Deep End-to-End Time-of-Flight Imaging (Supplemental Material)

Shuochen Su, UBC
Felix Heide, Stanford University
Gordon Wetzstein, Stanford University
Wolfgang Heidrich, KAUST

Introduction

In this supplemental document, we provide additional full-size depth reconstruction results and implementation details of the datasets generation and network training, ensuring full reproducibility of the proposed approach.

Contents

1. Additional Results

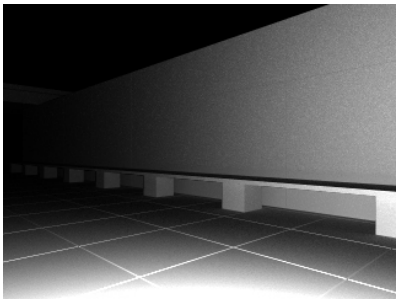
In this section, we provide full-size depth maps generated from different methods to evaluate the performance of the proposed end-to-end reconstruction framework for joint phase unwrapping, MPI correction and denoising. In the following figures, we denote

- SCENE: as the amplitude image of the ToF measurements;
- PHASE: as the phase map measured at $\omega_2 = 70$ MHz for visualization of phase ambiguity;
- CORRELATION: as the raw ToF correlation image measured at $\omega_2 = 70$ MHz and $\psi_1 = 0$ rad which is one of the four input images to our network;
- SINGLE: as the depth map directly computed from a single, lower modulation frequency, $\omega_1 = 40$ MHz;
- PHASOR: as the dual-frequency phase-unwrapped depth map [3, 1];
- SRA: as the phase-unwrapped and MPI compensated depth map [2];
- DEPTH2DEPTH (D2D): as the depth map generated from the depth postprocessing network similar to [5];
- OURS: as the depth map generated with our proposed method;
- GT: as the ground truth depth (available for synthetic data only).

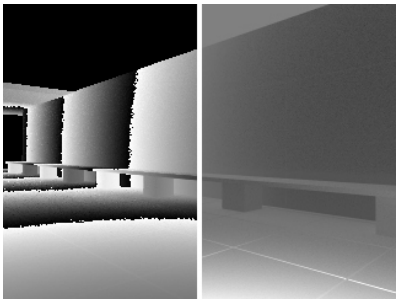
The resulting depth maps are best compared along with the depth scanlines shown at the bottom of each individual figure, where we use meter as the default unit of length. It is expected that SINGLE suffers here the most from both phase ambiguity and MPI distortion due to the lack of compensation for these measurement distortions. The optimization-based methods, PHASOR and SRA, reduce MPI and phase ambiguity, however they also introduce "flying pixels" in the presence of noise. The learning-based depth postprocessing framework, DEPTH2DEPTH, compensates for these ambiguities, yet introduces new, high-frequency artifacts when the input depth is of lower quality. Our results on the other hand, are consistently more accurate and robust to various capture scenarios.

1.1. Synthetic Results

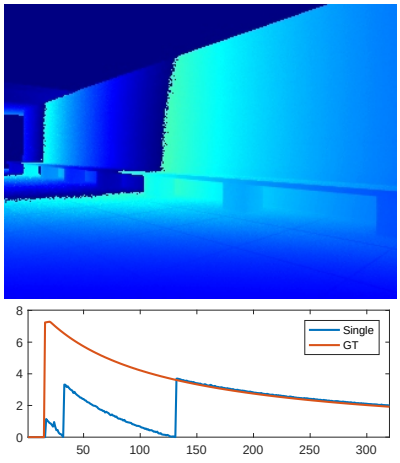
We demonstrate results on synthesized raw ToF measurements with known ground truth depth labels in Fig. 1, 2, 3, 4, 5 and 6. The depth scanlines below each subfigure are sampled from the middle row of the the corresponding depth maps, highlighted as red lines in the ground truth GT images in Fig. 1h, 2h, 3h, 4h, 5h and 6h.



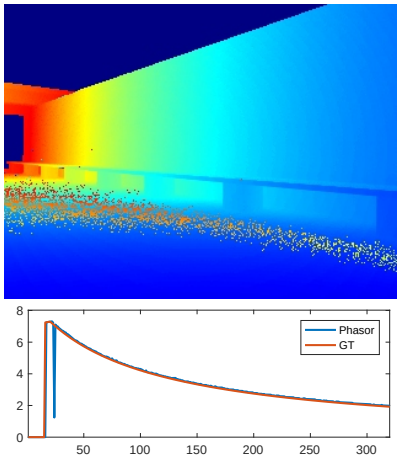
(a) Scene.



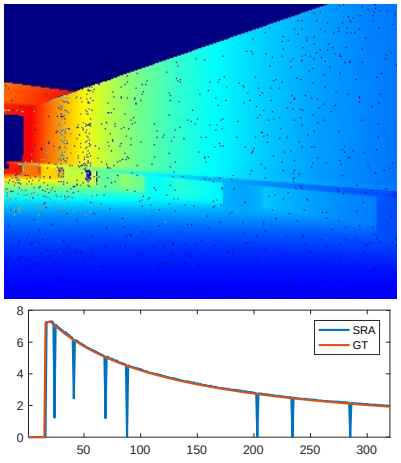
(b) Phase/Correlation.



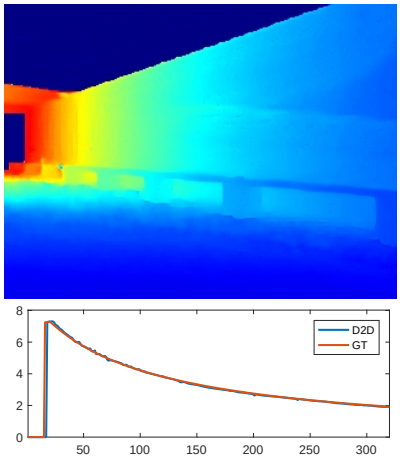
(c) Single.



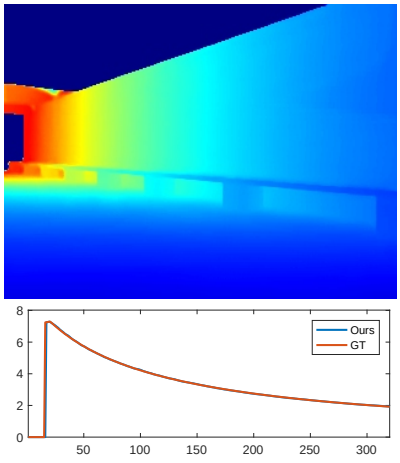
(d) Phasor.



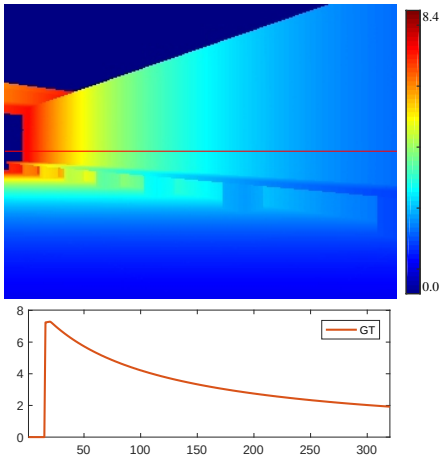
(e) SRA.



(f) Depth2Depth.

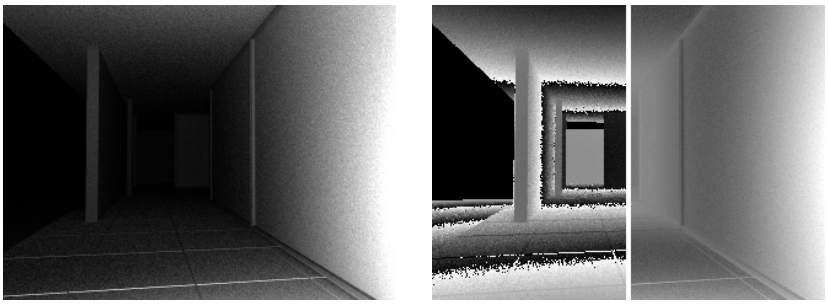


(g) Ours.

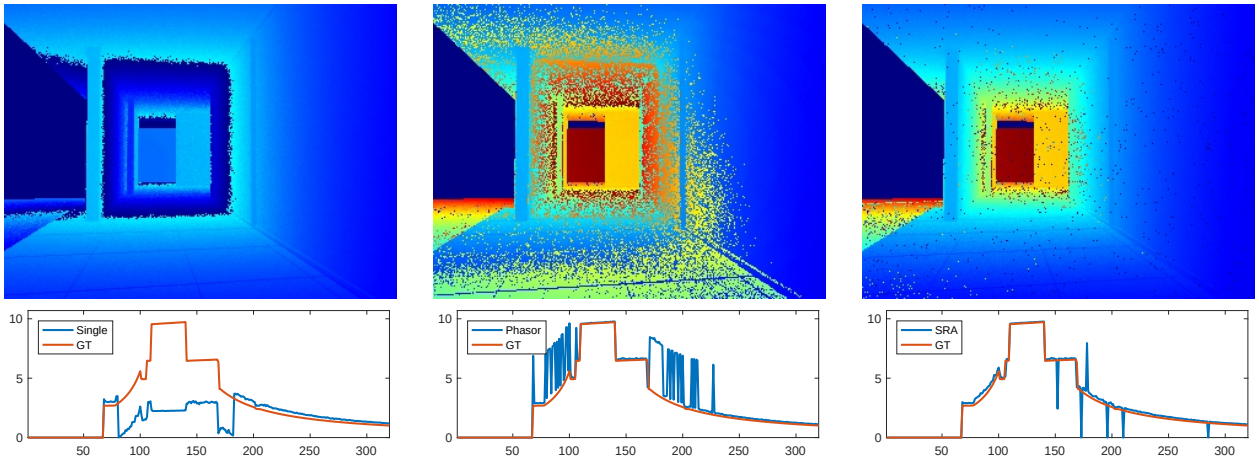


(h) GT.

