Supplementary file for "MD2E: Modeling Depth-to-Edge Cues for Monocular Metric Depth Estimation"

Chao Ning[1,2] Minghe Shen[3] Naoto Yokoya[1,2*]
[1]The University of Tokyo [2]RIKEN [3]University College London
{6575088851,yokoya}@edu.k.u-tokyo.ac.jp,
minghe.shen.24@ucl.ac.uk

1. Performanceof depth-to-edge modeling

In this section, we investigate the impact of our depth-to-edge algorithm on the sharpness of depth predictions. We first compare depth edges produced by our method with conventional edge maps extracted from RGB images. We then evaluate different forms of edge supervision. Here we only vary the edge supervision term within L_{MSE}.

1.1. Image edge vs. depth edge

We conduct a visualization comparison between image edges and depth edges. For image edges, we apply the Canny operator to the RGB images, while for depth edges we use our proposed depth-to-edge transformation on the depth annotations. For indoor scenes, as illustrated in the left group of Figure 1, depth-related edges are clearly misaligned with the edges detected from the images. Image edges contain a large number of texture-induced responses and are heavily contaminated by illumination-dependent noise, and they often exhibit broken and inconsistent boundaries. For outdoor scenes, as shown in the top group of Figure 2, image edges struggle to capture complex structures such as foliage and are sensitive to lighting and weather conditions. In contrast, depth edges contain far less noise and provide stable, fine-grained edge structures.

1.2. Edge supervision comparison

We compare three supervision settings: no edge supervision, supervision with image edges, and supervision with depth edges, and evaluate their impact on depth prediction quality. The qualitative results are shown in the right group of Figure 1 and the bottom group of Figure 2. Without edge supervision, the predicted depth maps tend to be blurry. When using image edge supervision, the predictions become sharper, but spurious texture edges appear that are inconsistent with the surrounding depth values, introducing noise into the depth maps. In contrast, supervision with depth edges sharpens structural details while avoiding such artifacts, yielding clean and coherent depth predictions.

2. Zero-shot comparison with state-of-the-art methods

2.1. Comparison with Metric3D v2

Metric3D v2 [3] achieves highly accurate depth predictions on DIODE [11], NYU-Depth V2 [10], and KITTI [2]. The qualitative comparisons on these three datasets are presented in Figure 3, Figure 4, and Figure 5, respectively. Although our MD2E is trained on a subset of the original Metric3D v2 training data without introducing any additional datasets, and the model architecture is largely aligned with Metric3D v2, our method produces depth maps with noticeably higher-quality details. For example, MD2E better resolves thin structures such as tree branches in DIODE, small objects such as faucets and shelves and gaps between chair legs in NYU-Depth V2, and fine structures such as trees in KITTI.

2.2. Comparison with UniDepthV2

UniDepthV2 [6] achieves highly accurate depth predictions on iBims-1 [5] and ETH3D [9], and the qualitative comparisons on these two datasets are presented in Figure 6 and Figure 7, respectively. UniDepthV2 enhances local detail by imposing strong supervision on edge-centered image patches, enforcing local consistency between the predicted depth and the ground-truth depth. This strategy leads to relatively sharp depth maps; however, the depth quality around fragile structures is still slightly inferior to that of our MD2E, for example on thin power lines, gaps between leaves, and small gaps at the top of chairs in iBims-1, and thin ropes in ETH3D.

2.3. Comparison with Depth Pro

Depth Pro produces sharp depth predictions on most datasets. Its strategy for obtaining sharp results is to upsample the input image to a fixed high resolution of

*Corresponding author

Method	HAMMER	Booster	FLSea
DepthPro [1]	0.294	0.584	0.106
UniDepthV2 [6]	0.645	**0.676**	**0.905**
Metric3D v2 [3]	0.653	0.091	0.175
MD2E (ours)	**0.872**	0.381	0.733

Table 1. Zero-shot δ_1 ↑ on HAMMER [4], Booster [7], and FLSea [8].

1024×1024 and to optimize the network architecture accordingly. In our experiments on NVIDIA A100 GPUs, this design leads to substantial memory consumption, and fine-tuning Depth Pro under our training setup results in out-of-memory errors, whereas our method can be trained without issue. Moreover, when the original image resolution exceeds 1024×1024, such as the 4032×6048 images in ETH3D, we observe pronounced grid artifacts in the predicted depth, as shown in Figure 7. In contrast, the other methods can handle depth estimation from the corresponding downsampled image inputs without exhibiting such artifacts.

3. Evaluation on Challenging Benchmarks

3.1. Comparison

We further evaluate on three challenging benchmarks with domain shifts and adverse imaging conditions: HAMMER [4], Booster [7], and FLSea [8]. We compare against UniDepthV2 [6] and Metric3D v2 [3] under the same zero-shot setting and report δ_1 ↑.

