
RSA: Resolving Scale Ambiguities in Monocular Depth Estimators through Language Descriptions

SUPPLEMENTARY MATERIALS

Ziyao Zeng¹ Yangchao Wu² Hyoungeob Park¹ Daniel Wang¹ Fengyu Yang¹
Stefano Soatto² Dong Lao² Byung-Woo Hong³ Alex Wong¹

¹Yale University ²University of California, Los Angeles ³Chung-Ang University

¹{ziyao.zeng, hyoungeob.park, daniel.wang.dhw33}@yale.edu

¹{fengyu.yang, alex.wong}@yale.edu

²wuyangchao1997@g.ucla.edu ²{soatto,lao}@cs.ucla.edu ³hong@cau.ac.kr

1 Evaluation Metrics

Following [5, 20, 45], we evaluate RSA and baseline methods quantitatively using mean absolute relative error (Abs Rel), root mean square error (RMSE), absolute error in log space (\log_{10}), logarithmic root mean square error (RMSE_{\log}) and threshold accuracy (δ_i). The evaluation metrics are summarized in Table 1.

Metric	Formulation
Abs Rel	$\frac{1}{N} \sum_{(i,j) \in \Omega_v} \frac{ y^*(i,j) - y(i,j) }{y^*(i,j)}$
RMSE	$\sqrt{\frac{1}{N} \sum_{(i,j) \in \Omega_v} (y^*(i,j) - y(i,j))^2}$
\log_{10}	$\frac{1}{N} \sum_{(i,j) \in \Omega_v} \log_{10}(y^*(i,j)) - \log_{10}(y(i,j)) $
RMSE_{\log}	$\sqrt{\frac{1}{N} \sum_{(i,j) \in \Omega_v} (\ln(y^*(i,j)) - \ln(y(i,j)))^2}$
δ	% of $y(i,j)$ s.t. $\max(\frac{y(i,j)}{y^*(i,j)}, \frac{y^*(i,j)}{y(i,j)}) < thr \in [1.25, 1.25^2, 1.25^3]$

Table 1: **Evaluation metrics.** y denotes predictions, y^* denotes ground truth, N denotes the valid number of pixels, Ω_v denotes the image space where the ground truth is valid, and (i, j) denotes pixel coordinate,

2 Datasets

We conduct experiments on five datasets in total. Details of each dataset are provided below.

KITTI [12, 13] contains 61 driving scenes with research in autonomous driving and computer vision. It contains calibrated RGB images with synchronized point clouds from Velodyne lidar, inertial, GPS information, etc. We used Eigen split [7], consisting of 23,488 training images and 697 testing images. We follow the evaluation protocol of [8] for our experiments.

VOID [36] comprises indoor (laboratories, classrooms) and outdoor (gardens) scenes with synchronized 640×480 RGB images and with ground truth depth maps captured by active stereo. For the purpose of simulating sequences observed in visual inertial odometry (VIO), e.g., XIVO

[10]), VOID contains 56 sequences with challenging 6 DoF motion captured on rolling shutter. 48 sequences (about 45,000 frames) are assigned for training and 8 for testing (800 frames). We follow the evaluation protocol of [36] where output depth is evaluated against ground truth within the depth range between 0.2 and 5.0 meters.

NYUv2 [28] consists of 24,231 synchronized 640×480 RGB images and depth maps for indoors scenes (household, offices, commercial areas), captured with a Microsoft Kinect. The official split consists of 249 training and 215 test scenes. We use the official test set. Following [2, 43], we remove samples without valid ground truth, leaving 652 valid images for testing. We evaluate on methods on NYU for depth range between 1×10^{-3} to 10 meters.

SUN-RGBD [30] is an indoor dataset consisting of around 10K images with high scene diversity collected with four different sensors. We use this dataset only for zero-shot evaluation of baselines and our model on the official test set of 5050 images. Evaluation is done on depth values up to 10 meters. Note that we do not use SUN-RGBD for training.

DDAD [16] comprise of diverse dataset of urban, highway, and residential scenes curated from a global fleet of self-driving cars. It contains 17,050 training and 4,150 evaluation frames with ground-truth depth maps generated from dense LiDAR measurements using the Luminar-H2 sensor. We use this dataset only for zero-shot evaluation of baselines and our model, where evaluations are done on depth values up to 80 meters. Note that we do not use DDAD for training.

3 Prompts for Natural Text Generation

To generate natural, free-form text that does not follow fixed templates and more closely resembles human descriptions, we use two visual question answering models LLaVA v1.6 Vicuna and LLaVA v1.6 Mistral [21]. To produce diverse, natural captions, we use five different prompts per model. These five prompts are listed in Table 2.