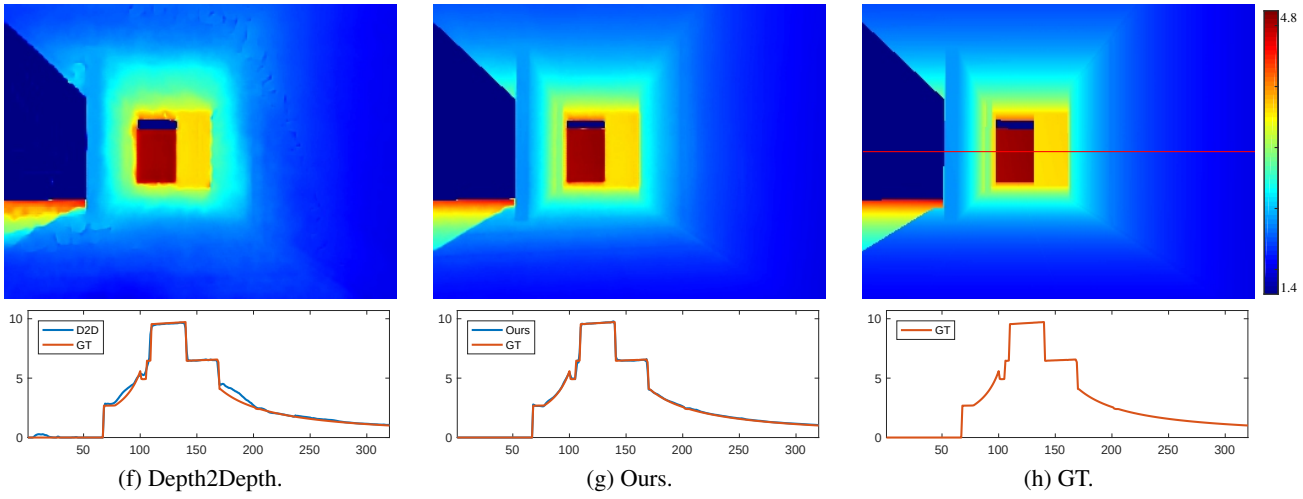
Figure 1: Synthetic results on a long-range scene with relatively low MPI and sensor noise.



(a) Scene. (b) Phase/Correlation.

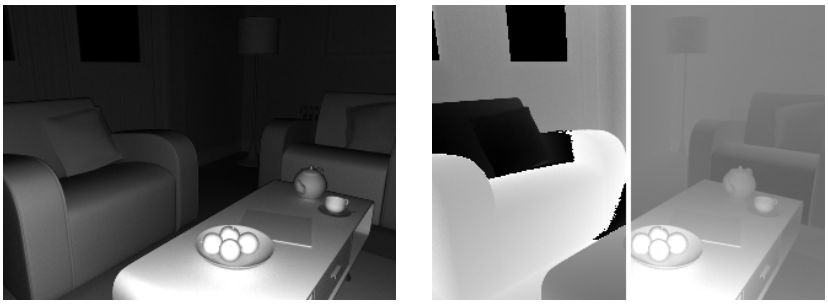


(c) Single. (d) Phasor. (e) SRA.

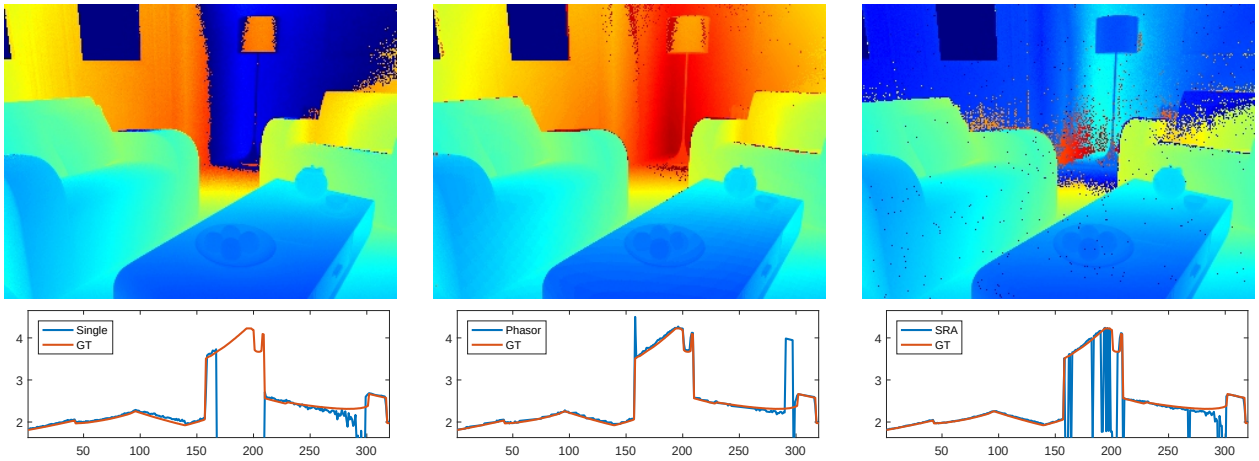


(f) Depth2Depth. (g) Ours. (h) GT.

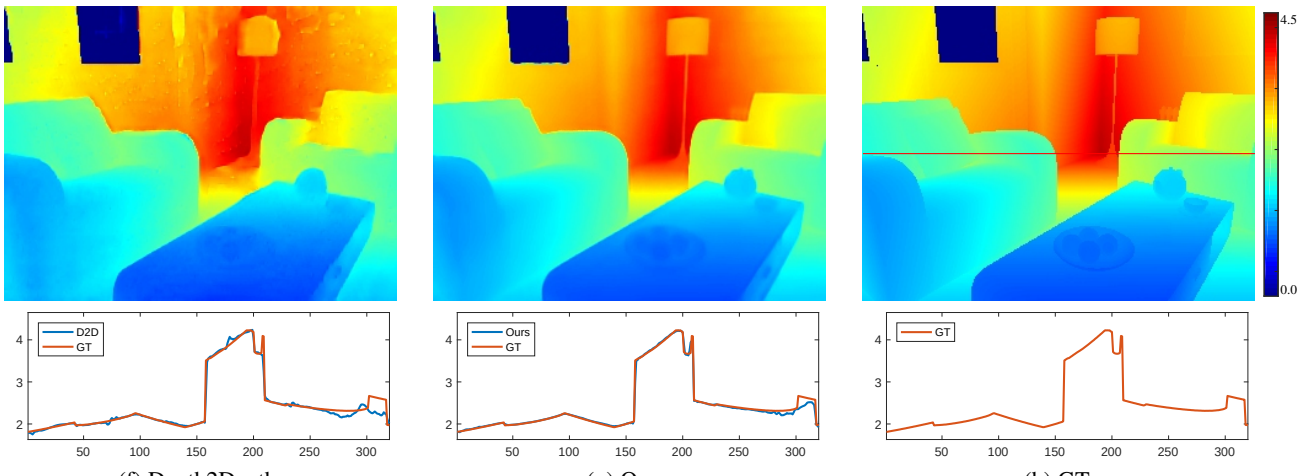
Figure 2: Synthetic results on a long-range scene with relatively high MPI and sensor noise.



(a) Scene. (b) Phase/Correlation.



(c) Single. (d) Phasor. (e) SRA.

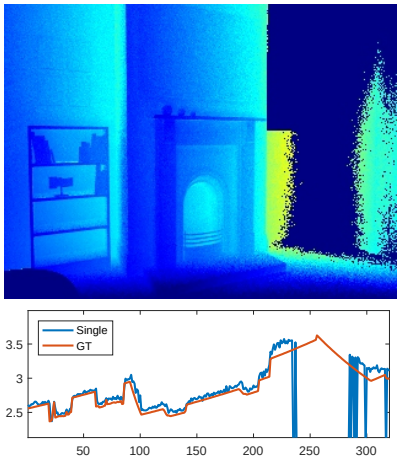


(f) Depth2Depth. (g) Ours. (h) GT.

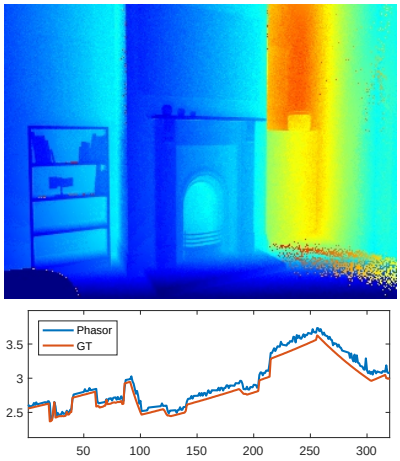
Figure 3: Synthetic results on a common indoor scene with relatively low MPI and sensor noise.

(a) Scene.

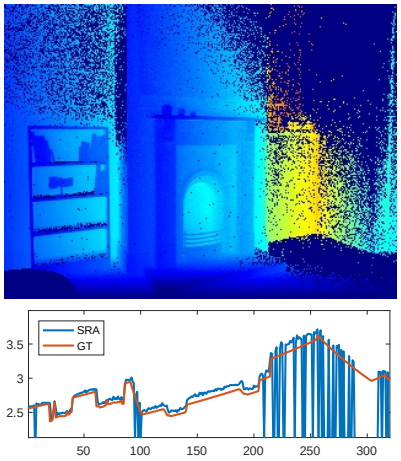
(b) Phase/Correlation.



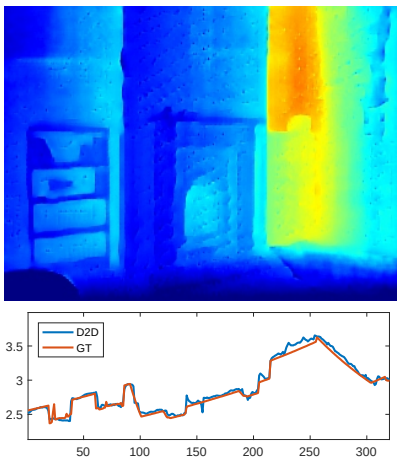
(c) Single.



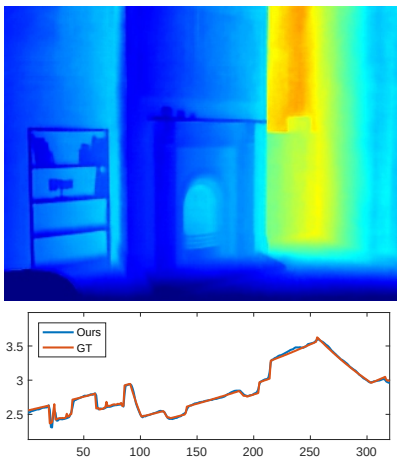
(d) Phasor.



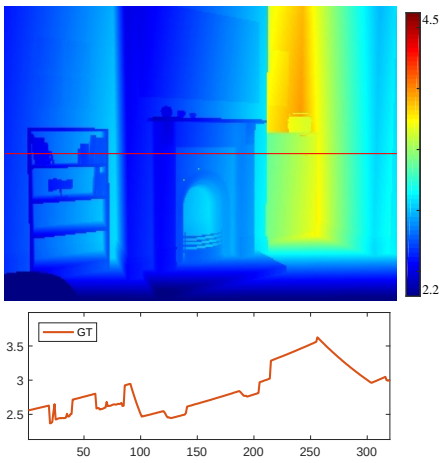
(e) SRA.



(f) Depth2Depth.

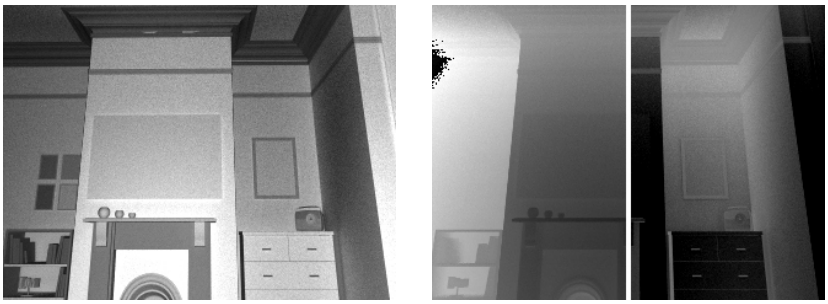


(g) Ours.



