Metric3D: Towards Zero-shot Metric 3D Prediction from A Single Image

Wei Yin[1*], Chi Zhang[2*], Hao Chen[3†], Zhipeng Cai[4], Gang Yu[2], Kaixuan Wang[1], Xiaozhi Chen[1], Chunhua Shen[3]

[1] DJI Technology [2] Tencent [3] Zhejiang University [4] Intel Labs

e-mail: [1]yvanwy@outlook.com, {halfbullet.wang, xiaozhi.chen}@dji.com;

[2]{johnczhang, skicyyu}@tencent.com;

[3]haochen.cad@zju.edu.cn, chunhua@me.com; [4]zhipeng.cai@intel.com

1. Details for Models

In our work, our encoder employs the convnext-large network, whose pretrained weight is from the official released ImageNet-22k pretraining. The decoder follows the [26]. The depth range is $[0.3m, 150m]$. We establish 4 flip connections from different levels of encoder blocks to the decoder to merge more low-level features.

2. Datasets and Training and Testing

We collect over $8M$ data from 11 public datasets for training. Dataset list is shown in 1. The autonomous driving datasets, including DDAD [11], Lyft [12], DrivingStereo [23], Argoverse2 [21], DSEC [9], and Pandaset [22], have provided LiDar and camera intrinsic and extrinsic parameters. We project the LiDar to camera image planes to obtain ground-truth depths. In contrast, Cityscapes [6], DIML [5], and UASOL [1] do not have ground truth depth, but are with calibrated stereo images. We use draftstereo [14] to achieve pseudo-ground truth. Mapillary PSD [15] dataset provides paired RGB-D, but the depth maps are achieved from a structure-from-motion method. The camera intrinsic parameters are estimated from the SfM. We believe that such achieved metric information is very noisy, so we do not enforce learning-metric-depth loss on this data, i.e. L_{silog}, to reduce the effect from noises. For the Taskonomy [28] dataset, we follow LeReS [25] to obtain the instance planes, which are employed in the pairwise normal regression loss. During training, we employ the training strategy from [24] to balance all datasets in each training batch.

The testing data information is listed in 1. All of them are captured by high-quality sensors. In our testing, we employ their provided camera intrinsic parameters to perform our proposed canonical space transformation. Datasets have different focal lengths.

Table 1 – Training and testing datasets used in experiments.

Datasets	Scenes	Label	Size	# Cam.
Training Data				
DDAD [11]	Outdoor	LiDar	∼80K	36+
Lyft [12]	Outdoor	LiDar	∼50K	6+
Driving Stereo (DS) [23]	Outdoor	Stereo†	∼181K	1
DIML [5]	Outdoor	Stereo†	∼122K	10
Arogoverse2 [21]	Outdoor	LiDar	∼3515K	6+
Cityscapes [6]	Outdoor	Stereo†	∼170K	1
DSEC [9]	Outdoor	LiDar	∼26K	1
Mapillary PSD [15]	Outdoor	SfM‡	750K	1000+
Pandaset [22]	Outdoor	LiDar	∼48K	6
UASOL [1]	Outdoor	Stereo†	∼137K	1
Taskonomy [28]	Indoor	LiDar	∼4M	∼1M
Testing Data				
NYU [18]	Indoor	Kinect	654	1
KITTI [10]	Outdoor	LiDar	652	4
ScanNet [7]	Indoor	Kinect	700	1
NuScenes (NS) [4]	Outdoor	LiDar	10K	6
ETH3D [16]	Outdoor	LiDar	431	1
DIODE [20]	In/Out	LiDar	771	1
7Scenes [17]	Indoor	Kinect	17k	1
iBims-1 [13]	Indoor	LiDar	100	1

† 'Stereo': we use RaftStereo [14] to retrieve the pseudo ground truth.
‡ 'SfM': pseudo ground truth is retrieved by structure from motion.

3. Details for Some Experiments

Evaluation of zero-shot 3D scene reconstruction. In this experiment, we use all methods' released models to predict each frame's depth and use the ground truth pose and camera intrinsic parameters to reconstruct point clouds. When evaluating the reconstructed point cloud, we employ the iterative closest point (ICP) [2] algorithm to match the predicted point clouds with ground truth by a pose transformation matrix. Finally, we evaluate the Chamfer l1 distance and F-score on the point cloud.

