FloatingFusion: Depth from ToF and Image-stabilized Stereo Cameras

Supplementary Document

Andreas Meuleman[1], Hakyeong Kim[1],
James Tompkin[2], and Min H. Kim[1]

[1] KAIST, South Korea
{ameuleman,hkkim,minhkim}@vclab.kaist.ac.kr
[2] Brown University, United States

A Appendix

A.1 Metrics

For depth evaluation, we use the following metrics:

- MAE: $\frac{1}{n}\sum|d-d^*|$
- RMSE: $\sqrt{\frac{1}{n}\sum(d-d^*)^2}$
- MAE Rel.: $\frac{1}{n}\sum|d-d^*|/d^*$
- RMSE Rel.: $\sqrt{\frac{1}{n}\sum\left((d-d^*)/d^*\right)^2}$
- Bad ratio: percentage of pixel with error $|d-d^*|$ above threshold

with d the estimated depth, d^* ground truth, n the number of valid pixels.

A.2 Implementation Details

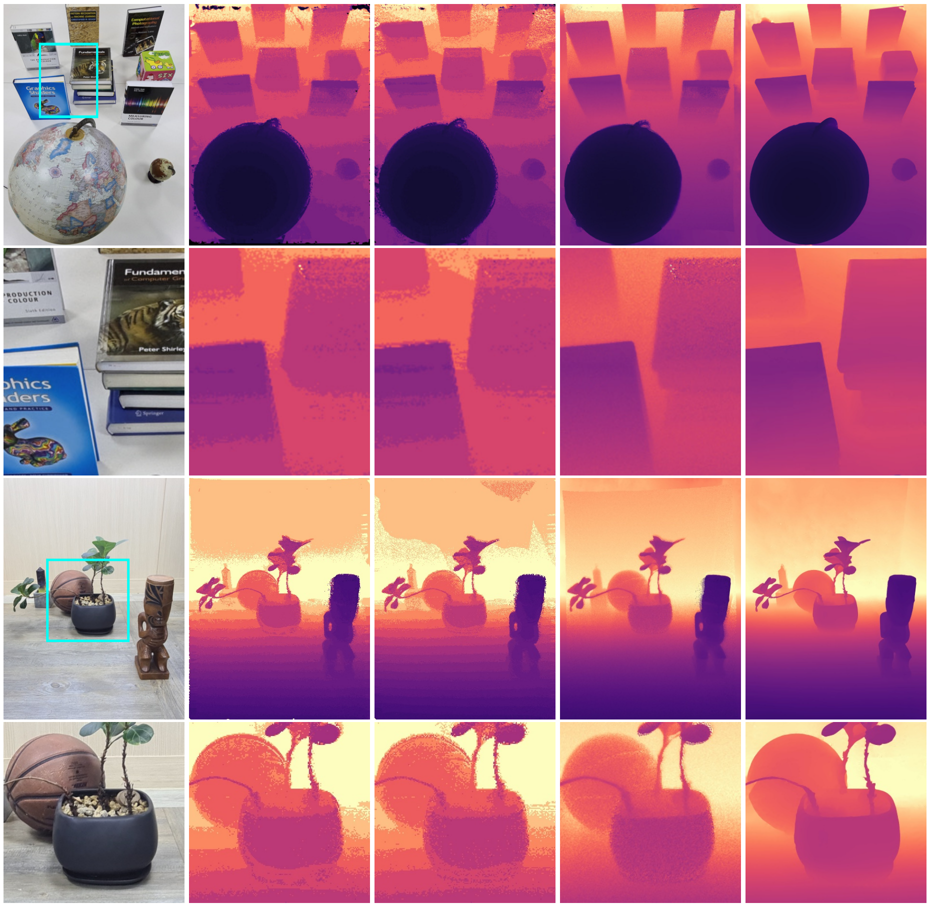
Fusion training. We initialize $\tau = 140$ and the RAFT-stereo weights using Middlebury checkpoints. For training, we run 10k iteration with a batch size of 4 and a crop size of 640×1440 pixels.

