LiteSense: Lifting Lightweight ToF with RGB for High-Resolution Metric Depth Estimation

Supplementary Material

6. More Details on CNH

Principle of Measurement CNH is a newly introduced form of raw measurement data provided by STM's latest multi-zone ToF sensors (*e.g.*, VL53L8CH). The sensing pipeline begins with active illumination using a 940 nm VCSEL[1]. The reflected infrared signal is captured by an array of SPADs[2], which record the arrival-time distribution of returned photons with picosecond-level temporal resolution, forming a complete histogram over multiple time bins.

A low-power on-chip MCU then compacts and normalizes the histogram in order to reduce data volume, suppress ambient-light interference, and maintain stable performance across frames. The resulting CNH output follows a highly structured format that preserves rich depth distribution and multi-path information, providing significantly more informative measurements than conventional single-value depth outputs while still supporting real-time operation.

Configuration of CNH The CNH data consist of 144 bins covering a distance range of 0 to 5.4 m, with each bin corresponding to 37.5 mm. However, due to bandwidth constraints, practical systems typically need to merge spatial regions, merge bins, or select specific bins to meet real-time processing requirements. In our work, we preserve the full 8×8 spatial layout to maximize spatial coverage. To retain long-range information as effectively as possible, the 144 bins are merged into 18 bins as illustrated in Fig. 8, resulting in a 300 mm interval per bin. This configuration enables a stable operating frequency of up to 15 Hz.

Before using the CNH data, several preprocessing steps are applied. Strong ambient illumination may cause negative values after ambient subtraction, which are physically meaningless for our method, so all negative values are clipped to zero. Additionally, the CNH histogram within each spatial region is normalized by its total sum.

7. More Ablation Studies

Baseline Selection and Enhancement We initially select LiteMono [56] as our baseline, which is a lightweight network with strong performance in relative MDE. However, when directly adapting it for metric depth estimation, the model generates depth maps with insufficient spatial resolution. To overcome this limitation, we design a new lightweight network. Compared with LiteMono, our model employs a more compact image encoder ($1.26M$ *vs.* $2.84M$), while moderately increasing the decoder capacity ($3.54M$ *vs.* $0.23M$) to better support high-resolution detail reconstruction.

[1] Vertical Cavity Surface Emitting Laser
[2] Single-Photon Avalanche Diodes

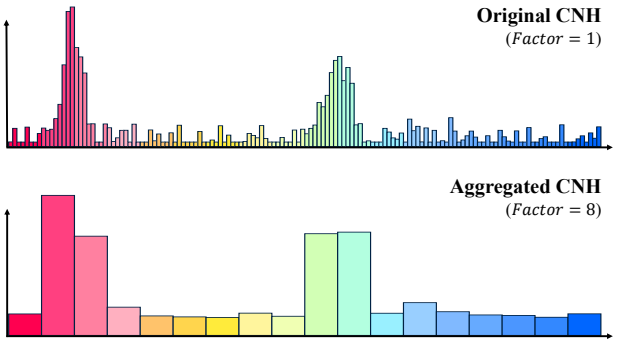

Figure 8. **Illustration of CNH aggregation in one zone.** As indicated by the different colors, every eight original bins are merged into one new bin, preserving the original covered distance range.

Table 5. **Results of applying the PCSI module at different spatial scales.** The last row corresponds to the configuration used in our full model, which performs the best.

Fusion Scale				$\delta_1 \uparrow$	RMSE(m) $\downarrow$	AbsRel $\downarrow$
1/2	1/4	1/8	1/16			
		✓	✓	0.981	0.216	0.033
✓	✓	✓	✓	0.977	0.238	0.039
✓	✓			**0.982**	**0.197**	**0.029**

Effect of Fusion at Different Scales As previously discussed, injecting CNH information at an appropriate spatial scale is crucial for recovering fine-grained depth distribution details. To further validate this observation, we design three experiments that fuse CNH with RGB-D features at: higher-resolution layers, lower-resolution layers, and all scales. For fusion at lower-resolution layers, the RGB-D feature map cannot be directly divided into 8×8 regions. Therefore, we interpolate it to the nearest multiple of eight, apply the PCSI module for fusion, and then resize it back to the original resolution.

