DEPTHOR: Depth Enhancement from a Practical Light-Weight dToF Sensor and RGB Image

Supplementary Material

This supplementary material provides additional details to complement the main paper. It includes introduction of dToF imaging (Sec. 1), detailed training settings (Sec. 2), descriptions of the adopted evaluation metrics (Sec. 3), introduction of dToF projection (Sec. 4), implementation details of the dToF simulation method (Sec. 5), additional ablation studies about simulation method (Sec. 6), and additional experimental results (Sec. 7).

1. Preliminary: dToF Imaging

We first briefly introduce the imaging principle of dToF. As shown in Fig. 1, a pulsed laser generates a short light pulse and emits it into the scene. The pulse scatters, and some photons are reflected back to the dToF detector. The depth is then determined by the formula $d = \Delta t \cdot c/2$, where Δt is the time difference between laser emission and reception, and c is the speed of light. Each dToF pixel captures all scene points reflected within its individual field-of-view (iFoV) using time-correlated single-photon counting (TCSPC). The iFoV is determined by the sensor's total field-of-view (FoV) and spatial resolution, returning the peak signal detected within that range. Interested readers are referred to [1, 5, 11] for more details.

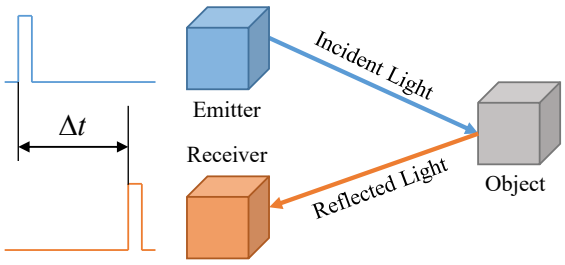

Figure 1. Imaging principle of direct Time-of-Flight sensor

2. Training Setting.

We implement our method in pytorch[6] and train it on 4 Nvidia RTX 3090 GPUs. We adopt AdamW[4] with 0.1 weight decay as the optimizer, and clip gradient whose l^2-norm is larger than 0.1. Our model is trained from scratch in roughly 230K iterations using the OneCycle[8] learning rate policy, setting the initial learning rate to 1/25 of the maximum learning rate and gradually reducing the learning rate to 1/100 of the maximum learning rate in the later stages of training. We set batch size as 12 and the largest learning rate as 0.0003.

3. Details on Evaluation Metrics

We present the precise definitions of the quantitative metrics reported in the main paper, which include δ_i, Rel, RMSE, $\log_{10}$, and edge-weighted mean absolute error (EWMAE). These metrics are defined as follows:

$$\text{Rel} = \frac{1}{|P|}\sum \frac{|y_p - x_p|}{y_p},$$

$$\text{RMSE} = \sqrt{\frac{1}{|P|}\sum (y_p - x_p)^2},$$

$$\text{EWMAE} = \frac{1}{|P|}\frac{\sum G_p \cdot |y_p - x_p|}{\sum G_p},$$

$$\delta_i = \frac{1}{|P|}\sum \left(\max\left(\frac{y_p}{x_p}, \frac{x_p}{y_p}\right) < 1.25^i\right),$$

$$\log_{10} = \frac{1}{|P|}\sum |\log_{10}(y_p) - \log_{10}(x_p)|$$

Here, x_p and y_p represent the predicted value and ground truth at valid pixel locations, respectively. The set P contains all pixels with valid ground truth, and $|P|$ denotes the total number of such pixels.

Following [3, 9, 10], we compute the weight coefficient G_p for a pixel p based on its intensity and directional gradients. First, the directional gradient $\nabla_D I(p)$ is calculated as:

$$\nabla_D I(p) = V_{pD} - V_p$$

where $D \in \{N, S, E, W\}$ represents the north, south, east, and west neighbors of pixel p and V_p is the depth of p. Using these gradients, we compute the reciprocals of directional conduction functions G_{D_p}, which is expressed as:

$$G_{D_p} = \frac{[\nabla_D I(p)]^2}{[\nabla_D I(p)]^2 + \kappa^2}$$

κ is a regularization constant. Finally, the weight coefficient G_p is obtained as the average of these directional coefficients:

$$G_p = \frac{G_{N_p} + G_{S_p} + G_{E_p} + G_{E_p}}{4}$$

Each pixel's weight G_p can be calculated based on the above formula. The weight approaches 0 when the pixel is in a homogeneous region and approaches 1 when the gradient in all four directions reaches a maximum.