As shown in Table 1, our method substantially improves performance on HAMMER, indicating stronger generalization to dense-geometry scenarios. On Booster (specular/transparent) and FLSea (underwater), the domain shift remains challenging; while UniDepthV2 achieves the best performance, our method remains competitive and consistently outperforms Metric3D v2.

3.2. Limitation

Our method has several limitations. While MD2E generalizes well to dense-geometry scenarios such as HAMMER, its performance is less robust under stronger domain shifts, including specular or transparent scenes (Booster) and underwater imagery (FLSea), as shown in Table 1. A possible reason is that these conditions weaken the edge–depth correspondence exploited by our method, making predicted edges less reliable for metric-depth calibration. Since MD2E depends on stable edge prediction for both supervision and scale reasoning, its advantage may diminish when edge cues are distorted by severe appearance changes. Improving the robustness of edge-based calibration in such settings is an important direction for future work.

4. Analysis of the Global Scale Cue

We analyze the learned global scale cue t and its relationship with camera and scene factors. Let f/W denote a focal-related term (focal length normalized by image width), and let $d_{\max}$ denote the per-sample maximum ground-truth depth (not a dataset-specific depth cap).

4.1. Controlling for depth variation

The marginal correlation between t and f/W can be confounded when f/W and depth statistics co-vary across samples. To isolate the association between t and f/W beyond depth variation, we compute the partial Pearson correlation between t_{pred} and $(f/W)^{0.1}$ while controlling for $(d_{\max})^{0.5}$. For completeness, the partial correlation is

$$\rho_{t_{\text{pred}},(f/W)^{0.1}|(d_{\max})^{0.5}} = \frac{\rho_{t_{\text{pred}},(f/W)^{0.1}} - \rho_{t_{\text{pred}},(d_{\max})^{0.5}}\,\rho_{(f/W)^{0.1},(d_{\max})^{0.5}}}{\sqrt{\left(1-\rho^2_{t_{\text{pred}},(d_{\max})^{0.5}}\right)\left(1-\rho^2_{(f/W)^{0.1},(d_{\max})^{0.5}}\right)}} \tag{1}$$

which yields $\rho_{t_{\text{pred}},(f/W)^{0.1}|(d_{\max})^{0.5}} = 0.476$. This indicates a moderate positive association between t_{pred} and the focal-related factor after accounting for depth variation.

4.2. Interpreting the marginal trend

A counter-intuitive marginal trend between t and f/W may occur because our supervision uses *depth-to-edge* cues derived from depth discontinuities, rather than RGB image edges; therefore t is unrelated to RGB sharpness/defocus. Moreover, t is a *global* statistic aggregated over the full image, so local edge thickness changes under focal/viewpoint variation do not determine the overall trend. When focal changes coincide with systematic shifts in scene layout or depth distribution, the marginal correlation can exhibit a sign flip even when the conditional relationship remains consistent. The positive partial correlation in Eq. (1) supports that such effects arise from confounding rather than from an invalid modeling assumption.

References

[1] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 2

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[3] 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. *arXiv preprint arXiv:2404.15506*, 2024. 1, 2

