Analyzing Results of Depth Estimation Models with Monocular Criteria

Jonas Theiner[1] Nils Nommensen[1] Jim Rhotert[1]
Matthias Springstein[1,2] Eric Müller-Budack[2] Ralph Ewerth[1,2]

[1]L3S Research Center, Leibniz University Hannover, Hannover, Germany
[2]TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
{theiner, ewerth}@l3s.de {eric.mueller, matthias.springstein}@tib.eu

Abstract

Monocular depth estimation is an essential but ill-posed (computer) vision task. While human visual perception of depth relies on several monocular depth clues, such as occlusion of objects, relative height, usual object size, linear perspective, deep learning models have to implicitly learn these cues from labeled training data to determine depth. In this paper, we investigate whether monocular depth criteria from human vision are violated for certain image instances given a model's predictions. We consider the task of depth estimation as a ranking problem, i.e., for a given pair of points, we estimate which point is nearer to the camera. In particular, we model four monocular depth criteria to automatically predict a subset of point pairs and infer their depth relation. Our experiments show that the implemented depth criteria achieve comparable performance to deep learning models. This allows the investigation of models with regard to the plausibility of predictions by finding image instances where the prediction is incorrect according to modeled human visual perception.

1. Introduction

Monocular depth estimation (MDE) is an essential computer vision task with many applications, such as augmented reality, robotics, or autonomous driving. Due to the natural lack of reliable stereoscopic visual relationships for monocular images, it is an ill-posed problem to regress depth in 3D space [13]. While human visual perception of depth relies on several monocular depth clues [5], such as occlusion of objects, relative height, usual object size, and linear perspective, deep learning models have to implicitly learn these cues from labeled training data to determine depth maps. The incorporation of additional constraints [15], e.g., from auxiliary tasks (e.g., semantic segmentation [22, 24]), regularizing constraints (e.g., occlusion [12], linear perspective [13]), and cross-dataset training [18], guide models to learn more robust representations. For specific image regions, e.g., from two visible objects, a human can judge whether one point is closer to the camera than the other. Motivated by this, some approaches learn a pairwise rank-based depth between selected point pairs sampled from ground-truth depth maps during training [10, 24, 25]. Other approaches incorporate metrics and loss functions to assess the reconstructed 3D scene [12, 21] via point cloud reconstruction [14] or depth discontinuities [9], and tackle uncertainty estimation [6] in addition to the traditional image-based metrics (like pixel-wise mean absolute error) [21]. However, whether monocular depth features are implicitly learned and whether the model's predicted depth maps are plausible remains unclear.

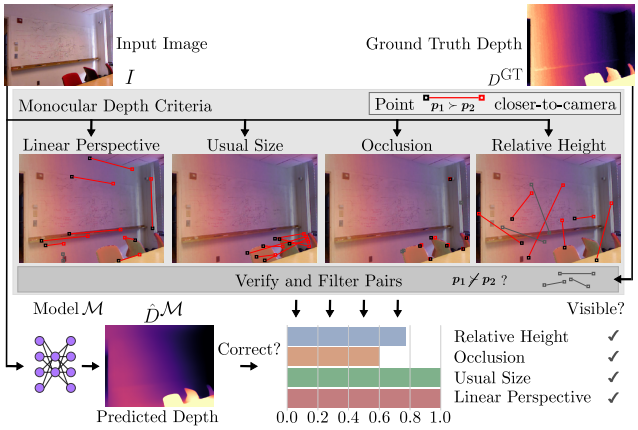

Figure 1. Consider the task of monocular depth estimation and a predicted depth map from a model. We automatically examine whether strong monocular depth cues are satisfied (if visible). For this purpose, a selection of point pairs with associated rank-based depth (*closer to camera*) is evaluated for each criterion.

This paper investigates whether monocular depth criteria from human vision are violated for certain image instances

given a model's prediction. To gain insights into this problem, we introduce *model sanity checks* [7, 8] for MDE. For example, humans often know whether a chair is in front or behind another through usual size or occlusion (as in Fig. 1). Consequently, we implement monocular depth criteria to find point pairs that rank the relative depth in relation to the camera. Point candidates are sampled based on criteria-based preconditions. Since these pairs should present strong depth cues according to human visual perception, we assume that an MDE model is supposed to be correct for these pairs, i.e., the relative ranking should be preserved. Experimental results show that the proposed approach is able to find image instances where this assumption is violated. This indicates that combining deep learning models with monocular depth criteria may lead to further improvements (e.g., plausible predictions and more comprehensible results).

