion, or object motion. Fortunately, this problem could be eased through an alignment process. Fig. 10 shows an example to clarify our viewpoint.

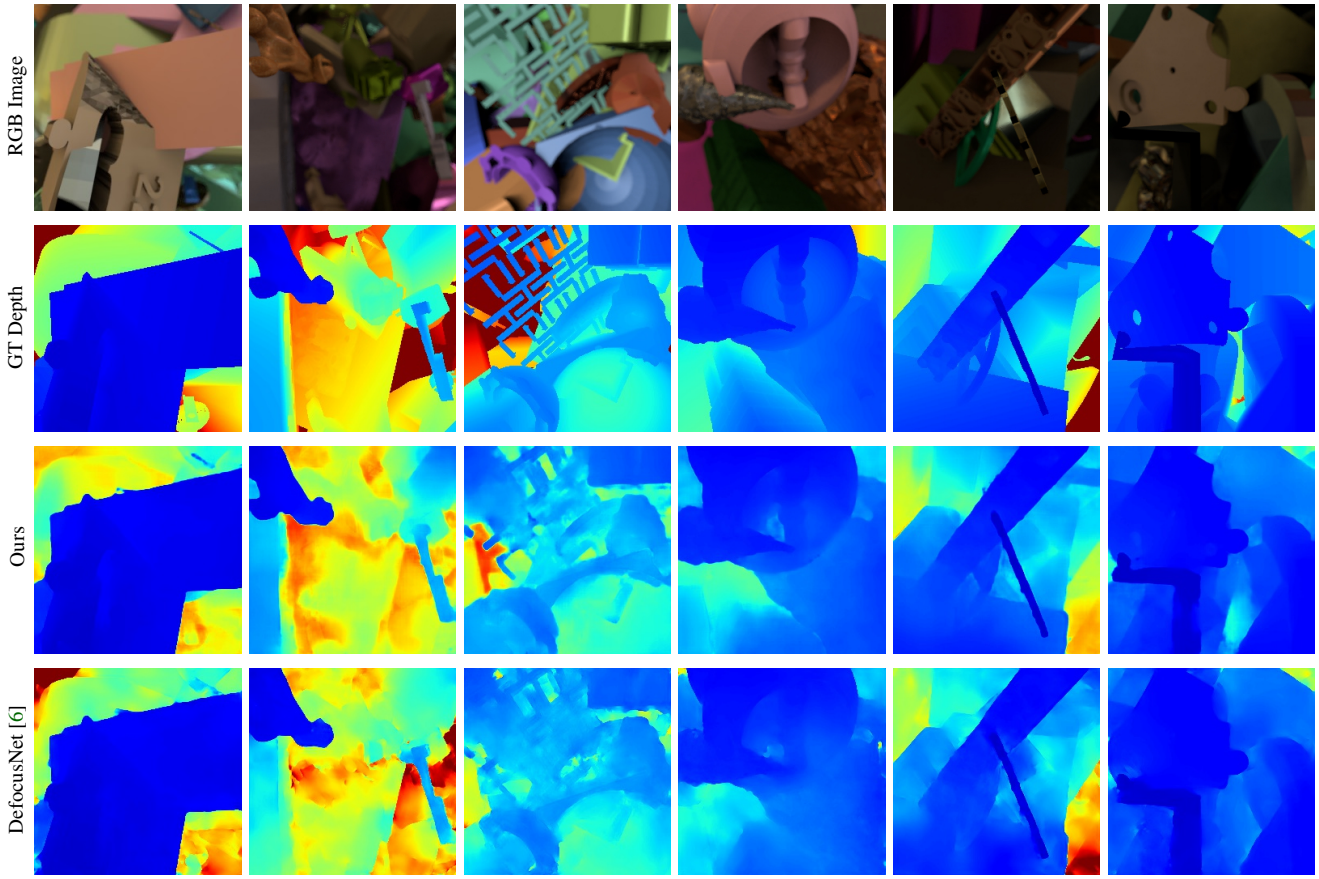

Figure 4: **Visual comparison on the DefocusNet [6] dataset.**

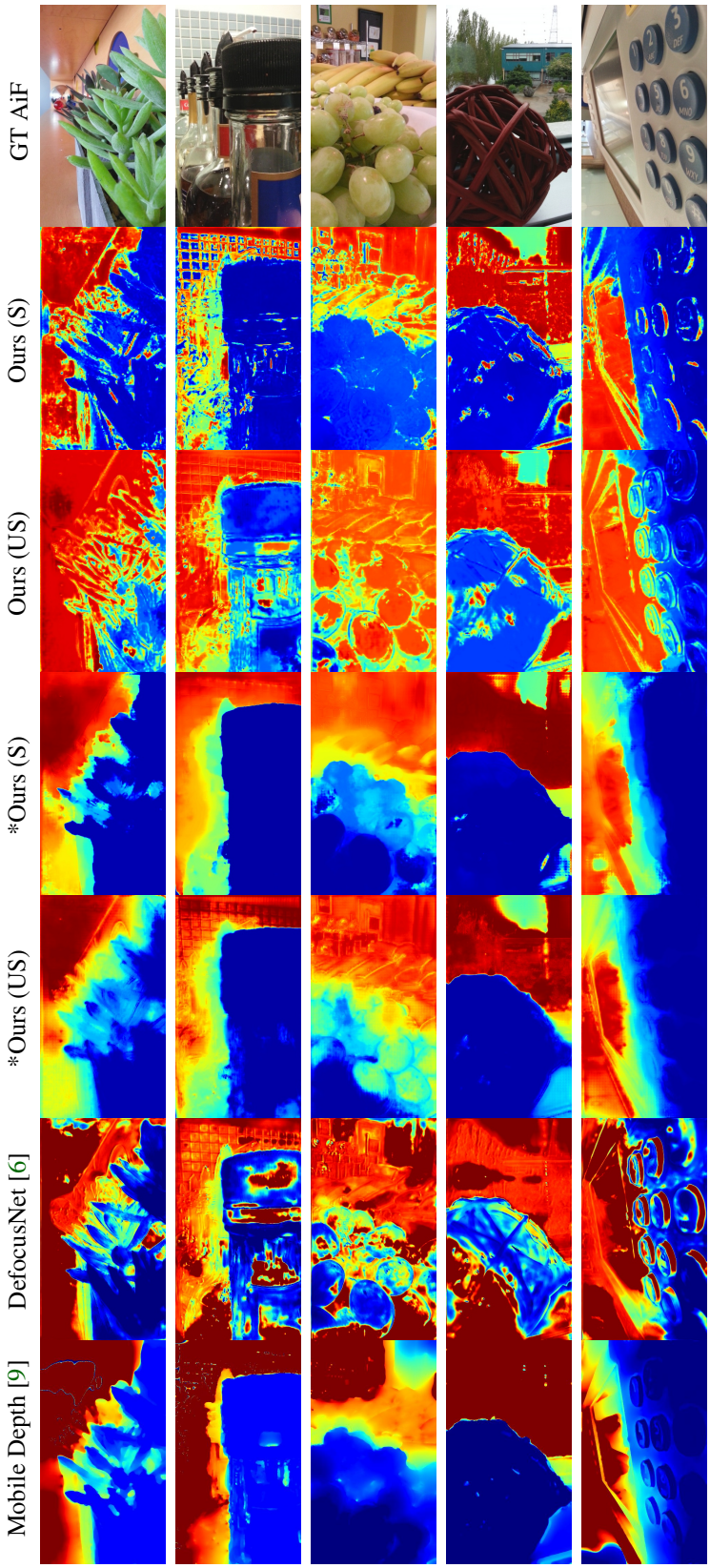