The quantitative results in Tab. 5 show that fusion only at higher-resolution scales yields the best performance, which is also the strategy adopted in our full model. This confirms that performing CNH-guided fusion at scales containing more spatial pixels enables the CNH cues to be more

effectively utilized. Consequently, the model can maximize high-resolution detail reconstruction while keeping computational overhead low.

8. More Implementation Details

Image Processing Due to the NYUv2 [40] official preprocessing procedure, the depth-completed images have a spatial resolution of 561×427. To ensure that the input size remains divisible by 32 for downsampling, we randomly crop 416×416 patches during training. For inference and fine-tuning, we use a fixed input resolution of 480×480.

Network Architecture Our full network architecture is detailed in Tab. 6. For RGB-D feature extraction, we adopt and modify the Universal Inverted Bottleneck (UIB) from the Small variant of MobileNetV4 [33]. And CNH features are extracted using an MLP implemented with 1D convolutions. The decoder follows a U-Net-style design, using transposed convolutions for upsampling, and a final Softplus layer to produce the metric depth output.

THDR3K Dataset Overview We use an Intel RealSense D435i and an STM VL53L8CH ToF sensor to collect paired RGB images, depth maps, and ToF measurements. The RealSense captures images and depth maps at a resolution of 640×480. Based on the approximately 45° vertical FOV shared by the ToF and RGB-D camera, a 480×480 region is selected as the valid overlapping area after calibration. The depth maps are directly produced by the RealSense internal registration without any completion. The ToF measurements include low-resolution depth maps and CNH data, along with additional metadata such as measurement status, target count, and ambient light intensity.

The sample composition of the constructed dataset is summarized in Tab. 7. To evaluate the generalization capability of our approach, we reserve 210 samples from 4 rooms for standalone testing, while the remaining 2528 samples are split into training and validation sets at a $3:1$ ratio for fine-tuning.

Fine-tuning on THDR3K Before transferring the model to real sensor data, we perform fine-tuning because the simulated CNH differs considerably from real measurements, and this discrepancy inevitably affects performance. To better demonstrate the model's generalization capability, we freeze the RGB-D feature extraction module and only fine-tune the components that rely on CNH data.

Definition of Evaluation Metrics We evaluate the prediction performance using three common metrics. Let $\hat{d}_i$ and d_i denote the ground-truth and predicted depth values at pixel i, and let M be the total number of valid pixels.

Table 6. **Network architecture details.** We list the primary operators of each layer and illustrates the tensor dimensions at each stage using the training configuration as an example.

Layer	Input	Operator	Output	Output Dim.
Preprocess	RGB, ToF-D	UpSample+Concat	#1	$416 \times 416 \times 4$
	ToF-CNH	-	#2	64×18
Spatial Blocks				
Block-1	#1	Conv2D+BN	#3	$208 \times 208 \times 32$
Block-2	#3	Conv2D+BN	#4	$104 \times 104 \times 64$
Block-3	#4	UIB	#5	$52 \times 52 \times 96$
Block-4	#5	UIB	#6	$26 \times 26 \times 960$
Histogram Blocks				
MLP-1	#2	Conv1D	#7	64×128
MLP-2	#7	Conv1D	#8	64×256
Fusion Module				
PCSI-1	#3, #7	CrossAttn+Reassemble	#9	$208 \times 208 \times 128$
PCSI-2	#4, #8	CrossAttn+Reassemble	#10	$104 \times 104 \times 256$
Depth Decoder				
UpSample-1	#6	Conv2D+BN+UpConv+BN	#11	$52 \times 52 \times 256$
UpSample-2	#5, #11	Conv2D+BN+UpConv+BN	#12	$104 \times 104 \times 128$
UpSample-3	#10, #12	Conv2D+BN+UpConv+BN	#13	$208 \times 208 \times 64$
UpSample-4	#9, #13	Conv2D+BN+UpConv+BN	#14	$416 \times 416 \times 32$
Head	#14	Conv2D+Softplus	Depth	$416 \times 416 \times 1$

– Threshold accuracy (δ_k with $k = 1, 2, 3$):

$$\delta_k = \sum_{i=1}^{M} \left(\max\left(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i} \right) < 1.25^k \right). \tag{9}$$

– Absolute Relative Error (AbsRel):

$$\text{AbsRel} = \frac{1}{M} \sum_{i=1}^{M} \left| \frac{d_i - \hat{d}_i}{d_i} \right|. \tag{10}$$

– Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{M} \sum_{i=1}^{M} (d_i - \hat{d}_i)^2}. \tag{11}$$

In addition, we report the number of parameters using PyTorch-OpCounter (THOP), and compute the FLOPs under an input resolution of 416×416.