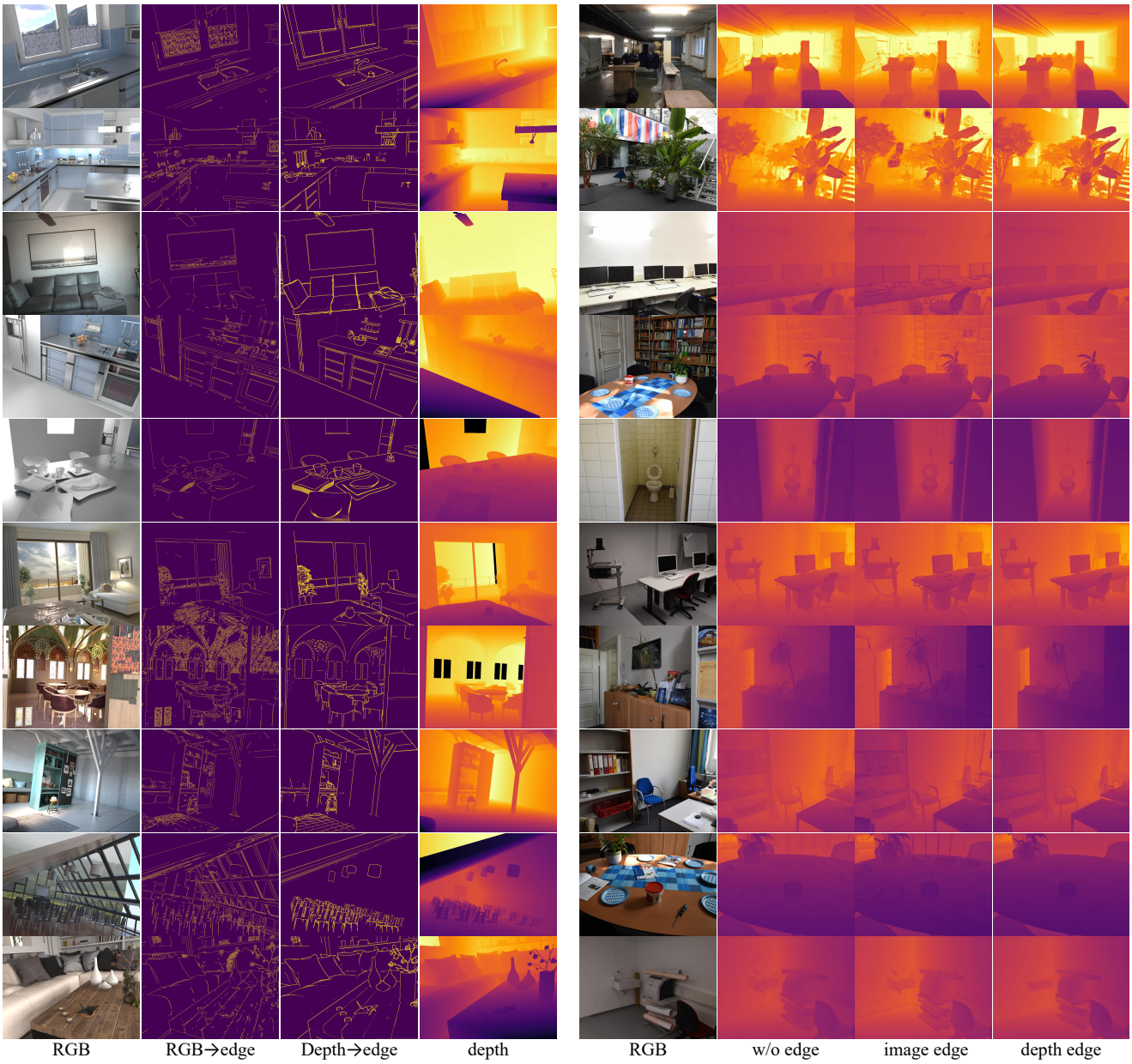

Figure 1. Left group: comparison between edges extracted from RGB images and edges derived from depth annotations on the training dataset. Right group: qualitative results on unseen dataset; from left to right: RGB input, model without edge supervision, model supervised by Canny image edges, and model supervised by our depth-to-edge edges.

[4] HyunJun Jung, Patrick Ruhkamp, Guangyao Zhai, Nikolas Brasch, Yitong Li, Yannick Verdie, Jifei Song, Yiren Zhou, Anil Armagan, Slobodan Ilic, Aleš Leonardis, Nassir Navab, and Benjamin Busam. On the importance of accurate geometry data for dense 3d vision tasks. In *CVPR*, pages 780–791, 2023. 2

[5] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of CNN-based single-image depth estimation methods. In *Computer Vision – ECCV 2018 Workshops*, pages 331–348, 2018. 1

[6] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025. 1, 2

[7] Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Open challenges in deep stereo: The booster dataset. In *CVPR*, pages 21168–21178, 2022. 2

[8] Yelena Randall and Tali Treibitz. Flsea: Underwater visual-inertial and stereo-vision forward-looking datasets. *arXiv*