Figure 5: **Visual comparison on the Mobile Depth dataset [9].**

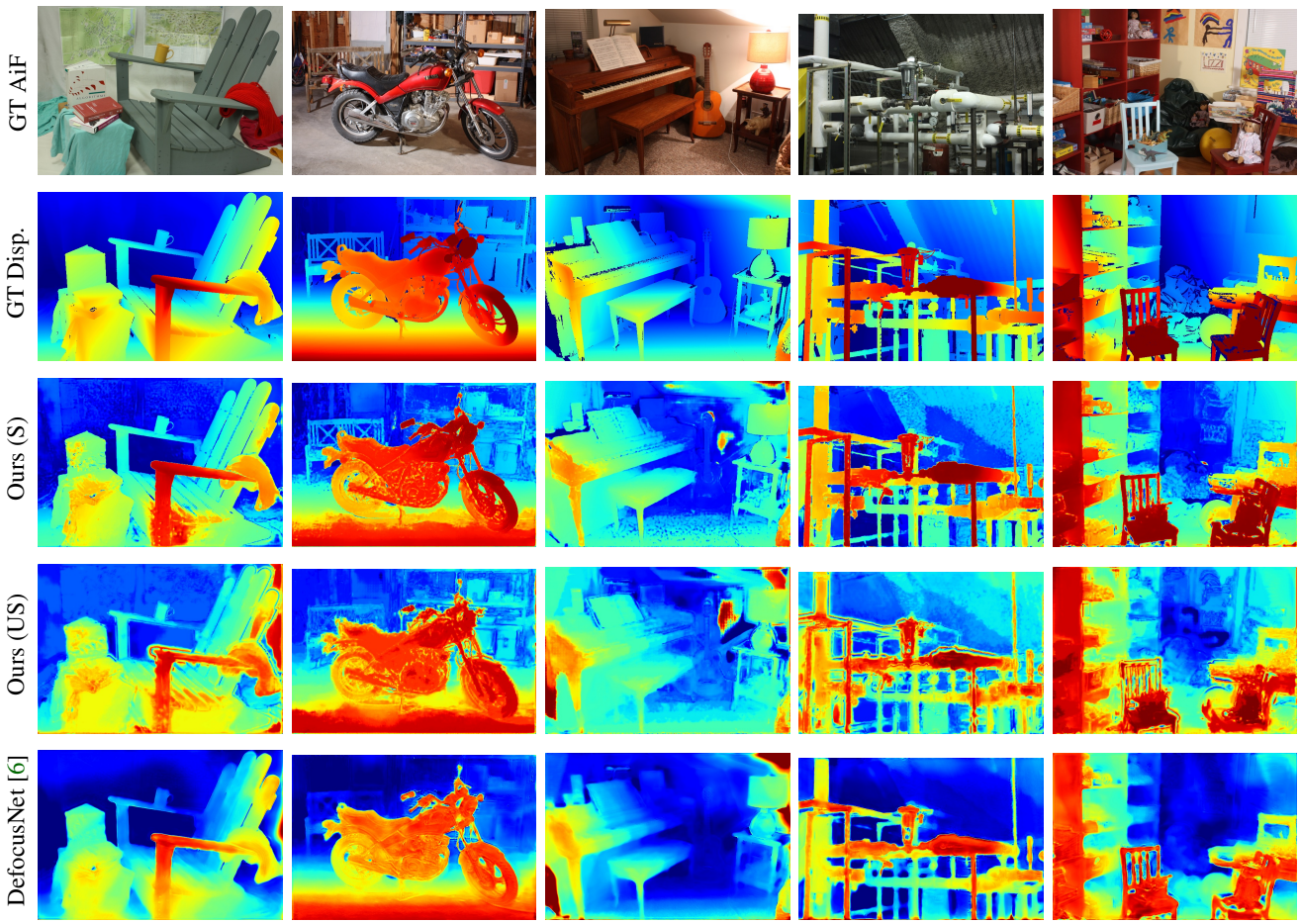

Figure 6: **Visual comparison on Middlebury Stereo Datasets [8].**

