Is AI superior to Multimodal 3D Sensor Technology for Transparent Objects?

Christina Junger[1,*], Benjamin A. Simon[1], and Gunther Notni[1,2,†]

[1]Technische Universität Ilmenau, Department of Mechanical Engineering, Group for Quality Assurance and Industrial Image Processing, Ilmenau, Germany
[2]Fraunhofer Institute for Applied Optics and Precision Engineering, Jena, Germany

[*]Corresponding author: Christina.Junger@tu-ilmenau.de
[†]Gunther.Notni@tu-ilmenau.de

Abstract – Transparent objects challenge 3D perception in robotics, especially in navigation and human-robot collaboration. Conventional 3D sensors in the visible or near-infrared spectrum often fail to detect transparent materials due to their optical properties. Collecting real-world datasets for deep learning is difficult and time-consuming because ground truth acquisition requires complex preparation. Multimodal 3D sensors like thermal 3D cameras can automate dataset creation but are costly and need restrictive safety setups. Combining standard 3D sensors or RGB cameras with zero-shot deep learning models offers a promising alternative, enabling recognition of unseen transparent objects without task-specific training. However, the accuracy and feasibility of such zero-shot methods for transparent object perception remain underexplored. This paper presents an initial investigation into their potential and limitations.

Keywords: Transparent Object Perception, 3D Sensor, Multimodal Sensor, Zero-Shot Capabilities, Dataset, Human-Robot Collaboration

1. Introduction

Accurate capture of visually uncooperative objects, especially transparent surfaces, is crucial for safe navigation and reliable collaborative object manipulation [1, 2, 3, 4]. Currently, only a few measurement methods are available that can reliably perceive transparent objects in 3D [5, 6, 7]. Commonly used low-cost 3D sensors operating in the visible or near-infrared spectrum face challenges detecting transparent objects due to their optical uncooperativeness [3, 8, 9, 10, 4, 11, 12]. Deep learning approaches show promise in overcoming these limitations, though 3D perception of transparent surfaces has been considered a special case, partly due to limited datasets and the complexity of acquiring non-synthetic data [10, 9]. Recent zero-shot deep learning models demonstrate strong generalization and qualitative results in reconstructing transparent objects [13, 14].

Foundation models like *CLIP* [15] enabled multimodal AI applications, while the *Segment Anything v2 Model* [16, 17] introduced zero-shot segmentation across diverse image domains. Building on this, specialized models such as *Depth Anything* [18] offer zero-shot 3D depth estimation. These advances open new possibilities for 3D perception of challenging materials, especially transparent objects. We present a preliminary investigation into a key research gap: applying foundation models to zero-shot 3D recognition of transparent objects.

2. Method Selection and Data

The selection of 3D depth estimation models was guided by stringent criteria: zero-shot capability, open-source availability, operation on visible (VIS) or near-infrared (NIR) images, suitability for transparent object perception, single-shot processing, and output of disparity or depth maps. Soft criteria included low computational complexity and hardware requirements, minimal model size, usability without fine-tuning, and the ability to capture fine structural details.

The following foundation models meet the stringent criteria: *Marigold* [20], *GeoWizard* [21], *Depth Pro* [19] (generative types) and *MiDaS* [22], *Metric3D v2* [23] and *Depth Anything v2* [18] (discriminative types). (While *Stereo Anything* [24] shows great promise, its code has not yet been released publicly.) Initial tests have shown that *Depth Pro* and *Depth-Anything-v2-giant* achieve the most accurate results; consequently, both models were selected for further analysis. Table 1 summarizes their differences according to the soft criteria. Both

Table 1: Comparison of soft criteria in the shortlist of monocular depth estimation methods.

Soft Criteria	***Depth Pro*** [19]	***Depth-Anything-v2-giant*** [18]
Computation speed	slower	faster (approx. 6 times)
Fine structures	very realistic	lower fidelity & smoother
Parameter model	504 M	1.3 B

models were evaluated without any additional fine-tuning on selected NIR images from the dataset, following the *TranSpec3D* methodology [12]. This dataset includes varying levels of complexity for transparent objects, featuring scenarios where these objects appear both in front of and behind transparent or opaque objects.