Reconstruction of in-the-wild scenes. We collect several photos from Flickr. From their associated camera metadata, we can obtain the focal length $\hat{f}$ and the pixel size δ. Ac-

*Equal contributions.
†Corresponding author.

cording to $\hat{f}/\delta$, we can obtain the pixel-represented focal length for 3D reconstruction and achieve the metric information. We use meshlab software to measure some structures' size on point clouds. More visual results are shown in 3.

Generalization of metric depth estimation. To evaluate our method's robustness of metric recovery, we test on 8 zero-shot datasets, i.e. NYU, KITTI, DIODE (indoor and outdoor parts), ETH3D, iBims-1, NuScenes, and 7Scenes. Details are reported in Tab. 1. We use the officially provided focal length to predict the metric depths. All benchmarks use the same depth model for evaluation. We don't perform any scale alignment.

Evaluation on affine-invariant depth benchmarks. We follow existing affine-invariant depth estimation methods to evaluate 5 zero-shot datasets. Before evaluation, we employ the least square fitting to align the scale and shift with ground truth [26]. Previous methods' performance is cited from their papers.

Dense-SLAM Mapping. This experiment is conducted on the KITTI odometry benchmark. We use our model to predict metric depths, and then naively input them to the Droid-SLAM system as an initial depth. We do not perform any finetuning but directly run their released codes on KITTI. With Droid-SLAM predicted poses, we unproject depths to the 3D point clouds and fuse them together to achieve dense metric mapping. More qualitative results are shown in 2.

4. More Visual Results

Reconstructing 360°NuScenes scenes. Current autonomous driving cars are equipped with several pin-hole cameras to capture 360°views. Capturing the surround-view depth is important for autonomous driving. We sampled some scenes from the testing data of NuScenes. With our depth model, we can obtain the metric depths for 6-ring cameras. With the provided camera intrinsic and extrinsic parameters, we unproject the depths to the 3D point cloud and merge all views together. See 4 for details. Note that 6-ring cameras have different camera intrinsic parameters. We can observe that all views' point clouds can be fused together consistently.

Reconstructing 360°DDAD scenes. In our provided supplementary videos, we show the reconstructed point cloud of 360°view of DDAD scenes. Videos show that our reconstructed point clouds don't have noticeable cross-view inconsistency issues.

Qualitative comparison of depth estimation. In Fig. 1, 5, 6, and 7, We show the qualitative comparison of depth maps with Adabins [3], NewCRFs [27], and Omnidata [8]. Our results have much less artifacts.

References

[1] Zuria Bauer, Francisco Gomez-Donoso, Edmanuel Cruz, Sergio Orts-Escolano, and Miguel Cazorla. Uasol, a large-scale high-resolution outdoor stereo dataset. *Scientific data*, 6(1):1–14, 2019. 1

[2] Paul Besl and Neil McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: Control Paradigms and Data Structures*, volume 1611, pages 586–606. Spie, 1992. 1

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4009–4018, 2021. 2

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 11621–11631, 2020. 1

[5] Jaehoon Cho, Dongbo Min, Youngjung Kim, and Kwanghoon Sohn. DIML/CVL RGB-D dataset: 2m RGB-D images of natural indoor and outdoor scenes. *arXiv: Comp. Res. Repository*, 2021. 1

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016. 1

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5828–5839, 2017. 1

[8] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 10786–10796, 2021. 2

[9] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 2021. 1

[10] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.*, 2013. 1

[11] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2020. 1

[12] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. Level 5 perception dataset 2020. https://level-5.global/level5/data/, 2019. 1

[13] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Eur. Conf. Comput. Vis. Worksh.*, pages 0–0, 2018. 1