4. Project dToF to Sparse Depth Map

Each dToF measurement provides a 3D point in the dToF sensor coordinate system:

$$P_{dToF} = (X_{dToF}, Y_{dToF}, Z_{dToF}, 1)^T. \tag{1}$$

The transformation from the dToF coordinate system to the RGB camera coordinate system is given by:

$$P_{RGB} = T_{dToF \to RGB} P_{dToF}, \tag{2}$$

where the transformation matrix is:

$$\begin{aligned} &T_{dToF \to RGB} = \begin{bmatrix} R_T & t_T \\ 0 & 1 \end{bmatrix}, \\ &R_T = R_{RGB} R_{dToF}^{-1}, \\ &t_T = t_{RGB} - R_{RGB} R_{dToF}^{-1} t_{dToF}. \end{aligned} \tag{3}$$

The transformed 3D point is then projected onto the RGB image using the intrinsic matrix K_{RGB} to get the homogeneous image coordinates:

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = K_{RGB} \begin{bmatrix} X_{RGB} \\ Y_{RGB} \\ Z_{RGB} \end{bmatrix}. \tag{4}$$

where

$$K_{RGB} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}. \tag{5}$$

The final pixel coordinates (u, v) are obtained via perspective division:

$$u = \frac{f_x X_{RGB}}{Z_{RGB}} + c_x, \quad v = \frac{f_y Y_{RGB}}{Z_{RGB}} + c_y. \tag{6}$$

Existing depth super-resolution methods typically compute the iFoV region coordinates for each measurement based on this central coordinate, resolution, and FoV. However, calibration errors can cause significant shifts in these depth points. Therefore, we approach this problem from the perspective of depth completion robustness.

5. Details of dToF Simulation Method

We trained our model on the Hypersim dataset. To reduce the impact of invalid data, we scaled some of the depth values that exceeded the sensor's detection limit. Similar to the approach of Sun *et al*. [11] on [12], if 60% or more of the depth values in an image exceed 6 meters, all depth values are halved. Additionally, we modified the parameters of our simulation method for each test dataset to match the characteristics of different dToF sensors.

ZJU-L5 Dataset. The resolution of the dToF sensor and the depth ground truth are 8×8 and 480×640, respectively. According to the calibration results provided by the authors, the FoV of the L5 sensor covers approximately 61% of the GT. The mean boundary values of its projected region on the GT are [-25, 405, 85, 535], corresponding to the upper (h_u), lower (h_l), left (w_l), and right (w_r) boundaries, respectively. Each dToF signal corresponds to an iFoV of approximately 52×56 pixels. Additionally, the maximum depth recorded by the L5 sensor is 4.1 m, whereas the maximum depth in the GT is 10 m.

Due to the low power of the L5 sensor, it typically exhibits signal loss in specific regions rather than returning incorrect depth values. Based on the dataset masks, the probability of signal loss is approximately 30%. As the authors performed strict calibration and no noticeable calibration errors were observed in the visualization results, we did not consider region shift in our simulation method.

Our Real-world Samples. The resolution of the dToF sensor and RGB camera are 40×30 and 912×684, respectively. To allow for 1/32 downsampling, we padded the images to 928×714. We used the internal parameters of the mobile phone to project the raw dToF signals; the FoV of the dToF sensor covers approximately 81% of the image. The mean boundary values of its projected region on the image are [30, 900, 40, 660]. Each dToF signal corresponds to an iFoV of approximately 21×21 pixels. Additionally, the maximum depth recorded by the dToF sensor is 6 m, whereas the theoretical detection limit is 8.1 m.

Due to the higher performance of the dToF sensor, it can still receive photons that pass through non-Lambertian surfaces and may return valid depth values even in low-reflectivity regions. As a result, the collected samples exhibit more complex anomalies. To address this, we set the probability of depth loss to 80% for pixels with a V-channel value below 40 in the HSV color space and assigned corresponding anomalies based on semantic labels.