Prompts
"Describe the image in one sentence."
"Provide a one-sentence description of the image, pay to attention object type."
"Capture the essence of the image in a single sentence, pay attention to object relationship."
"Condense the image description into one sentence, pay attention to object size."
"Express the image in just one sentence, pay attention to the overall layout."

Table 2: **Prompts for natural text generation.** We use LLaVA v1.6 Vicuna and LLaVA v1.6 Mistral with those 5 prompts to generate 10 sentences of different natural text that adhere to human input for each image.

4 Discussion

A single image does not afford metric depth estimation [2, 14, 15, 18, 31, 33, 38, 46]. While some methods [42, 27, 17, 26] opt to predict scaleless relative depth, our method offers a flexible alternative to using additional sensors, e.g., lidar [9, 22, 25, 34, 35, 36, 39, 40], radar [11, 19, 23, 24, 29], or multiple cameras, e.g., stereo [1, 4, 6, 32, 37, 41]. This is facilitated by the the presence of certain objects that tend to co-occur with certain scenes, which can be captured within text descriptions. As 3D scenes are continuous, a short description of objects populating the scene is sufficient to transfer relative depth estimates to metric scale.

Limitations. Our method makes the assumption that there exists an unknown scale in the estimated depth, as modeled by the parameters of a linear transformation. However, this is not always the case as the estimated depth may contain errors, and one may benefit from refinement of the relative depth maps, e.g., [3, 44]. To exploit the invariants of language in the context of depth estimation, future directions may include region-based or even pixel-wise transformations through the use of text descriptions. Also, our method has the advantage of allowing flexible descriptions as input for grounding depth estimate to metric scale. While it offers controllability of 3D reconstructions, it also opens our method to mis-use; malicious users may choose to provide adversarial descriptions to steer

predictions incorrectly. Additionally, barring malicious behaviors, uninformative captions regarding the 3D scene at hand, may also yield erroneous transformations.

5 RSA Model Architecture

Detailed architecture of RSA model is shown in Table 3. Given text descriptions $\mathbf{t} = \{t_1, t_2, \dots\}$, we first encode them into text embeddings and feed them into a 5-layer shared multi-layer perceptron (MLP) to project them into $k = 256$ hidden dimensions followed by two separate sets of 5-layer MLPs, one serves as the output head $\psi_{\hat{\alpha}} : \mathbb{R}^k \mapsto \mathbb{R}_+$ for scale parameter $\hat{\alpha}$ and the other as the output head $\psi_{\hat{\beta}} : \mathbb{R}^k \mapsto \mathbb{R}_+$ for shift $\hat{\beta}$ parameter. Scale and shift are passed through an exponential function so that they are assumed to be positive in favor of optimization.

Sub-network	Layer	Units	Activation
scene_feat_net	Linear	$1024 \rightarrow 512$	-
	LeakyReLU	-	LeakyReLU
	Linear	$512 \rightarrow 512$	-
	LeakyReLU	-	LeakyReLU
	Linear	$512 \rightarrow 512$	-
	LeakyReLU	-	LeakyReLU
	Linear	$512 \rightarrow 256$	-
	LeakyReLU	-	LeakyReLU
	Linear	$256 \rightarrow 256$	-
shift_net	Linear	$256 \rightarrow 256$	-
	LeakyReLU	-	LeakyReLU
	Linear	$256 \rightarrow 128$	-
	LeakyReLU	-	LeakyReLU
	Linear	$128 \rightarrow 128$	-
	LeakyReLU	-	LeakyReLU
	Linear	$128 \rightarrow 64$	-
	LeakyReLU	-	LeakyReLU
	Linear	$64 \rightarrow 1$	-
scale_net	Linear	$256 \rightarrow 256$	-
	LeakyReLU	-	LeakyReLU
	Linear	$256 \rightarrow 128$	-
	LeakyReLU	-	LeakyReLU
	Linear	$128 \rightarrow 128$	-
	LeakyReLU	-	LeakyReLU
	Linear	$128 \rightarrow 64$	-
	LeakyReLU	-	LeakyReLU
	Linear	$64 \rightarrow 1$	-

Table 3: Structure of RSA Model. Multi-Layer Perceptron: Layers, Units, and Activation Functions.