Reproducibility Experiment We conduct 10 runs with different random seeds to evaluate the reproducibility of our method. The results show that δ_1 remains stable at approximately 0.980 ($\pm 0.3\%$), RMSE at 0.211 ($\pm 7\%$), and AbsRel at 0.032 ($\pm 9\%$). The variations across runs are small, indicating strong reproducibility and stable performance.

9. More Performance Analysis

Degradation Beyond the ToF Coverage Area. Due to the input constraints of the model, only data within the aligned regions can be utilized. To evaluate performance

Table 7. **Summary of the THDR3K dataset.** The majority of the captured scenes originate from indoor locations on campus.

Subset	Scene	Room	Samples
Train & Validation	Book Store	#01	115
	Canteen	#02	150
	Classroom	#03-1	100
		#03-2	215
		#03-3	70
	Corridor	#04-1	40
		#04-2	100
	Kitchen	#05	50
	Library	#06-1	210
	Lobby	#07-1	50
		#07-2	80
		#07-3	40
	Meeting Room	#08-1	120
		#08-2	160
	Rest Space	#09-1	20
		#09-2	110
		#09-5	25
	Office	#10	60
	Sports Room	#11	100
	Stairs	#12-1	50
		#12-2	35
	Store	#13	100
	Study Room	#14-1	140
		#14-2	65
	Washroom	#15	80
	Others	#16	135
	Indoor Objects	#17	108
Test	Library	#06-2	25
	Rest Space	#09-3	50
		#09-4	70
	Study Room	#14-3	65

in RGB-only regions, we masked the marginal 1-2 columns of the ToF inputs. Fig. 9 demonstrates a clear performance degradation outside the ToF-covered areas. While RGB information alone can recover coarse structural outlines, the depth accuracy is significantly reduced. Moreover, the degradation becomes more severe as the distance from the ToF-covered regions increases. This performance degradation represents the primary limitation of the current model.

Long-Range and Outdoor Potential. The current sensor offers an effective range of 5 m, covering most indoor scenarios. To assess performance beyond this limit, we evaluated a simulated 10 m ToF inputs. Tab. 8 indicates that even outside the sensing range, the model effectively captures relative scale cues via RGB features.

Regarding outdoor applicability, most publicly available outdoor datasets are designed for autonomous driving scenarios, where depth distributions are significantly broader and extend far beyond the operational range of our target sensors and model design. Moreover, current ToF sensors are generally unable to capture reliable measurements in outdoor environments due to strong ambient interference. To better evaluate the model under outdoor conditions, we restrict our analysis to mid-range outdoor scenes by selecting samples with depths clipped to 25 m. Under this setting, the model achieves a δ_1 of 0.962, which is comparable to its indoor performance, thereby demonstrating its scalability and effectiveness in moderate-range outdoor scenarios.

Figure 9. **Qualitative results on RGB-only area.** Performance degrades progressively as the distance from the ToF-covered area increases. Regions farther from ToF support exhibit more pronounced errors, indicating a strong reliance on nearby depth cues.

Table 8. **Quantitative results across intervals.** The model achieves consistent performance across different intervals within the ToF range, while only exhibiting a slight increase in error beyond the sensing range. Overall, it maintains reliable accuracy. The metric is reported in AbsRel(↓) on NYUv2.