Mip-NeRF optimization. The ToF depth loss is scaled by an exponential decay $N_0 \exp^{-\lambda i/m}$ where $N_0 = 10$, $\lambda = 8$, i is the current iteration and m is the total number of iterations. We train Mip-NeRF for 200k iterations and a batch size of 4096 rays.

A.3 Additional Results

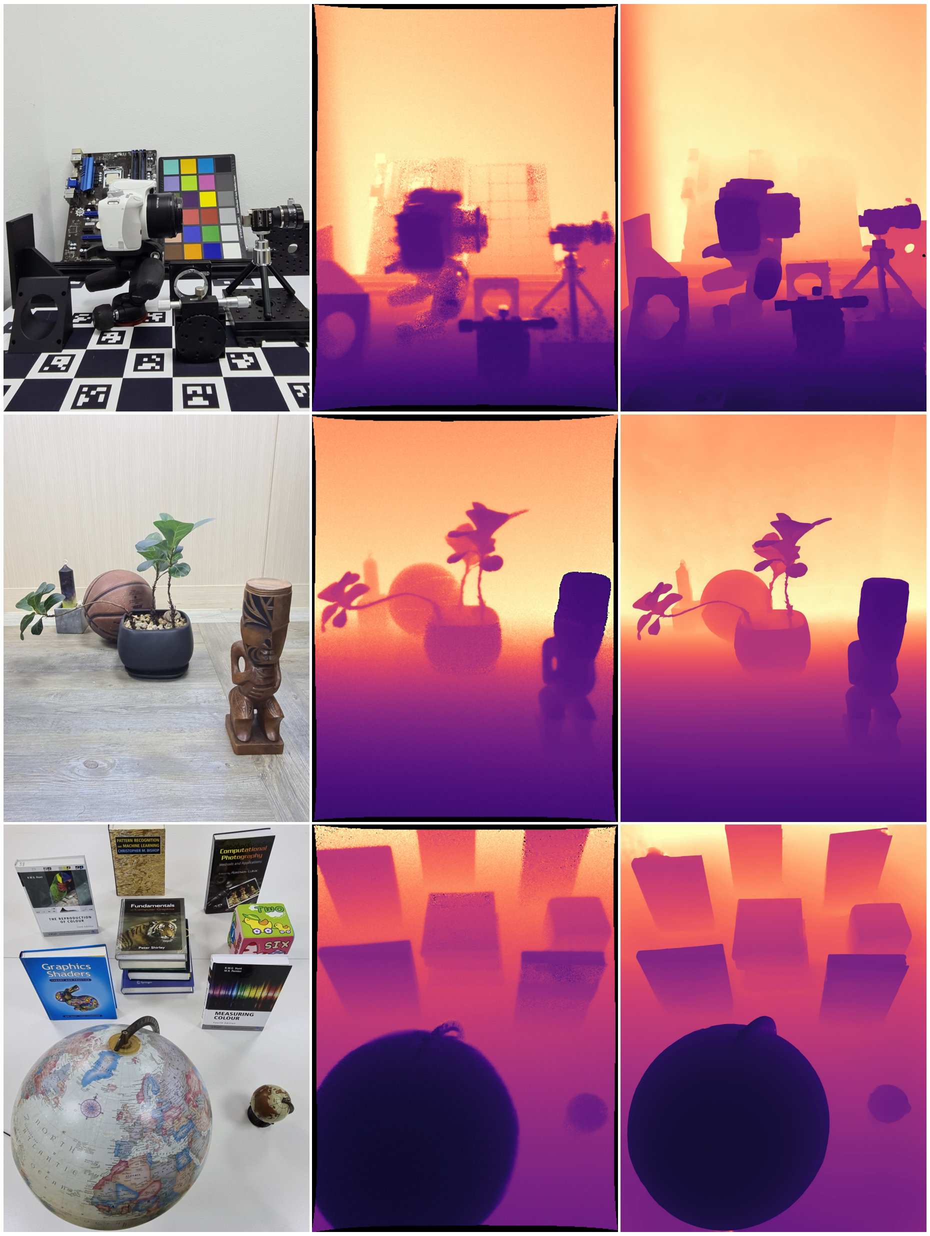
Additional Scenes. Figure 1 shows more ToF/stereo fusion results captured with our phone.