6. Ablation Studies of dToF Simulation

We demonstrated the effectiveness of certain components of our simulation method through quantitative results on the ZJU-L5 dataset in the main paper. Since the calibartion errors are not considered on the ZJU-L5, in this section, we provide additional visualizations on our collected data as a supplement. The results presented were obtained by training the lightweight PENet [2] on the Hypersim [7] dataset and evaluating its performance on real-world data.

During these experiments, we simulated signal loss in distant regions and applied supervision, which occasionally caused the model to predict areas with missing depth input as distant regions incorrectly, since PENet lack of global relationships provided by the MDE model.

Figure 2 illustrates the improvements in boundary pre-

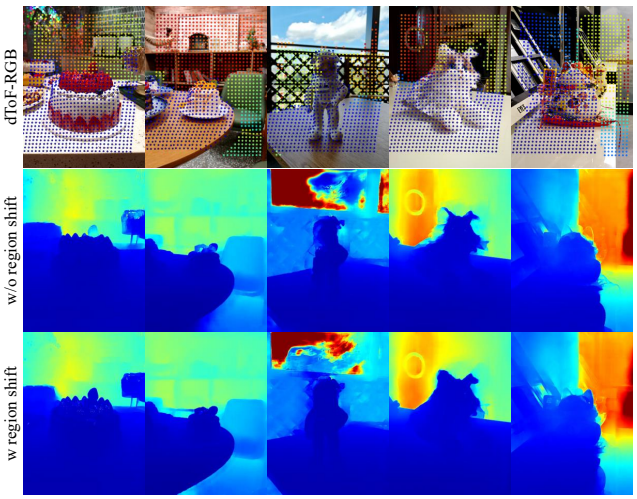

Figure 2. **Effect of simulating calibration errors.** Prediction results are generated by the lightweight PENet[2].

dictions achieved by incorporating region shifts. These include resolving foreground-background overlaps caused by calibration errors and correcting errors at object boundaries, where dToF depth points represent the regional peak value.

Figure 3 illustrates the results of simulating non-Lambertian surfaces. In cases of signal loss, the model utilizes surrounding information to predict values instead of directly assigning distant depths. Moreover, when photons pass through objects and return erroneous values, the model demonstrates the ability to partially correct these signals.

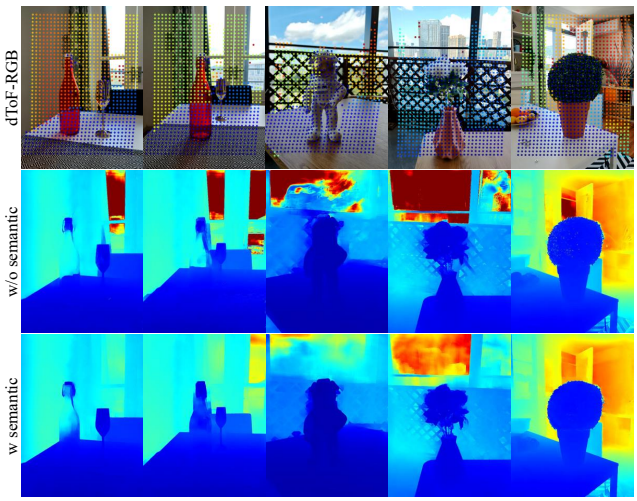

Figure 3. **Effect of simulating Non-Lambertian regions.** Prediction results are generated by the lightweight PENet[2].

7. Additional Experimental Results

Due to space limitations, we present additional experimental results here. Figure 5, Fig. 7 and Fig. 6 show the results on our dToF samples, the ZJU-L5 dataset and the NYUv2 dataset, respectively.

Figure 4 presents failure cases from real dToF data, primarily caused by excessive dToF anomalies, while our model shows some improvement in handling these issues, such as correcting the sculpture's arm in Fig. 4e and Fig. 4c, further refinement is needed. Additionally, the MDE model exhibited semantic errors when processing rotated images, failing to correct the anomaly in Fig. 4a. This issue can be resolved by converting the images to a normal perspective.