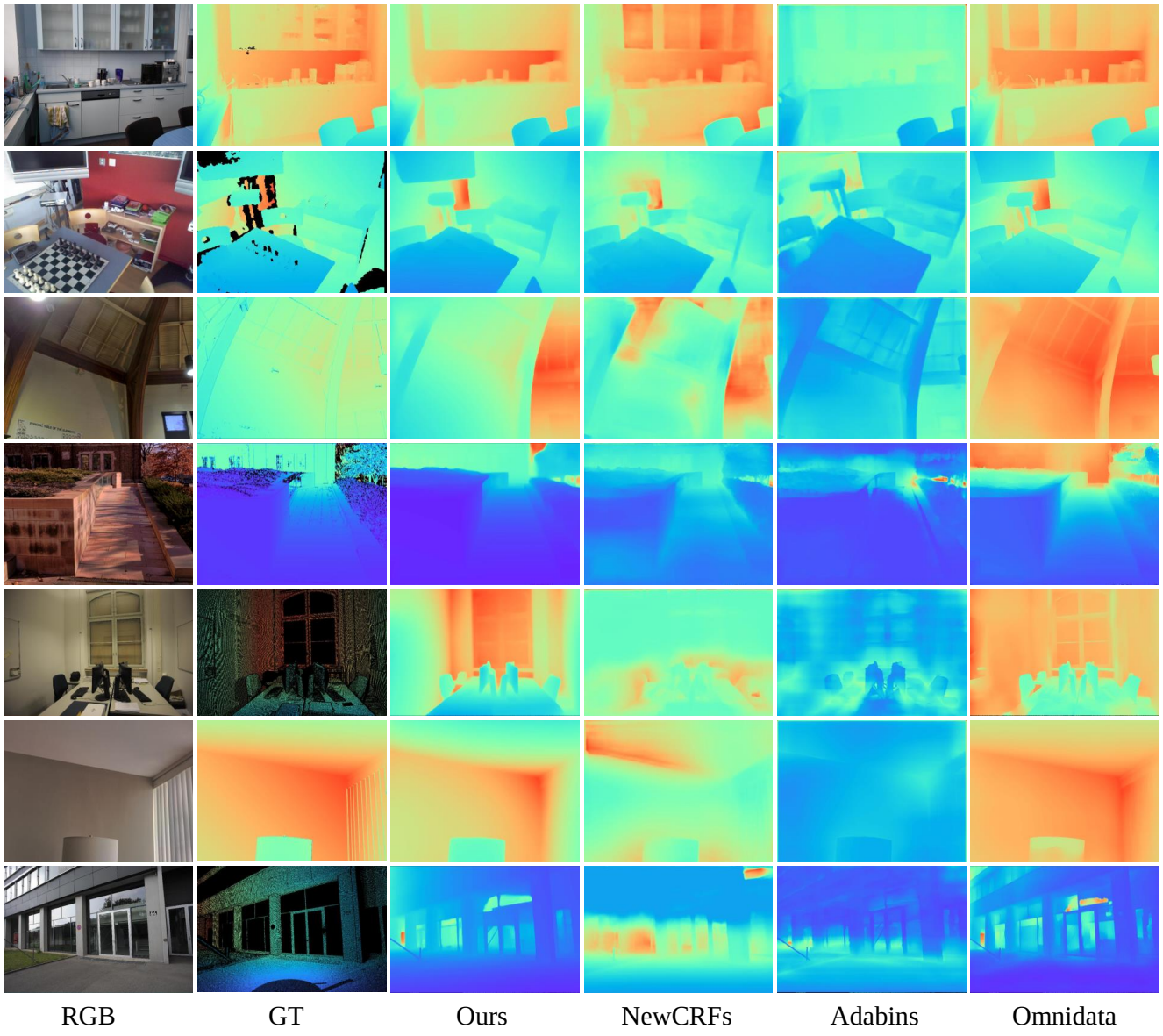

Figure 1 – Depth estimation. The visual comparison of predicted on iBims, ETH3D, and DIODE.

[14] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *Int. Conf. 3D. Vis.*, 2021. 1

[15] Manuel Lopez-Antequera, Pau Gargallo, Markus Hofinger, Samuel Rota Bulò, Yubin Kuang, and Peter Kontschieder. Mapillary planet-scale depth dataset. In *Proc. Eur. Conf. Comp. Vis.*, volume 12347, pages 589–604, 2020. 1

[16] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3260–3269, 2017. 1

[17] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2930–2937, 2013. 1

[18] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. Eur. Conf. Comp. Vis.*, pages 746–760. Springer, 2012. 1

[19] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. volume 34, pages 16558–16569, 2021. 4