(h) GT.

Figure 4: Synthetic results on a common indoor scene with moderate MPI and sensor noise.

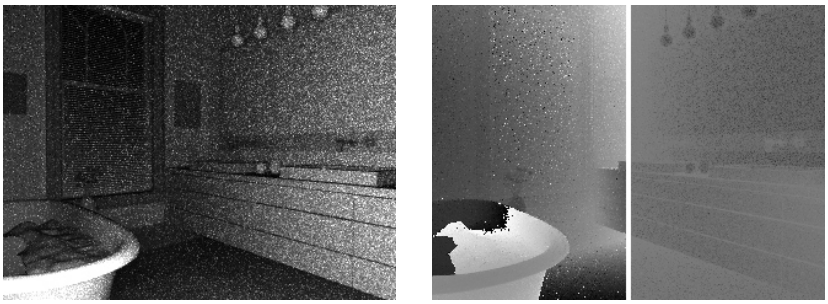


(a) Scene. (b) Phase/Correlation.

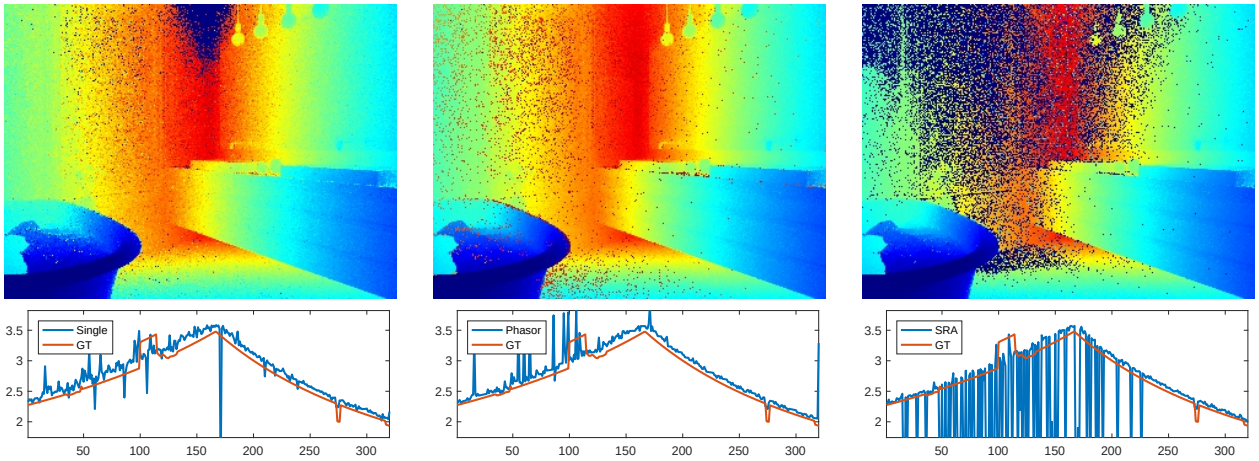
(c) Single. (d) Phasor. (e) SRA.

(f) Depth2Depth. (g) Ours. (h) GT.

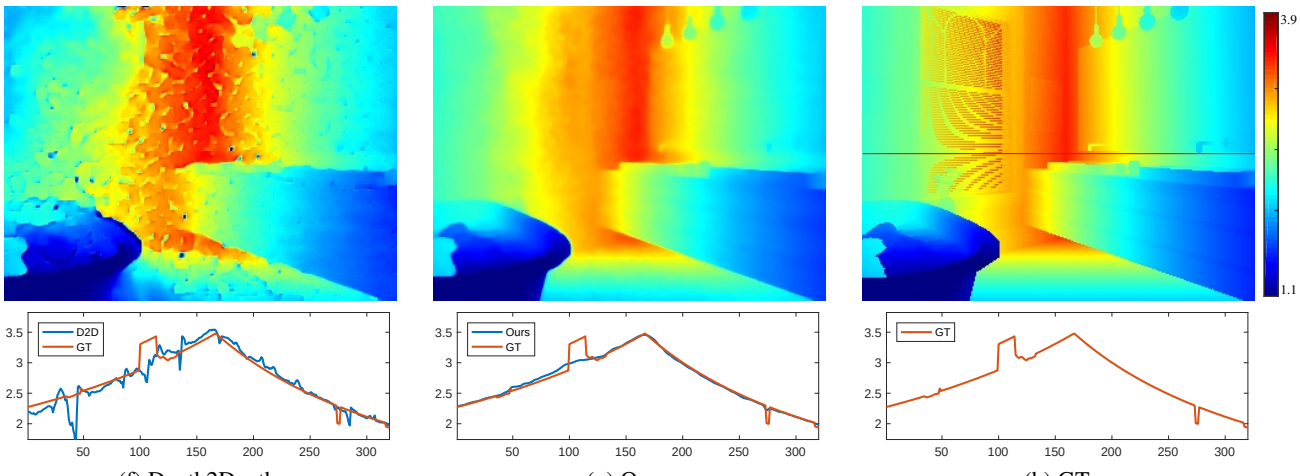
Figure 5: Synthetic results on a common indoor scene with moderate MPI and sensor noise.



(a) Scene. (b) Phase/Correlation.



(c) Single. (d) Phasor. (e) SRA.



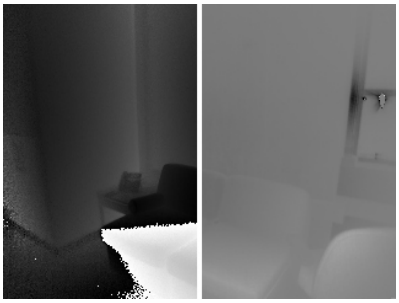
(f) Depth2Depth. (g) Ours. (h) GT.

Figure 6: Synthetic results on a common indoor scene with strong MPI and sensor noise.

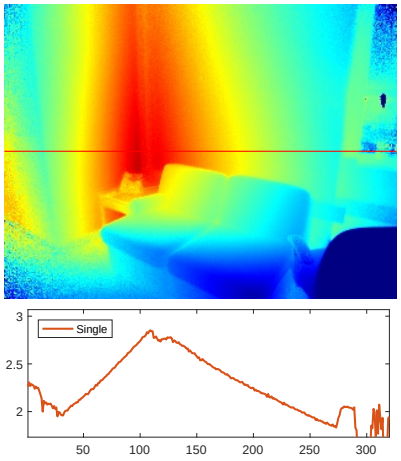
1.2. Real Results

We demonstrate results on experimentally captured raw ToF measurements in Fig. 7, 8, 9, 10, 11, 12, 13, 14, 15. The depth scanlines are sampled from the middle row of each result, highlighted as red lines in the baseline SINGLE outputs in Fig. 7c, 8c, 9c, 10c, 11c, 12c, 13c, 14c, and 15c.

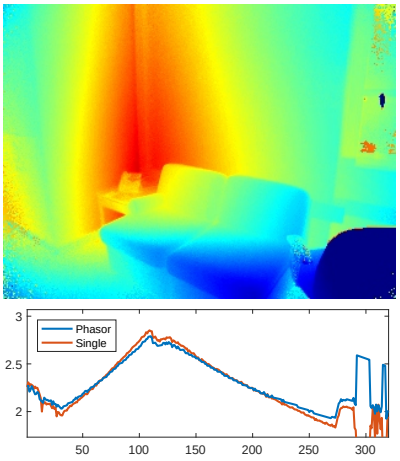
(a) Scene.



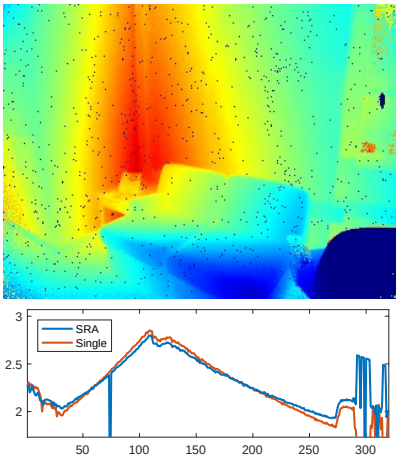
(b) Phase/Correlation.



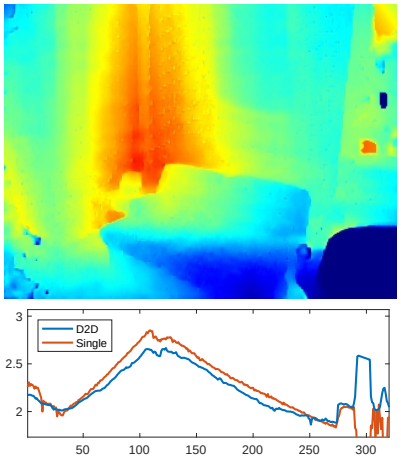
(c) Single.



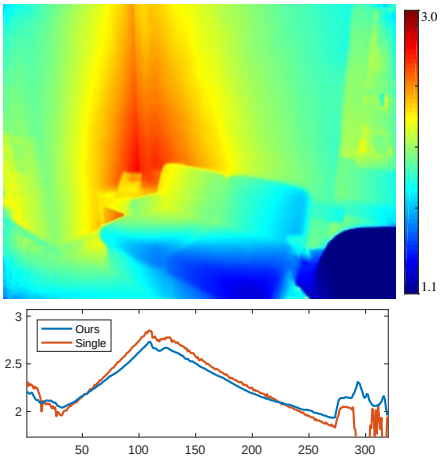
(d) Phasor.



(e) SRA.

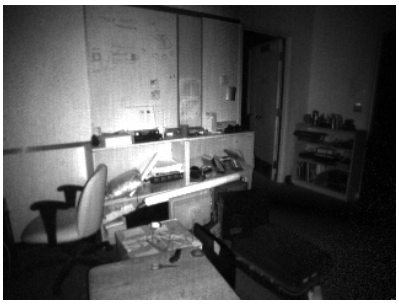


(f) Depth2Depth.