Input ToF. Figure 2 shows the depth estimated from the ToF sensor and our fusion results. Depth from ToF is much noisier and lower resolution, with blurry depth edges.



(a) RGB (b) Marin et al. (c) Agresti et al. (d) Gao et al. (e) Ours

Fig. 1: Additional mobile ToF/stereo fusion results.



(a) RGB (b) ToF only (e) Ours

Fig. 2: Input ToF depth and our fusion results.

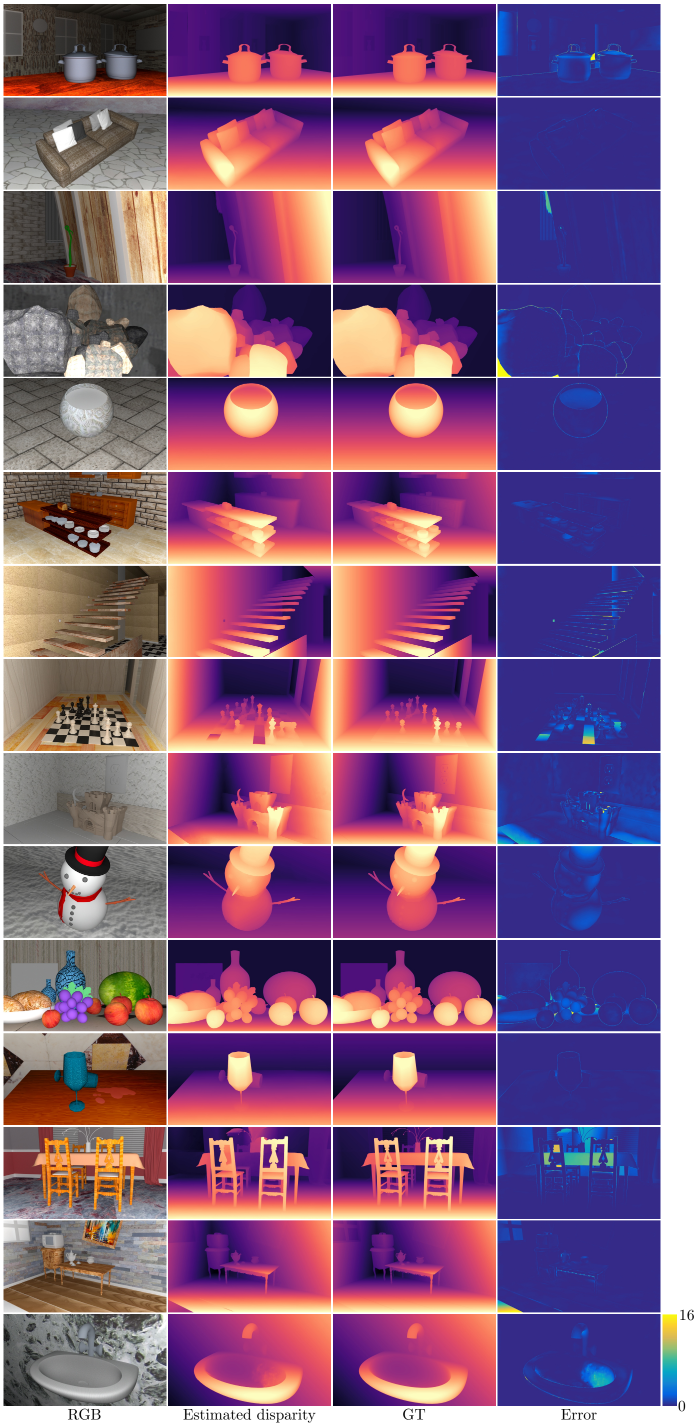


Fig. 3: Our ToF/stereo fusion results on the rendered SYNTH3 dataset [1].

