– Supplementary Material –
Monocular Depth Estimation via Listwise Ranking using the Plackett-Luce Model

1. Benchmark Dataset Characteristics

Table 1 shows the characteristics of the benchmark datasets, including their respective maximum depth values (capacity). We use these values to limit the recognized depth in the images as described in [9]. These values were also used to normalize the ground truth depths to get metric errors on a similar scale for all datasets.

Table 1. Properties of datasets being used within the zero-shot cross-dataset evaluation.

Dataset	# Test Instances	Sensor Type	Diversity	Depth Capacity (in m)
Ibims [5]	100	Laser	Indoor	50
Sintel [2]	1064	Synthetic	Animated	72
DIODE [12]	771	Laser	Indoor, outdoor	350
TUM [10]	1815	Motion Parallax	Indoor	10

Since the depth annotations of Sintel [2] are given in the inverse depth space, we did not apply the transformation $\frac{1}{D[l]+1}$ to calculate the nDCG scores as described in our paper (cf. Section 4.3). Instead, we directly used the already inverted depth values. Furthermore, in the case of TUM, we used the preprocessed dataset version as provided by the authors of [7]. Thereby, we considered the motion parallax (so-called "Plane-Plus-Parallax") depth map, which was constructed based on flow predictions between multiple views. We found these depth signals to provide a more precise and reliable source for calculating our error metrics compared to the originally provided Kinect sensor values.

2. Baseline Characteristics

Table 2 gives an overview of all baseline models considered within our empirical evaluation, together with the data being used for training. The diversity of the training data categorizes the individual datasets in terms of the variety of their captured scenes. MC represents a special case due to only incorporating images showing humans, but also indoor and outdoor.

Table 2. Baselines together with their training data considered within our empirical study.

Model	Loss Class	Training set	# Examples	Training Data Diversity
DenseDepth [1]	(Scale-invariant) Regression	NYU	50k	Low
MegaDepth [8]		MegaDepth	626k	High
BTS [6]		NYU	24k	Low
MC [7]		MC	136k	Medium
MiDaS [9]		HR-WSI	20k	High
MonoDepth2 [4]	Self-Sup.	KITTI	40k	Low
YouTube3D [3]	Relative	RW+DIW+YT3D	1219k	High
Xian 2020 [13]		HR-WSI	20k	High

3. Experimental Details

For the loss comparison (cf. Section 4.4.1), we compared our model on the ResNet-based architecture (PLDepthResNet) to the scale-invariant regression [9] and pairwise ranking [13] approach. Thereby, we optimized all models for 50 epochs with Adam and a batch size of 40 on four Nvidia Titan RTX. For the scale-invariant regression and our PL model, we used an initial learning rate of 0.001 multiplied by $\sqrt{0.1}$ after 25 epochs, while the pairwise ranking approach used an initial learning rate of 0.01 with the same learning rate schedule. In all three cases, input images were resized to 448×448 and data has been augmented by horizontally flipping with a 50% chance.

In the second experimental study, where we trained our model on the proposed EfficientNet-based architecture (PLDepthEffNet, cf. Fig. 3), we used a smaller initial learning rate of 0.0001 with the other parameters kepth the same. For each image, we sampled 100 rankings of size 5 per epoch. We further evaluated the scale-invariant regression variant on the same model architecture, where we used the same hyperparameters as for PLDepthEffNet, but with a higher learning rate of 0.001. The learning rate schedule was kept the same as for the previous experiments.

4. Additional Experiments on HR-WSI

As an additional experiment, we report the ordinal error, nDCG, RMSE and $\delta > 1.25$ as specified in the paper on the dataset HR-WSI [13]. Here, we compare only models being trained on this dataset, namely our PL model, MiDaS and Xian 2020. The presented results were computed on the separate validation set of 400 instances, which was used for model optimization of the mentioned approaches. Thereby, we consider the same trained models as used for the ordinal and metric error calculation in Section 4.4.2 and 4.4.3 of our paper.

