Supplementary Materials for Seeing and Seeing Through the Glass: Real and Synthetic Data for Multi-Layer Depth Estimation

Anonymous ICCV submission

Paper ID 2616

1. Additional Details

1.1. Benchmark

All images in our benchmark are released under the CC0 license. We will make the dataset publicly available; however, a subset of the ground truth annotations for the real-world images will be withheld for use on a public evaluation server. To validate our approach, we manually annotated 30 synthetic images with known depth ground truth. Our annotations matched the ground truth in 98% of cases, demonstrating the reliability. In total, we annotated 1,500 images, with 300 allocated for validation and 1,200 for testing. Our annotations include 5,406 monotonic depth lines and 38,392 relative depth points across 7 distinct layers, plus 3,011 fake depth lines and 6,025 fake relative depth points.

1.2. Baseline Training

We train all NeWCRFs baseline models from scratch on our synthetic dataset for 100 epochs. For the Multi-head strategy, we set the number of output depth channels from 1 to 4 to predict multi-layer depth. For the Index Concat strategy, we concatenate the image with an index channel of size $H \times W$, where all values are set to the corresponding layer index, and set the number of input channels to 4 (RGB plus Index). For the Recurrent strategy, we concatenate the image with the previous depth output and also set the number of input channels to 4 (RGB plus Depth). During each training step, a random layer is selected as the prediction target. To provide richer supervision, we utilize snapped layered depth: if a pixel lacks ground-truth depth at layer i, it inherits the depth value from layer $i-1$. For optimization, we use the Scale-Invariant Logarithmic loss [1].

1.3. Evaluation and Fine-tuning

All methods are evaluated using a single NVIDIA RTX 3090 GPU. When assessed on the synthetic validation dataset, both the ground-truth values and predictions are clipped to the range $(0.001, 30)$. Fine-tuning for Metric3D V2 [2] is performed using the publicly available code, with training for 100,000 steps.

2. Additional Results

2.1. Fine-tuning DepthAnything V2

The proposed multi-layer depth baseline method can be applied to any existing depth backbone. To demonstrate how recent models can benefit from our dataset, we further fine-tuned DepthAnything V2 [3] for multi-layer depth estimation using the multi-head strategy. To achieve this, we repli-

Method	All			Mixed			Layer 1			Layer 3			Layer 5			Layer 7		
	P	T	Q	P	T	Q	P	T	Q	P	T	Q	P	T	Q	P	T	Q
Multi-head (NeWCRFs)	63.42	42.55	25.97	74.72	46.13	26.38	65.66	39.93	24.21	56.58	43.36	30.77	52.17	38.66	29.94	38.40	35.65	33.10
Index Concat (NeWCRFs)	64.46	44.00	26.00	76.70	48.37	26.46	66.95	41.84	24.45	55.85	41.53	30.06	55.36	42.69	31.57	39.66	45.18	45.08
Recurrent (NeWCRFs)	62.36	41.88	24.64	73.51	45.27	25.35	68.08	44.46	26.12	49.47	31.53	21.25	45.51	30.06	21.29	30.09	34.10	27.59
Multi-head (DA v2)	**80.72**	**70.67**	**61.68**	**89.37**	**72.98**	**62.23**	**85.80**	**74.34**	**65.62**	**73.75**	**62.68**	**59.19**	**66.19**	**59.29**	**54.75**	**57.39**	**50.78**	**48.09**

Table 1. Baseline methods evaluated on our real-world benchmark via tuple-wise accuracy. Best scores are in **bold**. Second best underlined.

Method	Layer 1				Layer 3				Layer 5				Layer 7			
	AbsRel↓	RMS↓	$\delta 1$↑	$\delta 2$↑	AbsRel↓	RMS↓	$\delta 1$↑	$\delta 2$↑	AbsRel↓	RMS↓	$\delta 1$↑	$\delta 2$↑	AbsRel↓	RMS↓	$\delta 1$↑	$\delta 2$↑
Multi-head (NeWCRFs)	17.97	23.41	83.08	93.14	15.90	47.39	80.15	93.89	14.58	47.51	81.48	94.12	16.27	58.78	80.44	93.92
Index Concat (NeWCRFs)	17.26	23.70	83.02	93.27	16.25	48.19	79.63	93.72	15.03	48.11	80.61	93.97	16.49	59.03	80.26	93.76
Recurrent (NeWCRFs)	17.23	25.53	81.89	93.37	17.61	51.79	76.39	92.53	17.25	52.53	76.54	92.20	18.35	61.98	77.04	92.40
Multi-head (DA v2)	**8.24**	**15.21**	**93.43**	**97.73**	**10.68**	**37.97**	**88.76**	**96.83**	**11.15**	**40.67**	**87.51**	**96.40**	**13.84**	**54.63**	**84.79**	**95.13**

Table 2. Multi-layer baseline methods evaluated on our synthetic validation set. Values are scaled by 100 for clearer comparison. Best scores are in **bold**. Second best underlined.

cate the depth head four times, enabling the model to predict multi-layer depth.