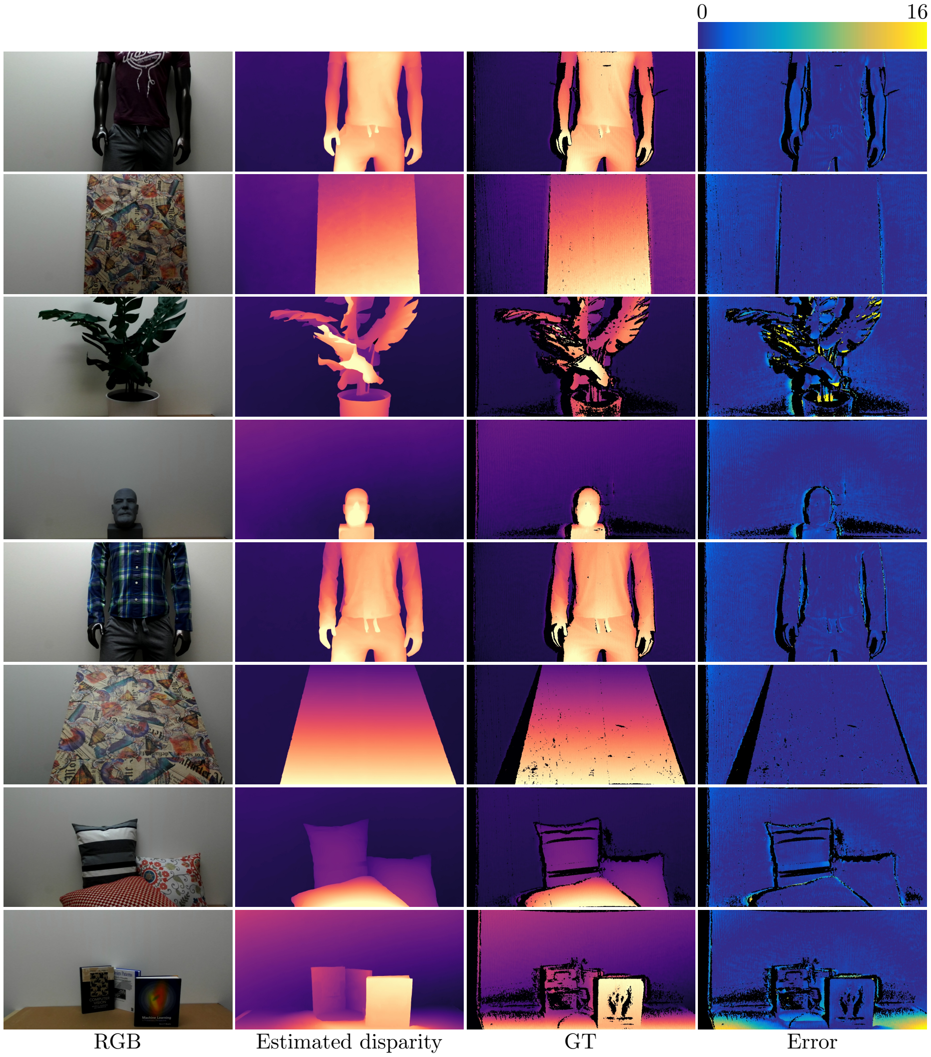


Fig. 4: Our ToF/stereo fusion results on REAL3 dataset [2].