Table 3 shows the averaged results for three runs. As can be seen, our model is superior with regard to most of the metrics. Only for the nDCG, the scale-invariant regression variant MiDaS turns out to be superior, although only slightly. Fig. 1 further shows exemplary predictions on HR-WSI.

Table 3. Results on the validation split of HR-WSI for the models being trained on the corresponding training split. The same experimental settings as for the model comparison apply here. ↓ refers to "lower is better", while ↑ denotes the opposite.

Model	Ord. Err. (↓)	nDCG (↑)	RMSE (↓)	$\delta > 1.25$ (↓)
MiDaS [9]	0.192	**0.839**	0.088	0.294
Xian 2020 [13]	0.166	0.838	0.154	0.558
PLDepthEffNet	**0.164**	0.837	**0.069**	**0.192**

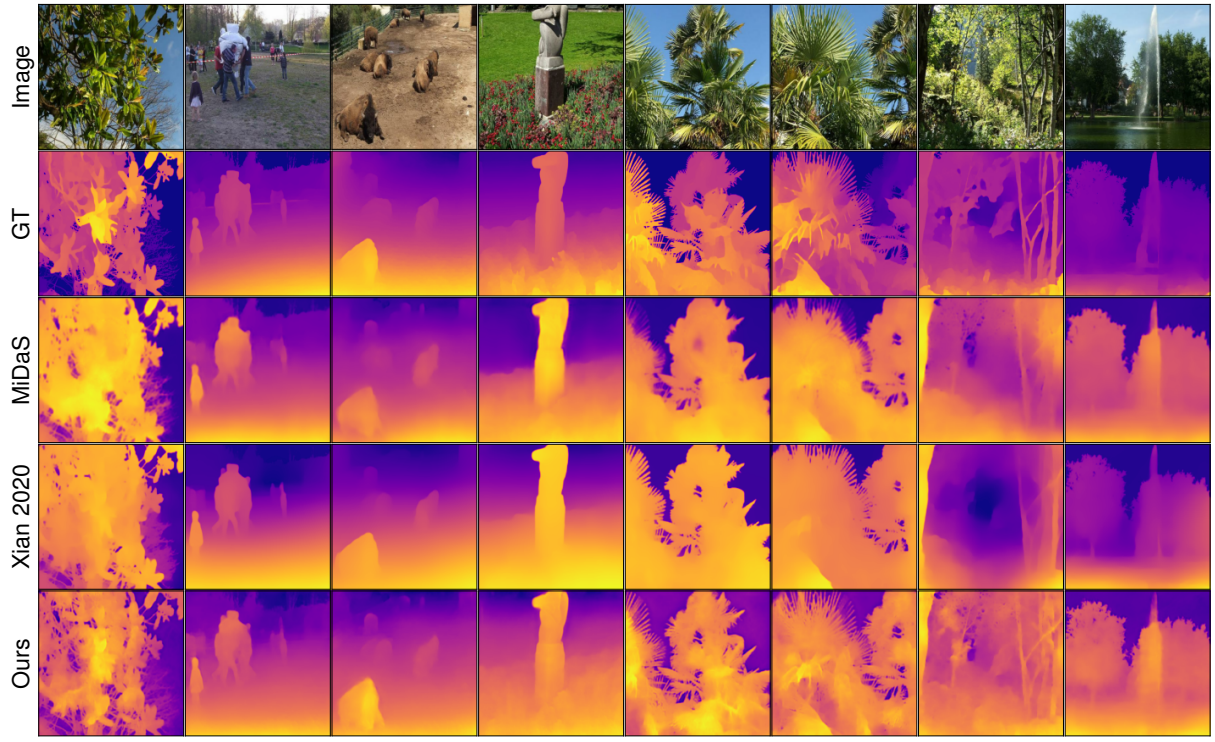

Figure 1. Exemplary predictions of the models trained on HR-WSI for validation set samples.