We describe the approach to perform this model-sanity check for MDE in Sec. 2 and provide details of the implemented depth criteria in Sec. 3. Experimental evaluation is reported in Sec. 4. Sec. 5 concludes the paper and outlines future research directions.

2. Model Sanity Check using Depth Criteria

As humans, we often know exactly whether a chair is in front or behind another, through *usual size* or *occlusion* (as in Fig. 1). In general, there are strong monocular depth features for humans that provide information about relative depth [5]. Inspired by Kang et al. [7, 8], we introduce a *model sanity check* for MDE. Does a depth estimation model that predicts depth maps violate certain relations and criteria? This section formally presents an approach to automatically detect potential violations.

Preliminaries: Consider an RGB image $\mathbf{I} \in \mathbb{R}^{h \times w \times 3}$ and depth information $\boldsymbol{D} \in \mathbb{R}^{h \times w}$ for each pixel, where h and w denote the image height and width. The task of an MDE model $\mathcal{M}$ is to predict the metric depth $\hat{\boldsymbol{D}}^{\mathcal{M}}$ given image $\mathbf{I}$ as input. Moreover, $\boldsymbol{D}[\boldsymbol{p}]$ denotes the depth information of a pixel or point $\boldsymbol{p} \in \{1, \dots, h\} \times \{1, \dots, w\}$ identified by a vertical and horizontal position. Without loss of generality, lower values $\boldsymbol{D}[\boldsymbol{p}]$ represent a lower distance to the camera.

2.1. Pairwise Rank-based Depth Information

Given a point pair $(\boldsymbol{p}_1, \boldsymbol{p}_2)$, we define its rank-based depth information such that $\boldsymbol{D}[\boldsymbol{p}_1] \succ \boldsymbol{D}[\boldsymbol{p}_2]$, i.e., $\boldsymbol{p}_1$ is "closer-to-camera" than $\boldsymbol{p}_2$ to evaluate the order relation of individual parts in an image $\mathbf{I}$. These rankings can be obtained from various sources, e.g., (incomplete) ground-truth (metric) depth maps, pseudo-depth maps, or even by humans, using pair sampling strategies including random [2], superpixel [25], and structure-guided sampling [10, 24]. Rank-based depth information can be sampled for a total number of $\texttt{nsamples} \ll \binom{h \times w}{2}$ point pairs per image.

2.2. Model Sanity Check using Depth Criteria

Given an image $\mathbf{I}$, we assume that the following inputs are provided: (1) a pseudo-depth maps for monocular depth criteria (DC) $\tilde{\boldsymbol{D}}^{\text{DC}}$, like *linear perspective* or *usual size*, (2) a predicted metric depth map $\hat{\boldsymbol{D}}^{\mathcal{M}}$ from an MDE model $\mathcal{M}$, and (3) a respective ground truth $\boldsymbol{D}^{\text{GT}}$.

1. We sample `nsamples` point pairs from each pseudo-depth map of the monocular depth criteria $\tilde{\boldsymbol{D}}^{\text{DC}}$ (if available), and order their depth ranking such that $\boldsymbol{D}[\boldsymbol{p}_1] \succ \boldsymbol{D}[\boldsymbol{p}_2]$. Since some depth criteria (Sec. 3) are only applicable to specific image areas, the objective is to select only those valid point pairs that can be described by the criterion.
2. To ensure that the ranking of the sampled points is correct, we verify the selected pairs derived from each pseudo depth map $\tilde{\boldsymbol{D}}^{\text{DC}}$ using the respective ground truth $\boldsymbol{D}^{\text{GT}}$. Hence, for all selected point pairs, such pairs are rejected. Please note that $\boldsymbol{D}^{\text{GT}}$ is only used to sample pairs with correct ranking for the model sanity check, the prediction of $\boldsymbol{D}^{\text{DC}}$ and $\hat{\boldsymbol{D}}^{\mathcal{M}}$ doe not use any ground-truth information.
3. For each verified pair $(\boldsymbol{p}_1, \boldsymbol{p}_2)$, we check whether model's $\mathcal{M}$ prediction is correct, i.e., $\hat{\boldsymbol{D}}^{\mathcal{M}}[\boldsymbol{p}_1] \succ \hat{\boldsymbol{D}}^{\mathcal{M}}[\boldsymbol{p}_2]$. For the set of verified pairs, the accuracy is computed as the fraction between the number of correct model predictions to the ground truth for each image and criterion.

