

Evaluating Monocular Depth Estimation Methods on Industrial Objects

Nazanin Padkan^{1,2}, Ziyang Yan^{1,3}, Fabio Remondino¹

¹ 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Web: <http://3dom.fbk.eu> - Email: <npadkan@fbk.eu><zyan@fbk.eu><remondino@fbk.eu>

² Dept. Mathematics, Computer Science and Physics, University of Udine, Italy

³ DISI, University of Trento, Italy

Keywords: Monocular Depth Estimation, 3D Reconstruction, Non-collaborative Surfaces

Abstract

Monocular Depth Estimation (MDE) has become a valuable tool in 3D reconstruction, especially when traditional methods such as photogrammetry are impractical. This study evaluates the performance of three state-of-the-art MDE algorithms, namely Depth Pro, Depth Anything V2 and Metric3Dv2, in estimating depth for challenging industrial objects with complex properties like reflectivity, transparency or lack of texture. Using both synthetic and real-world objects, algorithms' abilities to accurately estimate depth and generate 3D models are evaluated. Our findings show that Depth Pro outperforms Metric3Dv2 in handling these difficult scenarios, with significantly lower error rates and better handling of details such as edges and surfaces. The results demonstrate the potential of MDE in industrial applications, particularly where multi-camera systems or additional sensors are not feasible. However, while MDE offers promising solutions, further improvements are needed to fully address the unique challenges posed by industrial environments.

1. Introduction

Depth estimation plays a pivotal role in 3D imaging, enabling the 3D reconstruction of structures from 2D images. Among the various approaches, learning-based Monocular Depth Estimation (MDE) is gaining interest for various applications, despite facing still various challenges (Yan et al., 2025). MDE derives depth information from a single image, making it particularly useful in scenarios where multi-camera setups are unusable (Liao et al., 2025). MDE has made significant progress in recent years, but its performance on industrial objects remains largely unexplored. Among the most challenging objects for deep learning methods, including MDE, we can mention transparent and reflective surfaces, such as building windows, cars or glass and metallic items (Costanzino et al., 2023). Unlike typical datasets used in MDE research, manufacturing and industrial components often present extreme visual challenges, such as transparency, reflectivity, metallic and textureless surfaces as well as small or complex geometries. These characteristics make reliable depth prediction difficult, yet accurate 3D reconstruction is essential for applications like quality control, automated inspection and reverse engineering.

This work is driven by the need to address these challenges. By thoroughly evaluating state-of-the-art MDE methods on both synthetic and real-world industrial objects, we aim to uncover where these methods succeed, where they fail and how they could be improved. Ultimately, the goal is to help bridge the gap between current MDE techniques and the demanding requirements of industrial applications such as precision manufacturing and automated quality inspection.

2. Monocular Depth Estimation (MDE)

MDE using deep learning involves predicting depth maps from a single 2D color image through a deep neural network (Ming et al., 2021). A pioneering approach was introduced by Eigen et al. (2014). They proposed a coarse-to-fine framework, where the coarse network learned the global depth of the entire image to produce a rough depth map, while the fine network focused on local features to refine it. Since then, numerous researchers have advanced deep learning techniques for monocular depth estimation (Bhat et al., 2023; Birkel et al., 2023). The process of

deep learning-based MDE commonly employs an encoder-decoder network. In this setup, the input is simply an RGB image, and the output is a computed depth map (Ming et al., 2021). This depth map is often represented in inverse relative depth, where closer objects have higher values and distant pixels approach zero.

MDE remains a critical area of research despite advancements in stereo vision, Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) and Gaussian splatting (Kerbl Wang et al., 2025; Yan et al., 2024a; Yan et al., 2024b), due to its unique accessibility, flexibility and broad applicability. Unlike stereo vision, which depends on calibrated multi-camera setups, or NeRF and Gaussian splatting, which often require multiple views and substantial computational power, MDE works with a single RGB image. Its lightweight nature allows it to run on mobile and edge devices, making it ideal for applications like scene understanding (Schön et al., 2021), navigation (Dong et al., 2022), object detection (Yu et al., 2021), cultural heritage (Zhu et al., 2024) and medical imaging (He et al., 2024). MDE also complements multi-view approaches by providing depth information when multi-view data is unavailable or impractical. Its ability to address these challenges underlines its importance in advancing depth understanding and 3D reconstruction for a wide range of real-world applications. Despite their utility, industrial objects often present unique challenges for depth estimation. Features such as reflectivity, transparency, shininess, or small surface areas can severely affect the performance of conventional algorithms (Remondino et al., 2023). These complex optical properties often render traditional photogrammetry techniques, which depend on feature matching, ineffective. Such limitations highlight the necessity for more advanced and adaptable depth estimation solutions (Liang et al., 2023; Weibel et al., 2023; Chen et al., 2023 etc).

