OASIS: A Large-Scale Dataset for Single Image 3D in the Wild Supplementary Material

Weifeng Chen[1,2] Shengyi Qian[1] David Fan[2] Noriyuki Kojima[1] Max Hamilton[1] Jia Deng[2]
[1]University of Michigan, Ann Arbor [2]Princeton University
{wfchen,syqian,kojimano,johnmaxh}@umich.edu dfan@alumni.princeton.edu,jiadeng@princeton.edu

A1: Surface Normal Annotation UI

The surface normal annotation UI is shown in Fig. I.

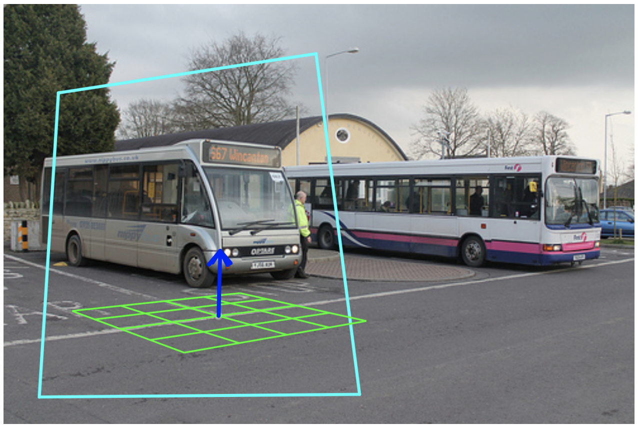

Figure I. Surface normal annotation UI. The surface normal is visualized as a blue arrow originating from a green grid, rendered in perspective projection according to the known focal length.

A2: Additional Examples from OASIS

Additional human annotations are shown in Fig. II.

A3: Comparison with Other Datasets

Tab. I compares OASIS and other datasets.

A4: Planar versus Curved Regions

Tab. II measures the annotation quality separately for planar regions and curved regions.

A5: Additional Depth Experiments

Sec 6.1 of the main paper trains and evaluates variants of the Hourglass [3] and ResNetD [15] that predict a metric depth map and a focal length on OASIS. Here we also provide results of Hourglass and ResNetD predicting only metric depth but not focal length. Tab. III shows the results.

A6: Additional Qualitative Outputs

Qualitative predictions presented in both Fig. III and Fig. 5 of the main paper are produced as follows: Depth predictions are produced by a ResNetD [15] network trained on OASIS + ImageNet [7]. Surface normal predictions are produced by an Hourglass [5] network trained on OASIS alone. Occlusion boundary and fold predictions are produced by an Hourglass [3] network trained on OASIS alone. Planar instance segmentations are produced by a PlanarReconstruction [17] network trained on Scannet [6] + OASIS.

A7: Evaluating Fold and Occlusion Boundary Detection

This section provides details on evaluating fold and occlusion boundary detection. As discussed in Sec 6.3 of the main paper, our metric is based on the ones used in evaluating edge detection [1, 16].

The input to our evaluation pipeline consists of (1) the probability of each pixel being on edge (fold or occlusion) p_e, and (2) a label of each pixel being occlusion or fold. By thresholding on p_e, we first obtain an edge map E_τ at threshold τ. We denote the occlusion pixels as O and the fold pixels as F. We find the intersection $O \cap E_\tau$ and use the same protocol as [1] to compare it against the ground-truth occlusion O^* and obtain true positive count TF_o, false positive count FP_o and false negative count FN_o. We follow the same protocol to compare $F \cap E_\tau$ against ground-truth fold F^* and obtain TF_f, FP_f and FN_f.

We then calculate the joint counts TF, FP and FN: $\text{TP}=\text{TF}_o+\text{TF}_f$, $\text{FP}=\text{FP}_o+\text{FP}_f$ and $\text{FN}=\text{FN}_o+\text{FN}_f$.

We iterate through different τ to obtain the joint counts TF, FP and FN at each threshold to obtain the final ODS/OIS F-score and AP.