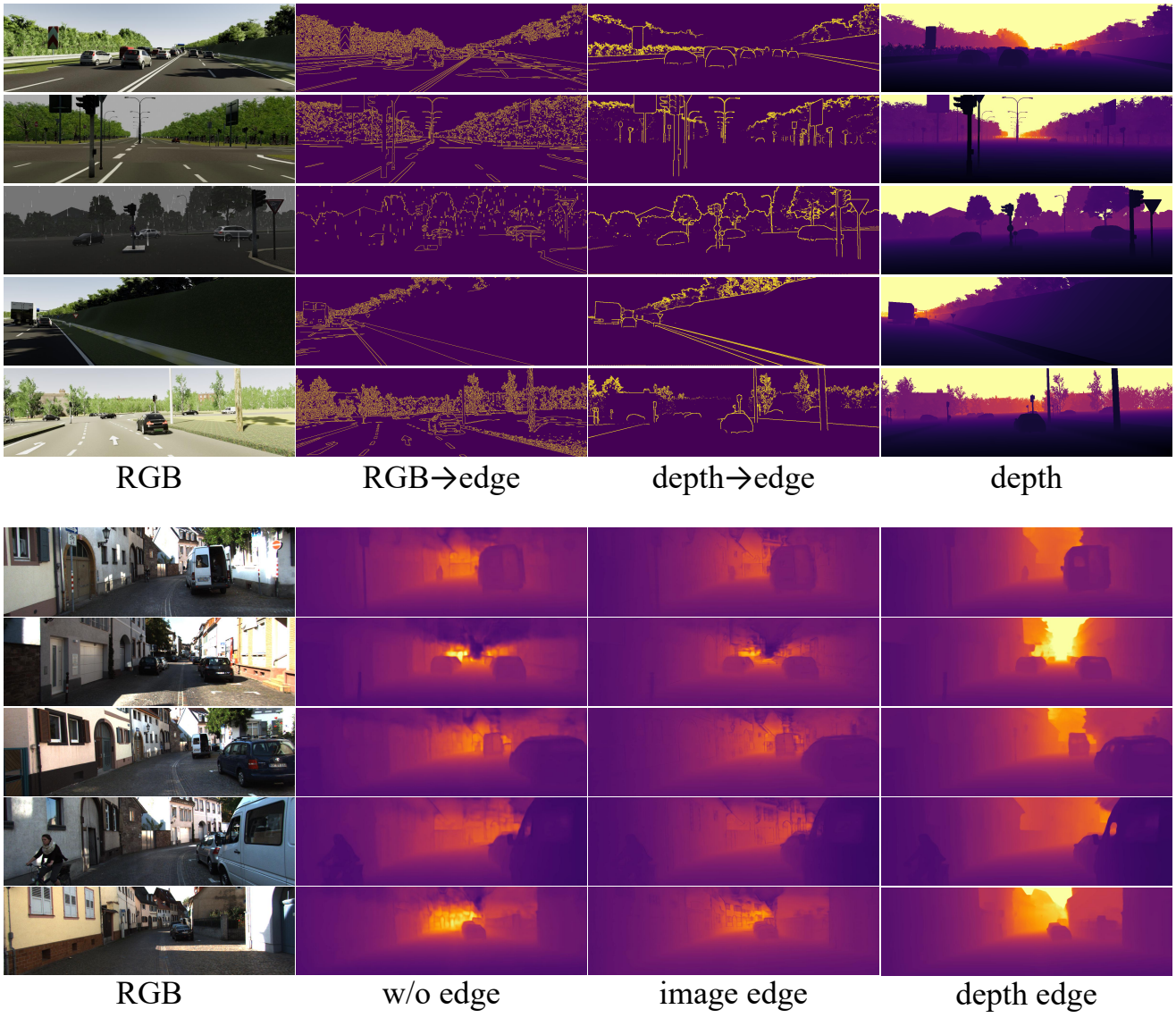

Figure 2. Top group: comparison on the training set between edges extracted from RGB images and edges derived from depth annotations. Bottom group: qualitative depth predictions on unseen datasets; from left to right: RGB input, model without edge supervision, model supervised by Canny image edges, and model supervised by our depth-to-edge edges.

preprint arXiv:2302.12772, 2023. 2

[9] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, pages 3260–3269, 2017. 1

[10] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer, 2012. 1

[11] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, and Gregory Shakhnarovich. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 1