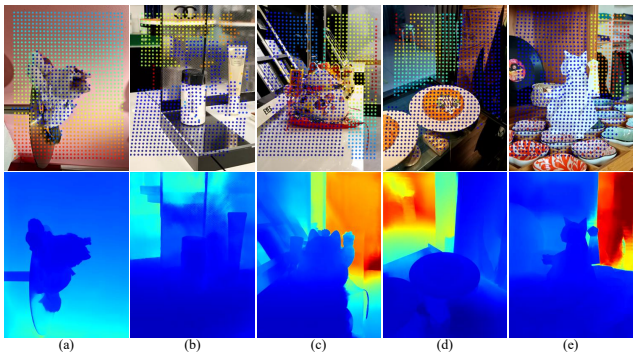

Figure 4. Failure example of real dToF data.

References

[1] E Charbon. Single-photon imaging in complementary metal oxide semiconductor processes. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2012):20130100, 2014.

[2] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Penet: Towards precise and efficient image guided depth completion. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13656–13662. IEEE, 2021. 2, 3

[3] Javier López-Randulfe, César Veiga, Juan J Rodríguez-Andina, and José Farina. A quantitative method for selecting denoising filters, based on a new edge-sensitive metric. In *2017 IEEE International Conference on Industrial Technology (ICIT)*, pages 974–979. IEEE, 2017. 1

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[5] Desmond O'Connor. *Time-correlated single photon counting*. Academic press, 2012.

[6] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[7] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In