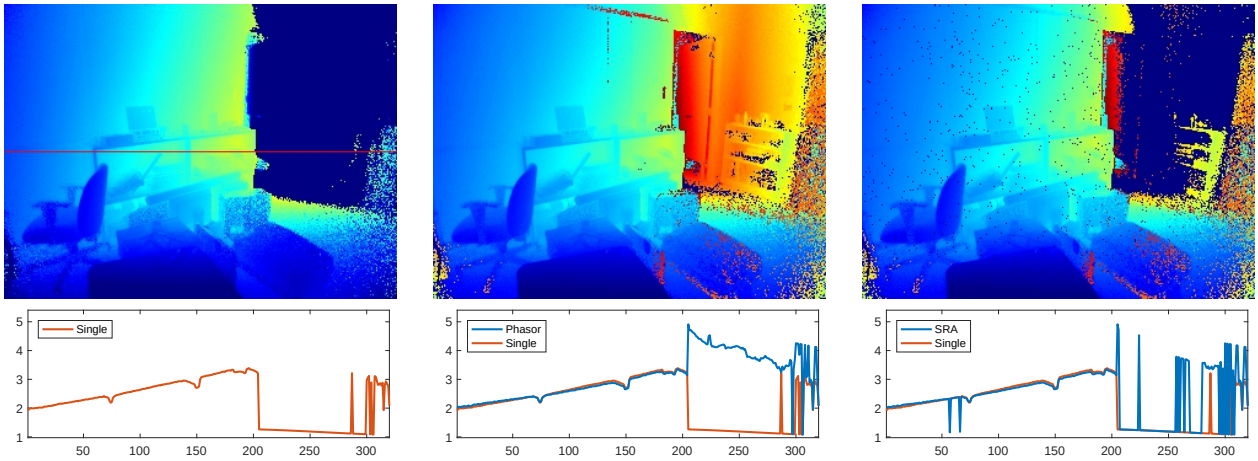
(g) Ours.

Figure 7: Real results from an office scene consisting of a concrete wall corner, sofa, chair and glass window. Our method generates piecewise smooth depth map while compensating for MPI distortion and glass reflectance.

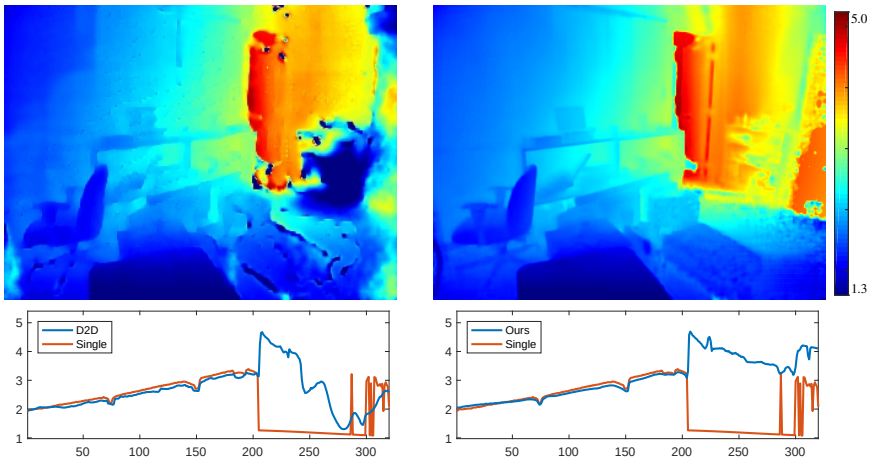


(a) Scene.

(b) Phase/Correlation.



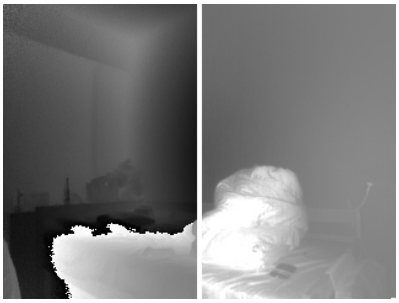
(c) Single. (d) Phasor. (e) SRA.



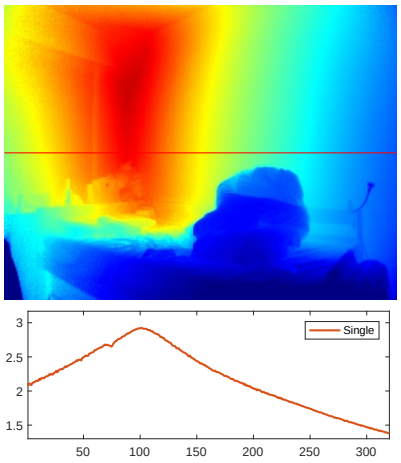
(f) Depth2Depth. (g) Ours.

Figure 8: Real results from a cluttered, long-range office scene consisting of a painted wall, bookshelves, cardboard boxes and dark carpet. Our method generates piecewise smooth depth map, compensates for MPI distortion, and removes noise in the low reflective regions.

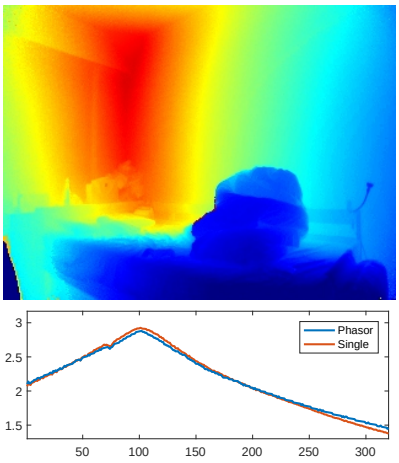
(a) Scene.



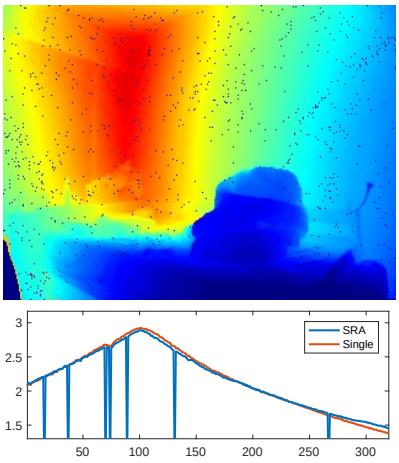
(b) Phase/Correlation.



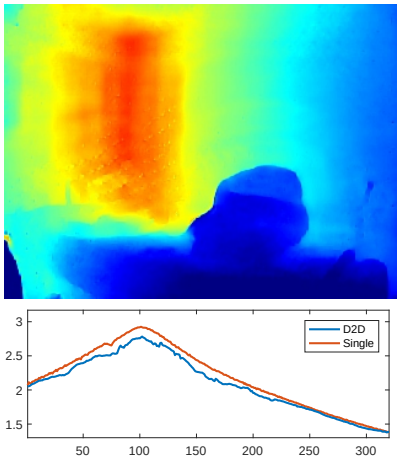
(c) Single.



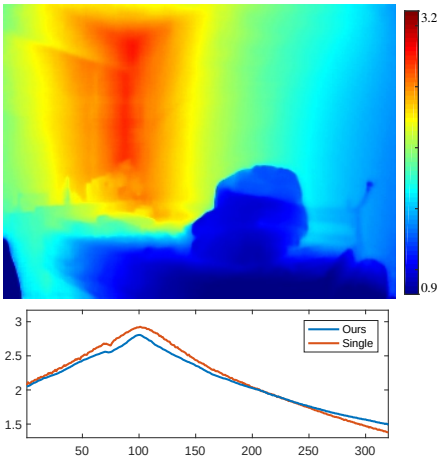
(d) Phasor.



(e) SRA.



(f) Depth2Depth.

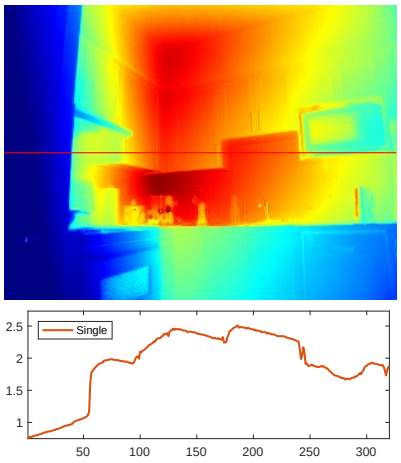


(g) Ours.

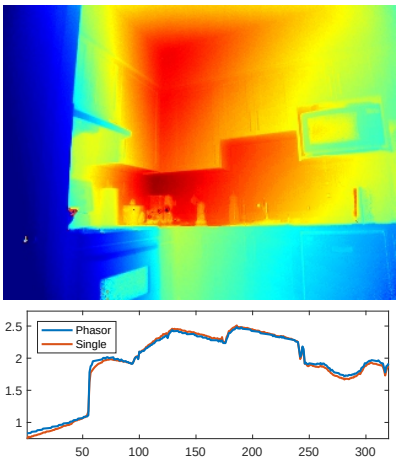
Figure 9: Real results from a bedroom scene consisting of a painted wall, bed and duvets. Our method generates a piecewise smooth depth map while compensating for MPI distortion. Notice the sharp corner reconstruction from our method in the scanline of Fig. 9g compared to the rounded output from the baseline approach as a result of strong multipath distortion near the wall corner.

(a) Scene.

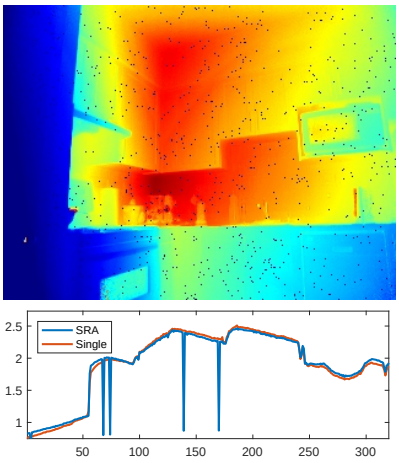
(b) Phase/Correlation.



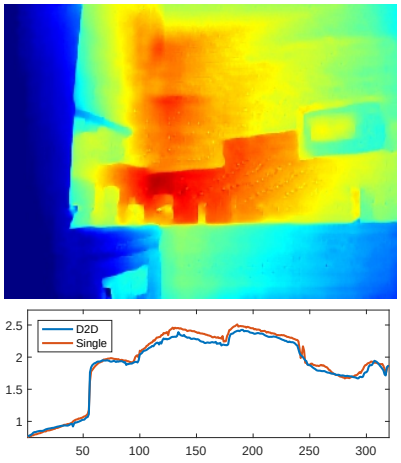
(c) Single.



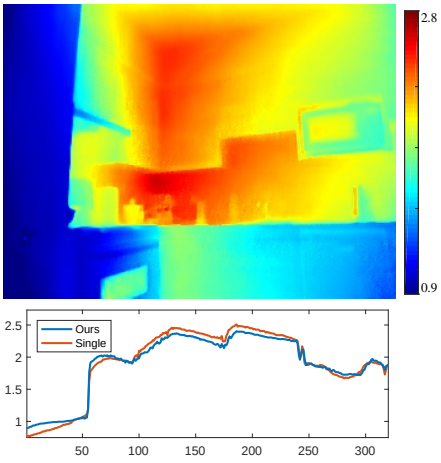
(d) Phasor.



(e) SRA.