[20] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv:*

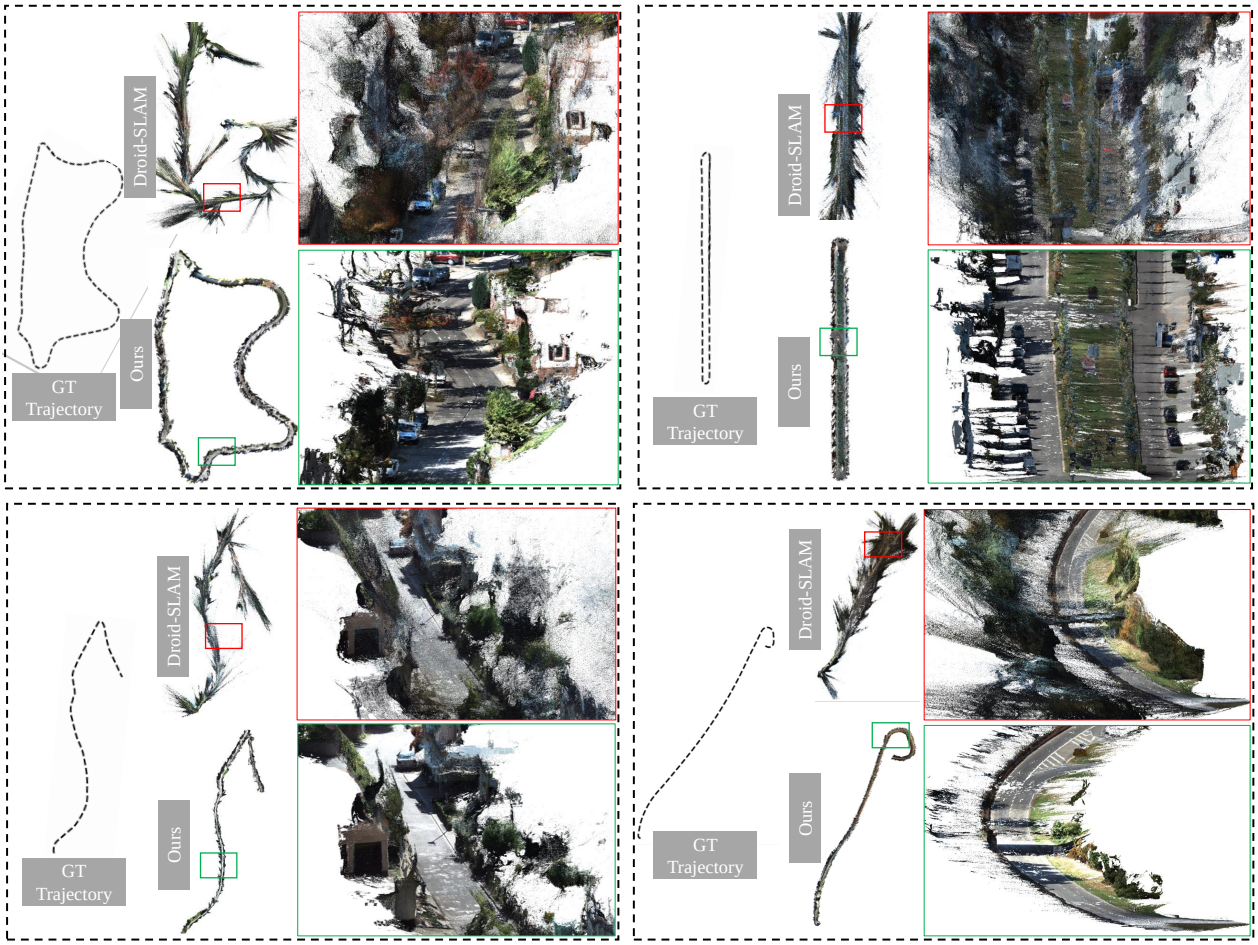

Figure 2 – Dense-SLAM Mapping. Existing SOTA mono-SLAM methods usually face scale drift problems in large-scale scenes and unable to achieve the metric scale. We show the ground-truth trajectory and Droid-SLAM [19] predicted trajectory and their dense mapping. Then, we naively input our metric depth to Droid-SLAM, which can recover a much more accurate trajectory and perform the *metric* dense mapping.