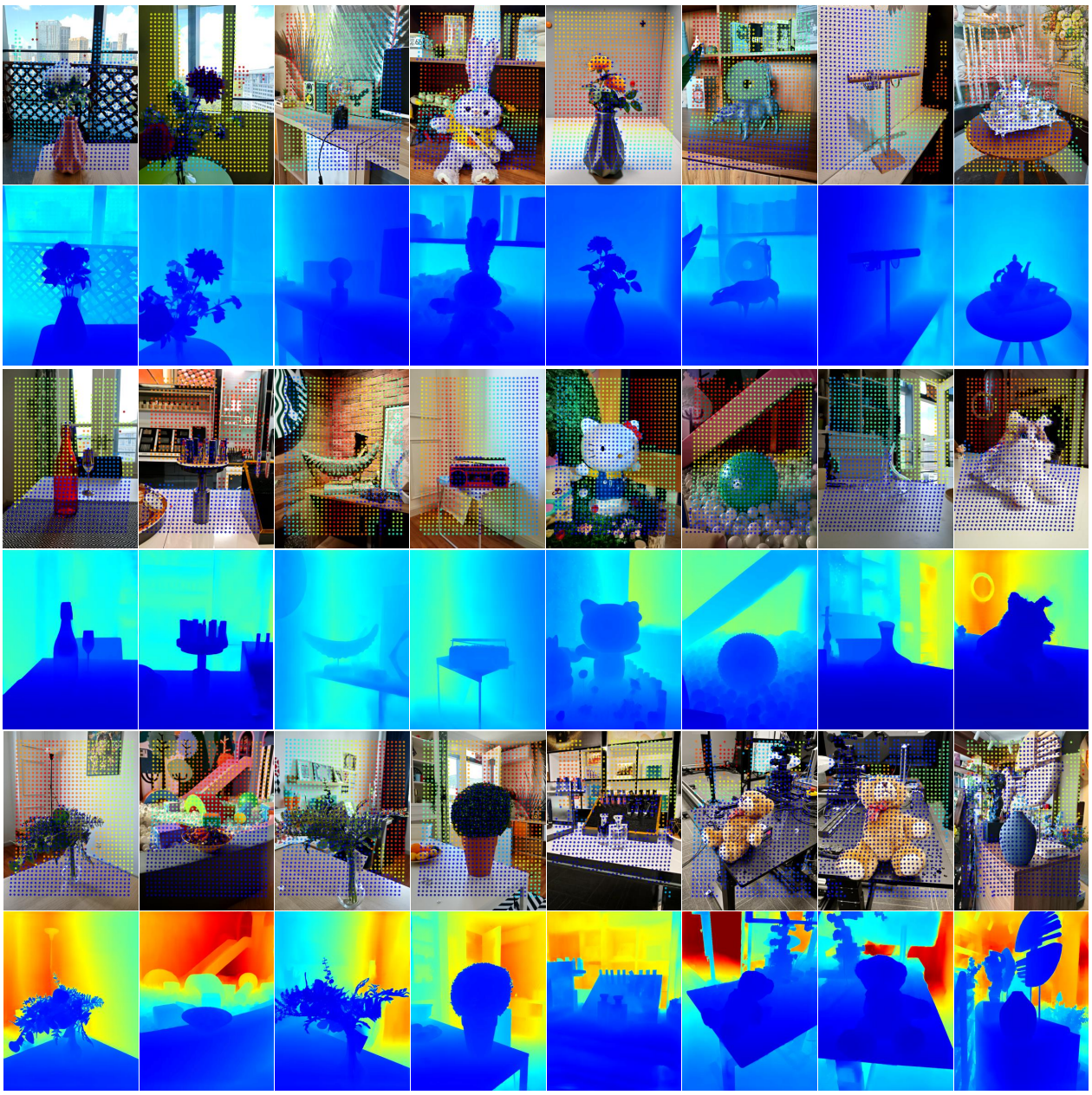

Figure 5. Additional qualitative results on real-world dToF samples.

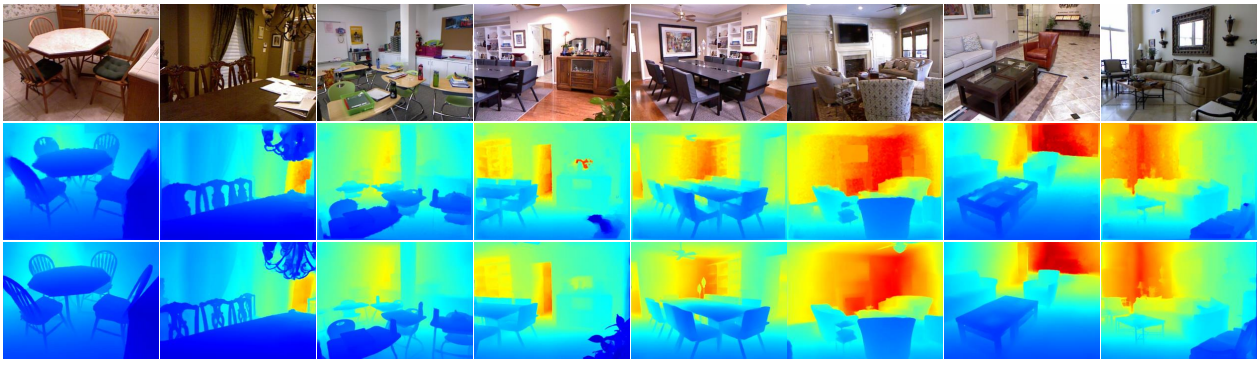

Figure 6. Additional qualitative results on NYUv2 dataset. From top to bottom: RGB, GT, Our results