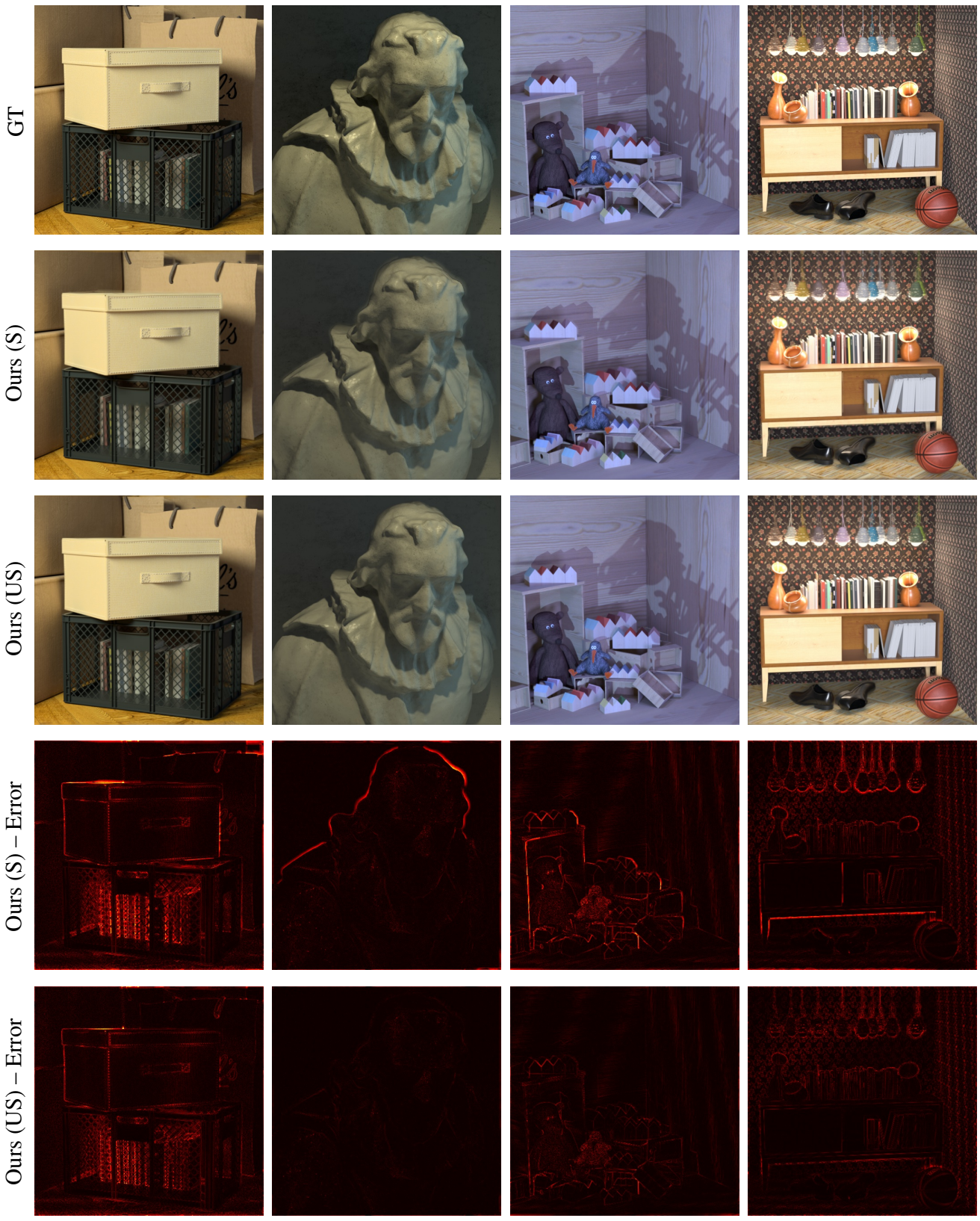

Figure 7: **Visual comparison for AiF image reconstruction on 4D Light Field Dataset [5].**