References

- [1] Zachary Berger, Parth Agrawal, Tian Yu Liu, Stefano Soatto, and Alex Wong. Stereoscopic universal perturbations across different architectures and datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2022.
- [2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021.
- [3] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.
- [4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.
- [5] Wenjie Chang, Yueyi Zhang, and Zhiwei Xiong. Transformer-based monocular depth estimation with attention supervision. In *32nd British Machine Vision Conference (BMVC 2021)*, 2021.
- [6] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4384–4393, 2019.
- [7] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.
- [8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.
- [9] Vadim Ezhov, Hyungseob Park, Zhaoyang Zhang, Rishi Upadhyay, Howard Zhang, Chethan Chinder Chandrappa, Achuta Kadambi, Yunhao Ba, Julie Dorsey, and Alex Wong. All-day depth completion. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024.
- [10] Xiaohan Fei, Alex Wong, and Stefano Soatto. Geo-supervised visual depth prediction. *IEEE Robotics and Automation Letters*, 4(2):1661–1668, 2019.
- [11] Stefano Gasperini, Patrick Koch, Vinzenz Dallabetta, Nassir Navab, Benjamin Busam, and Federico Tombari. R4dyn: Exploring radar for self-supervised monocular depth estimation of dynamic scenes. In *2021 International Conference on 3D Vision (3DV)*, pages 751–760. IEEE, 2021.
- [12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.
- [13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.
- [14] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.
- [15] 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*, pages 3828–3838, 2019.
- [16] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2485–2494, 2020.
- [17] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. *arXiv preprint arXiv:2312.02145*, 2023.
- [18] Dong Lao, Fengyu Yang, Daniel Wang, Hyungseob Park, Samuel Lu, Alex Wong, and Stefano Soatto. On the viability of monocular depth pre-training for semantic segmentation. In *European Conference on Computer Vision*. Springer, 2024.

- [19] Juan-Ting Lin, Dengxin Dai, and Luc Van Gool. Depth estimation from monocular images and sparse radar data. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10233–10240. IEEE, 2020.
- [20] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. Va-depthnet: A variational approach to single image depth prediction. *arXiv preprint arXiv:2302.06556*, 2023.
- [21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.
- [22] Tian Yu Liu, Parth Agrawal, Allison Chen, Byung-Woo Hong, and Alex Wong. Monitored distillation for positive congruent depth completion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 35–53. Springer, 2022.
- [23] Chen-Chou Lo and Patrick Vandewalle. Depth estimation from monocular images and sparse radar using deep ordinal regression network. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3343–3347. IEEE, 2021.
- [24] Yunfei Long, Daniel Morris, Xiaoming Liu, Marcos Castro, Punarjay Chakravarty, and Praveen Narayanan. Full-velocity radar returns by radar-camera fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16198–16207, 2021.
- [25] Hyungseob Park, Anjali Gupta, and Alex Wong. Test-time adaptation for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20519–20529, 2024.
- [26] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.
- [27] 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*, 44(3):1623–1637, 2020.
- [28] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *European conference on computer vision*, pages 746–760. Springer, 2012.
- [29] Akash Deep Singh, Yunhao Ba, Ankur Sarker, Howard Zhang, Achuta Kadambi, Stefano Soatto, Mani Srivastava, and Alex Wong. Depth estimation from camera image and mmwave radar point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9275–9285, 2023.
- [30] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.
- [31] Rishi Upadhyay, Howard Zhang, Yunhao Ba, Ethan Yang, Blake Gella, Sicheng Jiang, Alex Wong, and Achuta Kadambi. Enhancing diffusion models with 3d perspective geometry constraints. *ACM Transactions on Graphics (TOG)*, 42(6):1–15, 2023.
- [32] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14203, 2021.
- [33] Alex Wong, Safa Cicek, and Stefano Soatto. Targeted adversarial perturbations for monocular depth prediction. *Advances in neural information processing systems*, 33:8486–8497, 2020.
- [34] Alex Wong, Safa Cicek, and Stefano Soatto. Learning topology from synthetic data for unsupervised depth completion. *IEEE Robotics and Automation Letters*, 6(2):1495–1502, 2021.
- [35] Alex Wong, Xiaohan Fei, Byung-Woo Hong, and Stefano Soatto. An adaptive framework for learning unsupervised depth completion. *IEEE Robotics and Automation Letters*, 6(2):3120–3127, 2021.
- [36] Alex Wong, Xiaohan Fei, Stephanie Tsuei, and Stefano Soatto. Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters*, 5(2):1899–1906, 2020.
- [37] Alex Wong, Mukund Mundhra, and Stefano Soatto. Stereopagnosia: Fooling stereo networks with adversarial perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2879–2888, 2021.

- [38] Alex Wong and Stefano Soatto. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5644–5653, 2019.
- [39] Alex Wong and Stefano Soatto. Unsupervised depth completion with calibrated backprojection layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12747–12756, 2021.
- [40] Yangchao Wu, Tian Yu Liu, Hyungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Augundo: Scaling up augmentations for monocular depth completion and estimation. In *European Conference on Computer Vision*. Springer, 2024.
- [41] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020.
- [42] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024.
- [43] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3916–3925, 2022.
- [44] Ziyao Zeng, Daniel Wang, Fengyu Yang, Hyungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Worddepth: Variational language prior for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9708–9719, 2024.
- [45] Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can language understand depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6868–6874, 2022.
- [46] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.