Comp. Res. Repository, page 1908.00463, 2019. 1

[21] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proc. Advances in Neural Inf. Process. Syst.*, 2021. 1

[22] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, Yunlong Wang, and Diange Yang. Pandaset: Advanced sensor suite dataset for autonomous driving. In *IEEE Int. Intelligent Transportation Systems Conf.*, 2021. 1

[23] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019. 1

[24] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv: Comp. Res. Repository*, page 2002.00569, 2020. 1

[25] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen. Towards accurate reconstruction of 3d scene shape from a single monocular image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 1

[26] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021. 1, 2

[27] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New CRFs: Neural window fully-connected CRFs for monocular depth estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2022. 2

[28] Amir Zamir, Alexander Sax, , William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* IEEE, 2018. 1

Figure 3 – 3D metric reconstruction of in-the-wild images. We collect several Flickr images and use our model to reconstruct the scene. The focal length information is collected from the photo's metadata. From the reconstructed point cloud, we can measure some structures' sizes. We can observe that sizes are in a reasonable range.

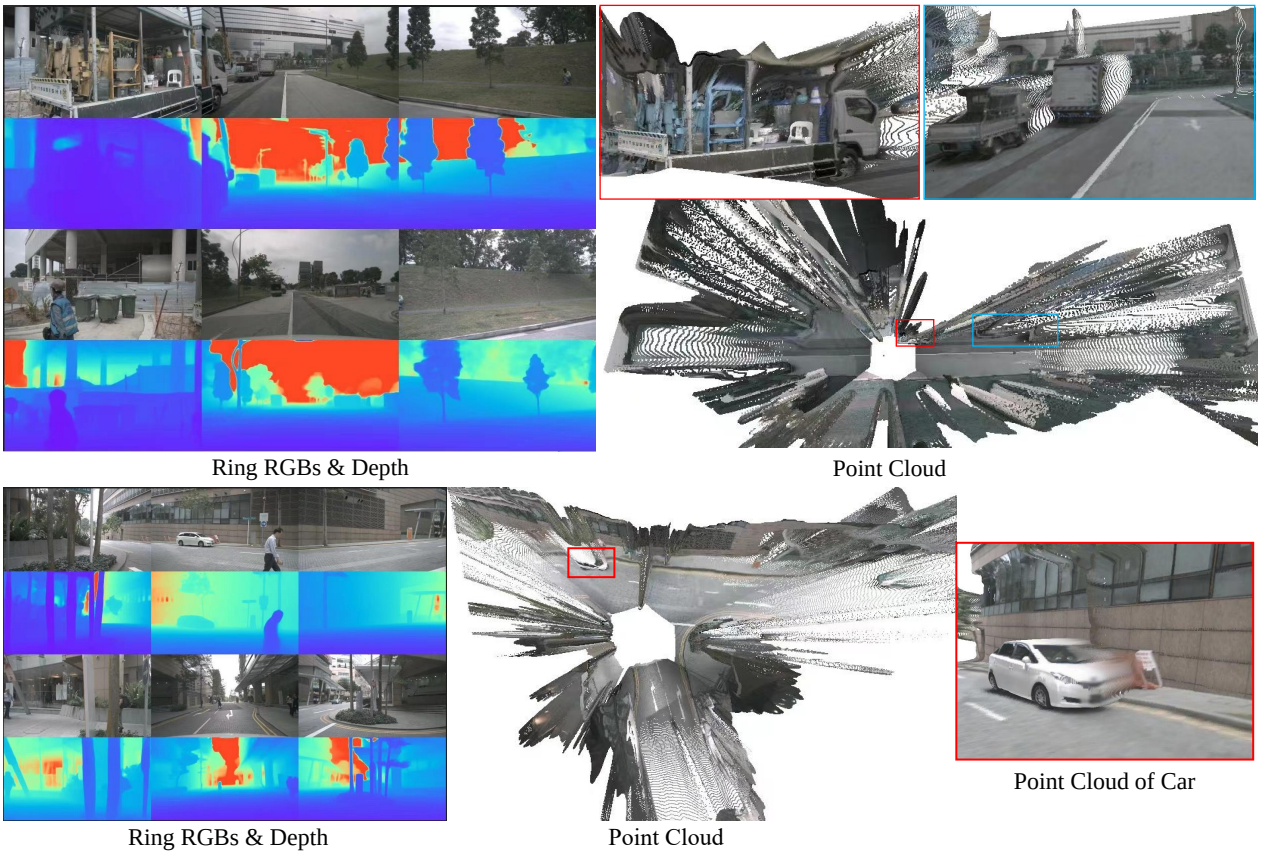

Figure 4 – 3D reconstruction of 360° views. Current autonomous driving cars are equipped with several pin-hole cameras to capture 360° views. With our model, we can reconstruct each view and smoothly fuse them together. We can see that all views can be well merged together without scale inconsistency problems. Testing data are from NuScenes. Note that the front view camera has a different focal length from other views.

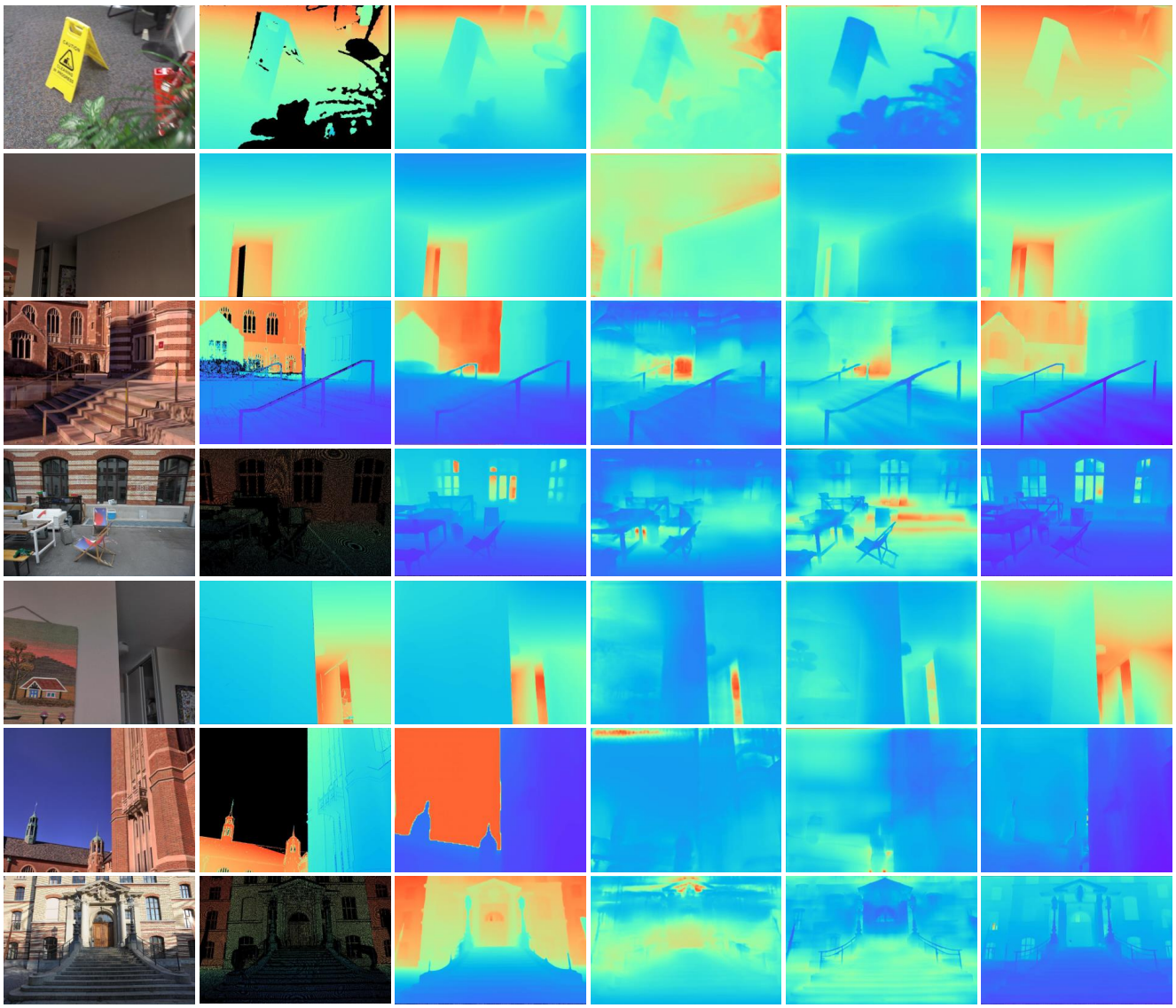

Figure 5 – Depth estimation. The visual comparison of predicted on iBims, ETH3D, and DIODE.