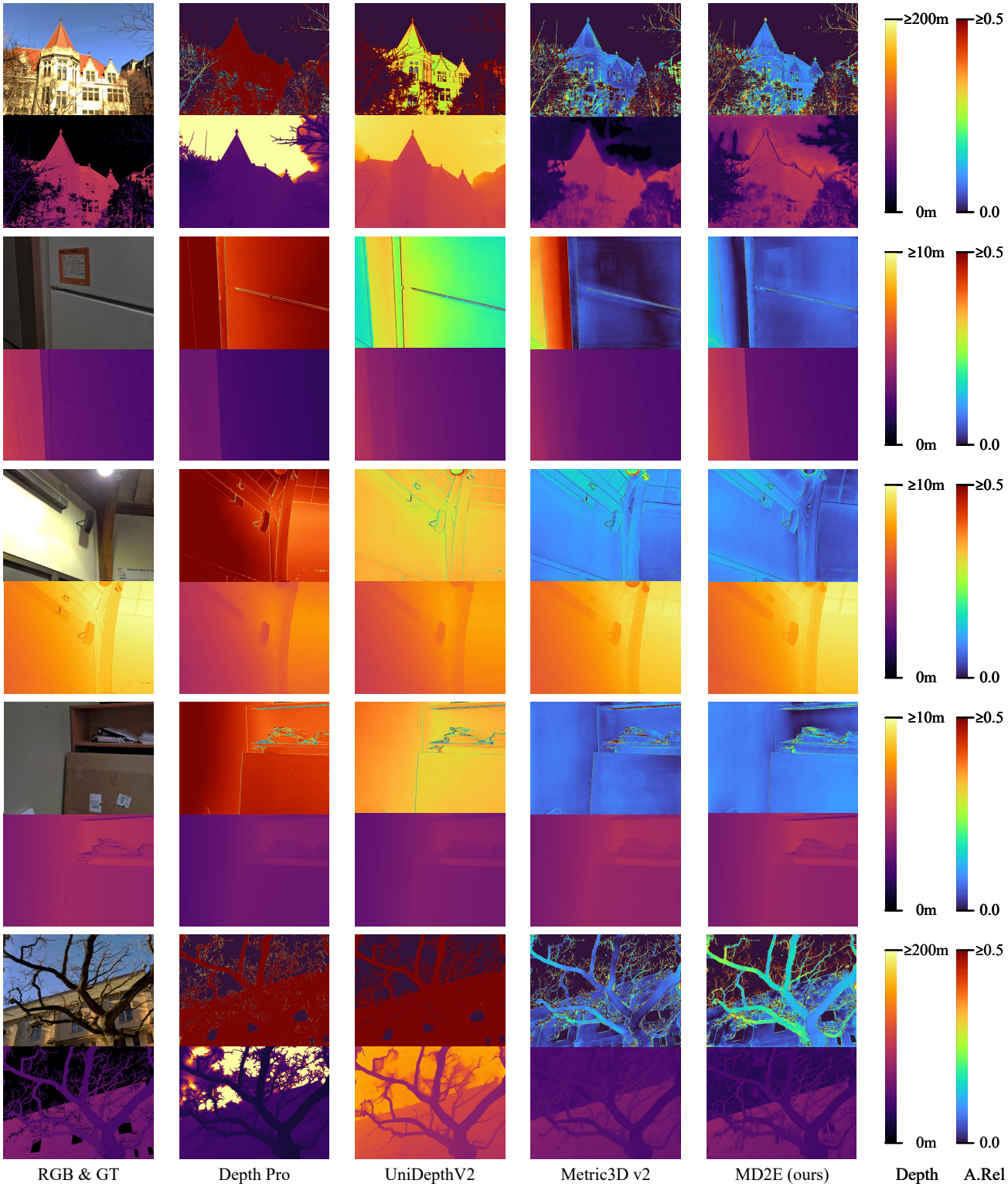

Figure 3. Zero-shot qualitative comparison on DIODE.

