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

Wei Yin1*, Chi Zhang2*, Hao Chen3†, Zhipeng Cai4, Gang Yu2, Kaixuan Wang1, Xiaozhi Chen1, Chunhua Shen3
1 DJI Technology 2 Tencent 3 Zhejiang University 4 Intel Labs
e-mail: 1yvanwy@outlook.com, {halfbullet.wang, xiaozhi.chen}@dji.com; 2{johnczhang, skicyyu}@tencent.com; 3haochen.cad@zju.edu.cn, chunhua@me.com, 4zhipeng.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 draft-stereo [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. $L_{silag}$, 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.

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 11 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 $f$ and the pixel size $\delta$. Ac-

| Table 1 – Training and testing datasets used in experiments. |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| **Datasets**    | **Scenes**      | **Label**       | **Size**        | **# Cam.**      |
| Training Data   |                 |                 |                 |                 |
| DDAD [11]       | Indoor          | LiDar           | ~80K            | 36+             |
| Lyft [12]       | Outdoor         | LiDar           | ~50K            | 6+              |
| Driving Stereo (DS) [23] | Outdoor | Stereo   | ~181K          | 1               |
| Argoverse2 [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]     | Outdoor         | 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               |

*Equal contributions.
†Corresponding author.
‡‘SfM’: pseudo ground truth is retrieved by structure from motion.
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 naïvely 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

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

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


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.


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.