Among many MDE algorithms, in this paper we focused on three zero-shot metric depth algorithms, Depth Pro, Depth Anything V2 and Metric3Dv2. Firstly, they are inferring metric depth and secondly they showed better performance in metrics.

Depth Pro (Bochkovskii et al., 2024) is a powerful AI tool developed by Apple that estimates depth from a single image with impressive accuracy. It stands out because it can measure real-world distances without needing extra information such as camera parameters. The model is designed to quickly produce

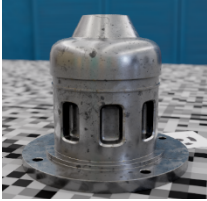

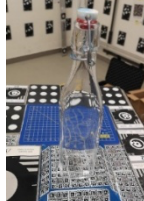

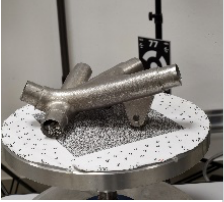
Name	Synthetic Metallic	Synthetic Glass	Bottle	Industrial_A	Industrial_B
View					
Image resolution	1080x1920 px	1080x1920 px	1080x1920 px	1080x1920 px	6016x4016 px
Ground Truth (GT)	Synthetic data	Synthetic data	Photogrammetry on powdered surface	Triangulation laser scanner	Triangulation laser scanner
Approx size	ca 18 cm height	ca 12 cm height	ca 30 cm height	ca 6 cm height	ca 20 cm width
Characteristics	Textureless, Complex, Reflective	Transparent, Highly refractive	Complex shape, Highly refractive	Textureless, Small, Complex	Textureless, Complex, Reflective

Table 1: The analysed synthetic and real objects from NeRFBK dataset¹ with their properties.

high-quality, detailed depth maps. By combining real-world and synthetic data during training, it excels at capturing fine details, especially around edges, making it ideal for applications like 3D modeling and augmented reality. Depth pro network processes an input image by downsampling it across multiple scales. At each scale, the image is divided into patches, which are encoded using a ViT-based (Dosovitskiy, A., 2020) patch encoder with shared weights across all scales. These patches are then merged to form feature maps, which are subsequently upsampled and fused using a DPT decoder (Ranftl et al., 2021). Additionally, predictions are grounded by a separate image encoder that provides global context, ensuring that the output benefits from both local and holistic information. The input image should have the 1536x1536 resolution and if the input image is not at network size, it is resized to 1536x1536 px and the estimated depth is resized to the original image resolution.

Metric3D v2 (Hu et al., 2024; Yin et al., 2023) is introduced as a geometric foundation model for zero-shot metric depth and surface normal estimation from a single image, crucial for accurate 3D reconstruction. While depth and surface normals are complementary, each presents distinct challenges. Current monocular depth methods generalize well but only produce affine-invariant depths, lacking real-world metric precision. Similarly, surface normal estimation struggles in zero-shot scenarios due to limited labelled data. To address these challenges, solutions for both metric depth and surface normal estimation are proposed. For metric depth, the model resolves ambiguities from different camera models through a canonical camera space transformation module, which integrates easily with existing monocular models. For surface normal estimation, a joint depth-normal optimization module is introduced, utilizing metric depth data to improve normal estimation beyond label constraints. With these modules, the model is trained on over 16 million images from diverse camera setups, enabling strong zero-shot generalization on unseen data.

Depth Anything V2 (Yang et al., 2024a,b) is one of the powerful MDE models which produce better estimation compared to the old version (Depth Anything V1). This model has three key features for delivering better estimations which are usage of synthetic images instead of real ones: scaling up the teacher model and leveraging large-scale pseudo-labelled real images for student training.

3. NERFBK dataset

The NeRFBK dataset¹ (Table 1; Yan et al., 2023) offers a much more suitable benchmark for industrial applications compared to

more general datasets such as NYU Depth V2, KITTI or ScanNet. While those datasets are widely used in the depth estimation community, they mainly focus on indoor or outdoor scenes with opaque and textured objects. Such data are less representative of challenges present in industrial environments. NeRFBK, on the other hand, is designed to include problematic industrial objects for monocular depth estimations and 3D reconstructions, including ground truth data, obtained through photogrammetry, triangulation-based laser scanning or synthetic rendering. The dataset contains transparent and shiny objects, reflective metallic surfaces, as well as textureless and small components with complex shapes. These characteristics make it especially valuable for testing how well algorithms cope with situations where conventional photogrammetry or stereo vision methods often fail.

For the paper aims, five NeRFBK objects are selected:

- *Synthetic Metallic*, which features reflective surfaces;
- *Synthetic Glass*, characterized by transparent surface;
- *Bottle*, featuring transparency and reflections;
- *Industrial A*, with textureless surface;
- *Industrial B*, with a reflective metallic surface.