	0-3 m	3-5 m	5-8 m	8-10 m	Overall (0-10 m)
ToF Range @ 5 m	0.027	0.031	0.083	0.198	0.029
ToF Range @ 10 m	0.029	0.028	0.048	0.075	0.025

Adaption Across Different Configurations. Our model is adaptable to various ToF configurations. Tab. 9 presents the performance under several configurations supported by current hardware, as well as extended settings without practical constraints. With a fixed sensing range, increasing the number of zones and CNH bins leads to improved model performance. This trend is further supported by the announced roadmap of next-generation STM's devices, which are expected to offer extended sensing ranges and a larger number of zones. Our model's adaptability makes it well-suited for seamless deployment, enabling straightforward transfer and application without architectural modifications.

10. More Discussion

As the only existing work utilizing lightweight ToF inputs, DELTAR serves as an important baseline for comprehensive comparison. Beyond the performance gains brought

Table 9. **Examples of different ToF-input configurations.** The model can adapt to different input configurations and demonstrates strong performance. The metrics are evaluated on NYUv2.

	RGB Cover	Zone	Range	CNH (Bins)	$\delta_1 \uparrow$	RMSE $\downarrow$ (m)
Practical Config	416×416	4×4	5.0 m	72	0.962	0.295
	416×416	8×8	5.0 m	18	0.982	0.197
Theoretical Config	416×416	8×8	10.0 m	18	0.990	0.146
	544×416	17×13	5.0 m	18	0.988	0.183

Table 10. **Additional comparison with DELTAR.** The same input from NYUv2 is used for both methods to ensure fairness.

Method	Prior Type	$\delta_1 \uparrow$	RMSE $\downarrow$ (m)
DELTAR	ToF-D Samp.	0.952	0.311
LiteSense	ToF-D Samp.	0.888	0.453
DELTAR	CNH	0.866	0.553
LiteSense	CNH	0.926	0.359
LiteSense	ToF-D + CNH	**0.982**	**0.197**

by different input representations, we further investigate the impact of model design. Although DELTAR is compatible with CNH inputs, Tab. 10 shows that it suffers from a significant performance drop, indicating that its feature extraction and fusion modules are not well-suited for CNH data.

11. More Qualitative Results

We present more quantitative results across multiple datasets and compare with several strong methods, including DA-V2 [53], ZoeDepth [5], M3D-V2 [16], and DELTAR [22]. Fig. 11 shows the results on NYUv2 [40], Fig. 13 reports the zero-shot performance on SUN RGB-D [42], and Fig. 12 illustrates the results on the real sensor dataset, THDR3K. In addition, Fig. 10 provides the quantitative results of our ablation studies. In summary, our method delivers accurate and stable metric scale estimation.

12. Limitations and Future Work

Currently, our fusion design is simple and efficient, but it is mainly effective within the spatial region covered by the ToF sensor and cannot fully generalize to areas that rely solely on RGB information. In future work, we plan to strengthen the network's ability to comprehensively reconstruct RGB features. By integrating vision foundation models and knowledge distillation strategies, we aim to extend accurate depth prediction to the entire RGB image without increasing model complexity, thereby producing clearer object boundaries. At the same time, the depth scale can remain consistent under ToF constraints, making the approach suitable for edge applications such as AR and SLAM. Additionally, We aim to enable the framework to directly extend to outdoor environments as more powerful lightweight ToF sensors become available in the future.