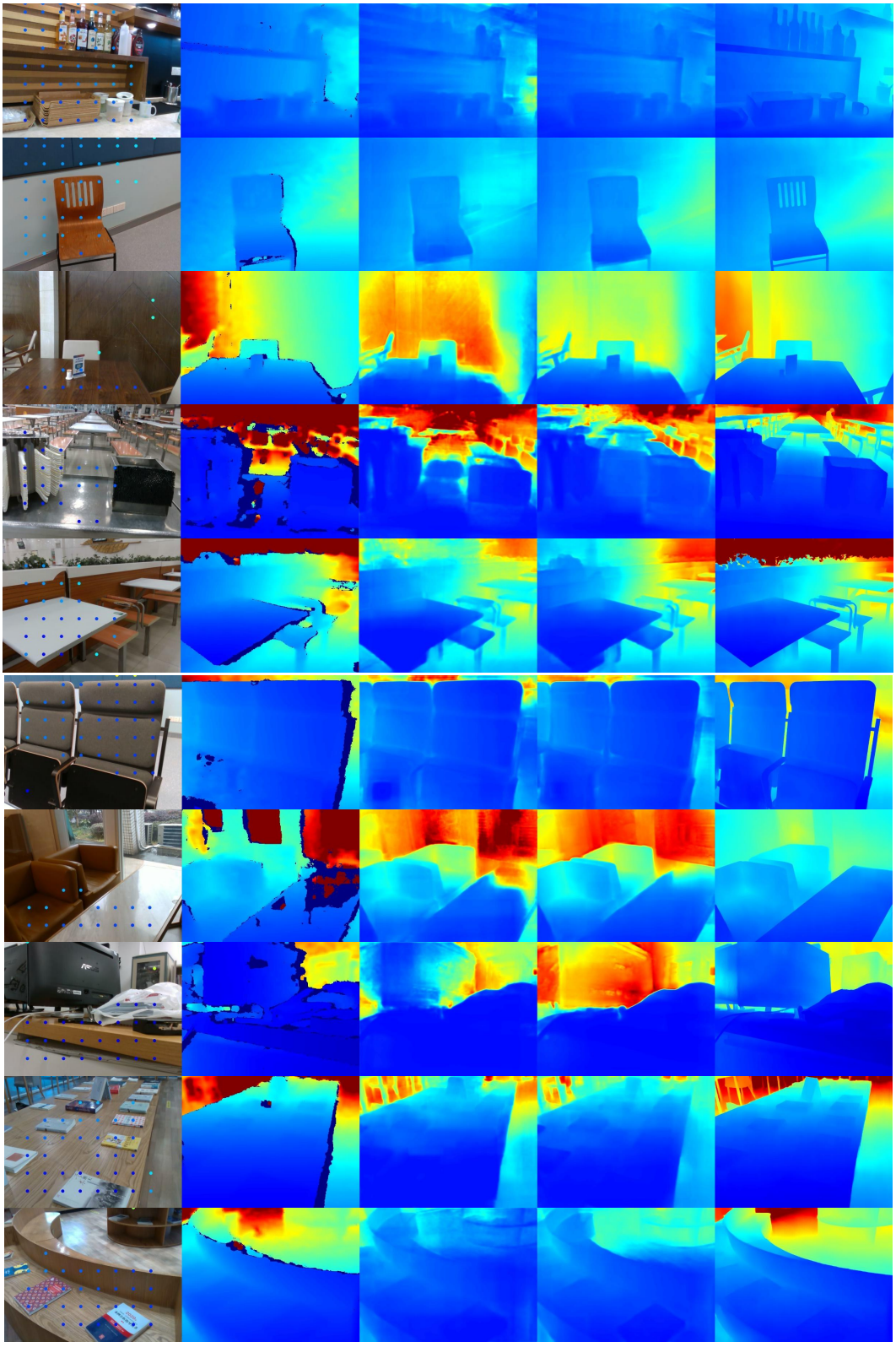

Figure 7. Additional qualitative results on ZJU-L5. From left to right, RGB-dToF, GT, Deltar, CFPNet, Our results.

Proceedings of the IEEE/CVF international conference on computer vision, pages 10912–10922, 2021. 2

[8] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019.

[9] Qianhui Sun, Qingyu Yang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, Yuekun Dai, Wenxiu Sun, Qingpeng Zhu, Chen Change Loy, Jinwei Gu, et al. Mipi 2023 challenge on rgbw remosaic: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2878–2885, 2023. 1

[10] Wenxiu Sun, Qingpeng Zhu, Chongyi Li, Ruicheng Feng, Shangchen Zhou, Jun Jiang, Qingyu Yang, Chen Change Loy, Jinwei Gu, Dewang Hou, et al. Mipi 2022 challenge on rgb+ tof depth completion: Dataset and report. In *European Conference on Computer Vision*, pages 3–20. Springer, 2022. 1

[11] Zhanghao Sun, Wei Ye, Jinhui Xiong, Gyeongmin Choe, Jialiang Wang, Shuochen Su, and Rakesh Ranjan. Consistent direct time-of-flight video depth super-resolution. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 5075–5085, 2023. 1, 2

[12] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020.