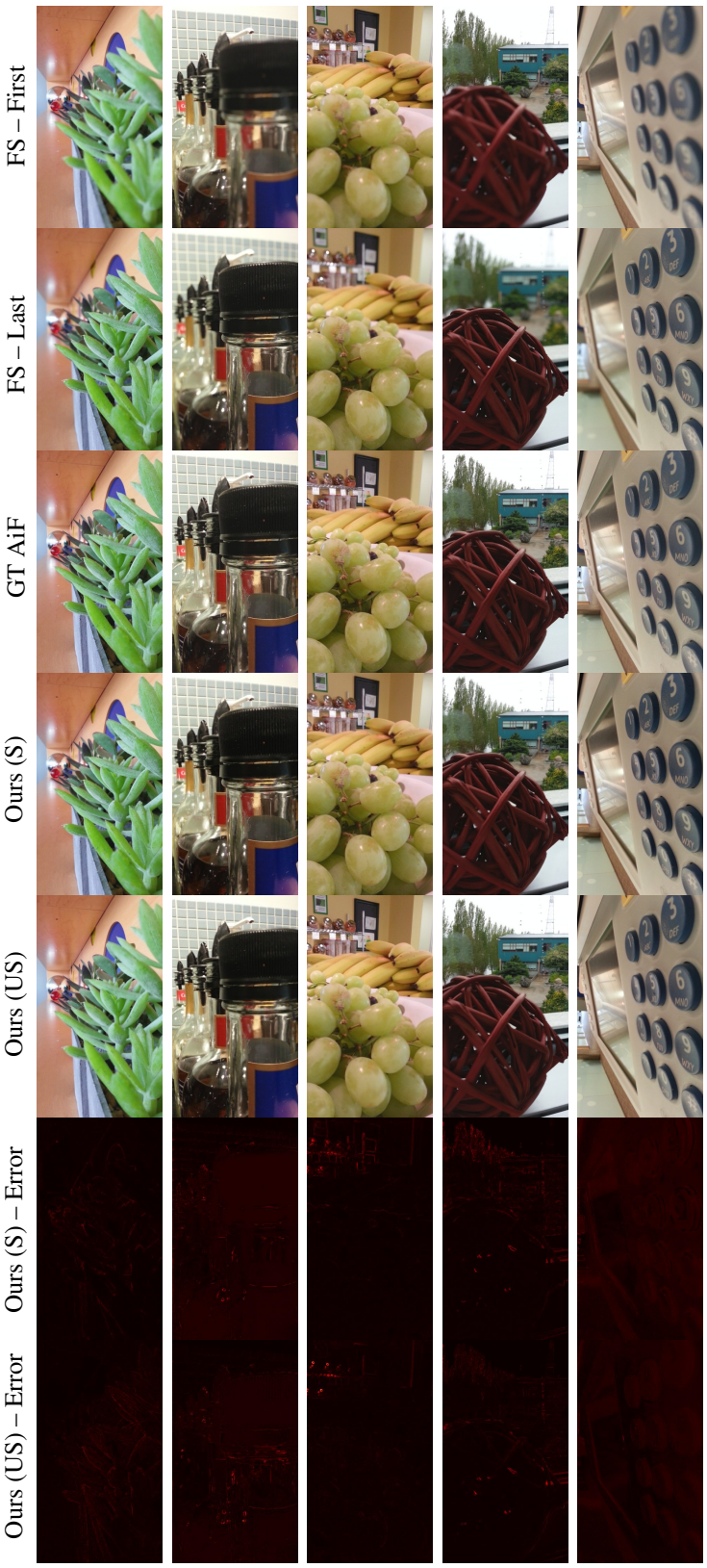

Figure 8: **Visual comparison for AiF image reconstruction on the Mobile Depth dataset [9].**