Results evaluated on the real world benchmark are shown in Tab. 1. Results evaluated on the synthetic validation set are shown in Tab. 2. Qualitative comparison on real benchmark can be found in Fig. 1 and Fig. 2. Qualitative comparison on synthetic validation set can be found in Fig. 3. These results show that DepthAnything V2 not only aligns more closely with the training domain but also generalizes exceptionally well after fine-tuning solely on our synthetic data. This underscores the power of our synthetic data generator for multi-layer depth estimation and transparent-object understanding. Nevertheless, some regions around object boundaries and in deeper layers still exhibit artifacts, leaving rooms for improved model designs in future work.

2.2. Additional Qualitative Results

To help readers better visually assess the effectiveness of our datasets, we provide additional qualitative results on real world benchmark and synthetic validation set, which can be found in Fig. 4 and Fig. 5, respectively.

References

[1] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 1

[2] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[3] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2025. 1

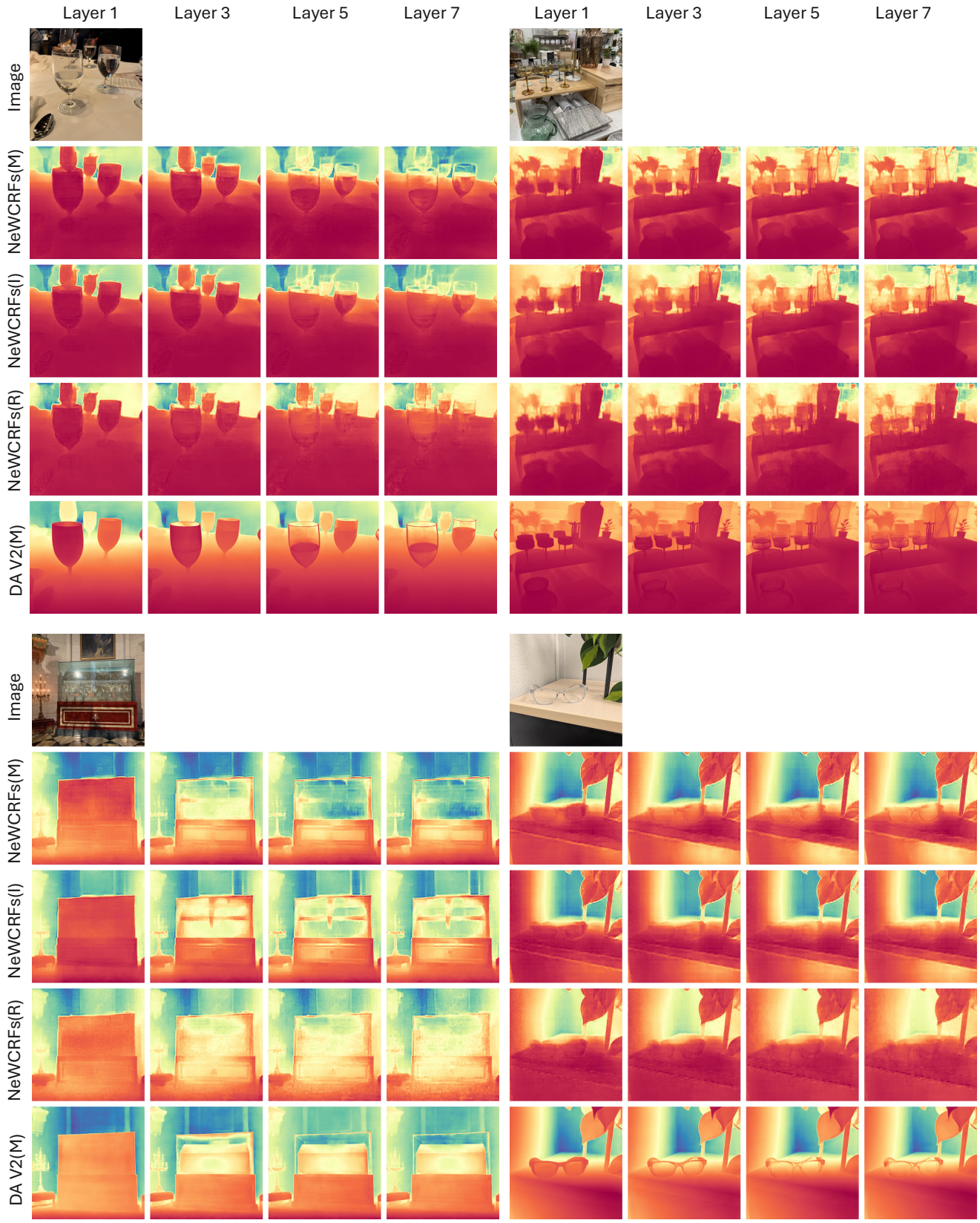

Figure 1. Additional qualitative comparison of multi-layer depth baselines on real benchmark.