References

[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image seg-

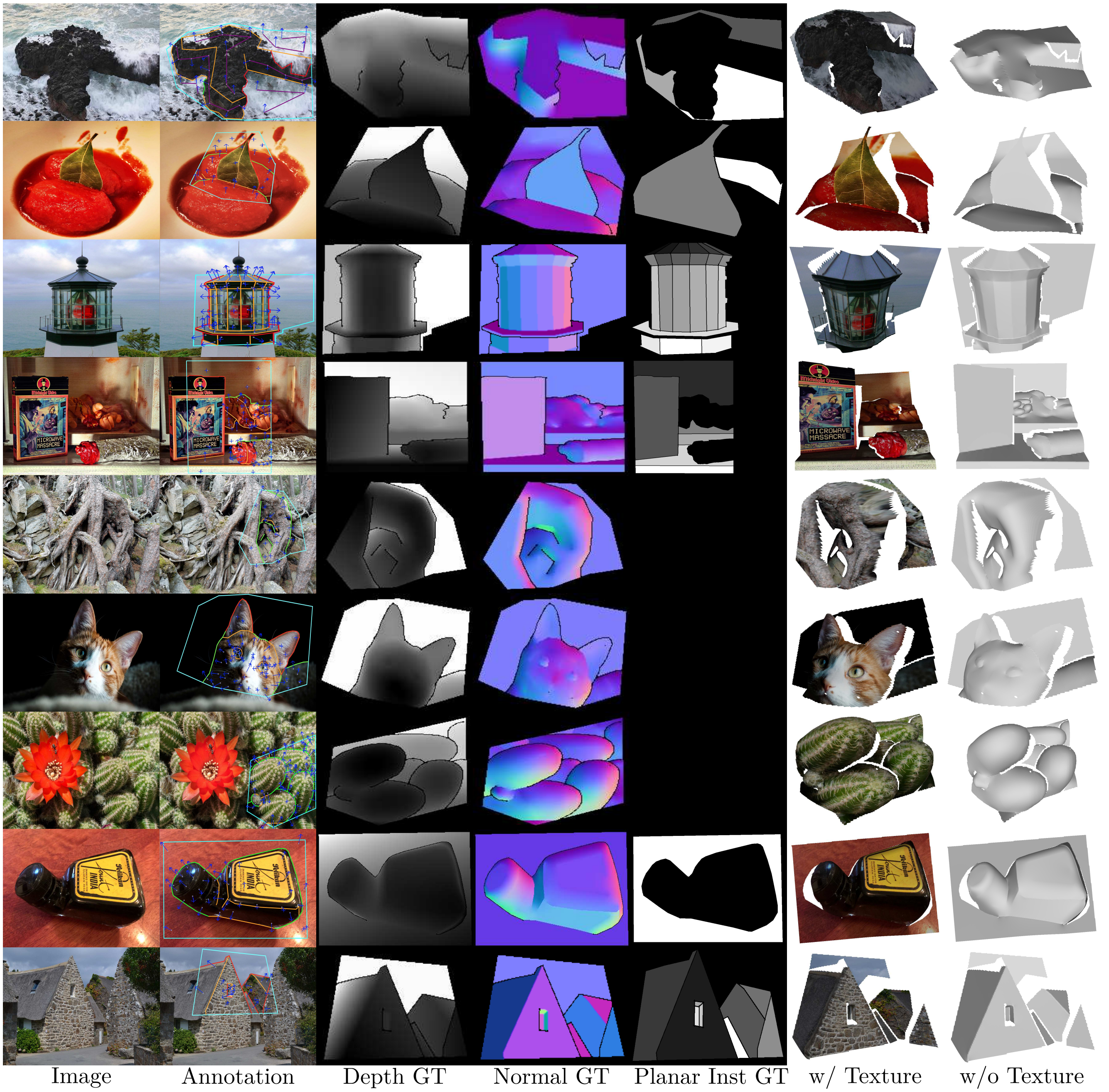

Figure II. Additional human annotations from OASIS. Note that each planar instance has a different color.

mentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2011. 1

[2] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Trans. on Graphics (SIGGRAPH)*, 32(4), 2013. 3

[3] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Advances in Neural Information Processing Systems*, pages 730–738, 2016. 1, 3

[4] Weifeng Chen, Shengyi Qian, and Jia Deng. Learning single-image depth from videos using quality assessment networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5604–5613, 2019. 3

[5] Weifeng Chen, Donglai Xiang, and Jia Deng. Surface normals in the wild. In *Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy*, pages 22–29, 2017. 1, 3

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas A Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes.

Dataset	In the Wild	Acquisition	Depth	Normals	Occlusion & Fold	Relative Normals	Planar Inst Seg	# Images
OASIS	✓	Human annotation	Metric (up to scale)	Dense	✓	✓	✓	140K
NYU Depth V2 [13]	-	Kinect	Metric	Dense	-	-	-	407K
KITTI [8]	-	LiDAR	Metric	-	-	-	-	93K
DIW [3]	✓	Human annotation	Relative	-	-	-	-	496K
SNOW [5]	✓	Human annotation	-	Sparse	-	-	-	60K
MegaDepth [11]	✓	SfM	Metric (up to scale)	-	-	-	-	130K
ReDWeb [15]	✓	Stereo	Metric (up to scale)	-	-	-	-	3.6K
3D Movie [10]	✓	Stereo	Metric (up to scale)	-	-	-	-	75K
OpenSurfaces [2]	-	Human annotation	-	Dense	-	-	-	25K
CMU Occlusion [14]	✓	Human annotation	-	-	Occlusion Only	-	-	538

Table I. Comparison between OASIS and other 3D datasets. *Metric (up to scale)* denotes that the depth is metrically accurate up to scale.

	NYU Depth [13]	
	Human-Human	Human-Sensor
Planar Regions	0.079m	0.091m
Curved Regions	0.077m	0.102m

Table II. Depth difference between different humans (Human-Human) and between humans and depth sensors (Human-Sensor) in planar and curved regions. The results are averaged over all human pairs. The mean of depth in tested samples is 2.471 m, the standard deviation is 0.754 m.

Prediction	Method	Training Data	LSIV_RMSE	WKDR
Depth	FCRN [9]	ImageNet [12] + NYU [13]	0.67	39.94%
	Hourglass [3, 11]	MegaDepth [11]	0.67	38.37%
	Hourglass [3, 11]	OASIS	0.65	42.80%
	ResNetD [15, 4]	ImageNet [12] + YouTube3D [4] + ReDWeb [15] + DIW [3]	0.66	34.03%
	ResNetD [15, 4]	ImageNet [12] + OASIS	0.63	40.08%
Depth & Focal	ResNetD [15]	ImageNet [12] + OASIS	0.37	32.04%
	ResNetD [15]	OASIS	0.47	38.79%
	Hourglass [3]	OASIS	0.47	39.64%

Table III. Depth estimation performance of different networks on OASIS (lower is better). For networks that do not produce a focal length, we use the best focal length leading to the smallest error.

In *CVPR*, volume 2, page 10, 2017. 1

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[8] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 3

[9] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 3

[10] Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019. 3

[11] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 3

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3

[13] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012. 3

[14] Andrew N Stein and Martial Hebert. Occlusion boundaries from motion: Low-level detection and mid-level reasoning. *International journal of computer vision*, 82(3):325, 2009. 3

[15] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 311–320, 2018. 1, 3

[16] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 1

[17] Zehao Yu, Jia Zheng, Dongze Lian, Zihan Zhou, and Shenghua Gao. Single-image piece-wise planar 3d reconstruction via associative embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1029–1037, 2019. 1