4. Analysis and evaluation methodology

For the evaluation, four partially overlapping views of the same object are considered and given to the MDE algorithms for estimating the depth. Depths and known camera parameters (including focal length) are then used to generate scaled point clouds. All point clouds are finally co-registered using an Iterative Closest Point (ICP) method to the available ground truth 3D data. To evaluate MDE algorithms' performances, various metrics - i.e. Root Mean Square Error (RMSE), Standard Deviation (STD) and Mean Absolute Error (MAE) - are used to quantify error rates with respect to the ground truth.

5. Experimental results

The processing and analyses are performed using an Intel E5-1650 @3.2GHz, 32 GB RAM and NVIDIA GeForce GTX 1050 Ti GPU.

Visual and quantitative results for the Synthetic Metallic object are reported in Figure 1 and Table 2. Depth Pro demonstrates superior performance in capturing details (Figure 2), such as the holes on the surface of the object and provides better estimations along the edges, exhibiting greater consistency on reflective surfaces that lack holes or edges. Depth Pro achieves lower error

¹ <https://github.com/3DOM-FBK/NeRFBK>

rates compared to Metric3Dv2, with the lowest standard deviation (STD) recorded at 1.57 mm, whereas Depth Anything V2 and Metric3Dv2 have an STD of 2.57 mm and 3.51 mm, respectively. Additionally, the point clouds generated from four different viewpoints are merged to create a complete 3D representation of the object (Figure 3). Using CloudCompare, a Mean Distance of -8.75 mm is achieved, which represents the average of all signed distances between the reconstructed point cloud and the ground truth. Such value indicates good overall alignment between the two data, with the negative sign suggesting MDE results being slightly smaller with respect to the ground truth. The Standard Deviation resulted 2.34 mm.

Visual and quantitative results for the Glass object are reported in Figure 4 and Table 3. All three algorithms can estimate the depth of the transparent object. However, Depth Pro provides a more detailed 3D point cloud and shape of the object compared to Metric3Dv2 and Depth Anything V2. In terms of metrics, Depth Pro achieves lower error rates (RMSE, MAE and STD) than the other methods. Figure 5 shows the 3D point clouds derived using the depth inferred from the first viewpoint of the glass. It is clear that Metric3Dv2 performs poorly in comparison to Depth Pro. The glass surface shows significant deformation, and the handle is misaligned. In contrast, Depth Pro provides a considerably better visualization of both the glass and the handle, with the curves on the glass surface appearing almost complete.