In other words, the proposed procedure can be seen as an additional metric for MDE, which is more comprehensible for humans since it involves monocular depth criteria.

3. Modeling Monocular Depth Criteria

In this section, we present implementations for different monocular depth criteria. All criteria are designed such that for a selection of point pairs, it can indicate whether one point is closer to the camera than the other (see Sec. 2.1).

Relative Height: Sometimes referred as *elevation* criterion, it is assumed that points at the top of an image are typically further away from the camera than points at the bottom, in particular for outdoor images. Of course, this (simple) criterion is not correct for all point pairs, e.g., if a point is above the horizon, but can be considered as a weak feature [3], i.e., often much better than random guessing. The relative ranking for a point pair is computed according to their vertical positions $\boldsymbol{D}[\boldsymbol{p}_1] \succ \boldsymbol{D}[\boldsymbol{p}_2]$ if $\boldsymbol{p}_1[0] < \boldsymbol{p}_2[0]$.

Linear Perspective: One possibility to model linear perspective is to detect vanishing lines and corresponding vanishing points in an image. Since some lines meet in the

same vanishing point, all points on a line towards the vanishing point increase in depth. Instead, 2D planes in the image can be detected, which can be viewed as a generalization of the vanishing point problem (all lines on one plane meet in the same vanishing point). We apply the *PlaneRCNN* model [11] to detect planes, instance masks, and corresponding normal vectors. Given two pixels $(\boldsymbol{p}_1, \boldsymbol{p}_2)$ on a detected plane with the predicted normal vector $\boldsymbol{n}$, a 3D point $\boldsymbol{x}$ can be described by $\boldsymbol{n}\boldsymbol{x} - d = 0$, where d is the (unknown) distance of the plane to the origin. Given that any 3D point is projected to a 2D point according to $\mathbf{p} = \boldsymbol{K}\boldsymbol{x} = (\hat{x}, \hat{y}, 1)^T$ where $\boldsymbol{K}$ is the intrinsic camera matrix, we can rewrite this equation, s.t.:

$$\mathbf{n}\boldsymbol{K}^{-1}\mathbf{p} - d = \begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix} \begin{pmatrix} \hat{x}z \\ \hat{y}z \\ z \end{pmatrix} - d = 0 \tag{1}$$

Resolving Eq. (1) for z obtains $z = d/(n_1\hat{x} + n_2\hat{y} + n_3)$. Finally, given two randomly selected points $\boldsymbol{p}_1, \boldsymbol{p}_2$ on the same plane, we know $\boldsymbol{D}[\boldsymbol{p}_1] \succ \boldsymbol{D}[\boldsymbol{p}_2]$ if $z_1 < z_2 \iff n_1\hat{x}_2 + n_2\hat{y}_2 < n_1\hat{x}_1 + n_2\hat{y}_1$.

Occlusion: Recognizing clear boundaries of two objects and being able to tell that one object is occluding the other gives a strong sense of the depth ordering of the scene. The *P2ORM* occlusion model [16] predicts occlusion masks, orientation angles [16], and confidence scores for each pixel $\boldsymbol{p}$. We randomly choose $k = 400$ points using the top-k confidence scores to provide enough candidates $k >$ `nsamples` for sampling (Sec. 2.1). Given the rotation angle, the second point is calculated to be directly orthogonal on both sides of the occlusion boundary, and hence, in combination with the occlusion mask, it can be inferred whether $\boldsymbol{p}_1$ is closer to the camera than $\boldsymbol{p}_2$.