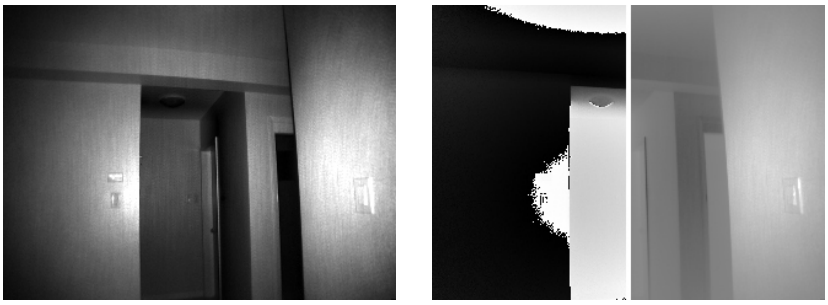


(f) Depth2Depth.

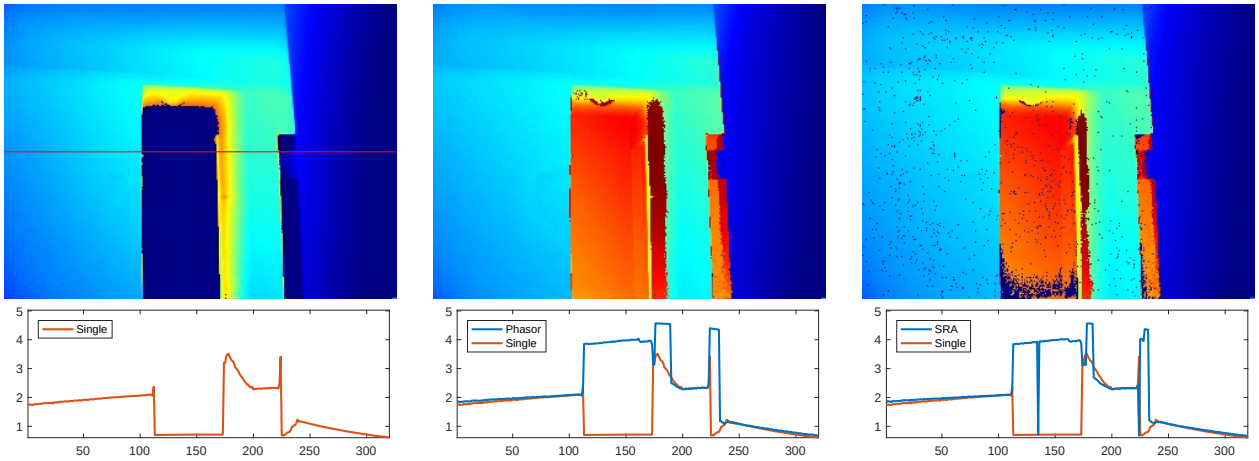


(g) Ours.

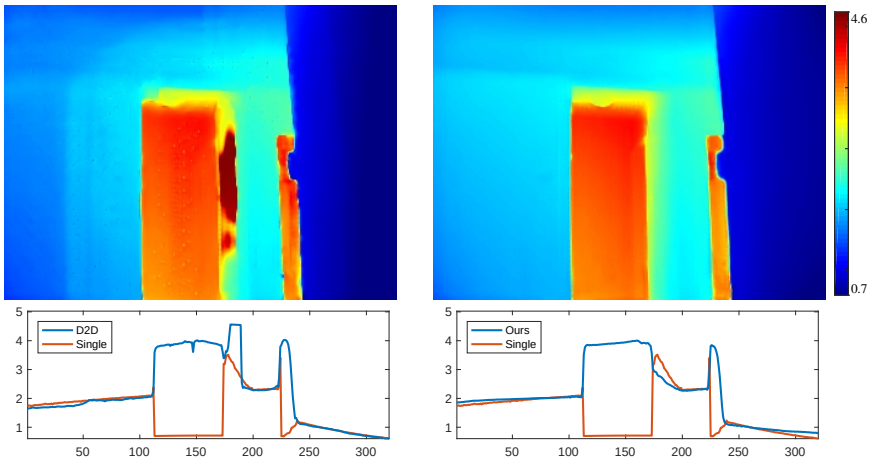
Figure 10: Real results from a kitchen scene consisting of cabinets and kitchenware. Our method generates a piecewise smooth depth map while preserving depth details and compensating for MPI distortion.



(a) Scene. (b) Phase/Correlation.



(c) Single. (d) Phasor. (e) SRA.

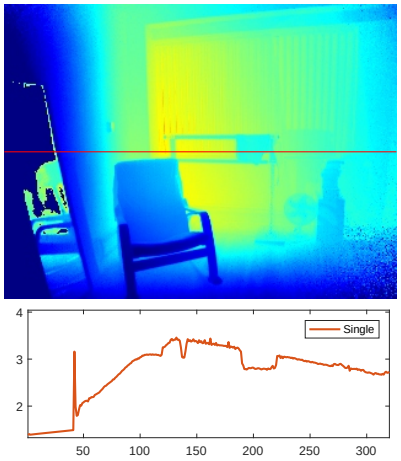


(f) Depth2Depth. (g) Ours.

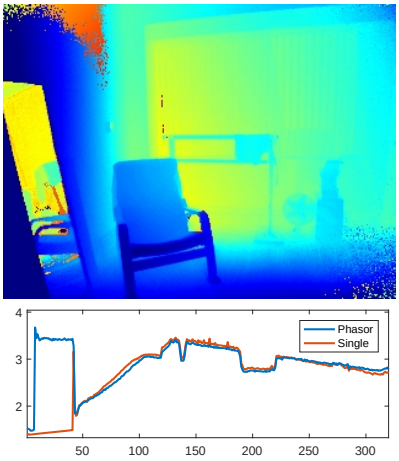
Figure 11: Real results from a long-range indoor scene consisting of planar walls and mirrors of different orientations and distances. Our method generates a piecewise smooth depth map while resolving phase ambiguities and compensating for MPI distortion.

(a) Scene.

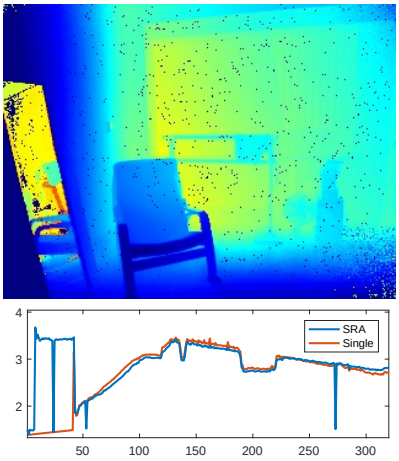
(b) Phase/Correlation.



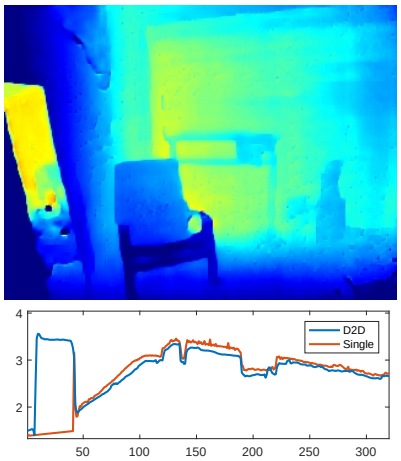
(c) Single.



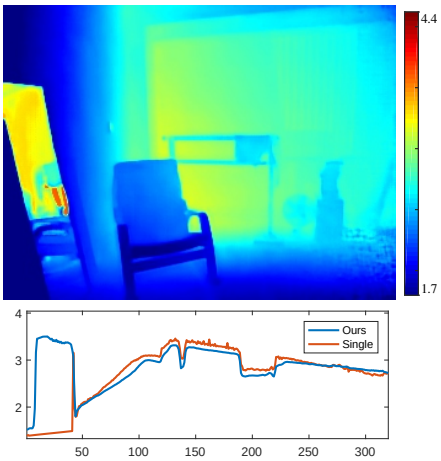
(d) Phasor.



(e) SRA.



(f) Depth2Depth.

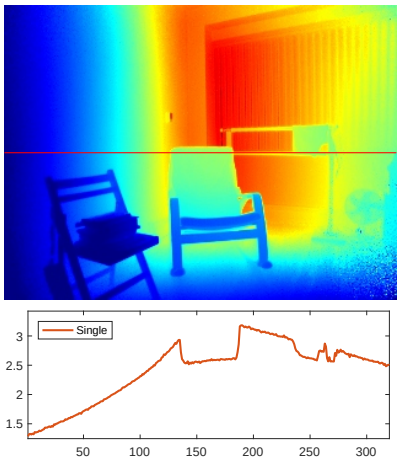


(g) Ours.

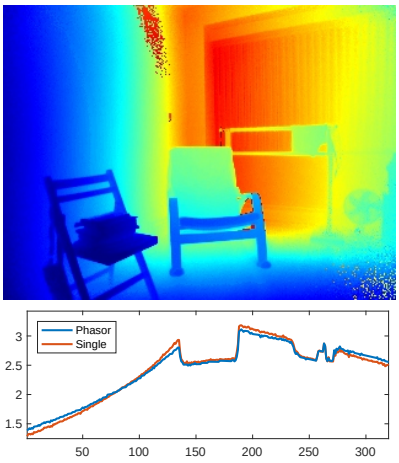
Figure 12: Real results from a cluttered living room scene consisting of painted walls, window blinds, hanger, a rocking chair and mirror. Our method generates a piecewise smooth depth map while resolving phase ambiguities and compensating for MPI distortion.

(a) Scene.

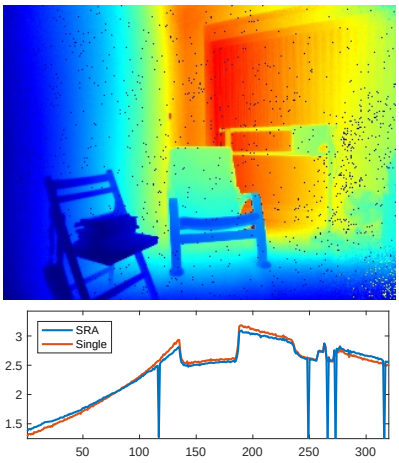
(b) Phase/Correlation.



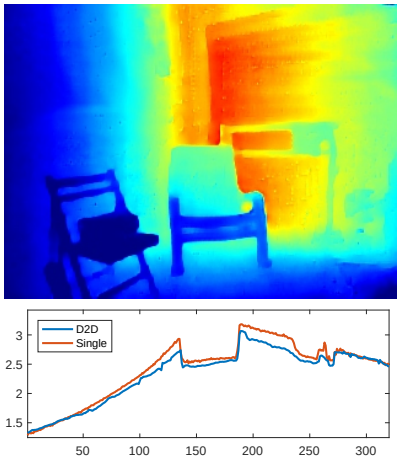
(c) Single.