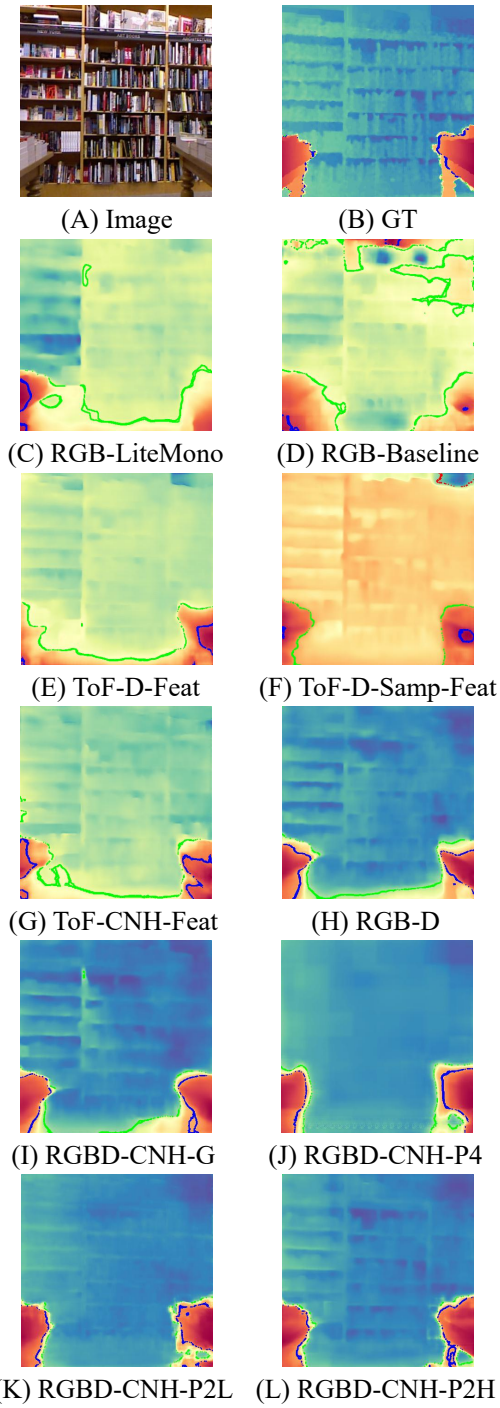

Figure 10. **Qualitative example of ablation study results.** In particular, for the fusion-related results, subfigure [*I*] corresponds to the experiment without using the PCSI module, [*J*] applies PCSI at all scales, [*K*] applies PCSI only at low-resolution scales, and [*L*] applies PCSI at high-resolution scales. Subfigure [*L*] also represents our final full model, which achieves the best prediction quality among all variants. Contour lines are overlaid to indicate equal depth levels, where **black**, **blue**, **green**, and **red** correspond to 1.0 m, 2.0 m, 3.0 m, and 5.0 m, respectively.