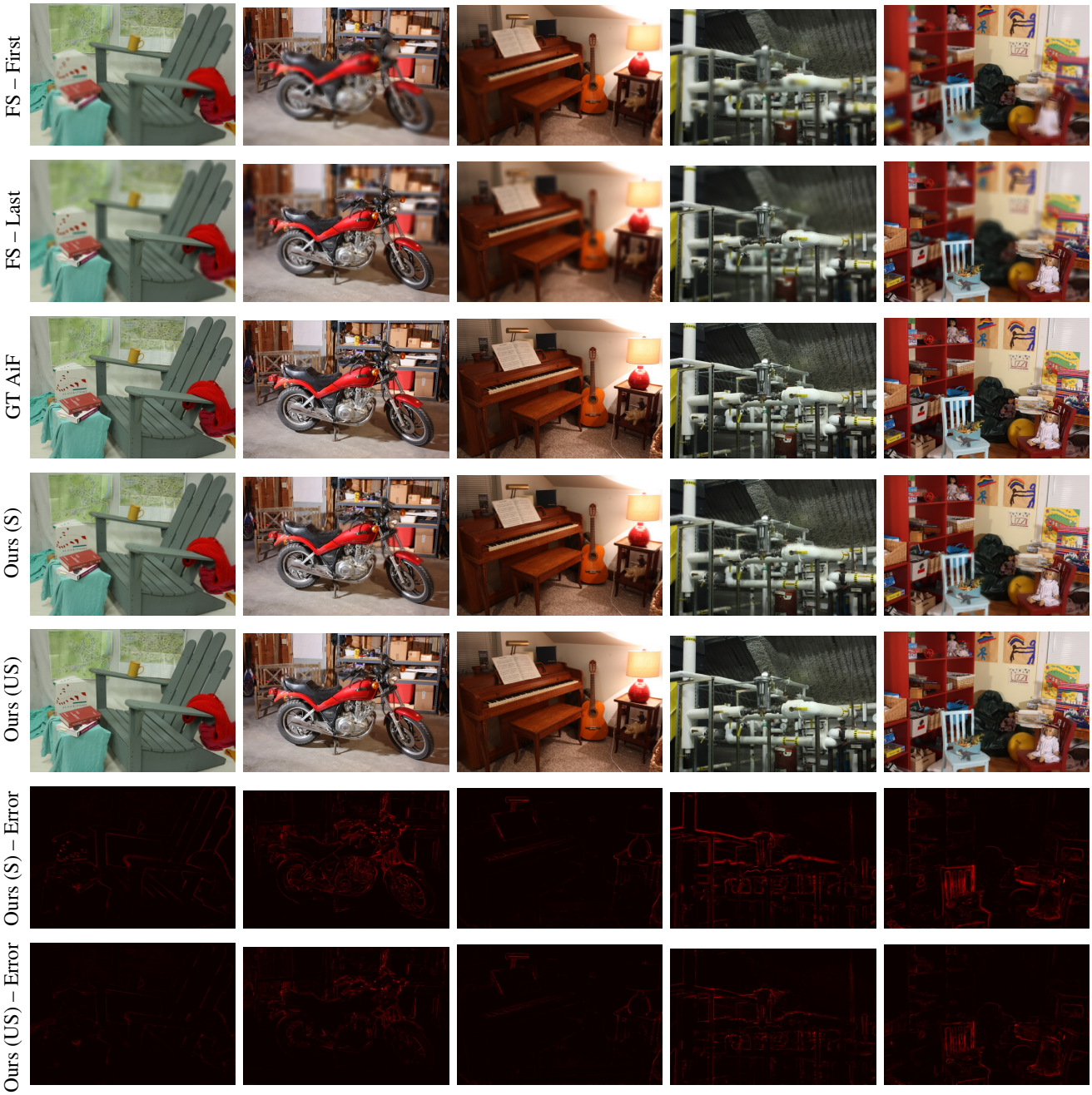

Figure 9: **Visual comparison for AiF image reconstruction on Middlebury Stereo Datasets [8].**

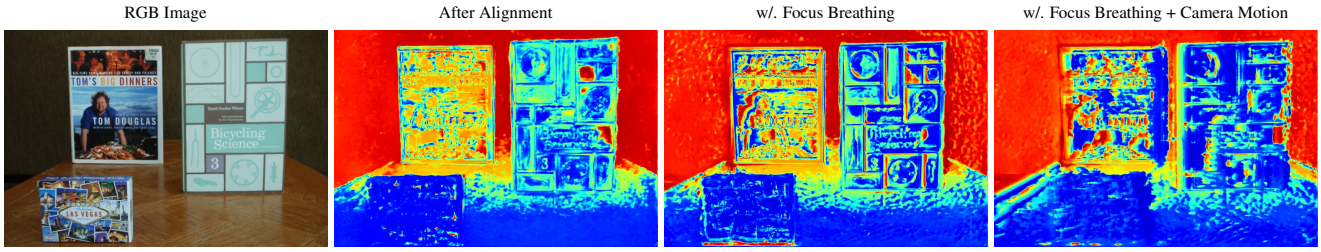

Figure 10: **Limitations of our method.** Not considered in our method are those factors that would make the images of the focal stack out of alignment. However, an alignment process could remedy the quality degradation that comes from focus breathing w/. and w/o. camera motion for depth estimation on the Mobile Depth dataset [9].

References

[1] 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

[2] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 1

[3] Caner Hazirbas, Sebastian Georg Soyer, Maximilian Christian Staab, Laura Leal-Taixé, and Daniel Cremers. Deep depth from focus. In *Asian Conference on Computer Vision*, pages 525–541. Springer, 2018. 1, 5

[4] 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. 5

[5] 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. 1, 2, 5, 9

[6] 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. 1, 2, 5, 6, 7, 8

[7] Michael Möller, Martin Benning, Carola Schönlieb, and Daniel Cremers. Variational depth from focus reconstruction. *IEEE Trans. Image Process.*, 24(12):5369–5378, 2015. 5

[8] 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. 2, 5, 8, 11

[9] 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. 2, 5, 7, 10, 12

[10] 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. 5