Usual Size: Humans have learned the usual size of objects in the world and apply the perceived size of those objects to infer the actual distance to that object. This knowledge can be modeled by determining instance segmentation masks for all objects in an image and by collecting prior statistics about the 3D size of each object instance. We apply *DETR* [1] to automatically infer instance segmentation masks. For each object class, we sample all instance masks for a set of images with *known* depth maps. We then calculate each object instance's average depth $\hat{z}$ from the respective ground-truth depth values. We then define the size in 3D as $s \approx \mathbf{g}(z) \approx \frac{z^2}{f^2}a$ from the 2D object size a (sum of all pixels) and focal length f, which can be interpreted as the area of the 2D region of the image plane projected into the scene by a factor of z. Having collected M measurements for an object class, we then compute the approximate probability density function $\rho_s(s)$ [19]. Next, given $\rho_s(s)$ of an object, a measurement of the 2D size a and the focal length f, we calculate the probability of z using the change of variable method:

$$\rho_z(z) = \rho_s(\mathbf{g}(z))\mathbf{g}'(z) = \rho_s(\mathbf{g}(z))2z\frac{a}{f^2} \tag{2}$$

We aim to compute the probability that *object 1* is closer to the camera than *object 2*. Therefore, we define a random variable $r = \frac{z_1}{z_2}$ where r follows the ratio distribution

$$\rho_r(r) = \int_{z_2=z_{\min}}^{z_2=z_{\max}} |z_2|\, \rho_{z_1}(rz_2)\, \rho_{z_2}(z_2)\, dz_2 \tag{3}$$

and thus, for two points $\boldsymbol{p}_1 \in z_1$, and $\boldsymbol{p}_2 \in z_2$, hold $\boldsymbol{D}[\boldsymbol{p}_1] \succ \boldsymbol{D}[\boldsymbol{p}_2]$ if:

$$\int_{r=0}^{r=1} \int_{z_2=z_{\min}}^{z_2=z_{\max}} |z_2|\, \rho_{z_1}(rz_2)\, \rho_{z_2}(z_2)\, dz_2 dr < 0.5$$

Please note that the actual focal length is only required to compute the actual object size and has no effect on the relative depth ranking.

4. Experiments

In this section, we experimentally verify the proposed approach of modeling sanity checks for MDE (Sec. 2) and evaluate the implemented monocular depth criteria (Sec. 3). The experimental setup is described in Sec. 4.1 whereas experimental results are presented in Sec. 4.2.

4.1. Experimental Setup

We apply several pre-trained state-of-the-art models[1] including Dense Prediction Transformers (DPT) [17] with cross-dataset training (*MiDaS*) [18] and different backbones (*SwinFormer, BEiT, ResNet*), namely `DPT_Large`, `midas_DPT_SwinV2_L_384`, `midas_DPT_BEiT_L_512`. Due to the multi-dataset training, we assume that general depth cues, i.e., monocular depth criteria, were learned, which is less likely with training on individual benchmark datasets [18]. Evaluation is performed on the *NYUv2* [20], *KITTI* [4], and *HRWSI* [24] dataset. A subset of 1000 randomly selected images is taken for evaluation from the respective training split for each dataset.

4.2. Experimental Results

Do models violate monocular depth criteria? For each image, we check whether the selected pairs, for which depth criteria should be valid, the model's prediction is also correct according to the procedure in Sec. 2.2. Results are presented in Fig. 2 for one model $\mathcal{M}$ since similar conclusions can be drawn from the other two tested models.

[1] https://github.com/isl-org/MiDaS

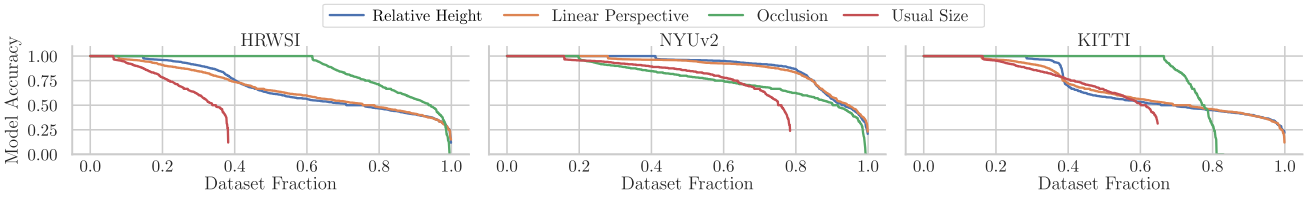

Figure 2. Performance of model $\mathcal{M}$ (`midas_DPT_BEiT_L_512`) on individual images with respect to the compliance of monocular depth criteria: For each image, point pairs per criteria are sampled (`nsamples` $= 32$) if the criterion is present and averaged over these pairs (accuracy). The graph shows the sorted accuracy of the model $\mathcal{M}$ over different fractions of a dataset containing 1000 images.