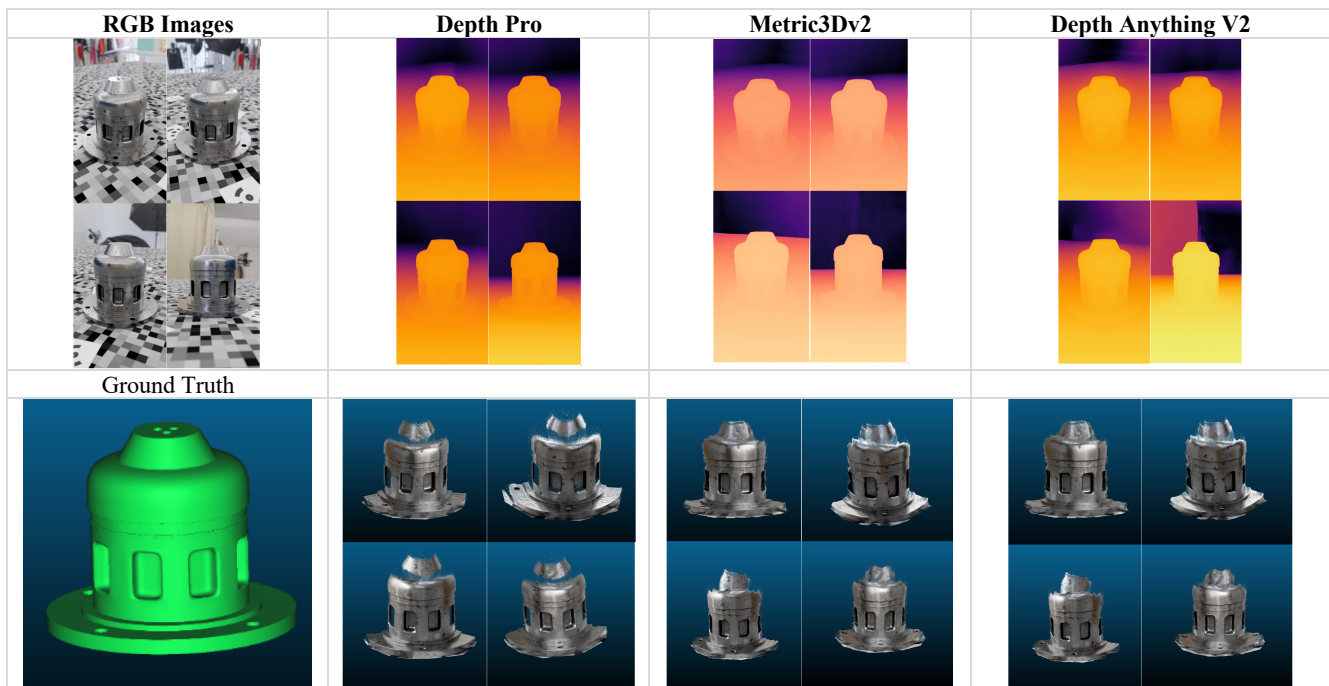


Figure 1. Recovered depths and 3D point clouds per view for the object “Synthetic Metallic”.

Method	Depth Pro			Metric3Dv2			Depth Anything V2		
Metric	RMSE	MAE	STD	RMSE	MAE	STD	RMSE	MAE	STD
View_01	2.18	1.67	1.40	4.89	3.59	3.33	4.25	3.24	2.76
View_02	2.78	2.04	1.89	3.77	2.89	2.42	3.86	3.04	2.38
View_03	2.59	1.97	1.68	2.92	2.43	1.62	4.22	3.21	2.73
View_04	1.55	0.85	1.29	10.53	8.14	6.66	3.801	2.94	2.40
Average	2.28	1.63	1.57	5.52	4.26	3.51	4.03	3.11	2.57

Table 2. Metrics [mm] for the cloud-to-mesh comparisons of examined MDE methods applied to the object “Synthetic Metallic”.



Figure 2. Geometric details in the generated point clouds for each algorithm on one view.

Figure 3. Merged point clouds of the object “Synthetic Metallic” for Depth Pro (left) and co-registration on the ground truth model (right).

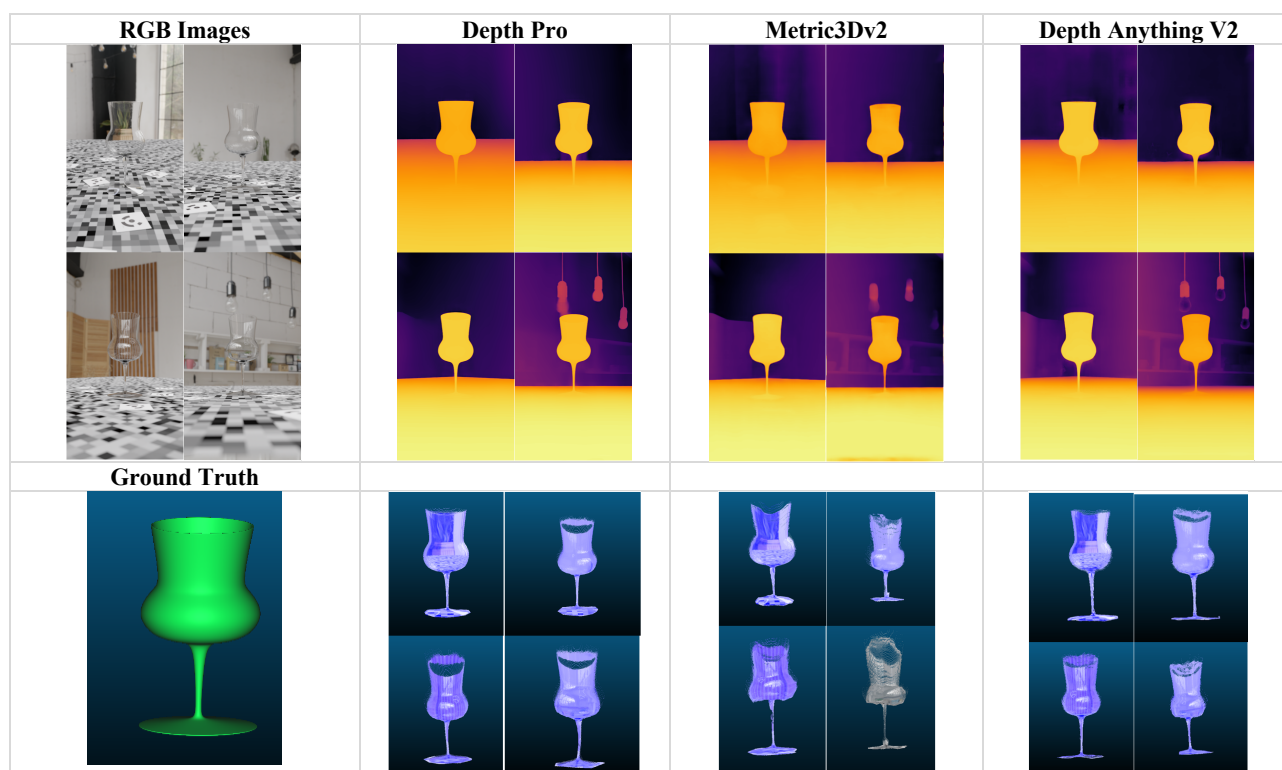


Figure 4. Recovered depths and 3D point clouds of the object “Glass”.