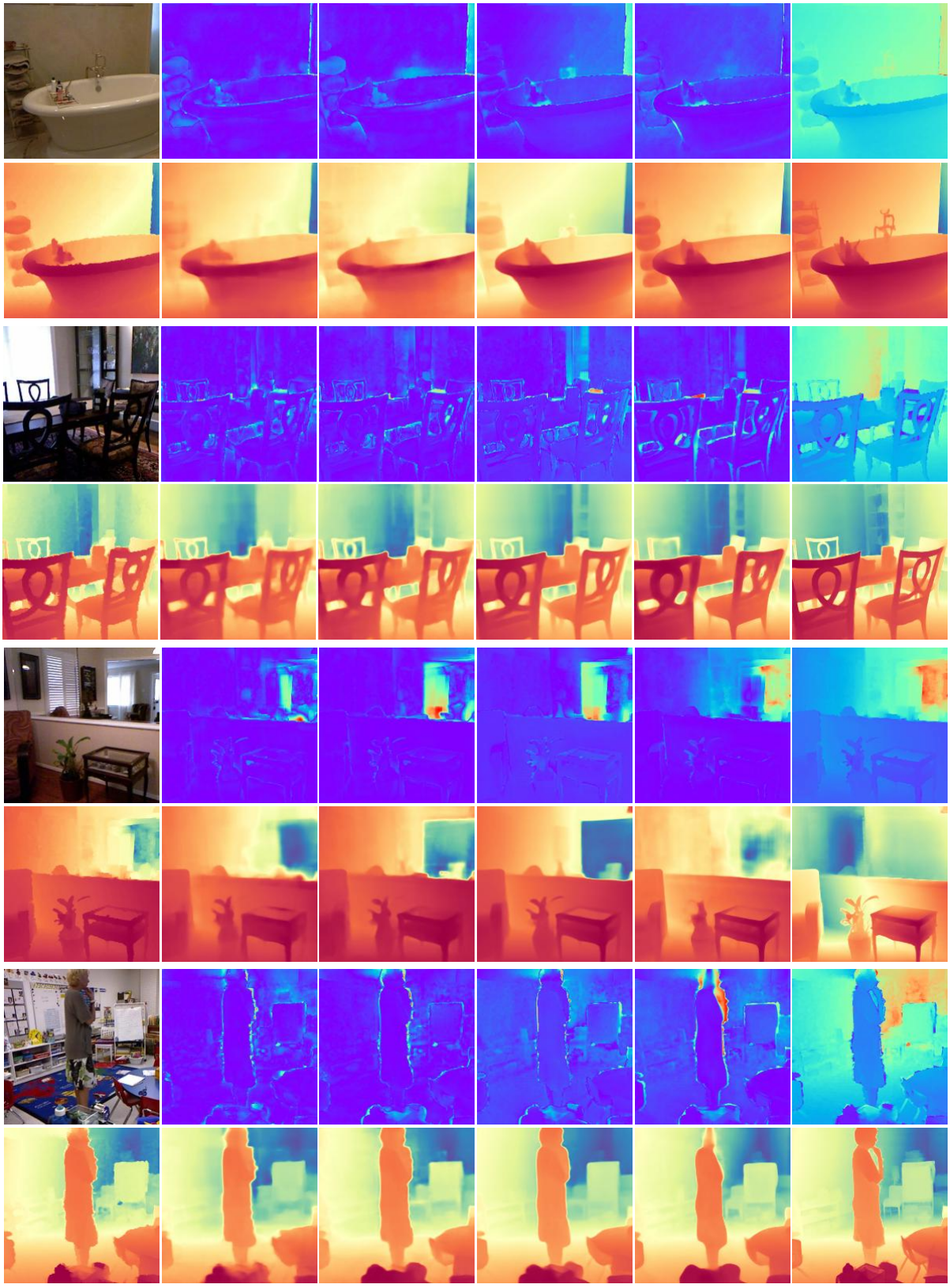

Figure 11. **More quantitative results on NYU Depth V2.** The first row in each group shows the error maps.

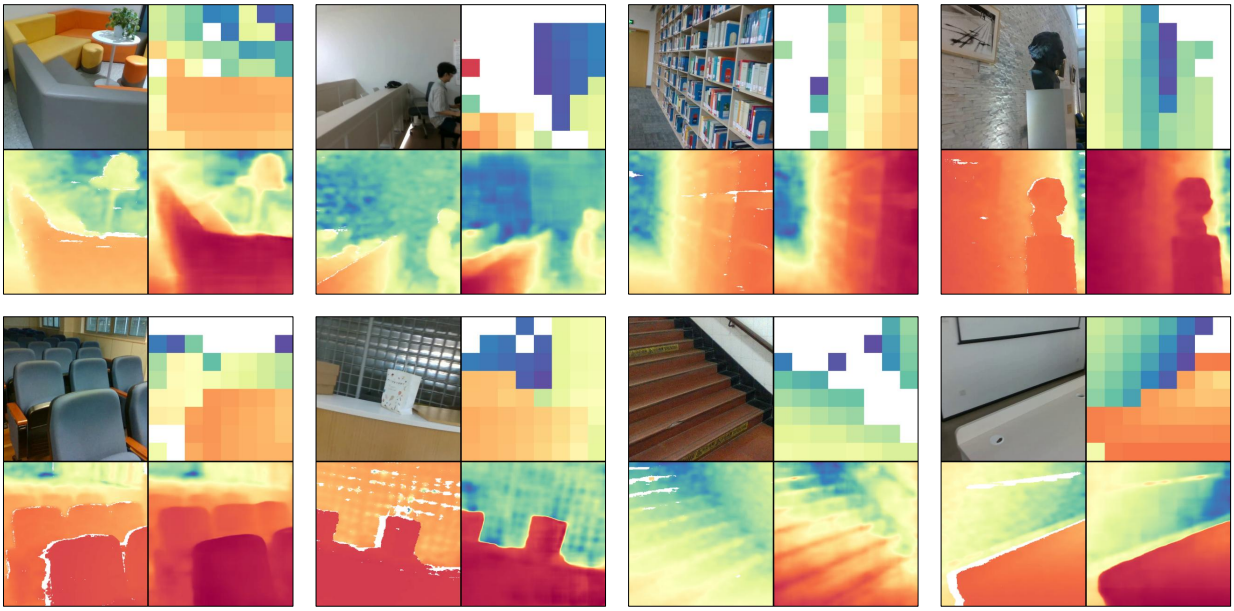

Figure 12. **More quantitative results on THDR3K.** For each example, the bottom-left one shows the ground-truth, and the bottom-right one presents the result produced by our method. As shown, our predictions closely match the depth quality obtained from the RealSense.