The accuracy per criterion and image is independently ordered per dataset. The performance drops of model $\mathcal{M}$ for individual depth criteria indicate its violation for certain image instances. Additionally, some variance across different datasets can be observed. Consequently, individual images can be inferred where the model is wrong according to a single monocular depth criterion.

Quality of implemented depth criteria: To assess the performance of all implemented depth criteria (Sec. 3), a comparison to the respective ground-truth ranking is conducted for selected point pairs per criteria and image. Fig. 3 shows how often the prediction of each criterion is correct compared to three MDE models on the same selected pairs.

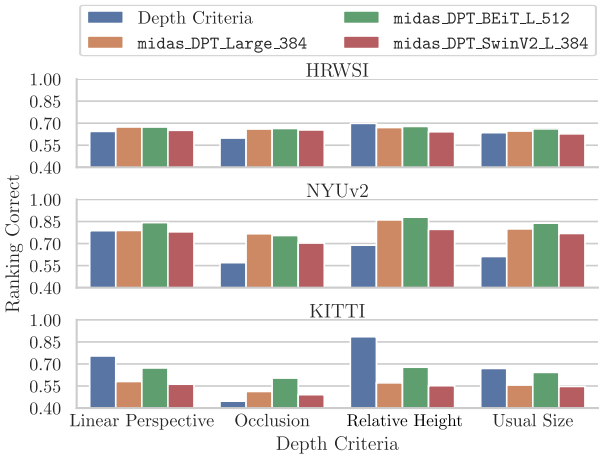

Figure 3. Pairwise relative depth ranking accuracy for all criteria compared to different models on three dataset. The same pairs per image (`nsamples` $= 32$) are selected for all models.

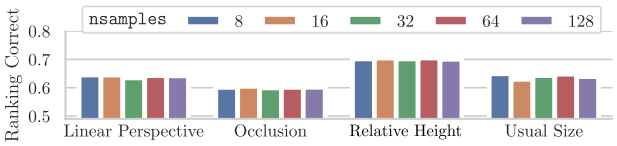

Figure 4. Pairwise relative depth ranking accuracy while varying the number of selected pairs per image (`nsamples`) on *HRWSI*.

The overall quality is similar to or slightly lower than the tested depth estimation models. All selected pairs for depth criteria are better than random depth ranking (correct $= 0.5$) with the exception of *occlusion* on *KITTI*. *Relative height* is a strong depth cue, in particular for outdoor images.

Does the number of sampled pixel pairs matter? While varying the number of selected pairs per criteria, the performance is almost unaffected by the number of point pairs selected according to Fig. 4. This indicates that there are always enough pairs to sample if the monocular depth criterion is visible for a given image.

5. Conclusion

In this paper, we have investigated to what extent we can analyze the plausibility of depth estimation results by modeling and exploiting (human) monocular depth criteria. Similar to related work, we consider depth prediction as a ranking task of two or more points in the scene. Particularly, we have modeled monocular depth criteria, namely *usual size, linear perspective, occlusion*, and *relative height* (most of them are only applicable to specific parts of an image). Their implementation allows to sample a subset of point pairs and to infer their depth relation. Experimental results indicate that the proposed approach is able to find image instances where these criteria are violated. This knowledge can help to debug and improve these models [8]. For example, during training of ranking models, the incorrect ranking of point pairs could be weighted where depth criteria are assumed to be correct, or this knowledge may help for user-guided depth estimation [23]. One drawback of the sampling based on the implemented depth criteria is that some sampled point pairs may not reflect the human perception of the underlying depth criteria. However, an evaluation of this issue requires human annotations, which we leave open as future work along with a more in-depth analysis (e.g., dataset bias) of the proposed sanity check.

