Supplementary Material
Structure-Guided Ranking Loss for Single Image Depth Prediction

1. Sky mask generation for RGB-D data postprocessing

As the sky region often has erroneous disparity predictions from the optical flow algorithm, we use a sky segmentation model to detect sky regions and set the disparity to the minimum disparity value (farthest away) on a map.

We use two models: 1) a scene parsing model from [16] and 2) a sky segmentation model. The scene parsing model can more robustly detect sky regions if there is any, and the sky segmentation can provide accurate sky region segmentation. The sky segmentation model adopts a densenet [4] backbone and a two-branch structure described in [15]. The model is trained on sky regions from COCO-stuff plus an internally collected dataset of 2K high-res sky images.

For a given image to be post-processed, we run the scene parsing model and the sky segmentation model to get two sky region masks. If their IOU is above a threshold of 0.75, we will use the output from the sky segmentation model as the final sky mask of the image. Otherwise, we choose the output from the scene parsing model. Example sky segmentation results are shown in Fig. 1.

Figure 1. Examples of our sky masks. The sky regions are shown by green masks.

2. Qualitative results of monodepth models

As shown in Fig. 2, Fig. 3, and Fig.4, we note that our method has the most accurate depth discontinuities when compared to related work.

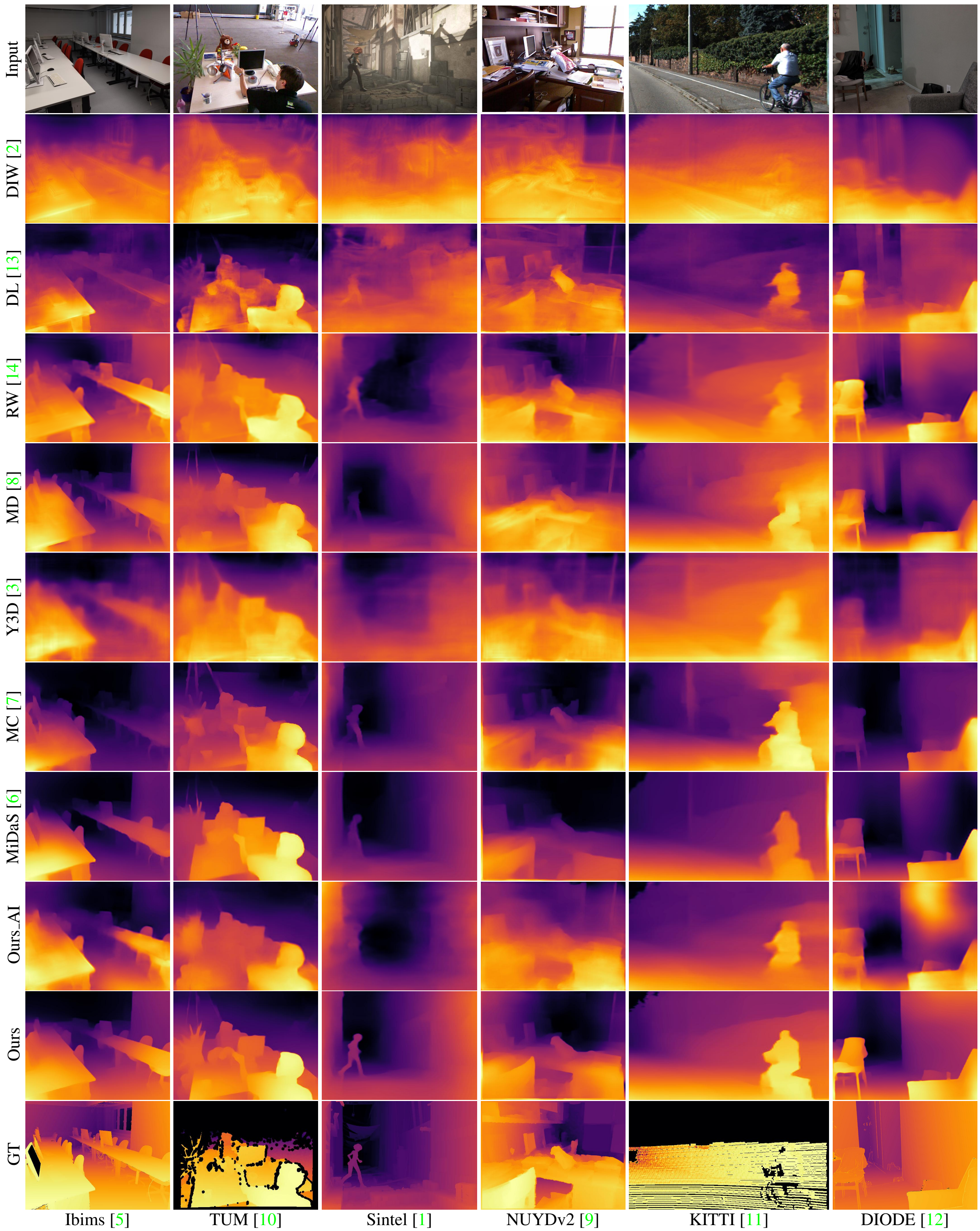

Figure 2. Additional qualitative results of single image depth prediction methods applied to different datasets.