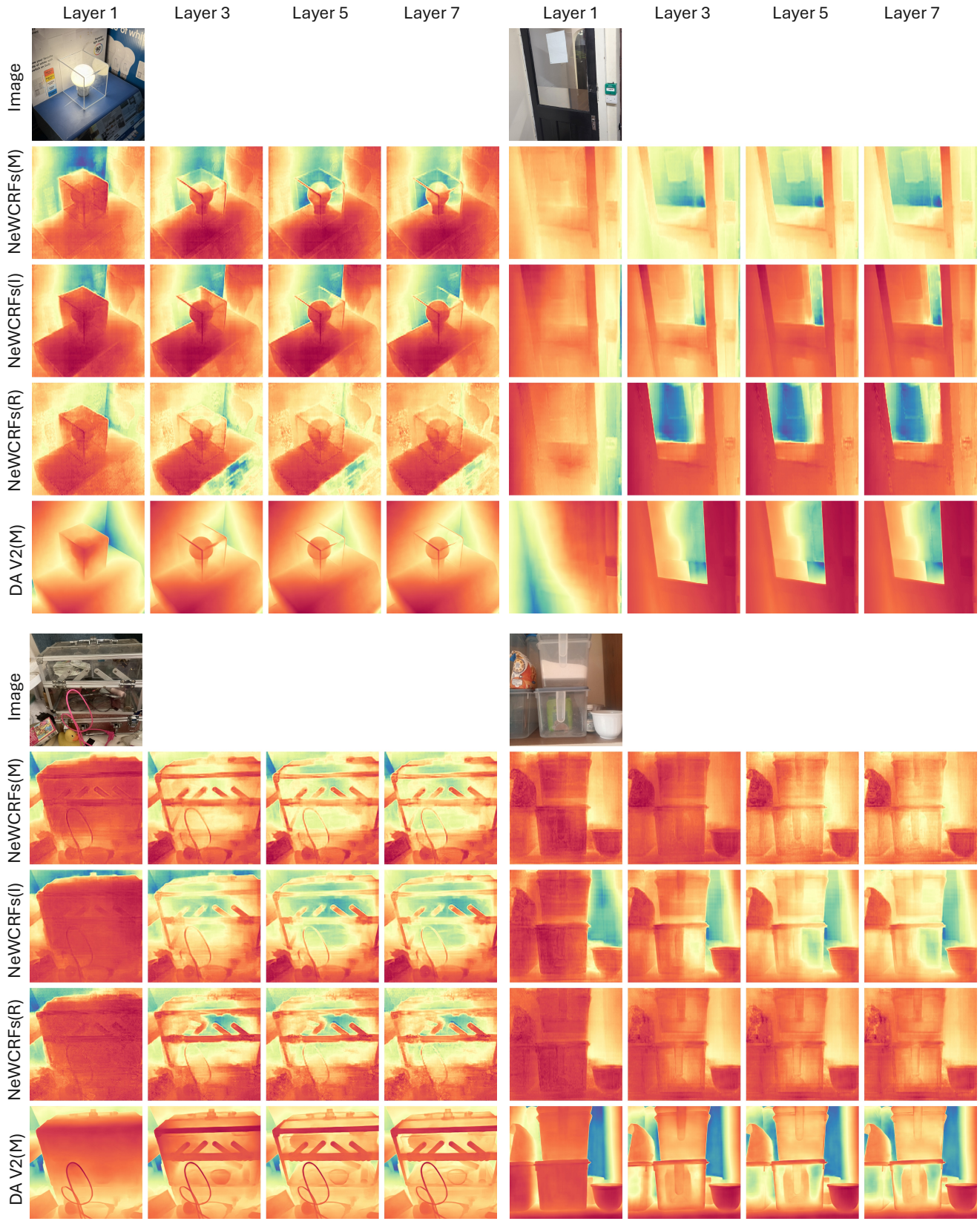

Figure 2. Additional qualitative comparison of multi-layer depth baselines on real benchmark.