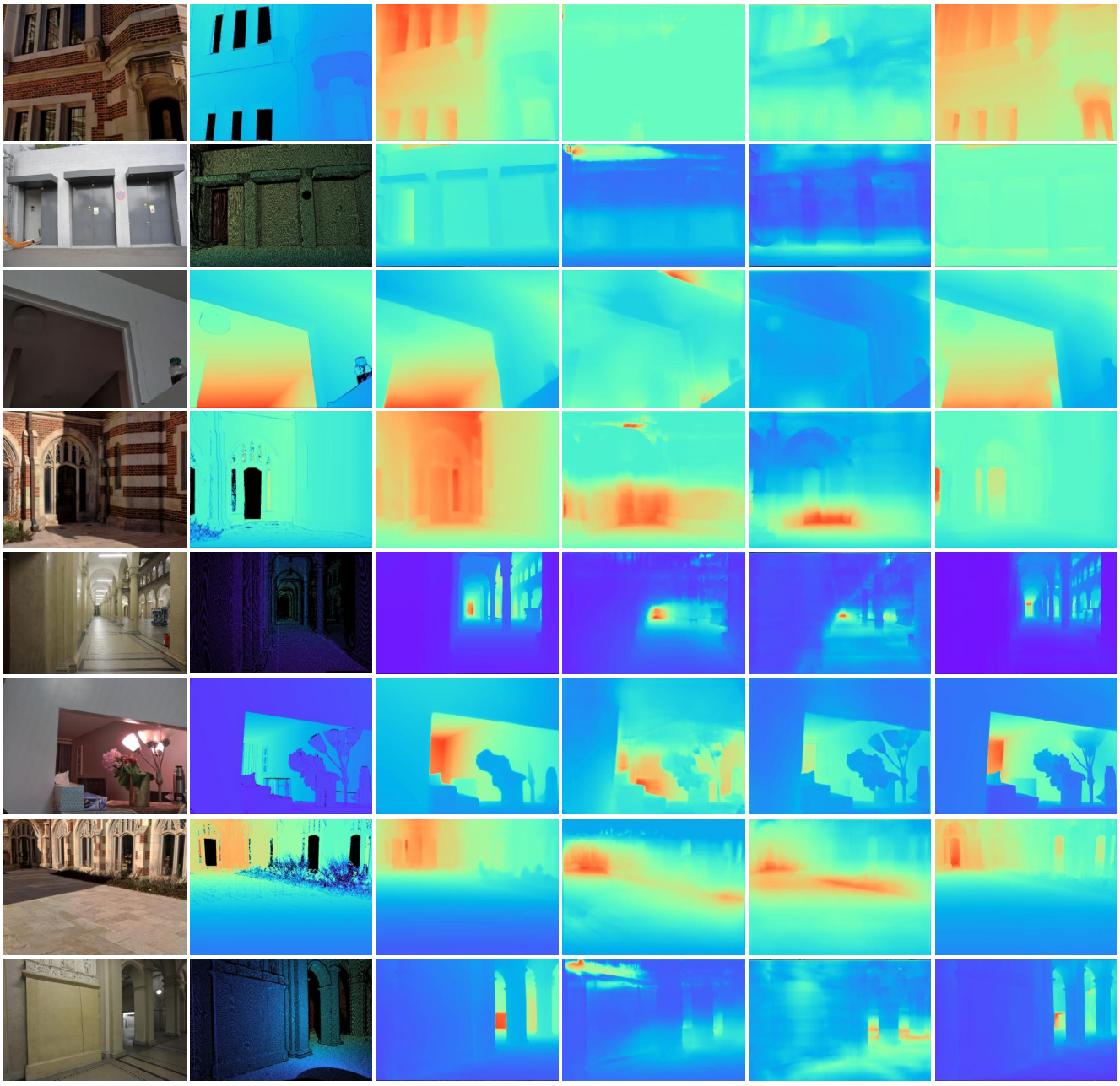

Figure 6 – Depth estimation. The visual comparison of predicted on iBims, ETH3D, and DIODE.