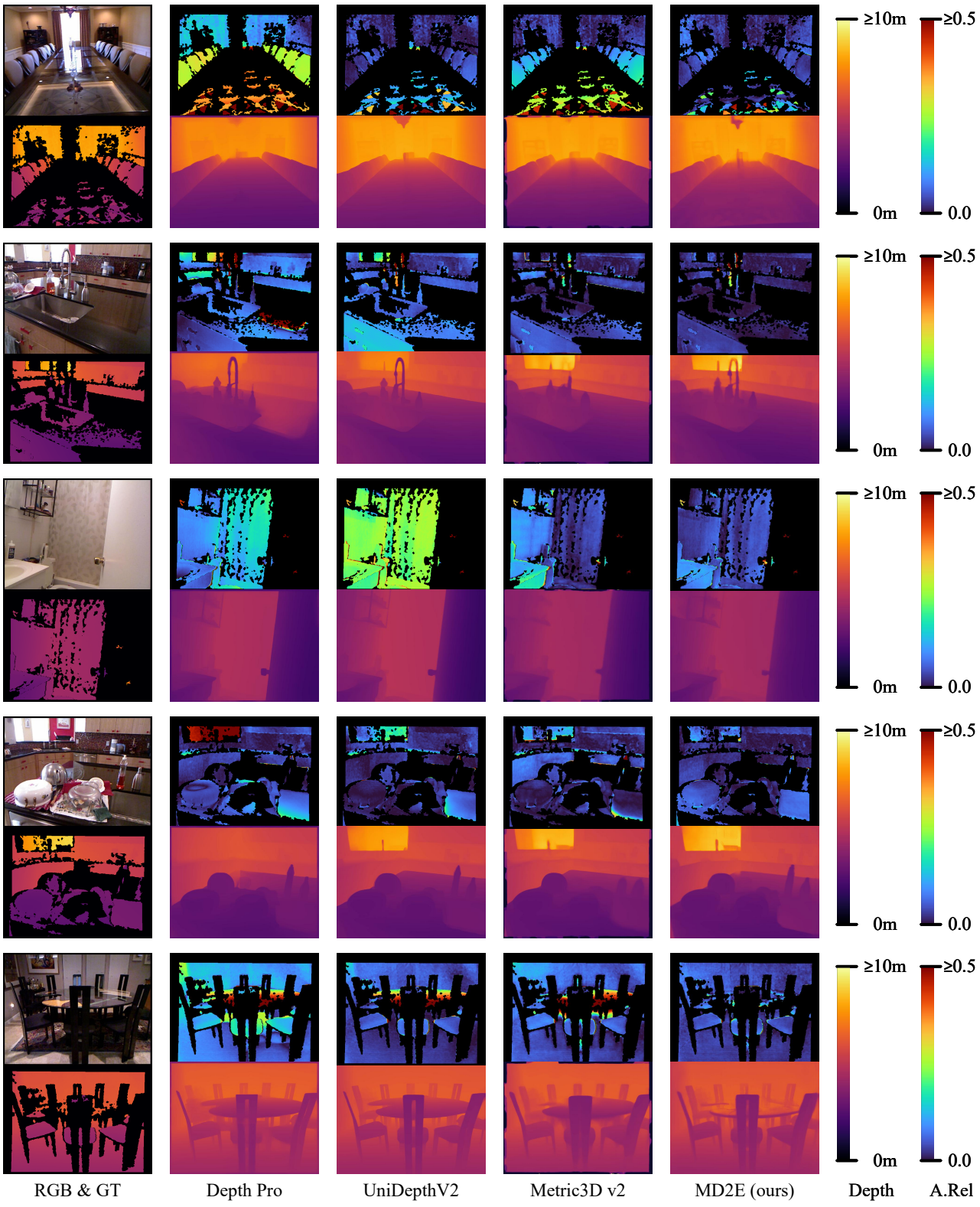

Figure 4. Zero-shot qualitative comparison on NYUv2.