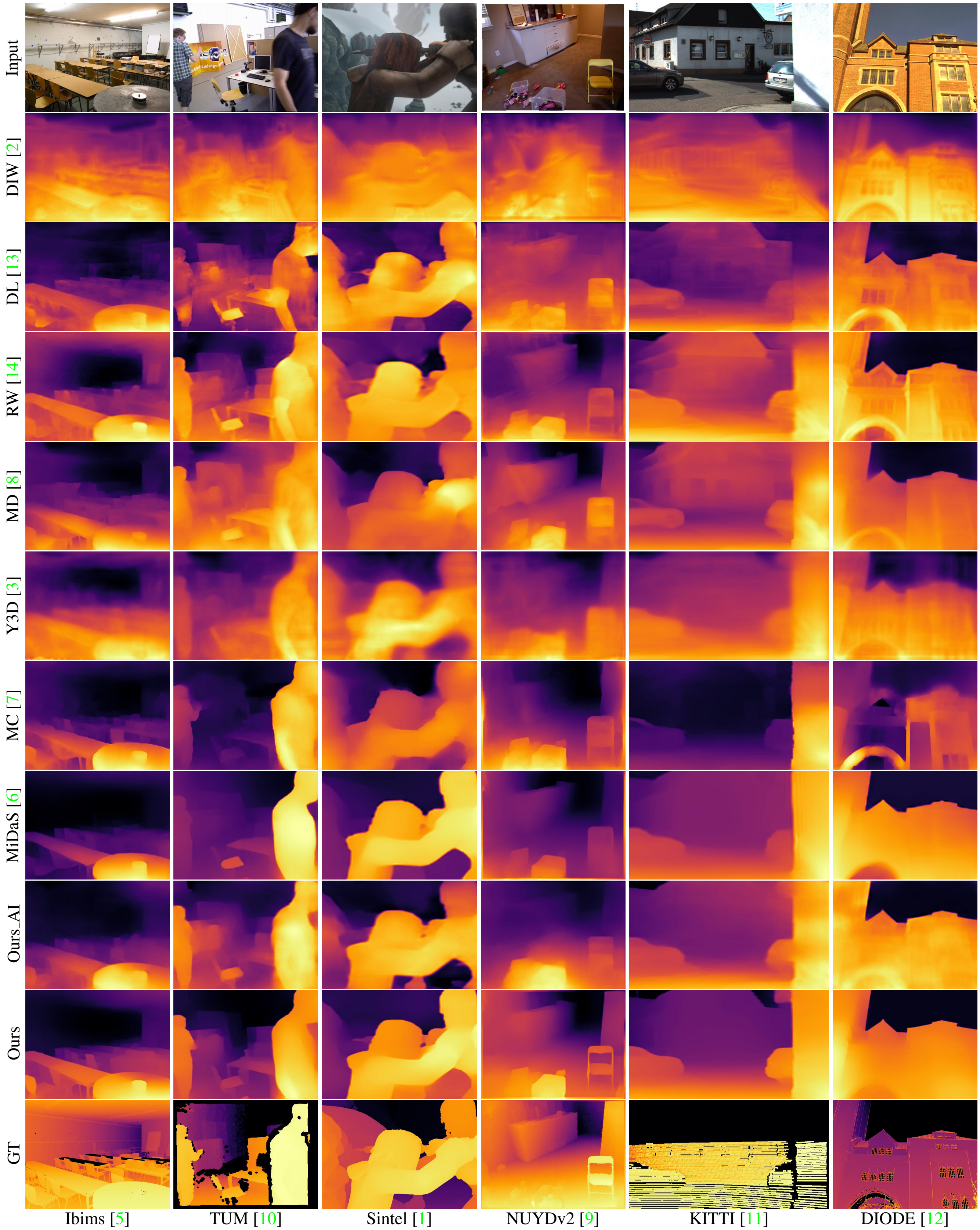

Figure 3. Additional qualitative results of single image depth prediction methods applied to different datasets.