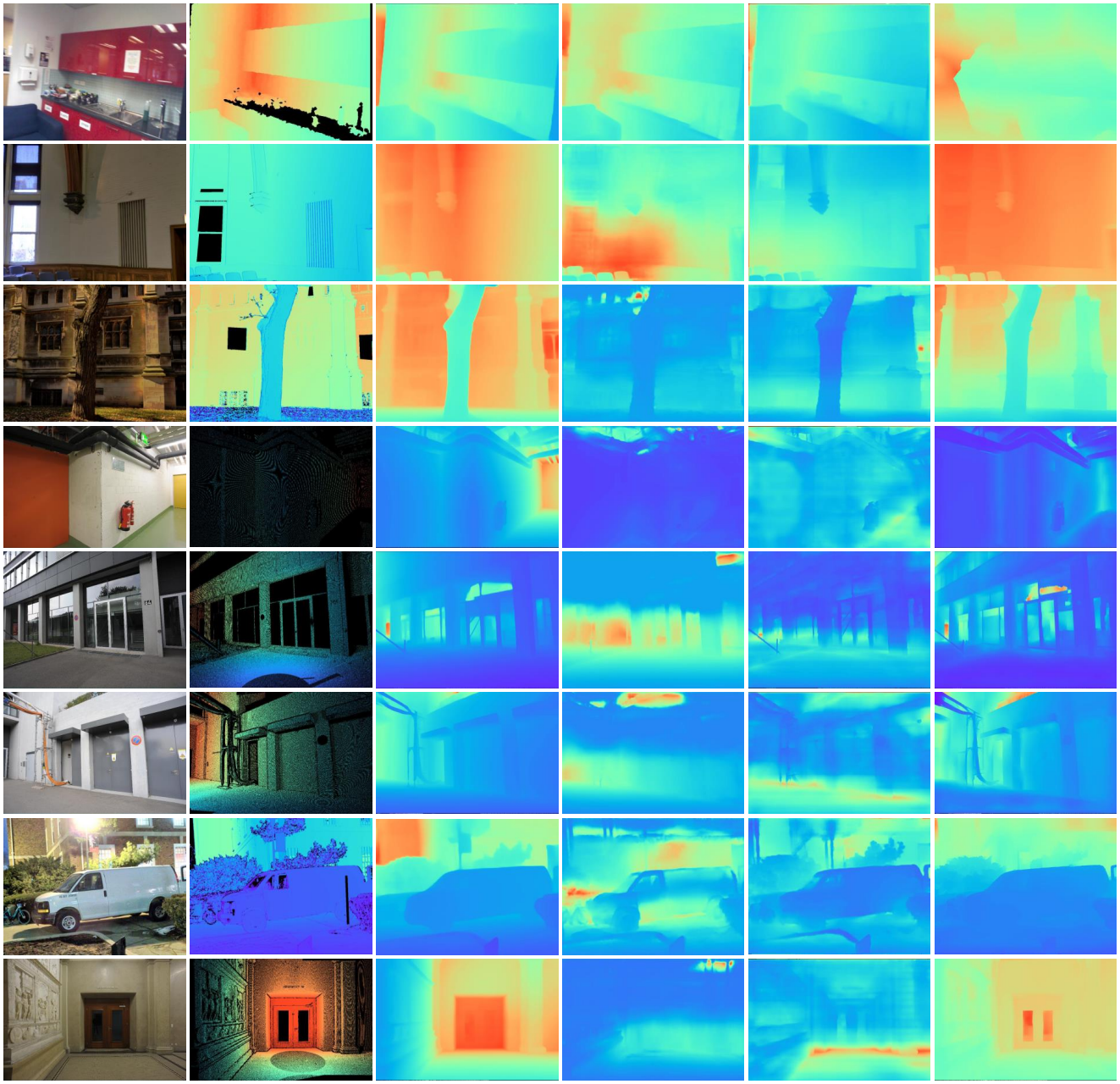

Figure 7 – Depth estimation. The visual comparison of predicted on iBims, ETH3D, and DIODE.