Method	Depth Pro			Metric3Dv2			Depth Anything V2		
Metric	RMSE	MAE	STD	RMSE	MAE	STD	RMSE	MAE	STD
View_01	3.24	2.56	1.99	12.49	9.66	7.92	5.69	4.12	3.93
View_02	6.18	4.10	4.61	8.18	5.72	5.85	12.26	8.64	8.69
View_03	6.83	3.84	5.64	10.71	7.64	7.50	10.71	7.64	7.50
View_04	8.59	5.86	6.28	13.73	8.95	10.41	12.53	8.99	8.72
Average	6.21	4.09	4.13	11.78	8.49	7.92	10.29	7.35	7.21

Table 3. Metrics [mm] for the cloud-to-mesh comparisons of examined MDE methods applied to the object “Glass”.

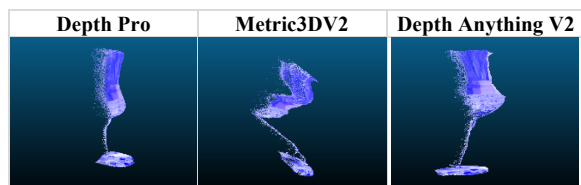


Figure 5. Generated point clouds of the object “Glass” from a side view depth (View_01).

For the Bottle object (Figure 6, Table 4), Depth Pro demonstrates better performance. In contrast, Metric3Dv2 exhibits some inaccuracies in depth estimation, resulting in incomplete point clouds. Depth Pro metrics has lower error rates compared to the other methods, with an average MAE of 2.64 mm in the four views.

For Industrial A (Figure 7 and Table 5), Depth Pro provides better estimations for the edges and finer details in the generated point clouds. Although Table 5 indicates minimal differences in terms of error rates, Depth Pro still demonstrates better performances. Finally, the Industrial B object (Figure 8, Table 6), which features high-resolution images (6016×4016 px), reveals that Depth Pro produces also in this case more pronounced details in the estimated depth, resulting in cleaner and defined point clouds.

6. Conclusions

The paper reported and compared three state-of-the-art Monocular Depth Estimation (MDE) algorithms - Depth Pro, Metric3Dv2, and Depth Anything V2 - on various challenging industrial objects, including shiny, small and transparent ones. To evaluate the performance of these algorithms, we used several metrics and compared the generated point clouds to ground truth data. The assessment on the considered industrial objects reveals that generally Depth Pro infers better depth maps with lower error rates.

Given that MDE algorithms require only a single RGB image, unlike other methods that need multiple images or additional sensors, they present a promising complement for inspections and 3D reconstruction applications. Although the results are often satisfactory, particularly in non-industrial settings, these algorithms still face challenges with industrial objects. Nevertheless, with the continued advancements in MDE algorithms, we can expect further improvements in their performance in the near future. We aim to specifically re-train MDE models with larger industrial datasets to improve results and enhance industrial applications, exploring ways to overcome still open challenges related to transparency, reflections and texture-less surfaces.

