Recovering Dense Metric Depth in Indoor Scenes from Monocular Depth Foundation Models and 2D LiDARs

Rémi Marsal, Alexandre Chapoutot, Philippe Xu, David Filliat

Recovering dense metric depth in indoor scenes from monocular depth foundation models and 2D LiDARs

Rémi Marsal, Alexandre Chapoutot, Philippe Xu, and David Filliat

U2IS, ENSTA Paris, Institut Polytechnique de Paris, Palaiseau, France
firstname.surname@ensta-paris.fr

Abstract. Recently, the first foundation models for monocular depth estimation such as Depth Anything have emerged. However, by being trained to make affine-invariant predictions, these methods rely on fine-tuning for making metric depth predictions and therefore perform poorly on zero-shot metric depth estimation. In a real use case, the fine-tuning stage is costly because a dedicated dataset with ground truth depth must be created and used as a training set. Additionally, fine-tuning can compromise the model's generalization ability. This paper proposes to leverage 2D LiDARs to rescale Depth Anything's predictions in the context of indoor scenes so as to prevent expensive fine-tuning or harming the model capacity. Our experiments demonstrate similar performance with fine-tuned approaches and enhanced results over zero-shot metric depth estimation methods.

Keywords: Zero-shot metric monocular depth estimation

1 Introduction

While 3D sensors like stereo cameras or time-of-flight are becoming more common in indoor robotics, monocular cameras remain a crucial and cost-effective option. Associated with depth prediction models, they provide denser outputs and a wider depth range and are more robust to challenging viewing conditions like direct sunlight or untextured surfaces. The emergence of foundation models for monocular depth estimation [1] was made possible by the abundance of publicly available datasets for monocular depth estimation combined with advanced neural network architectures such as Vision Transformers [2].

Nevertheless, metric or absolute depth cannot be predicted from a single image due to the inherent scale ambiguity. This is why a model trained on images captured with a unique camera calibration struggles to generalize with different cameras. Moreover, this issue makes the training of monocular metric depth estimation models difficult on several datasets at once. The solution leveraged by [3,1,4] consists in learning an affine-invariant depth or disparity. These models must be fine-tuned on a dedicated dataset to perform end-to-end metric depth estimations for this domain. However, creating a new and large enough dataset that includes the depth annotations and the fine-tuning is costly and may reduce the generalization abilities of the fine-tuned model.

In this paper, we propose a zero-shot monocular depth estimation method that leverage 2D LiDARs to rescale affine-invariant disparity maps provided by a foundation model

such as [1]. Here, the term *rescale* means recovering the accurate affine transformation to get the metric depth. In practice, the 3D points returned by the 2D LiDAR are used as metric depth reference to regress the rescaling parameters. This approach offers two key benefits. First, it can be applied to recent foundation models for monocular depth estimation including Depth Anything [1] which exhibit strong generalization capabilities so that no additional fine-tuning is required. Second, 2D LiDARs are quite affordable sensors that are already very used in indoor robotics applications. Extensive experiments highlight the effectiveness of our approach on standard benchmarks for monocular metric depth estimation.

2 Related work

2.1 Monocular depth estimation

First, [5] leverage Markov Random Fields for monocular depth estimation before the advent of neural networks with convolutions [6], transformers [7] then diffusion [8]. Early monocular depth estimation approaches were learned as a regression problem [6] while more recent ones adopt a classification strategy for better results [9]. Recently, depth models have been trained on diverse datasets [3], unlocking zero-shot monocular depth estimation. These approaches address the issue of the different camera calibrations depending on the dataset by learning affine-invariant depth or disparity [3,7,1]. Our work focuses on recovering the metric depth for any monocular depth estimation model that produces metric disparity maps up to an affine transformation.

2.2 Scale recovery for monocular depth

Monocular depth prediction faces the challenge of scale ambiguity, where the true size of objects cannot be reliably determined from a single image. Models trained for specific camera calibrations often produce inaccurate depth when used with a different camera. The usual solution is to fine-tune the model on data from the target camera, but this process is costly as it requires creating a new dataset and retraining. Thus, [10] trained an additional module to estimate a metric depth where the true scale is provided by a visual-inertial odometry. The method [11] learns to transfer the scale factors from a source domain with ground truth annotation to a target domain without ground truth. In contrast, we propose a simpler approach that does not require any fine-tuning and is adaptable to any intrinsic camera parameters by using 3D points from a 2D LiDAR as a reference to rescale affine-invariant disparity maps.

2.3 Zero-shot metric monocular depth estimation

The first zero-shot metric depth estimation approach, ZoeDepth [12], fine-tunes two metric bin modules on top of a pretrained MiDaS architecture [3] for indoor and outdoor scenes. Then, ScaleDepth [13] directly trains a model that predicts relative depth and a module for scale estimation. Another method in [14] proposes an architecture that incorporates the intrinsic parameters alongside the image. In contrast, [15] projects images into the same camera domain before predicting metric depth. UniDepth [16] relies

on a map representing the camera calibration that is predicted from a single image. Such approaches have limitations: they are more computationally expensive at inference and require image calibration during training hindering the use of datasets with unknown image calibration.