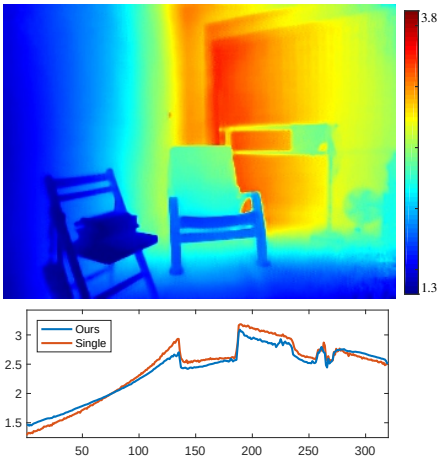
(d) Phasor.



(e) SRA.



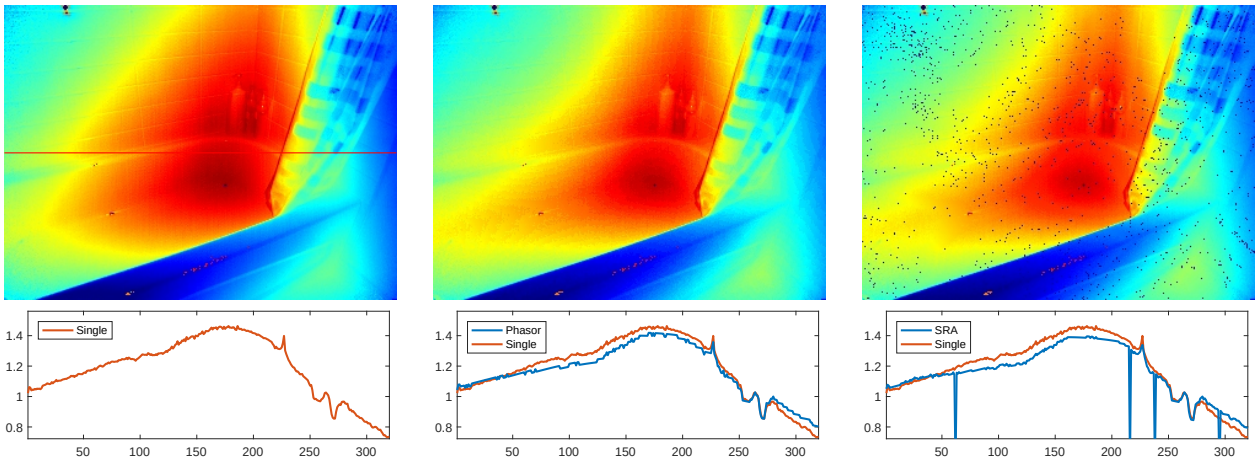
(f) Depth2Depth.



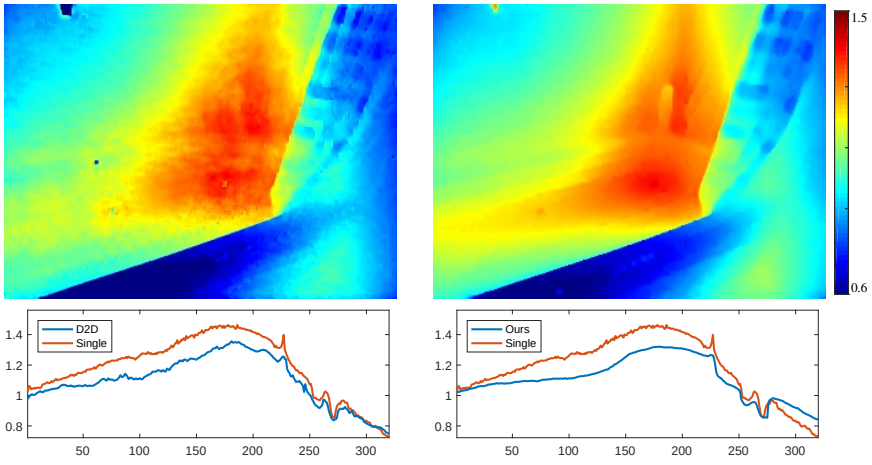
(g) Ours.

Figure 13: Real results from a cluttered living room scene consisting of painted walls, window blinds, hanger, and chairs. Our method generates a piecewise smooth depth map while compensating for MPI distortion.

(a) Scene. (b) Phase/Correlation.

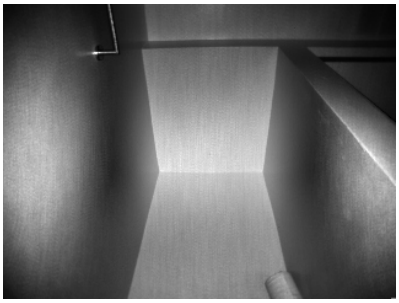


(c) Single. (d) Phasor. (e) SRA.

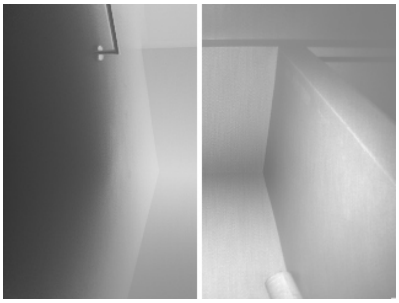


(f) Depth2Depth. (g) Ours.

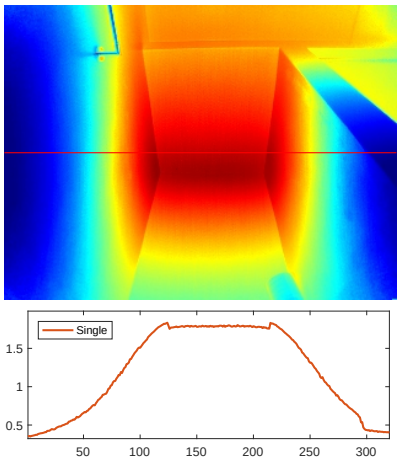
Figure 14: Real results from a bathroom scene consisting of a ceramic bathtub and plastic curtains, with strong MPI leading to 10-20cm depth distortions. Our method generates piecewise smooth depth map while compensating for MPI distortions. Notice that, similar to other approaches, the proposed framework fails in the curtain area, treating the tile boundaries as depth edges. This is an limitation of our method which we will discuss in Sec. 4.



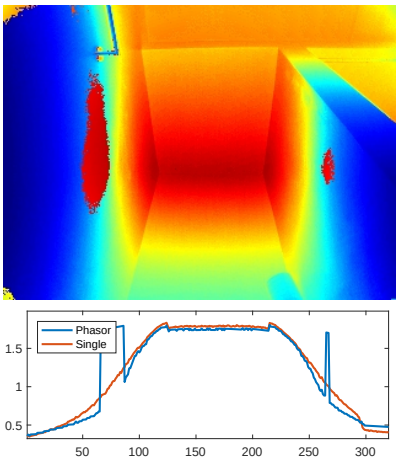
(a) Scene.



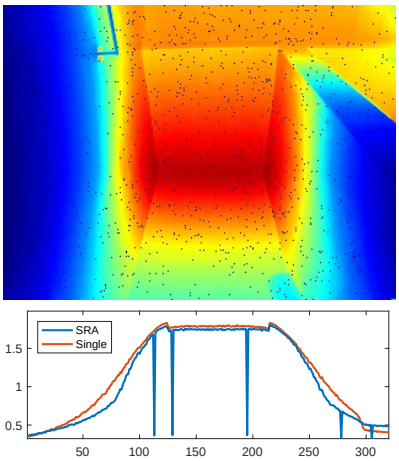
(b) Phase/Correlation.



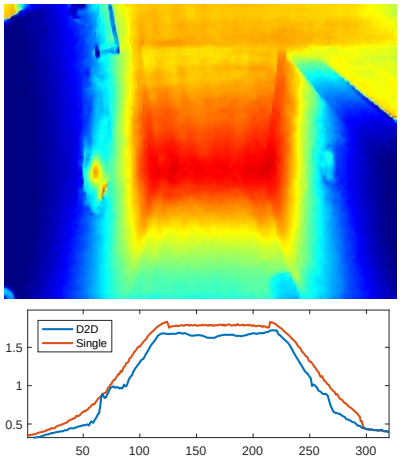
(c) Single.



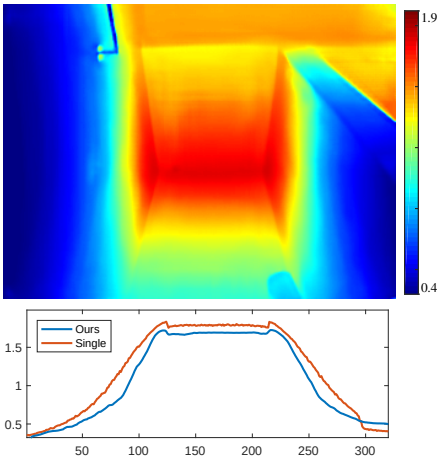
(d) Phasor.



(e) SRA.



(f) Depth2Depth.



(g) Ours.

Figure 15: Real results from an indoor corner scene consisting of orthogonal painted walls. Our method generates a piecewise smooth depth map while compensating for MPI distortion.

2. Additional Experiments

In this section we demonstrate the temporal consistency of our method, its robustness to scene albedo, and the effect of loss functions.

2.1. Temporal Consistency

In Fig. 16, we demonstrate that the proposed method produces temporally consistent depth prediction results. Specifically, we capture a short image burst with our ToF camera and document the reconstructed depth maps for each individual frame, without other frames being seen by our framework. The proposed method generates temporally consistent depth maps with reduced MPI distortion, which are also apparent in the scanline visualization at the bottom of the figure.