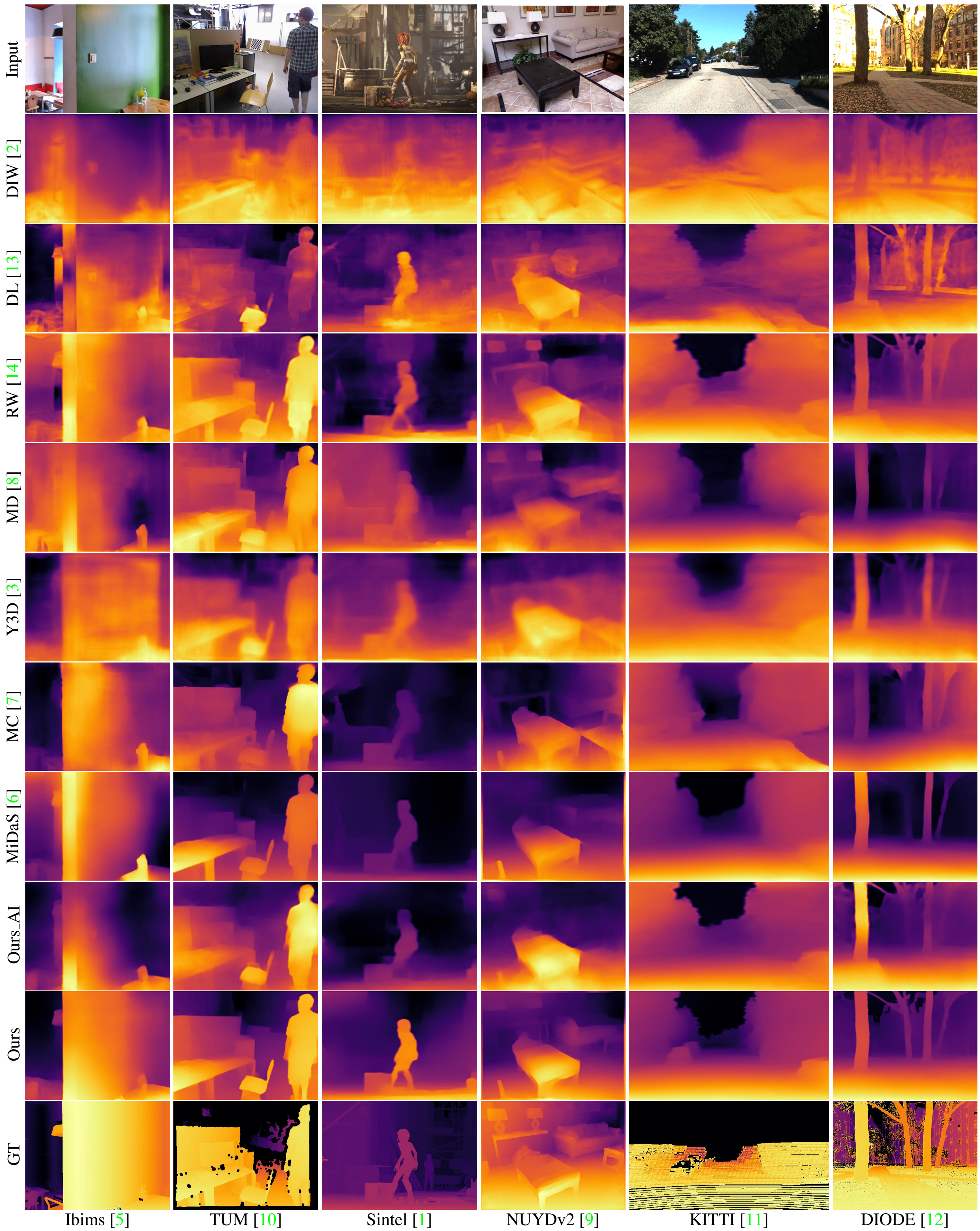

Figure 4. Additional qualitative results of single image depth prediction methods applied to different datasets.

3. Quantitative comparisons of monodepth models

In addition to ordinal error (see Table.1 in our main paper), we also report metric depth error scores (*i.e.*, rel and $\delta > 1.25$) in Table 1. Although we use less training data, our full model still achieves competitive results under these metrics. To demonstrate the effectiveness of our loss, we also report the scores of a baseline model with the affine-invariant loss [6]. One can observe that the model trained with our loss performs better under this setting.

Methods	Training Datasets	Ibims		TUM		Sintel		NYUDv2		KITTI		DIODE		Avg. Ranking
		$\delta > 1.25$	rel	$\delta > 1.25$	rel	$\delta > 1.25$	rel	$\delta > 1.25$	rel	$\delta > 1.25$	rel	$\delta > 1.25$	rel	
DIW [2]	DIW	39.30	0.232	37.42	0.270	56.21	0.405	36.85	0.210	51.45	0.306	42.25	0.307	10.00
DL [13]	ID	34.75	0.211	25.26	0.205	48.20	0.407	32.71	0.196	45.32	0.271	40.04	0.311	8.50