3 Method

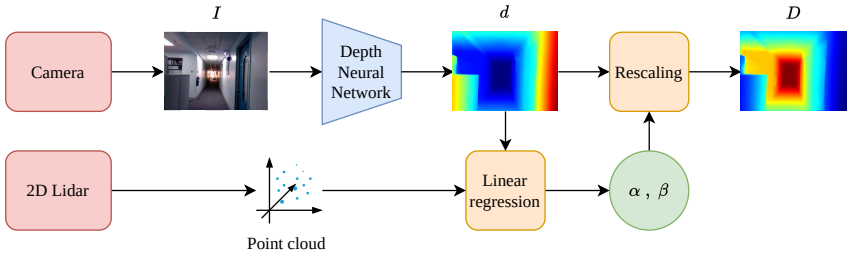

Fig. 1. Illustration of our approach. A depth neural network such as Depth Anything [1] takes as input an image I and returns an affine-invariant disparity map d. The scaling parameters α and β are regressed using a point cloud P provided by a 2D lidar. The metric depth D is obtained by applying these parameters to the disparity map d.

Consider an input image $I \in \mathbb{R}^{H\times W\times 3}$ where H and W are the height and the width and a neural network Φ that takes I as input and returns the corresponding affine-invariant disparity map $d \in \mathbb{R}^{H\times W}$ such as [3,1]. This means that Φ has been trained so there exists unknown parameters $\alpha \in \mathbb{R}_*^+$ and $\beta \in \mathbb{R}$ such that the relation between d and the absolute or metric disparity $D^{-1} \in \mathbb{R}^{H\times W}$, *i.e.*, the inverse of the absolute or metric depth D is given by:

$$D^{-1} = \alpha d + \beta, \tag{1}$$

Our approach is depicted in Fig. 1. It focuses on estimating the parameters α and β given a set of N reference 3D points $P \in \mathbb{R}^{N\times 3}$ and the disparity map d to obtain the metric depth map D. Each reference 3D point of P is projected on the image plane so the corresponding disparity value in d can be extracted. Then, a linear regression is performed to obtain α and β. Finally, these coefficients are applied to the affine-invariant map d to recover the metric disparity and the metric depth.

In practice, only two 3D reference points are necessary, but due to prediction errors by Φ and measurement errors by the sensor, more points need to be considered. The 3D reference points are given by an extra sensor, in this paper, we consider 2D lidars because they are affordable.

4 Experiments

4.1 Datasets

We evaluate our method using standard metrics [6] on standard monocular depth estimation benchmark of indoor scenes. This includes NYU V2 [17], SUN-RGBD [18], IBIMS-1 [19] and DIODE indoor [20]. To compare the performance relative to zero-shot metric depth estimation methods and fine-tuned approaches, we leverage datasets that do not belong to Depth Anything training set.

4.2 Implementation details

In our experiments, the affine-invariant disparity maps are predicted by Depth Anything [1] using a ViT Large backbone. We leverage their settings and their checkpoint without fine-tuning. Although these datasets do not contain 2D LiDAR annotations, we simulate them with a horizontal line extracted from the ground truth depth.

4.3 Results

Table 1. Quantitative results on the NYU V2 benchmark [17] where (ft) means the model has been fine-tuned on NYU v2 dataset and (zs) stands for Zero-Shot, *i.e.*, the model has not been trained on NYU v2 images.

Methods	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	AbsRel $\downarrow$	RMSE $\downarrow$	log10 $\downarrow$
ZeroDepth [14] (zs)	0.926	0.986	–	0.081	0.338	–
Metric3D (zs)	0.944	0.986	0.995	0.083	0.310	0.035
Metric3D V2 [15] (zs)	0.975	0.994	0.998	0.063	0.251	0.028
Unidepth [16] (zs)	0.984	–	–	0.058	0.201	0.073
ZeroDepth [14] (ft)	0.954	0.995	**1.000**	0.074	0.269	0.103
ZoeDepth [12] (ft)	0.955	0.995	0.999	0.075	0.270	0.032
Metric3D V2 [15] (ft)	**0.989**	**0.998**	**1.000**	**0.047**	**0.183**	**0.020**
ScaleDepth [13] (ft)	0,957	0,994	0.999	0,074	0,267	0,032
Depth Anything [1] (ft)	0.984	**0.998**	**1.000**	0.056	0.206	0.024
Depth Anything V2 [4] (ft)	0.984	**0.998**	**1.000**	0.056	0.206	0.024
Ours (zs)	0.960	0.992	0.999	0.056	0.567	0.024