Acknowledgment This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 420493178.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision, ECCV*, pages 213–229. Springer, 2020. 3

[2] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Advances in Neural Information Processing Systems, NIPS*, pages 730–738, 2016. 2

[3] Ralph Ewerth, Matthias Springstein, Eric Müller, Alexander Balz, Jan Gehlhaar, Tolga Naziyok, Krzysztof Dembczynski, and Eyke Hüllermeier. Estimating relative depth in single images via rankboost. In *IEEE International Conference on Multimedia and Expo, ICME*, pages 919–924. IEEE Computer Society, 2017. 2

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *European Conference on Computer Vision, ECCV*, pages 3354–3361. IEEE Computer Society, 2012. 3

[5] E. Bruce Goldstein. *Sensation and perception*. Cengage Learning, 2009. 1, 2

[6] Julia Hornauer and Vasileios Belagiannis. Gradient-based uncertainty for monocular depth estimation. In *European Conference on Computer Vision, ECCV*, pages 613–630. Springer, 2022. 1

[7] Daniel Kang, Deepti Raghavan, Peter Bailis, and Matei Zaharia. Model assertions for debugging machine learning. In *NeurIPS MLSys Workshop*, page 10, 2018. 2

[8] Daniel Kang, Deepti Raghavan, Peter Bailis, and Matei Zaharia. Model assertions for monitoring and improving ML models. In *Machine Learning and Systems, MLSys*. mlsys.org, 2020. 2, 4

[9] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of CNN-based single-image depth estimation methods. In *ECCV Workshops*, pages 331–348. Springer, 2018. 1

[10] Julian Lienen, Eyke Hüllermeier, Ralph Ewerth, and Nils Nommensen. Monocular depth estimation via listwise ranking using the Plackett-Luce model. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 14595–14604. Computer Vision Foundation / IEEE, 2021. 1, 2

[11] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. PlaneRCNN: 3d plane detection and reconstruction from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 4450–4459. Computer Vision Foundation / IEEE, 2019. 3

[12] Jaime Spencer Martin, Chris Russell, Simon Hadfield, and Richard Bowden. Deconstructing self-supervised monocular reconstruction: The design decisions that matter. *arXiv preprint*, abs/2208.01489, 2022. 1

[13] Yue Ming, Xuyang Meng, Chunxiao Fan, and Hui Yu. Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438:14–33, 2021. 1

[14] Evin Pinar Örnek, Shristi Mudgal, Johanna Wald, Yida Wang, Nassir Navab, and Federico Tombari. From 2d to 3d: Re-thinking benchmarking of monocular depth prediction. *arXiv preprint*, abs/2203.08122, 2022. 1

[15] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3Depth: Monocular depth estimation with a piecewise planarity prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1600–1611. IEEE, 2022. 1

[16] Xuchong Qiu, Yang Xiao, Chaohui Wang, and Renaud Marlet. Pixel-pair occlusion relationship map (P2ORM): Formulation, inference and application. In *European Conference on Computer Vision, ECCV*, pages 690–708. Springer, 2020. 3

[17] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 12159–12168. IEEE, 2021. 3

[18] 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 Trans. Pattern Anal. Mach. Intell.*, 44(3):1623–1637, 2022. 1, 3

[19] David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley, 1992. 3

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

[21] Jaime Spencer, C. Stella Qian, Chris Russell, Simon Hadfield, Erich W. Graf, Wendy J. Adams, Andrew J. Schofield, James H. Elder, Richard Bowden, Heng Cong, Stefano Mattoccia, Matteo Poggi, Zeeshan Khan Suri, Yang Tang, Fabio Tosi, Hao Wang, Youmin Zhang, Yusheng Zhang, and Chaoqiang Zhao. The monocular depth estimation challenge. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV*, pages 623–632. IEEE, 2023. 1

[22] Lijun Wang, Jianming Zhang, Oliver Wang, Zhe Lin, and Huchuan Lu. SDC-Depth: Semantic divide-and-conquer network for monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 538–547. Computer Vision Foundation / IEEE, 2020. 1

[23] Zhihao Xia, Patrick Sullivan, and Ayan Chakrabarti. Generating and exploiting probabilistic monocular depth estimates. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 62–71. Computer Vision Foundation / IEEE, 2020. 4

[24] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 608–617. Computer Vision Foundation / IEEE, 2020. 1, 2, 3

[25] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T. Freeman. Learning ordinal relationships for mid-level vision. In *IEEE International Conference on Computer Vision, ICCV*, pages 388–396. IEEE Computer Society, 2015. 1, 2