RW [14]	RW	30.46	0.220	25.16	0.200	45.46	0.410	28.86	0.178	31.32	0.207	38.27	0.320	6.92
MD [8]	MD	31.31	0.200	26.86	0.226	53.56	0.422	29.69	0.182	36.32	0.238	39.03	0.323	8.83
YT3D [3]	RW+DIW+YT3D	26.02	0.174	26.36	0.230	47.50	**0.329**	23.13	0.153	30.20	0.185	36.48	**0.279**	5.25
MC [7]	MC	21.53	**0.152**	26.06	0.204	44.85	0.476	23.70	0.159	48.02	0.280	39.29	0.337	6.58
MiDaS [6]	RW+MD+MV	**21.51**	0.153	20.44	0.201	**39.73**	0.341	**21.38**	**0.148**	26.84	**0.175**	35.12	0.296	**1.92**
Ours_AI	HRWSI	27.14	0.180	22.12	0.195	46.91	0.396	25.41	0.163	29.86	0.192	36.49	0.293	4.33
Ours†	RW	29.16	0.199	23.58	0.209	44.46	0.414	27.90	0.174	34.69	0.220	37.96	0.316	6.50
Ours_R	HRWSI	25.46	0.192	21.24	0.197	47.93	0.450	25.71	0.165	28.45	0.192	36.40	0.341	5.58
Ours	HRWSI	23.09	0.170	**19.41**	**0.194**	44.84	0.402	23.50	0.157	**25.40**	0.179	**34.44**	0.301	2.42

Table 1. **Zero-shot cross-dataset evaluation.** We use $\delta > 1.25$ and rel as our additional metrics for model evaluation. The lowest error is boldfaced and the second lowest is underlined.