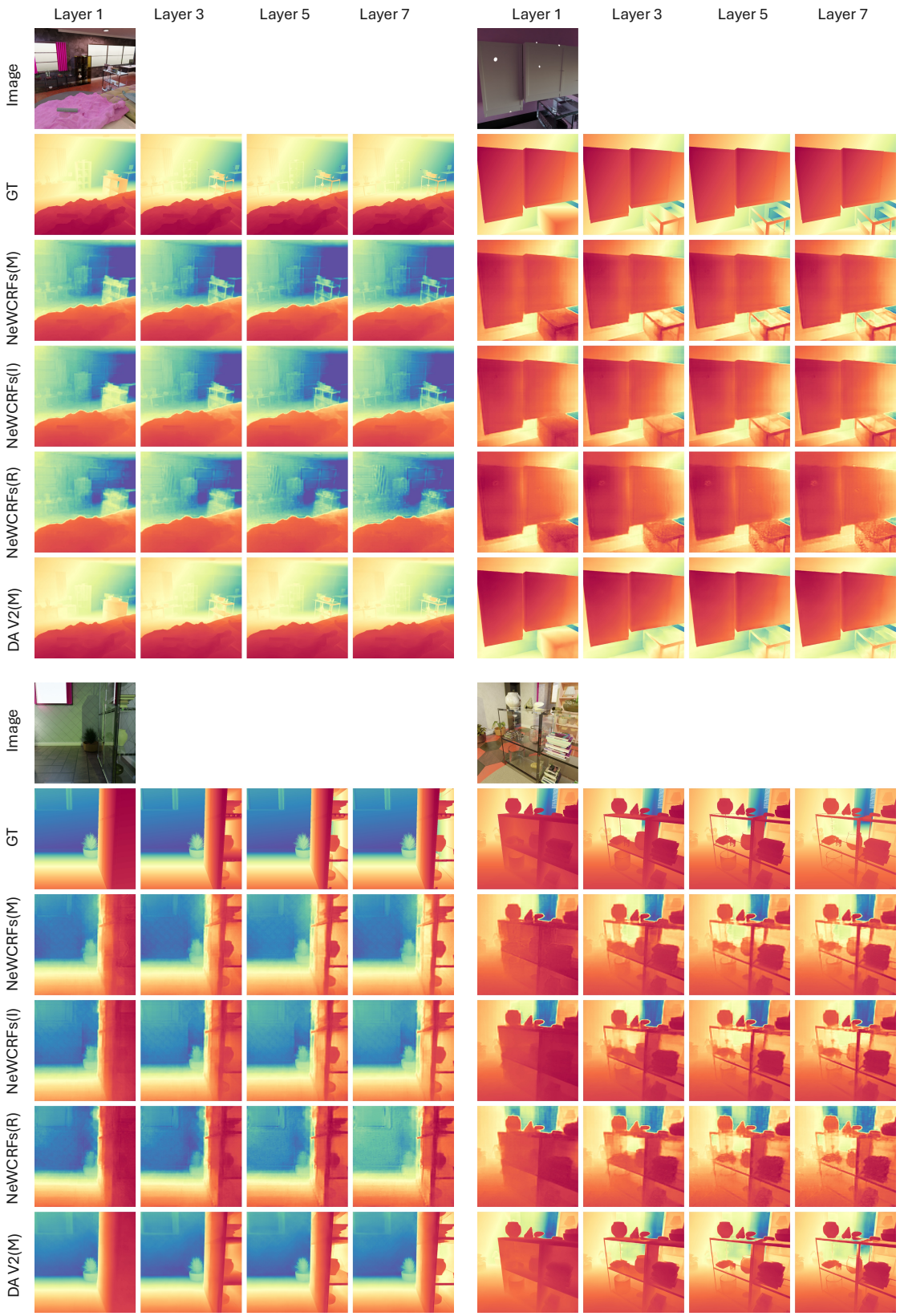

Figure 3. Additional qualitative comparison of multi-layer depth baselines on synthetic validation set.