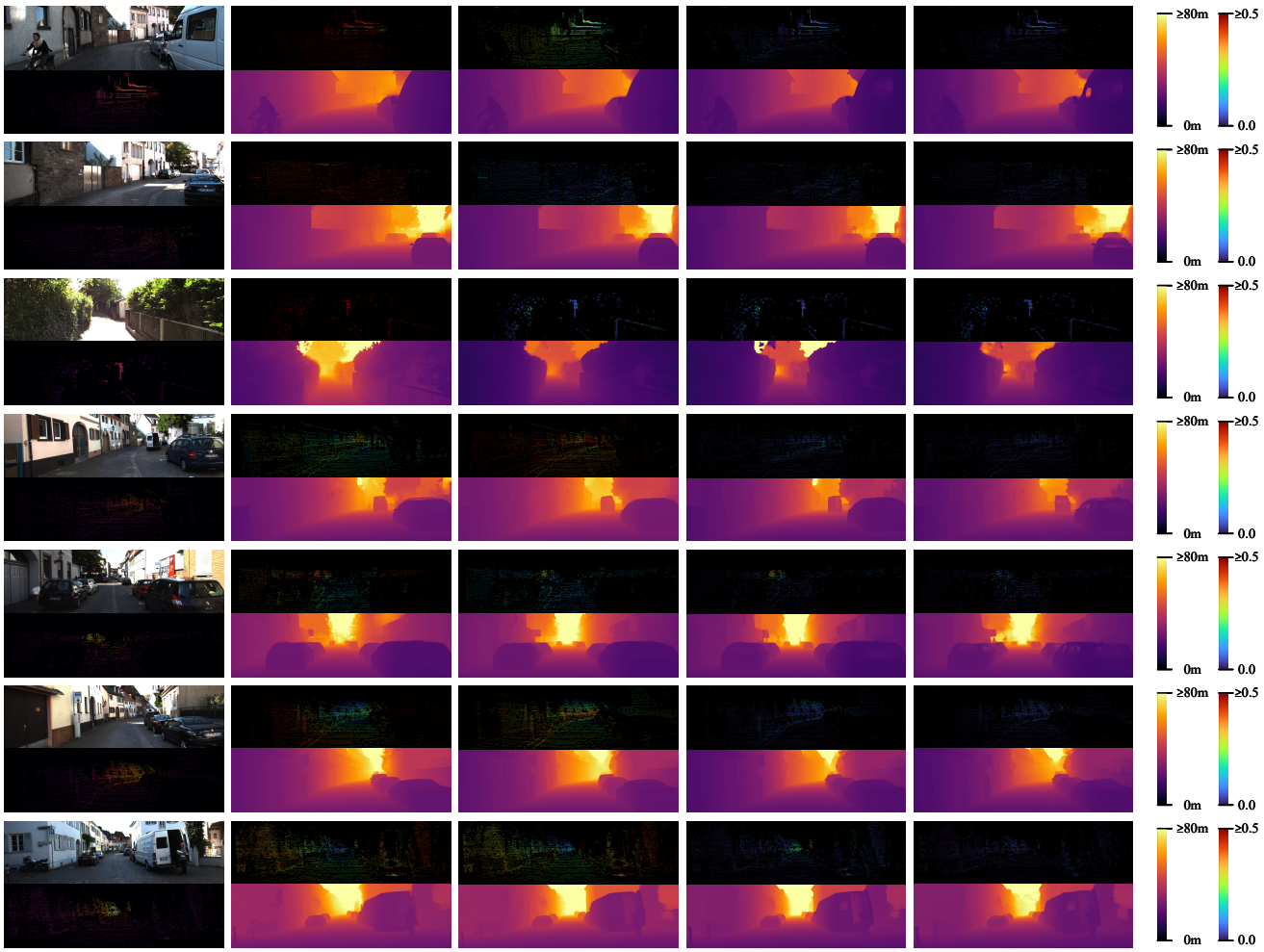

Figure 5. Zero-shot qualitative comparison on KITTI.

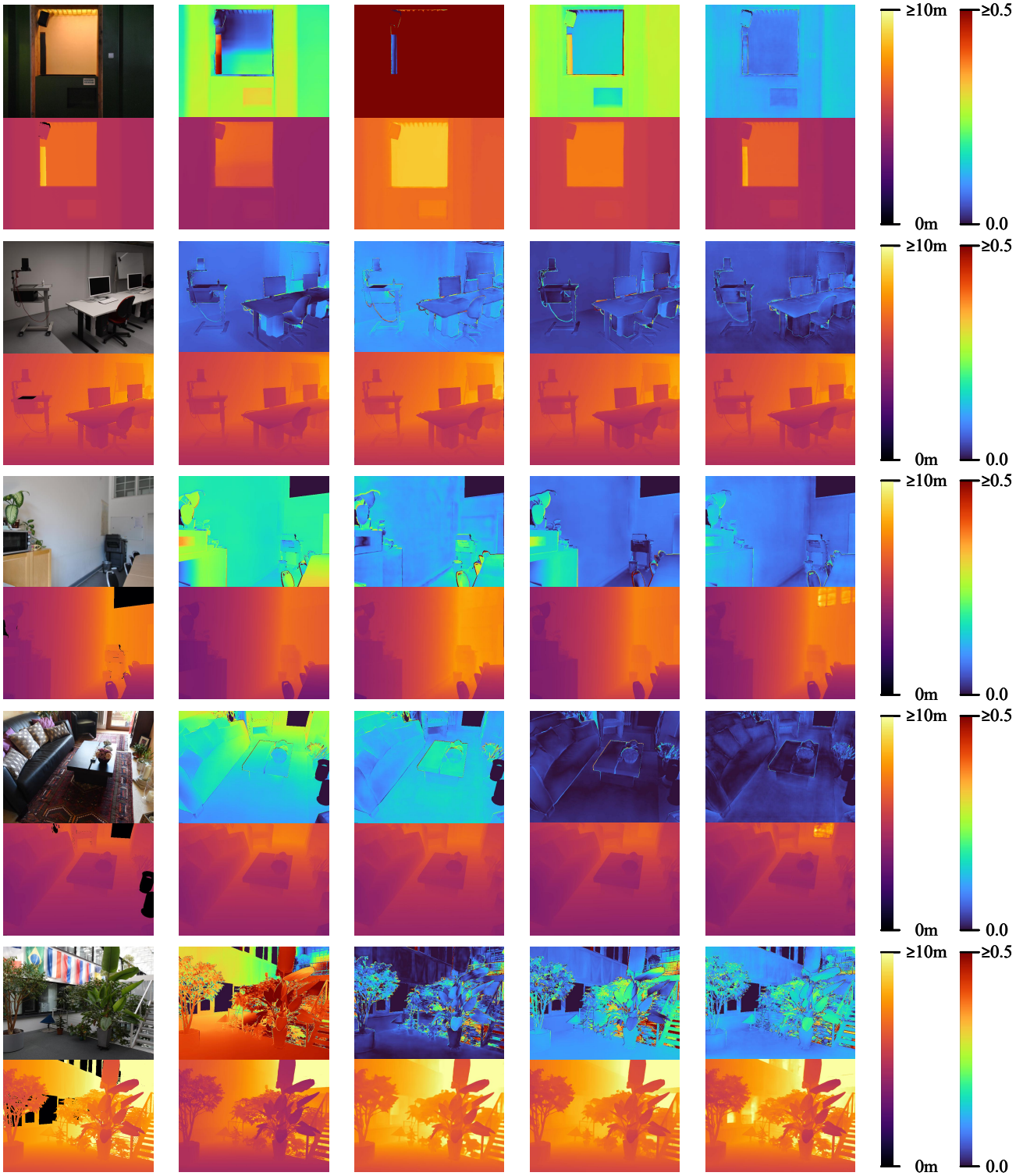

Figure 6. Zero-shot qualitative comparison on iBims-1.