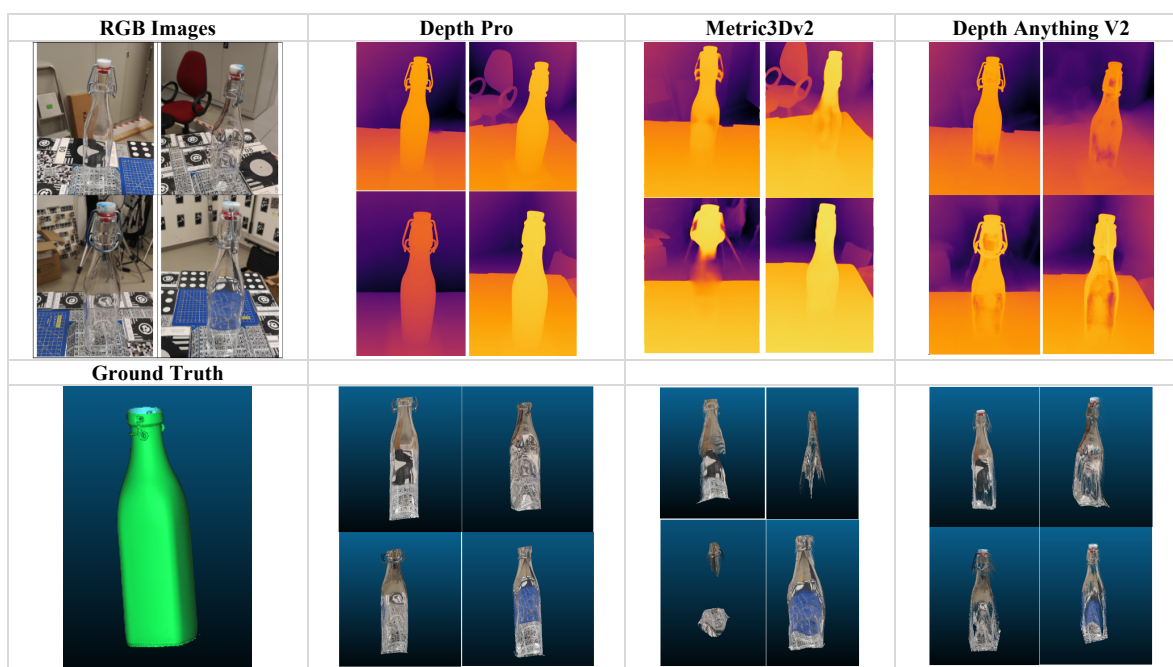


Figure 6. Recovered depths and 3D point clouds for the object “Bottle”.

Method	Depth Pro			Metric3Dv2			Depth Anything V2		
Metric	RMSE	MAE	STD	RMSE	MAE	STD	RMSE	MAE	STD
View_01	3.68	2.09	3.03	8.12	5.90	5.57	5.61	3.95	3.98
View_02	3.93	2.40	3.11	11.31	7.90	8.08	11.32	8.23	7.77
View_03	3.82	3.04	2.32	15.31	11.32	10.30	16.48	11.82	11.48
View_04	4.07	3.02	2.73	15.17	12.16	9.07	16.66	11.27	12.26
Average	3.88	2.64	2.79	12.48	9.57	8.26	12.52	8.82	8.88

Table 4. Metrics [mm] for the cloud-to-mesh comparisons [mm] of examined MDE methods applied to the object “Bottle”.

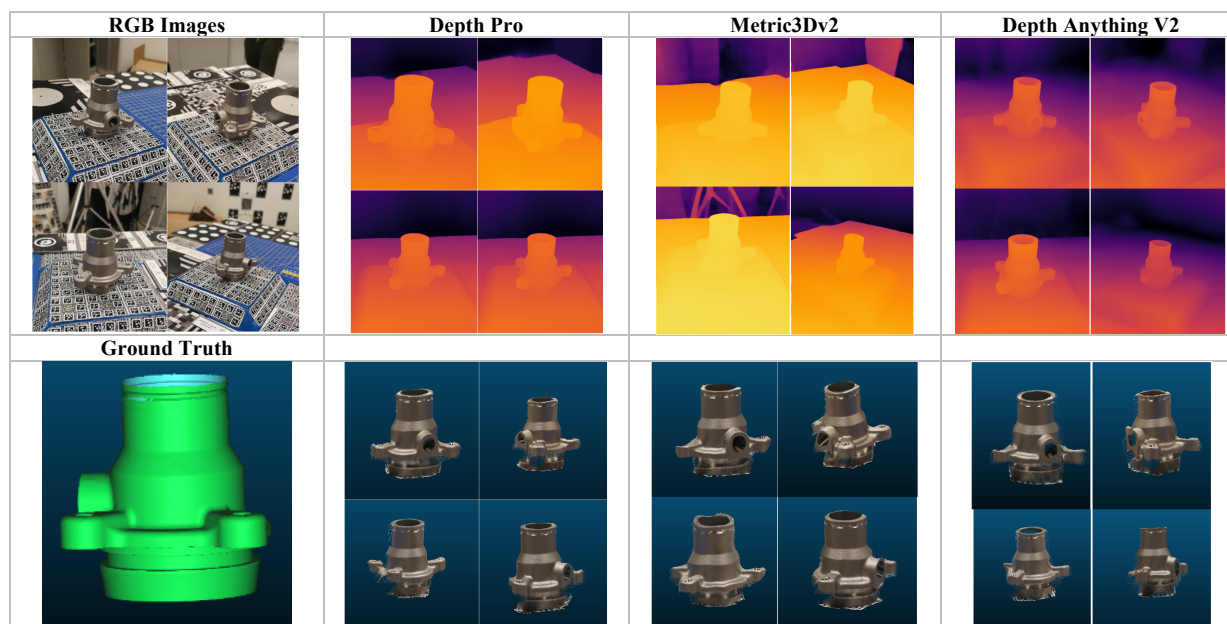


Figure 7. Recovered depths and 3D point clouds for the “Industrial A” object.