Figure 4. Additional qualitative comparison of state-of-the-art single layer depth methods on our benchmark.

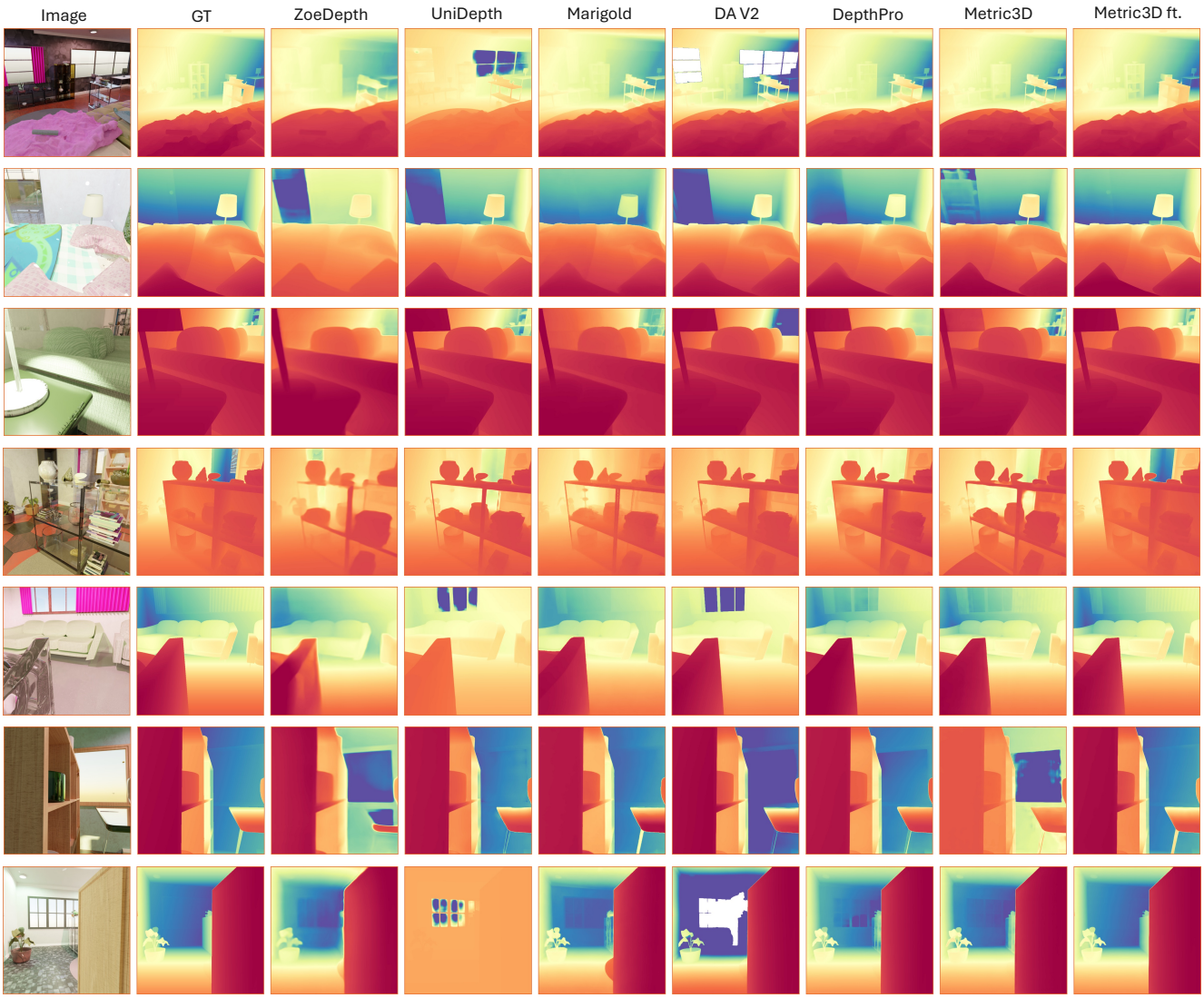

Figure 5. Additional qualitative comparison of state-of-the-art single layer depth methods on our synthetic validation set.