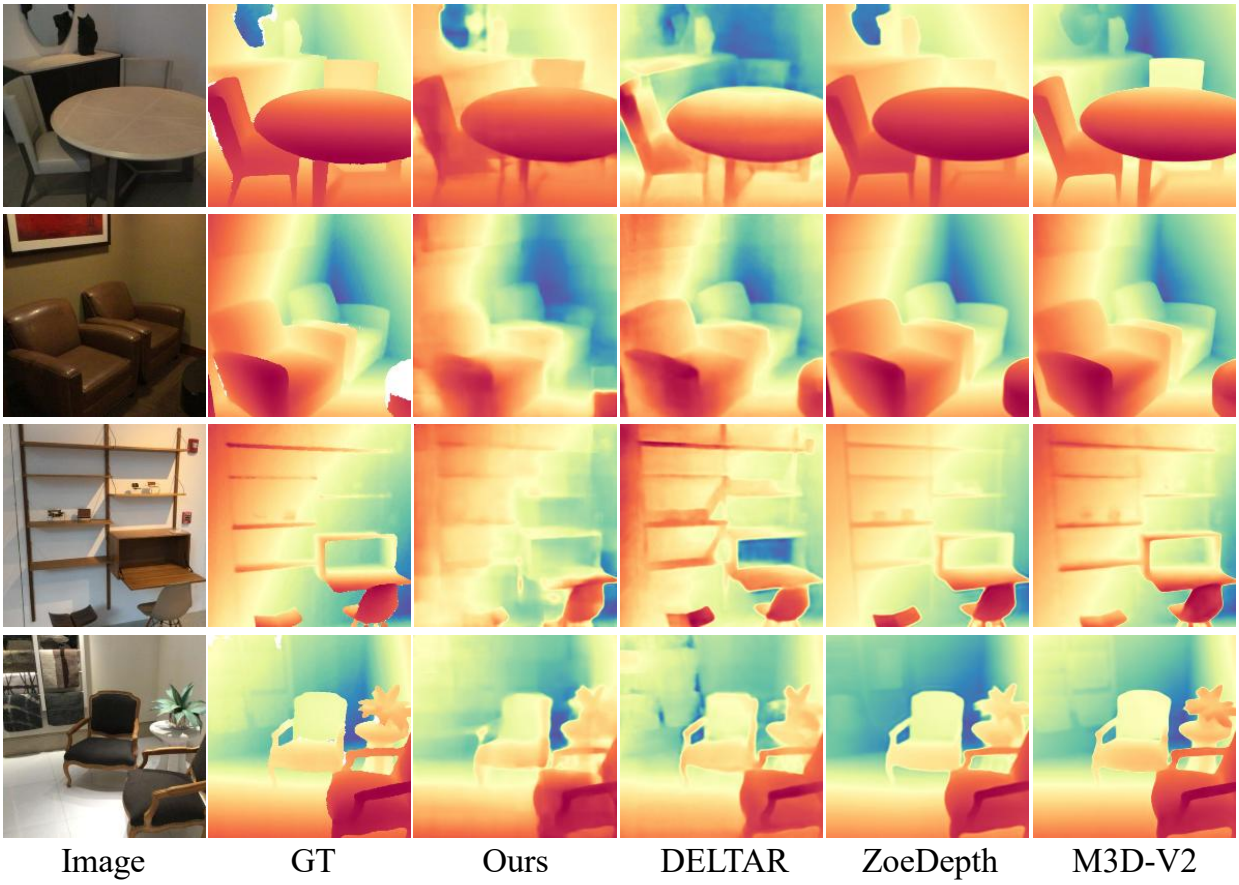

Figure 13. **More quantitative results on SUN RGB-D.** This zero-shot result demonstrates the generalization capability of our method. The predicted depth preserve accurate metric scale and recover fine structural details, achieving performance comparable to larger models.