Comparisons between our rescaling with 2D LiDAR and other monocular metric depth estimation methods either zero-shot (zs) or fine-tuned (ft) on the dataset NYU V2 [17] are provided in Tab. 1 for NYU V2 only and in Tab. 2 for other indoor datasets ($\uparrow$ means "higher is better" and $\downarrow$ stands for "lower is better"). We show in our experiments that metric depth obtained by rescaling affine-invariant depth predictions using 2D LiDAR compares favorably with end-to-end zero-shot metric depth estimation methods. More specifically, we observe an average improvement relative to other zero-shot methods of 28%, 42% and 8% for the δ_1, AbsRel and RMSE metrics, respectively. The lower performance on the SUN-RGBD dataset [18] relative to the [12,13,1] approaches can be

Table 2. Quantitative results on different zero-shot indoor benchmarks.

Methods	SUN-RGBD			IBIMS-1			DIODE Indoor		
	$\delta_1 \uparrow$	AbsRel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	RMSE $\downarrow$
Metric3D	–	–	–	–	0.144	–	–	0.252	–
Metric3D V2 [15]	–	–	–	–	0.185	0.592	–	**0.093**	**0.389**
Unidepth [16]	**0.966**	–	–	0.797	–	–	0.774	–	–
ZoeDepth [12]	0.864	**0.119**	**0.346**	0.658	0.169	0.711	0.4	0.324	1.581
ScaleDepth [13]	0.864	0.127	0.360	0.788	0.156	0.601	0.455	0.277	1.35
Depth Anything [1]	0.658	0.500	0.616	0.714	0.150	0.593	0.303	0.325	1.476
Ours	0.897	0.204	0.363	**0.960**	**0.050**	**0.275**	**0.909**	0.106	0.439

explained by the fact that these approaches rely on a fine-tuning on NYU V2 [17]: the domain of both datasets is close since they have been collected using the same camera: a Kinect. We notice that the results of in-domain fine-tuned methods are overall better than ours, however, such methods are much more costly, requiring the creation of a dedicated dataset and additional training.

5 Conclusion

This paper presents a low-cost method for estimating metric depth using Depth Anything [1] affine-invariant disparity predictions, rescaled with 3D reference points from a 2D LiDAR. Our approach is adaptable to any camera calibration and more affordable than traditional scale recovery methods, as it avoids costly fine-tuning of a monocular depth estimation model. By leveraging publicly available Depth Anything weights, the method generalizes well across diverse image domains. Experiments on standard indoor scene benchmarks demonstrate the competitiveness of our method compared to zero-shot monocular depth estimation techniques. In real cases, the 2D LiDAR and the camera cannot be perfectly calibrated. This implies that mismatches between the reference 3D points or the corresponding pixels of the image can introduce outliers that may harm the regression. For this reason, robust regression techniques such as RANSAC must be considered. Future work will explore more diverse sensors to improve performance and adapt our approach to outdoor scenes.

Acknowledgement

This research was funded in whole or in part by the French National Research Agency (ANR) under the "ANR-23-MOXE-0003" project.

References

1. L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024.

2. A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
3. R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.
4. L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv preprint arXiv:2406.09414*, 2024.
5. A. Saxena, S. Chung, and A. Ng, "Learning depth from single monocular images," *Advances in neural information processing systems*, vol. 18, 2005.
6. D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, 2014.
7. R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.
8. B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9492–9502, 2024.
9. S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4009–4018, 2021.
10. D. Wofk, R. Ranftl, M. Müller, and V. Koltun, "Monocular visual-inertial depth estimation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2023.
11. A. Dana, N. Carmel, A. Shomer, O. Manela, and T. Peleg, "Do more with what you have: Transferring depth-scale from labeled to unlabeled domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4440–4450, 2024.
12. S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.
13. R. Zhu, C. Wang, Z. Song, L. Liu, T. Zhang, and Y. Zhang, "Scaledepth: Decomposing metric depth estimation into scale prediction and relative depth estimation," *arXiv preprint arXiv:2407.08187*, 2024.
14. V. Guizilini, I. Vasiljevic, D. Chen, R. Ambruș, and A. Gaidon, "Towards zero-shot scale-aware monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9233–9243, 2023.
15. M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen, "Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," *arXiv preprint arXiv:2404.15506*, 2024.
16. L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu, "Unidepth: Universal monocular metric depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10106–10116, 2024.
17. P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.
18. S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.
19. T. Koch, L. Liebel, F. Fraundorfer, and M. Korner, "Evaluation of cnn-based single-image depth estimation methods," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.
20. I. Vasiljevic, N. Kolkin, S. Zhang, R. Luo, H. Wang, F. Z. Dai, A. F. Daniele, M. Mostajabi, S. Basart, M. R. Walter, and G. Shakhnarovich, "DIODE: A Dense Indoor and Outdoor DEpth Dataset," *CoRR*, vol. abs/1908.00463, 2019.