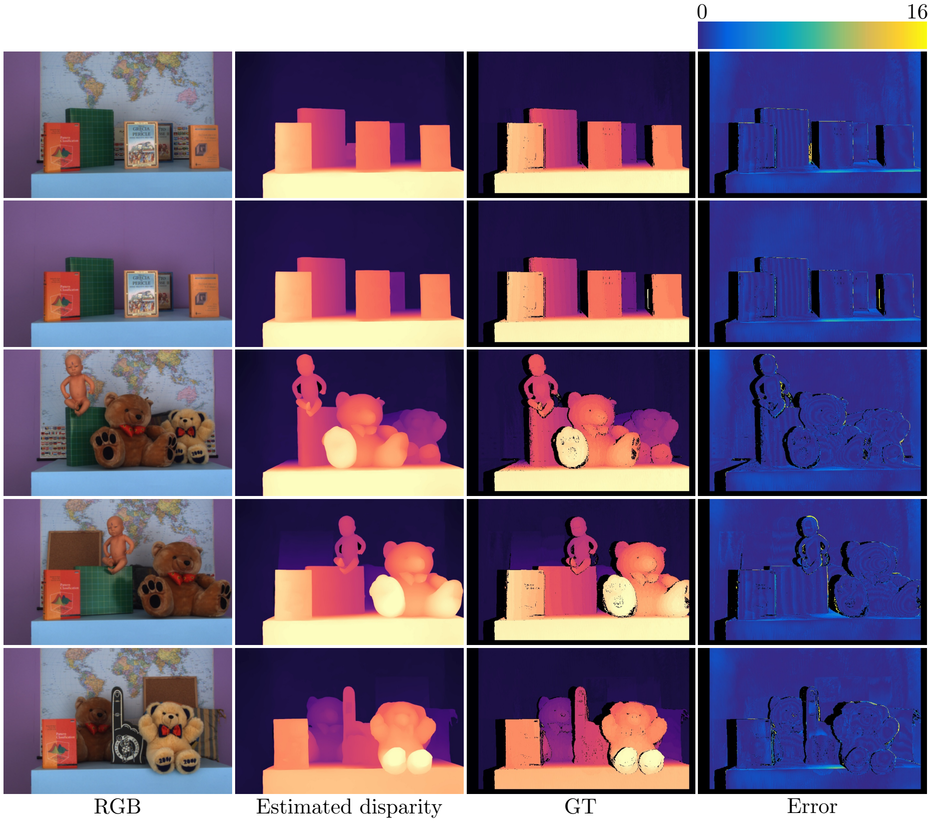


Fig. 5: Our ToF/stereo fusion results on LTTM5 dataset [5].

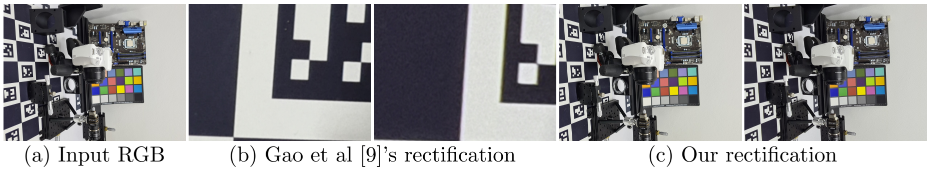


Fig. 6: Estimating innacurate calibration parameters can lead to poor rectification. (b) rectifying the stereo pair based on Gao et al. [9]'s approach aggressively crops the images and they are not horizontally aligned (see the vertical shift of the checkerboard corner). (c) Our calibration provides more sensible results and the images of stereo pair are aligned.

Table 1: Disparity error of stereo+ToF fusion techniques. Reported numbers are from respective works, except for † that have been evaluated in [2]. Our method outperforms others on the real-world datasets REAL3 [2] and LTTM5 [5]. We show the results of guided ToF upsampling as "Interpolated ToF" for reference.

	SYNTH3		REAL3		LTTM5	
	MAE	MSE	MAE	MSE	MAE	MSE
Interpolated ToF [2]	0.66	4.75	2.55	10.76	1.53	11.68
Marin et al. [10]†	0.64	4.20	2.19	8.82	1.15	7.67
Agresti et al. [2]	0.53	3.92	1.65	8.35	0.89	7.40
Deng et al. [8]	0.30	**1.84**	0.93	5.84	0.79	3.47
Dal Mutto et al. [6]	–	–	–	–	1.43	12.21
Dal Mutto et al. [5]	–	–	–	–	1.36	10.06
Ours	**0.26**	2.14	**0.67**	**4.10**	**0.45**	**1.52**

Table 2: Plenoxels shows good novel view synthesis, but depth maps from Mip-NeRF are more accurate. Evaluated on the NeRF synthetic datasets [12].