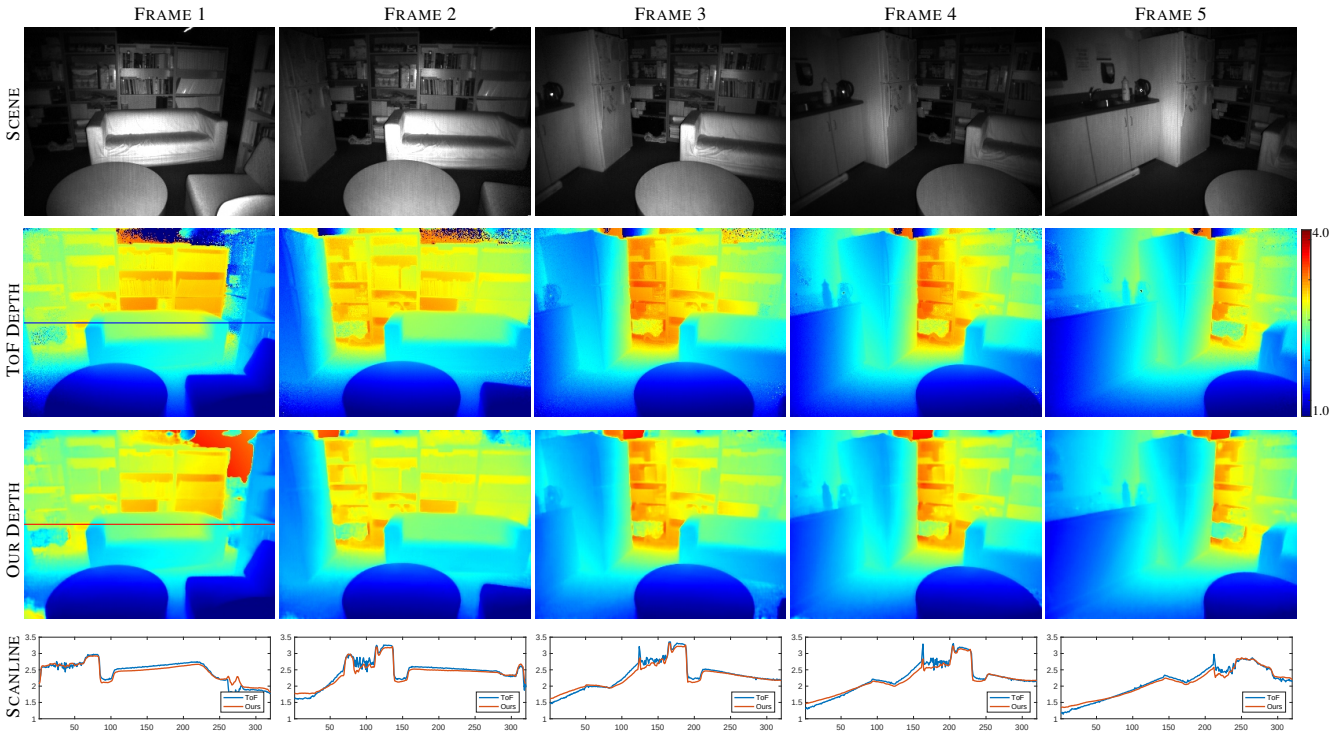


Figure 16: Evaluation of the temporal consistency with experimentally captured raw ToF measurements. We denote ToFDepth as the single frequency output measured at $\omega_1 = 40$ MHz, and OurDepth as the results of the proposed framework.

2.2. Robustness to Albedos

We demonstrate the robustness of our framework with respect to scene albedo variations in Fig. 17 and Fig. 18. Note that for these results only object reflectivity varies. Compared to competing methods, our results are substantially less affected by albedo variations, in addition to eliminating noise, phase ambiguity and MPI.

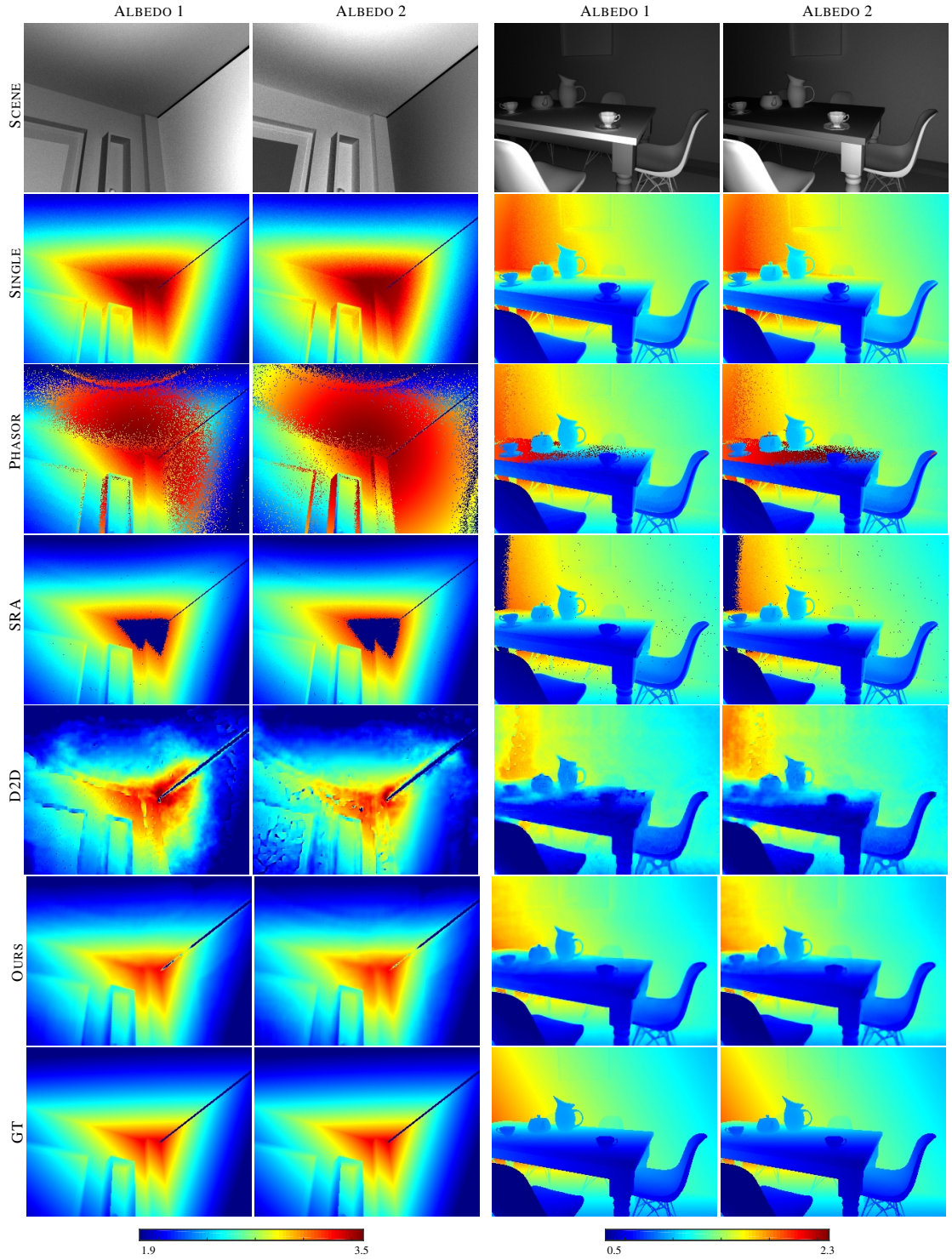


Figure 17: Robustness to albedos on two pairs of synthetic scenes.

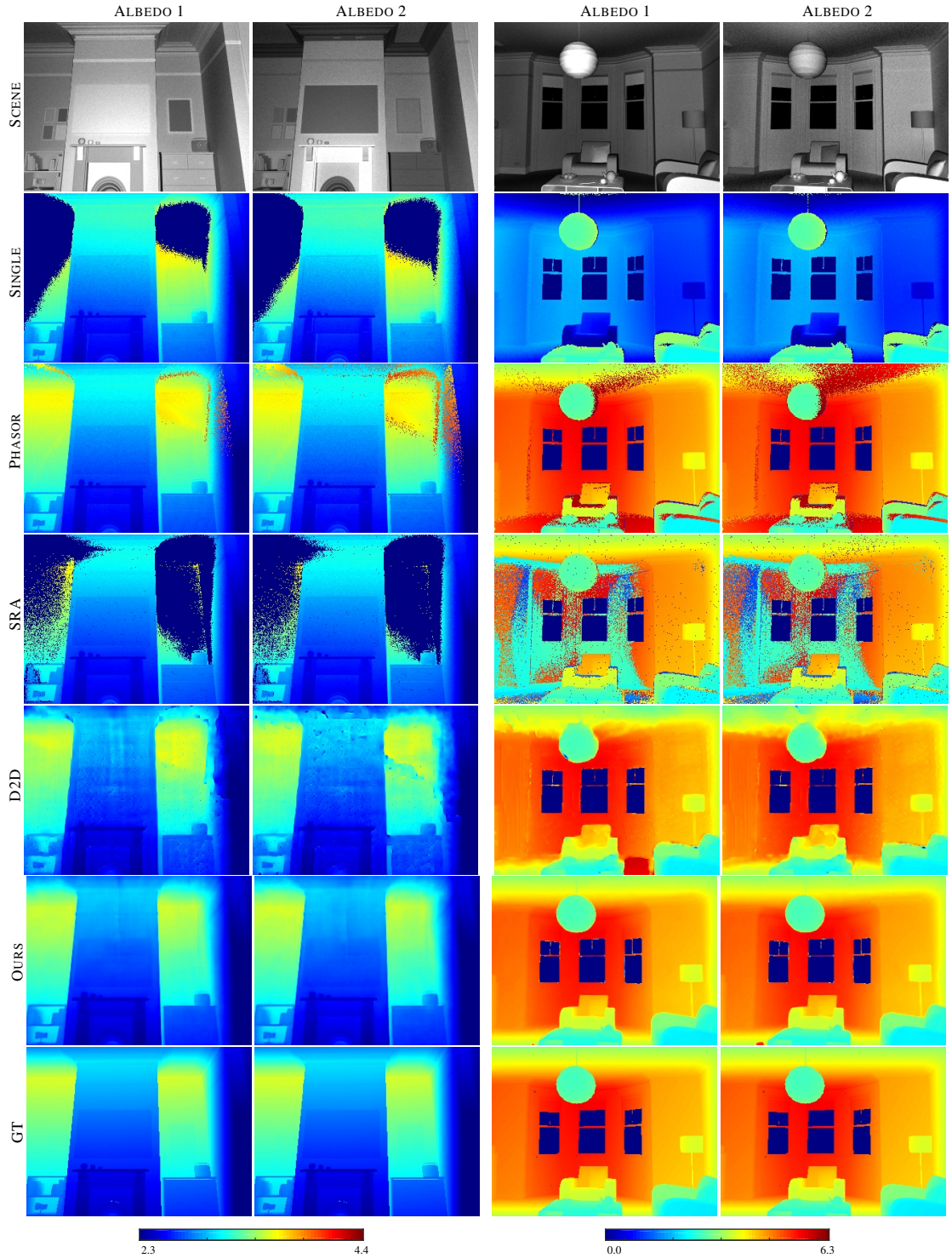


Figure 18: Robustness to albedos on two pairs of synthetic scenes.

2.3. Effect of λ_s and λ_a

We have trained variants of ToFNet with different pairs of λ_s and λ_a. See examples in Fig. 19 on results of real ToF measurements. With increasing λ_a, i.e. the weight of the adversarial loss, our network learns to generate sharper depth edges but it also hallucinates high-frequency noise (see column $\lambda_a = 0.03$ and $\lambda_a = 0.1$). When the weight on the smoothness term λ_s is increased, our network generates over-regularized "cartoon"-style depth maps with reduced noise in an edge-aware fashion. As discussed in the main paper, our final objective function combines the relative strengths of each loss.