5. Additional Model Predictions

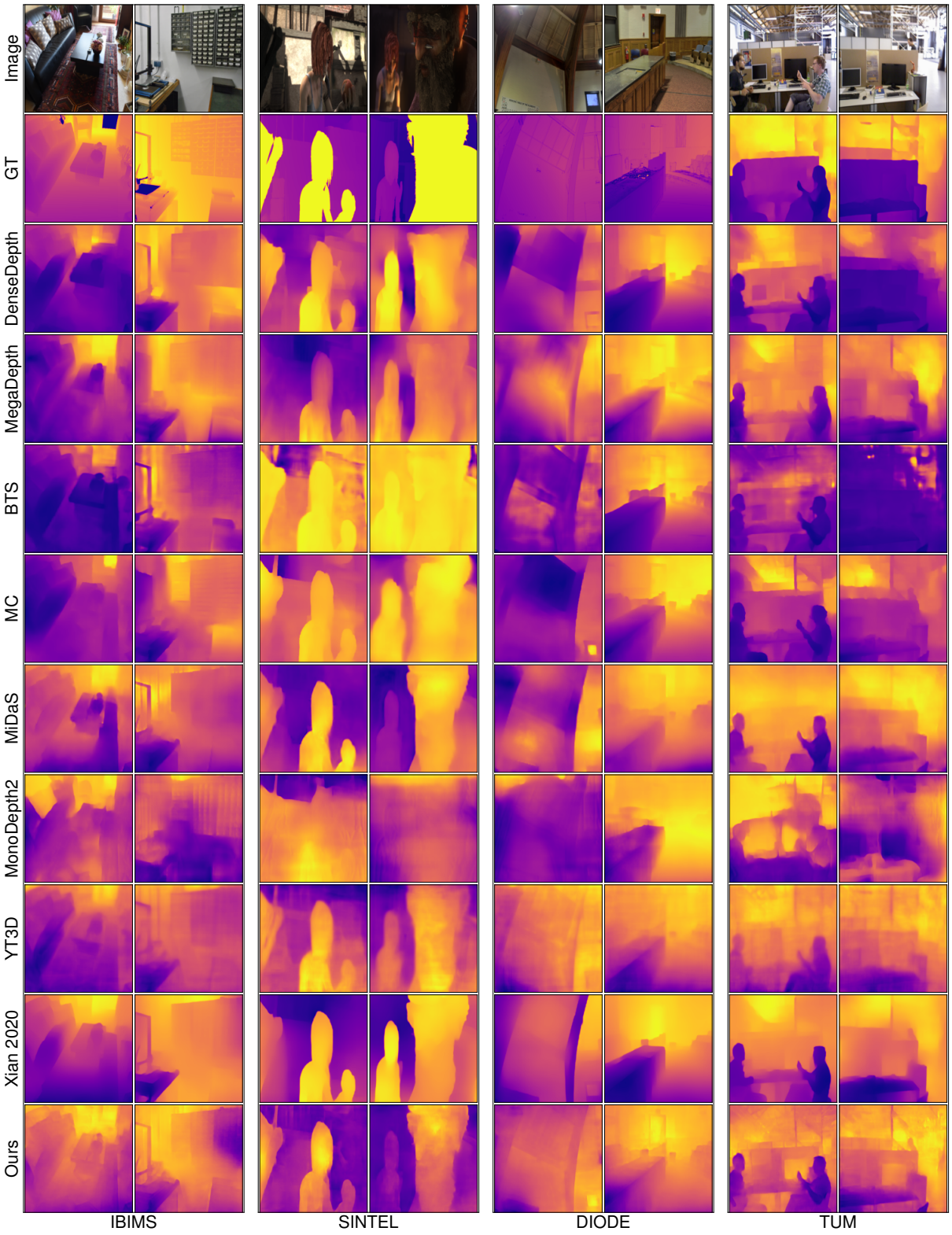

Figure 2. Model predictions for samples of the benchmark datasets (rescaled and shifted acc. to the ground truth depth values).