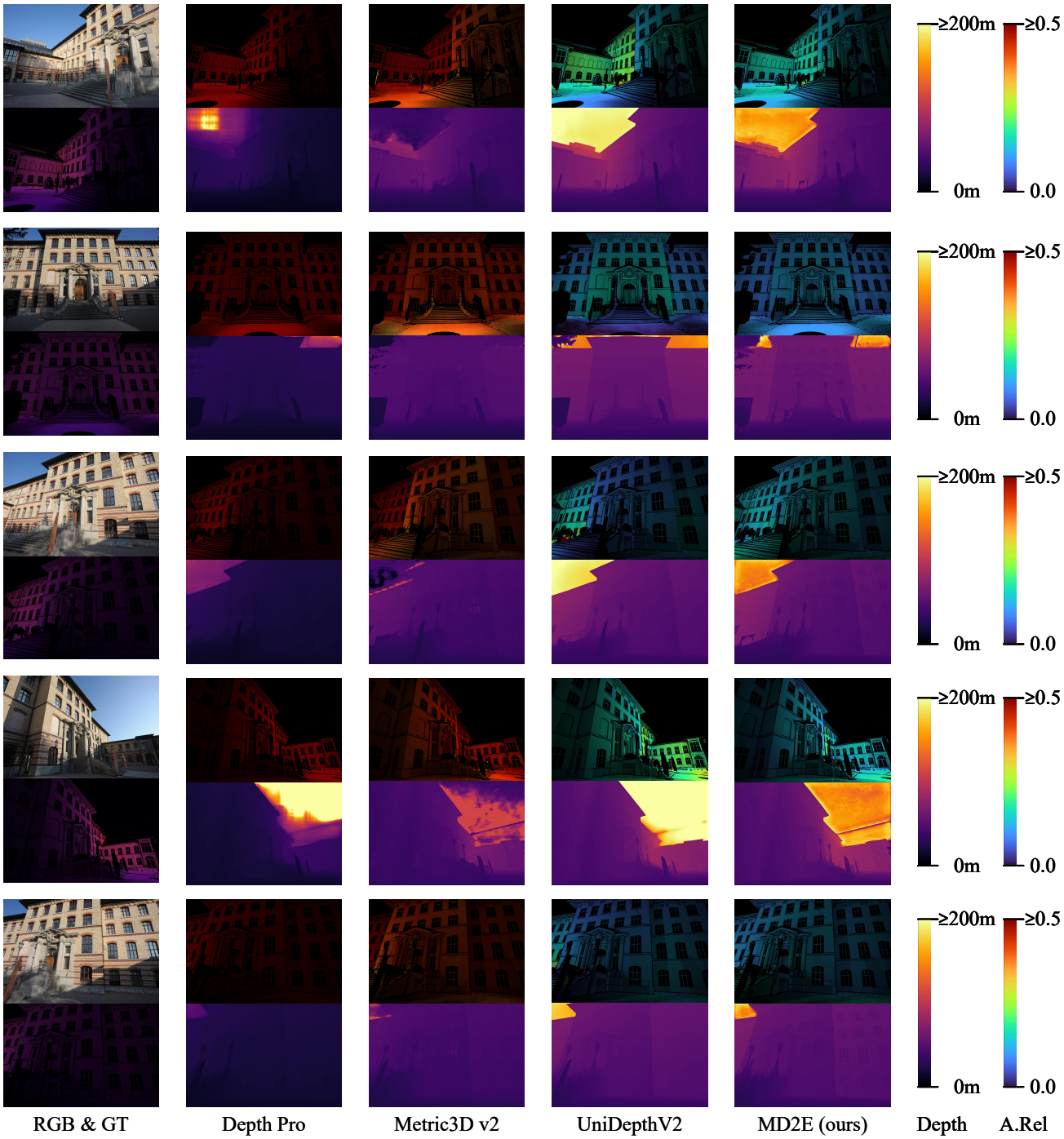

Figure 7. Zero-shot qualitative comparison on ETH3D.