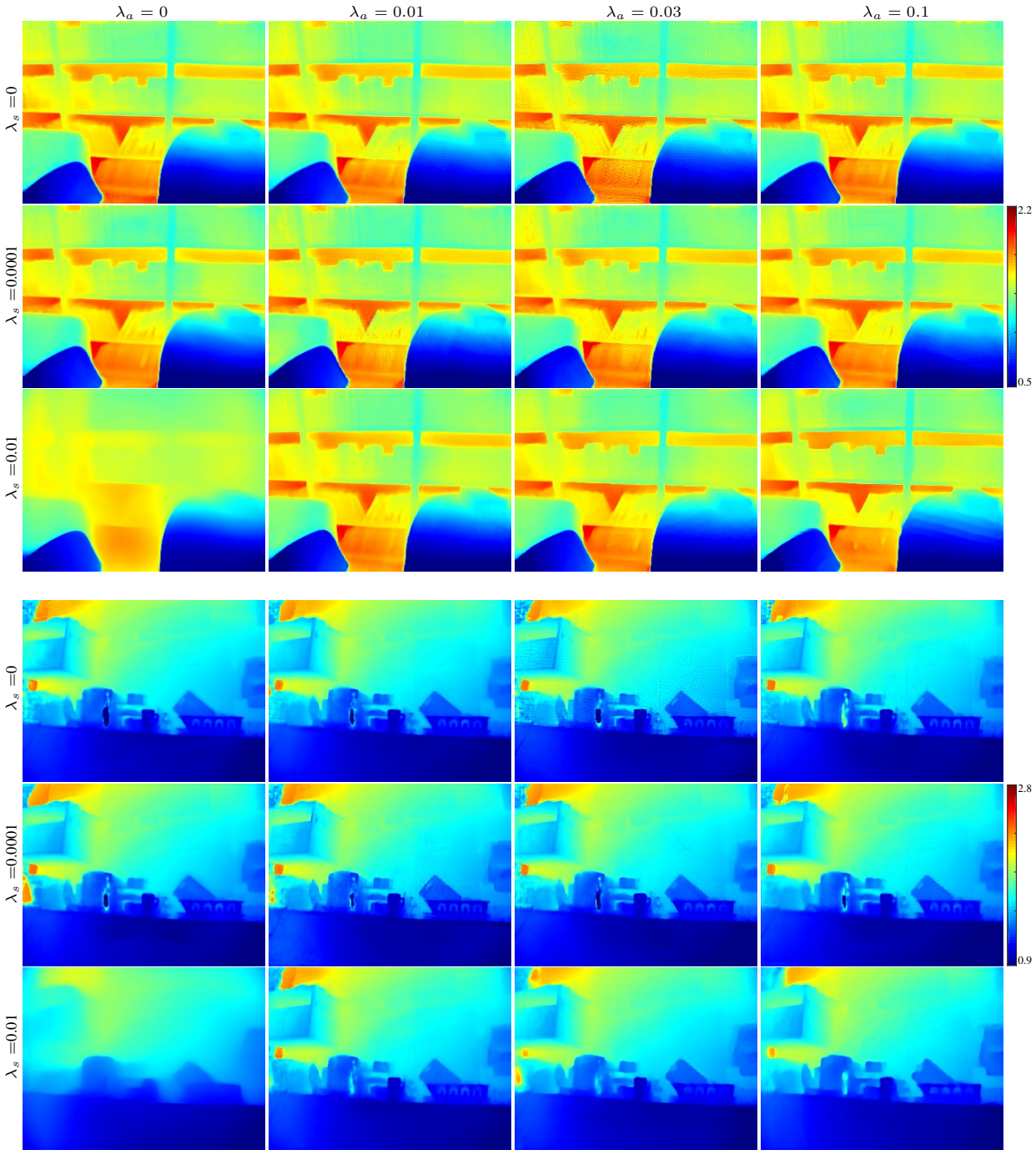


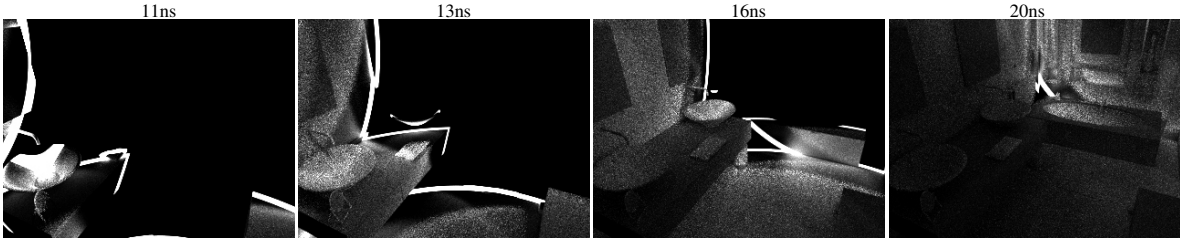
Figure 19: Effect of λ_a and λ_s variations on two example scenes, Bookshelf (top) and Kitchen (bottom). Results are best seen with a computer.

3. Implementation Details

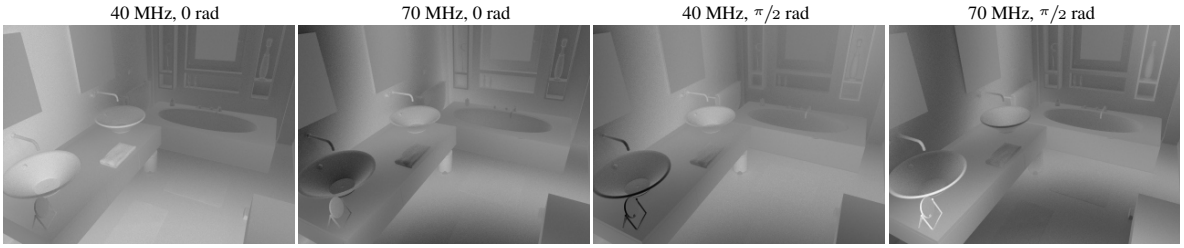
3.1. Dataset Generation

To generate the synthetic dataset for training and quantitative evaluations, we first simulate the transient images of a given scene model (Fig. 20a) with our time-resolved renderer. Fig. 20b shows a few synthesized transient images from this step, in which both direct peak and indirect global illumination are present. The correlation images of the scene can then be generated as described in the main manuscript, and these results are shown in Fig. 20c. Finally, we provide the full size phase and amplitude images in Fig. 20d.

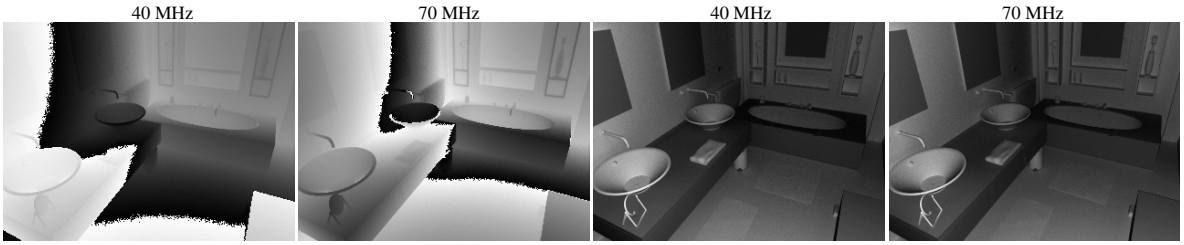
(a) 3D scene model in Blender.



(b) Transient frames.



(c) Correlation images.



(d) Phase and amplitude images.

Figure 20: Synthetic data generation steps.

3.2. Network Training

Tab. 1 provides layer specifications of the ToFNet model. Each convolutional layer in G is followed by ReLU, except those that are skip connected to deeper layers where the sum is rectified through a ReLU layer. At the end of the network a Tanh layer is applied to normalize the intensities. In D we use 0.2 LeakyReLU and, unlike traditional GAN methods, omit the final Sigmoid layer. We adopt the Torch implementation of `SpatialConvolution` and `SpatialFullConvolution` for down- and up-convolutional layers.

Generator G

Layer	Kernel Size	Stride	Output Size	Skip Connection
Input	-	-	$64\times H\times W$	to TV
F1_1	7×7	1×1	$64\times H\times W$	-
F1_2	3×3	1×1	$64\times H\times W$	to U2
D1	3×3	2×2	$128\times H/2\times W/2$	-
F2	3×3	1×1	$128\times H/2\times W/2$	to U1
D2	3×3	2×2	$256\times H/4\times W/4$	-
R1-R9	3×3	1×1	$256\times H/4\times W/4$	-
U1	4×4	$1/2\times 1/2$	$128\times H/2\times W/2$	from F2
F3	3×3	1×1	$128\times H/2\times W/2$	-
U2	4×4	$1/2\times 1/2$	$64\times H\times W$	from F1_2
F4_1	3×3	1×1	$64\times H\times W$	-
F4_2	3×3	1×1	$64\times H\times W$	-
TV	-	-	$1\times H\times W$	from Input*

Discriminator D

Layer	Kernel Size	Stride	Output Size
Input	-	-	$1\times H\times W$
D1	4×4	2×2	$64\times H/2\times W/2$
D2	4×4	2×2	$128\times H/4\times W/4$
D3	4×4	2×2	$256\times H/8\times W/8$
F1	4×4	1×1	$1\times H/4\times W/4$

Table 1: Specifications of the generator and discriminator networks of the proposed ToFNet model. We train on image patches with $H = 128$ and $W = 128$. *The amplitude image is extracted from the input correlation measurements and is skipped to the TV layer for calculating the edge-aware smoothness term.

We apply two recent successful techniques from [6] and [4] to stabilize the training procedure. In particular, we optimize for the least square loss during the D updates using a history of the generated depth maps. This strategy proves to be effective on reducing model oscillation and generating higher quality results in our experiment. Our code and datasets will be made publicly available in the future.

4. Limitations

As shown in the previous sections, our method gracefully degrades when the scene consists of saturated regions and/or severe noise as a result of low reflectivity and long distance light fall-off, see Fig. 14c. In the future, we plan to investigate this problem by considering the sensor's noise model and radiometric response function when simulating the training data, as well as devising unsupervised strategies [6] to improve the realism of the synthetic training results.

References

[1] A. A. Dorrington, J. P. Godbaz, M. J. Cree, A. D. Payne, and L. V. Streeter. Separating true range measurements from multi-path and scattering interference in commercial range cameras. In *Conference on the Three-Dimensional Imaging, Interaction, and Measurement*, volume 7864, pages 1–1. SPIE–The International Society for Optical Engineering, 2011. 2

[2] D. Freedman, Y. Smolin, E. Krupka, I. Leichter, and M. Schmidt. Sra: Fast removal of general multipath for tof sensors. In *European Conference on Computer Vision*, pages 234–249. Springer, 2014. 2

[3] M. Gupta, S. K. Nayar, M. B. Hullin, and J. Martin. Phasor imaging: A generalization of correlation-based time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 34(5):156, 2015. 2

[4] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *The IEEE International Conference on Computer Vision*, 2017. 23

[5] J. Marco, Q. Hernandez, A. Muoz, Y. Dong, A. Jarabo, M. Kim, X. Tong, and D. Gutierrez. Deeptof: Off-the-shelf real-time correction of multipath interference in time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 36(6), 2017. to appear. 2

[6] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *The IEEE conference on computer vision and pattern recognition*, 2017. 23