4. Qualitative results of different sampling strategies

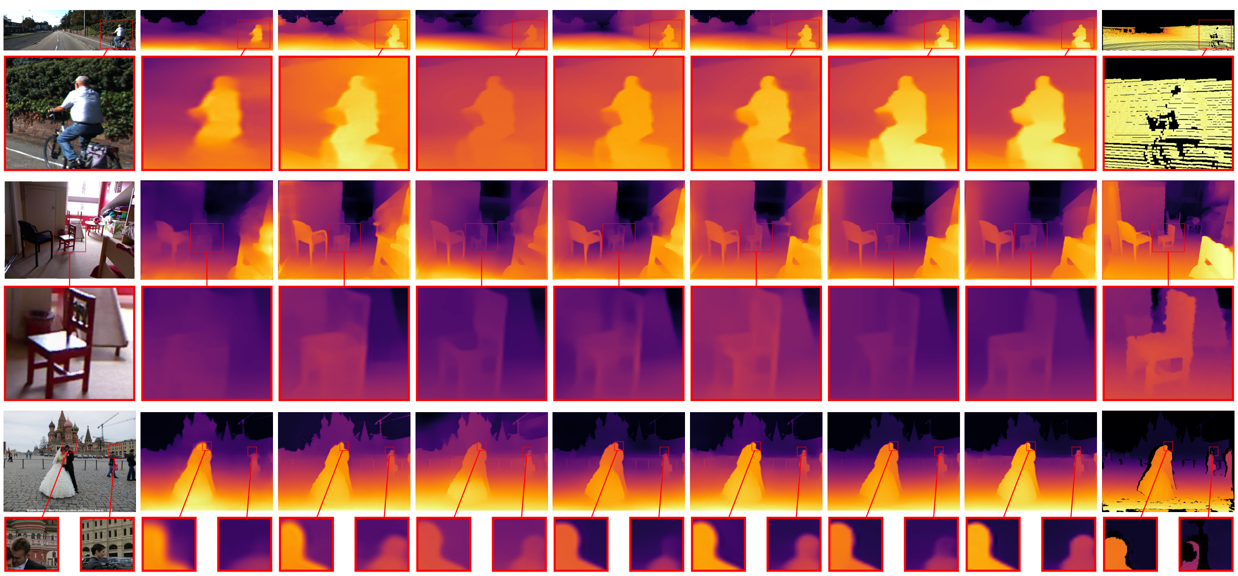

Figure 5. Additional qualitative evaluation of different sampling strategies and the affine-invariant loss. Best viewed zoomed in on-screen. Our full model trained with a combination of the structure-guide ranking loss and the multi-scale gradient matching loss generates a globally consistent depth map with sharp depth boundaries and detailed depth structures (*e.g.*, the basket, chair, and head).

References

[1] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 611–625, 2012. 2, 3, 4

[2] 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. 2, 3, 4, 5

[3] Weifeng Chen, Shengyi Qian, and Jia Deng. Learning single-image depth from videos using quality assessment networks. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 5604–5613, 2019. 2, 3, 4, 5

[4] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[5] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of cnn-based single-image depth estimation methods. In *Proc. European Conf. on Computer Vision Workshop (ECCV-WS)*, pages 331–348, 2018. 2, 3, 4

[6] Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv:1907.01341*, 2019. 2, 3, 4, 5

[7] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3, 4, 5

[8] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 4, 5

[9] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. European Conf. on Computer Vision (ECCV)*, 2012. 2, 3, 4

[10] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. Int. Conf. on Intelligent Robot Systems (IROS)*, 2012. 2, 3, 4

[11] Jonas Uhrig, Nick Schneider, Lucas Schneider, Uwe Franke, Thomas Brox, and Andereas Geiger. Sparsity invariant cnns. In *Proc. IEEE Int. Conf. on 3D Vision (3DV)*, 2017. 2, 3, 4

[12] 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: 1908.00463*, 2019. 2, 3, 4

[13] Lijun Wang, Xiaohui Shen, Jianming Zhang, Oliver Wang, Zhe Lin, Chih-Yao Hsieh, Sarah Kong, and Huchuan Lu. Deeplens: Shallow depth of field from a single image. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 37(6):6:1–6:11, 2018. 2, 3, 4, 5

[14] 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 *Proc. Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3, 4, 5

[15] Jianming Zhang. Generating image segmentation data using a multi-branch neural network, Apr. 18 2019. US Patent App. 15/784,918. 1

[16] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2017. 1