Method	Depth Pro			Metric3Dv2			Depth Anything V2		
Metric	RMSE	MAE	STD	RMSE	MAE	STD	RMSE	MAE	STD
View_01	1.34	0.94	0.95	1.35	0.92	0.98	1.01	0.74	0.69
View_02	1.76	1.13	1.35	2.09	1.43	1.52	1.18	0.81	0.86
View_03	1.65	1.14	1.18	1.32	0.97	0.90	1.12	0.78	0.80
View_04	0.69	0.51	0.47	1.72	1.25	1.18	0.93	0.69	0.62
Average	1.36	0.93	0.99	1.62	1.14	1.14	1.06	0.75	0.74

Table 5. Metrics [mm] for the cloud-to-mesh comparisons of examined MDE methods applied to the object “Industrial A”.

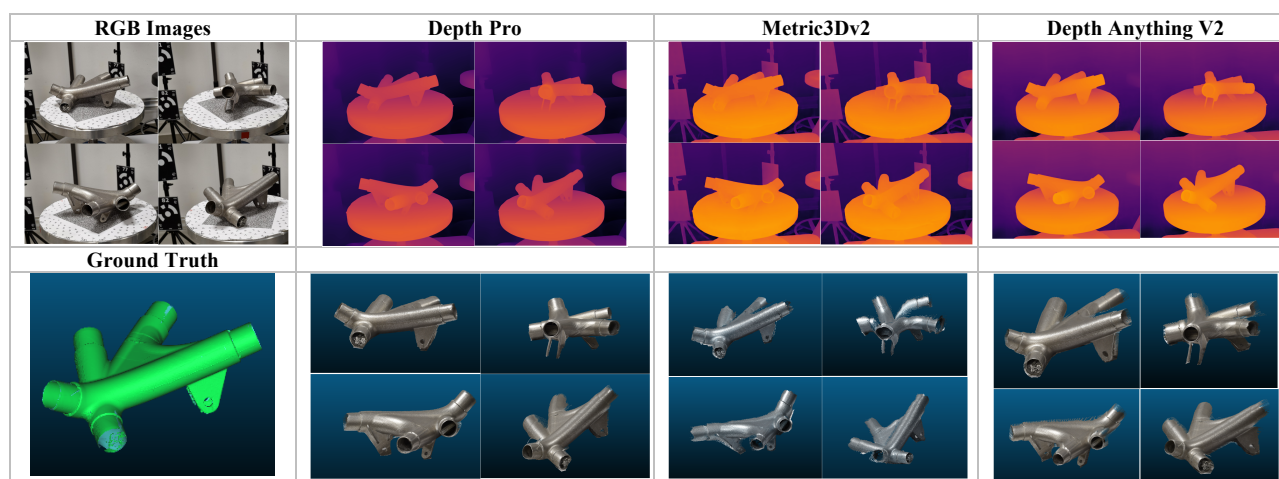


Figure 8. Recovered depths and 3D point clouds of the object “Industrial B”.

Method	Depth Pro			Metric3Dv2			Depth Anything V2		
Metric	RMSE	MAE	STD	RMSE	MAE	STD	RMSE	MAE	STD
View_01	6.76	4.54	5.01	8.68	5.26	6.90	10.61	7.21	7.79
View_02	4.19	3.20	2.69	8.35	5.80	6.01	13.18	8.65	9.94
View_03	4.11	3.14	2.66	6.08	4.57	4.02	12.35	7.62	9.72
View_04	4.54	3.38	3.03	4.77	3.51	3.22	6.73	4.39	5.11
Average	5.15	3.56	3.10	7.22	4.79	5.28	10.22	6.72	8.14

Table 6. Metrics [mm] for the cloud-to-mesh comparisons of examined MDE methods applied to the object “Industrial B”.

Acknowledgments

This study was partially carried out within the Interconnected Nord-Est Innovation Ecosystem (iNEST) and received funding from the European Union Next-GenerationEU (Piano Nazionale Di Ripresa e Resilienza (PNRR) – Missione 4, Componente 2, Investimento 1.5 – D.D. 1058 23/06/2022, ECS00000043).