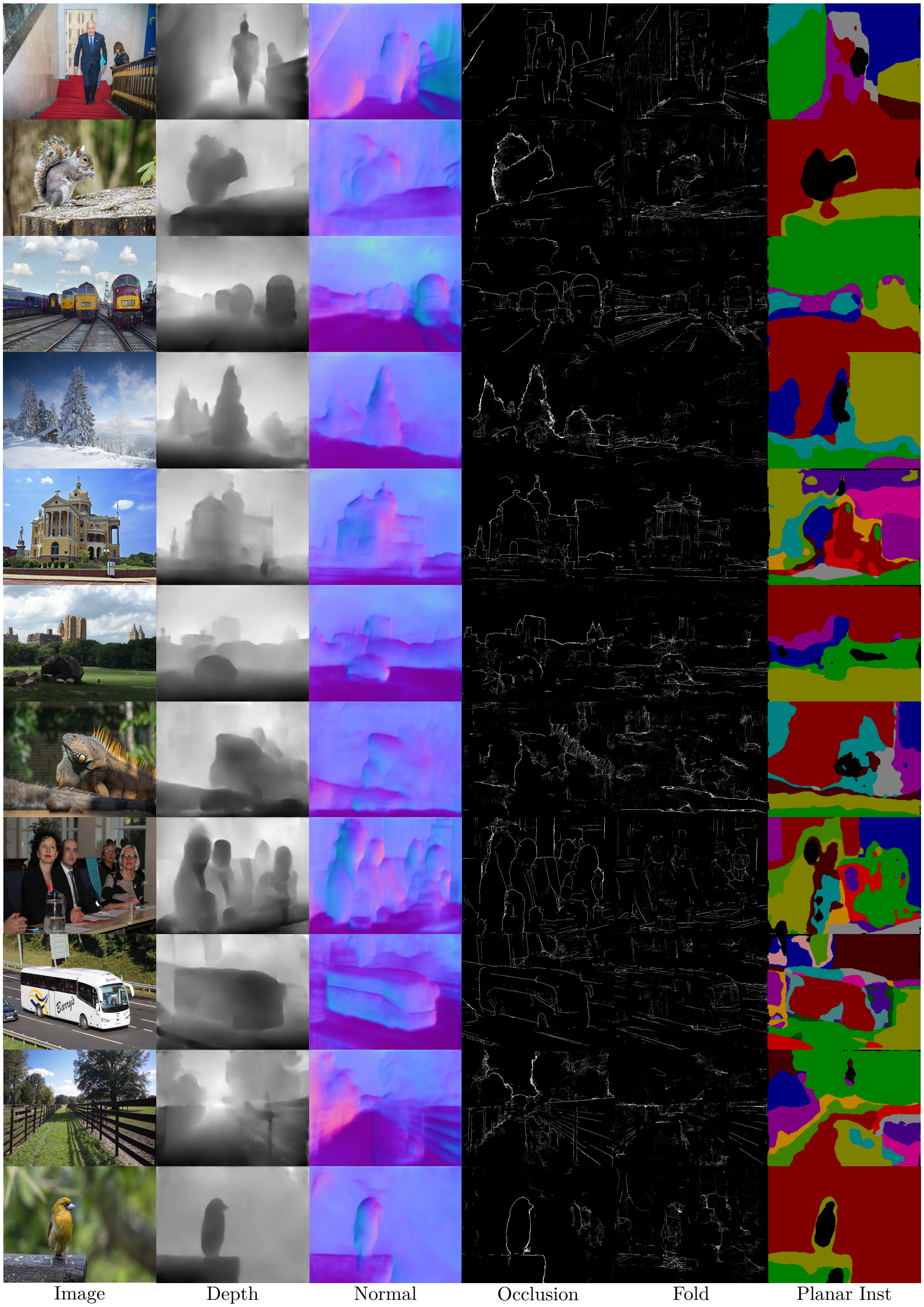

Figure III. Additional qualitative outputs from four tasks: (1) depth estimation, (2) normal estimation, (3) fold and occlusion boundary detection, and (4) planar instance segmentation.