6. PLDepthEffNet Model Architecture

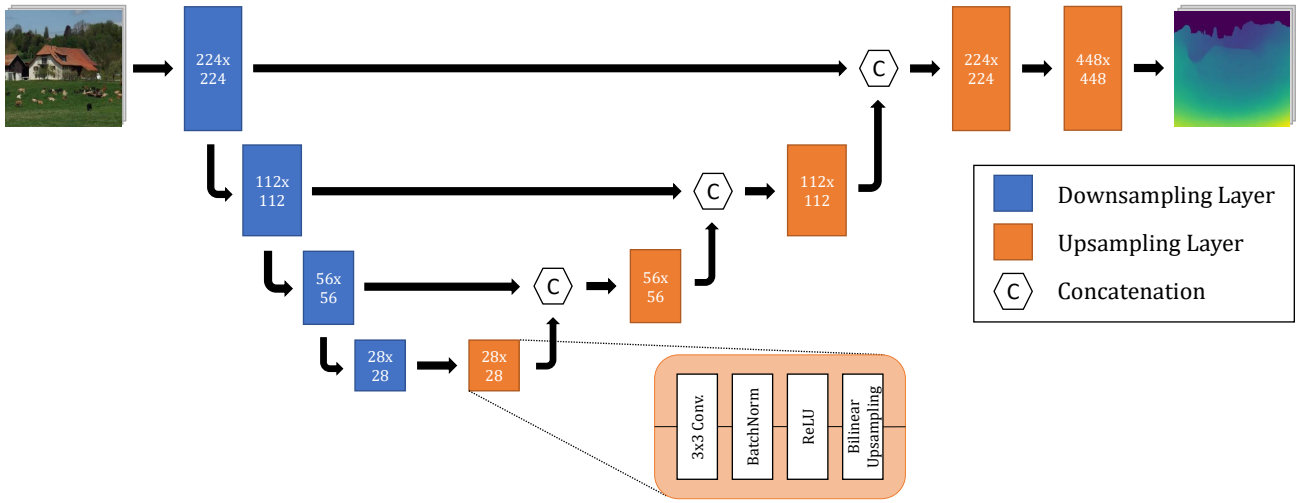

Figure 3. PLDepthEffNet U-net model architecture as proposed in the paper. The blue downsampling layers are specified by the used EfficientNet [11] backbone. The layer captions specify the corresponding output dimensionality of the respective layers.

References

[1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *CoRR*, abs/1812.11941, 2018.

[2] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Proceedings of the 12th European Conference on Computer Vision (ECCV), Part VI, October 7-13, 2012, Florence, Italy*, volume 7577 of *Lecture Notes in Computer Science*, pages 611–625. Springer, 2012.

[3] 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 (CVPR), June 16-20, 2019, Long Beach, CA, USA*, pages 5604–5613. Computer Vision Foundation / IEEE, 2019.

[4] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 27 - November 2, 2019, Seoul, Korea (South)*, pages 3827–3837. IEEE, 2019.

[5] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of CNN-based single-image depth estimation methods. In Laura Leal-Taixé and Stefan Roth, editors, *Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Part III, September 8-14, 2018, Munich, Germany*, volume 11131 of *Lecture Notes in Computer Science*, pages 331–348. Springer, 2018.

[6] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *CoRR*, abs/1907.10326, 2019.

[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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 16-20, 2019, Long Beach, CA, USA*, pages 4521–4530. Computer Vision Foundation / IEEE, 2019.

[8] 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 (CVPR), June 18-22, 2018, Salt Lake City, UT, USA*, pages 2041–2050. IEEE Computer Society, 2018.

[9] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[10] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 7-12, 2012, Vilamoura, Algarve, Portugal*, pages 573–580. IEEE, 2012.

[11] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning (ICML), 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019.

[12] Igor Vasiljevic, Nicholas I. 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. *CoRR*, abs/1908.00463, 2019.

[13] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA*, pages 608–617. IEEE, 2020.