References

- Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V., 2024. Depth pro: Sharp monocular metric depth in less than a second. *arXiv:2410.02073*.
- Besl, P.J. and McKay, N.D., 1992. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures*, Vol. 1611, pp. 586-606.
- Bhat, S.F., Birkel, R., Wofk, D., Wonka, P., Müller, M., 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv:2302.12288*.
- Birkel, R., Wofk, D. and Müller, M., 2023. Midas v3. 1--a model zoo for robust monocular relative depth estimation. *arXiv:2307.14460*.
- Chen, K., Wang, S., Xia, B., Li, D., Kan, Z. and Li, B., 2023, May. Tode-trans: Transparent object depth estimation with transformer. *Proc. ICRA*, pp. 4880-4886.
- Costanzino, A., Ramirez, P.Z., Poggi, M., Tosi, F., Mattoccia, S. and Di Stefano, L., 2023. Learning depth estimation for transparent and mirror surfaces. *Proc. ICCV*, pp. 9244-9255.
- Dosovitskiy, A., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*.
- Dong, X., Garratt, M.A., Anavatti, S.G. and Abbass, H.A., 2022. Towards real-time monocular depth estimation for robotics: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(10), pp.16940-16961.
- Eigen, D., Puhrsch, C. and Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27.
- He, Q., Feng, G., Bano, S., Stoyanov, D. and Zuo, S., 2024. MonoLoT: Self-Supervised Monocular Depth Estimation in Low-Texture Scenes for Automatic Robotic Endoscopy. *IEEE Journal of Biomedical and Health Informatics*.
- Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C. and Shen, S., 2024. Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-shot Metric Depth and Surface Normal Estimation. *arXiv preprint arXiv:2404.15506*.
- Liao, M., Dong, H. B., Wang, X., Ubul, K., Yan, Z., & Shao, Y. 2025. GM-MoE: Low-Light Enhancement with Gated-Mechanism Mixture-of-Experts. *arXiv:2503.07417*.
- Liang, Y., Deng, B., Liu, W., Qin, J. and He, S., 2023. Monocular depth estimation for glass walls with context: a new dataset and method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R. and Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), pp.99-106.
- Ming, Y., Meng, X., Fan, C. and Yu, H., 2021. Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438, pp.14-33.

- Padkan, N., Trybala, P., Battisti, R., Remondino, F., Bergeret, C., 2023. Evaluating Monocular Depth Estimation methods. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48(1), pp.137-144.
- Ranftl, R., Bochkovskiy, A. and Koltun, V., 2021. Vision transformers for dense prediction. *Proc. ICCV*, pp. 12179-12188.
- Remondino, F., Karami, A., Yan, Z., Mazzacca, G., Rigon, S., Qin, R., 2023. A critical analysis of NERF-based 3D reconstruction. *Remote Sensing*, 15(14), 3585.
- Schön, M., Buchholz, M. and Dietmayer, K., 2021. Mgnnet: Monocular geometric scene understanding for autonomous driving. *Proc. ICCV*, pp. 15804-15815.
- Tankovich, V., Hane, C., Zhang, Y., Kowdle, A., Fanello, S. and Bouaziz, S., 2021. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. *Proc. CVPR*, pp. 14362-14372.
- Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J. and Ding, G., 2024. YOLOv10: Real-time end-to-end object detection. *arXiv:2405.14458*.
- Wang, N., Chen, Y., Xiao, L., Xiao, W., Li, B., Chen, Z., ... & Zhao, H. 2025. Unifying Appearance Codes and Bilateral Grids for Driving Scene Gaussian Splatting. *arXiv:2506.05280*.
- Weibel, J.B., Sebeton, P., Thalhammer, S. and Vincze, M., 2023, October. Challenges of Depth Estimation for Transparent Objects. In *International Symposium on Visual Computing*, pp. 277-288.
- Yan, Z., Mazzacca, G., Rigon, S., Farella, E.M., Trybala, P. and Remondino, F., 2023. NeRFBK: a holistic dataset for benchmarking NeRF-based 3D reconstruction. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48(1), pp.219-226.
- Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J. and Zhao, H., 2024a. Depth anything: Unleashing the power of large-scale unlabeled data. *PORoc. CVPR*, pp. 10371-10381.
- Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J. and Zhao, H., 2024b. Depth Anything V2. *arXiv preprint arXiv:2406.09414*.
- Yan, Z., Padkan, N., Trybala, P., Farella, E. M., Remondino, F., 2025. Learning-Based 3D Reconstruction Methods for Non-Collaborative Surfaces - A Metrological Evaluation. *Metrology*, 5(2), 20.
- Yan, Z., Dong, W., Shao, Y., Lu, Y., Haiyang, L., Liu, J., Ma, Y. 2024a. Renderworld: World model with self-supervised 3D label. *arXiv preprint arXiv:2409.11356*.
- Yan, Z., Li, L., Shao, Y., Chen, S., Wu, Z., Hwang, J. N., Remondino, F., 2024b. 3dsceneeditor: Controllable 3d scene editing with gaussian splatting. *arXiv preprint arXiv:2412.01583*.
- Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X. and Shen, C., 2023. Metric3d: Towards zero-shot metric 3d prediction from a single image. *Proc. ICCV*, pp. 9043-9053.
- Yu, J. and Choi, H., 2021. YOLO MDE: Object detection with monocular depth estimation. *Electronics*, 11(1), p.76.
- Zhou, Q.Y., Park, J. and Koltun, V., 2018. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*.
- Zhu, B., Liu, G., Xia, H. and Zhang, L., 2024. Ancient depthnet: an unsupervised framework for depth estimation of ancient architecture. *31st Int. Conf. "Methodological aspects of education: achievements and prospects"*.