	Bad ratio (%)		Depth error			
	>0.2	>0.05	MAE Rel.	RMSE Rel.	MAE	RMSE
Plenoxels [3]	11.84	55.84	0.032	0.139	0.120	0.275
Mip-NeRF [4]	0.93	6.65	0.005	0.024	0.019	0.049

Stereo+ToF Fusion Evaluation on other Datasets. We evaluate our fusion on SYNTH3 [1], REAL3 [2] and LTTM5 [5]. Note that None of those datasets show the strong challenges encountered in the mobile environment, with higher power ToF modules and low distortion. In addition, their resolution is lower and raw ToF measurements are not provided, making the results not directly applicable to our phone. However, they allow us to compare against numerous methods without reimplementation.

SYNTH3 [1] dataset features 15 scenes rendered following [11]. REAL3 [2] features eight scenes captured with a ZED stereo camera and a Microsoft Kinect v2 ToF depth camera. Ground truth disparity is obtained with a line laser. LTTM5 [5] consists of five scenes captured with two BASLER scA1000 RGB cameras and a MESA SR4000 ToF camera. Ground truth is acquired by space-time stereo [13,7]. To accommodate for REAL3 and LTTM5's low RGB resolution, we linearly upsample the input stereo images two times in horizontal and vertical directions. Since ToF raw measurements are not available, we remove the confidence based on the difference between the estimated distance from the two frequencies. We also account for the intensity scale difference in the amplitude maps by setting $\sigma_A = 0.001$. We present our results on these datasets in Table 1 and on Figures 4 and 5.

Runtimes. On our test system equipped with an NVIDIA RTX 3090, our online calibration takes 4.4 seconds and fusion takes 1.4 seconds.

Comparison against Plenoxels. Table 2 shows that Mip-NeRF [4] is better than Plenoxels [3] at estimating depth maps. We therefore use MipNeRF as a basis to generate training data.

Calibration Failure. Figure 6 shows an example of total calibration failure for Gao et al. [9]. Due to this, we use our calibration when comparing our fusion method against others in the main paper.

References

1. Agresti, G., Minto, L., Marin, G., Zanuttigh, P.: Deep learning for confidence information in stereo and tof data fusion. In: ICCV Workshops (2017)
2. Agresti, G., Minto, L., Marin, G., Zanuttigh, P.: Stereo and tof data fusion by learning from synthetic data. Information Fusion (2019)
3. Alex Yu and Sara Fridovich-Keil, Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks (2021)
4. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: ICCV (2021)
5. Dal Mutto, C., Zanuttigh, P., Cortelazzo, G.M.: Probabilistic tof and stereo data fusion based on mixed pixels measurement models. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2015)
6. Dal Mutto, C., Zanuttigh, P., Mattoccia, S., Cortelazzo, G.M.: Locally consistent tof and stereo data fusion. In: ECCV Workshops (2012)
7. Davis, J., Nehab, D., Ramamoorthi, R., Rusinkiewicz, S.: Spacetime stereo: a unifying framework for depth from triangulation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2005)
8. Deng, Y., Xiao, J., Zhou, Z.: Tof and stereo data fusion using dynamic search range stereo matching. IEEE Transactions on Multimedia (2021)
9. Gao, Y., Esquivel, S., Koch, R., Keinert, J.: A novel self-calibration method for a stereo-tof system using a kinect V2 and two 4k gopro cameras. In: 3DV (2017)
10. Marin, G., Zanuttigh, P., Mattoccia, S.: Reliable fusion of tof and stereo depth driven by confidence measures. In: ECCV (2016)
11. Meister, S., Nair, R., Kondermann, D.: Simulation of Time-of-Flight Sensors using Global Illumination. In: Vision, Modeling & Visualization. The Eurographics Association (2013)
12. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
13. Zhang, L., Curless, B., Seitz, S.: Spacetime stereo: shape recovery for dynamic scenes. In: CVPR (2003)