3. Results and Discussion

Figure 1 shows results addressing the two key challenges of glass in front of another object and occlusion ordering.

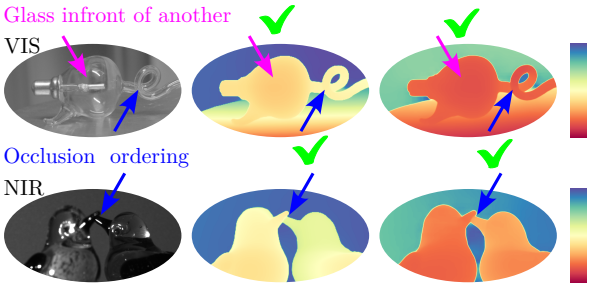

Figure 1: Input image (left), depth map based on *Depth Anything v2* (mid) and *Depth Pro* (right).

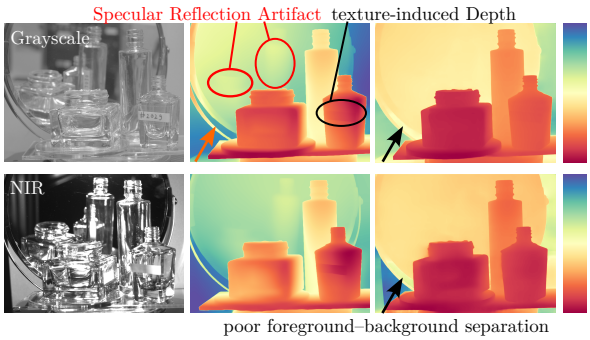

Figure 2: Input image (left), depth map based on *Depth Anything v2* (mid) and *Depth Pro* (right).

Figure 2 demonstrates that, although both models deliver impressive qualitative results in complex measurement scenarios, they exhibit notable artifacts: specular reflection artifacts on mirrored surfaces, texture-induced depth artifacts from opaque labels on glass objects, and poor foreground–background separation of fine details on optically uncooperative materials (e.g., metallic surfaces). Figure

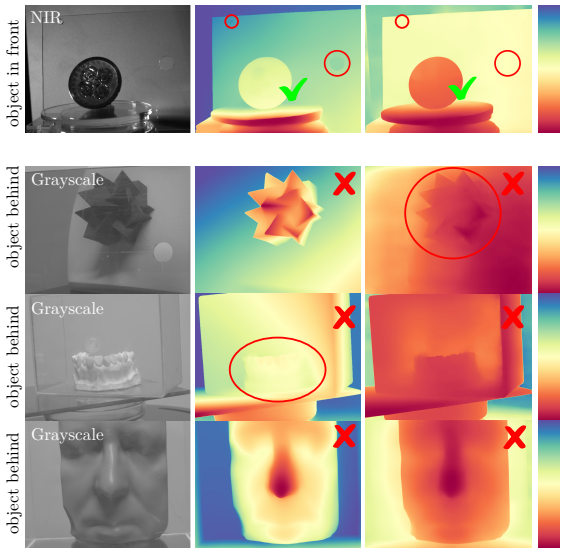

Figure 3: Objekt in front (top) and object behind transparent surface (bottom).

3 shows the challenge of a transparent polymer plate positioned behind (top) and in front of (bottom) an object. In particular, when the transparent surface lies in the foreground and object edges fall outside the image, no boundary cues are available. Thermal 3D sensors, such as the system developed by Landmann et al. [7], consistently capture the foremost surface, since this interface behaves as an optically cooperative surface.

4. Conclusion

Two selected zero-shot models deliver outstanding qualitative results for 3D detection of transparent surfaces, yet complex scenes remain challenging without fine-tuning. Future work will compare both their depth maps and resulting point clouds. Additional insights may be gained by systematically evaluating the models' performance under varying lighting and surface conditions. Future work will evaluate risk-based metrics (ISO 31000) combining failure probability and severity to assess suitability for navigation and human-robot collaboration. We suggest integrating multimodal inputs to improve transparent object perception, while AI-based depth estimation can complement physics-based sensors and emerging stereo modules.

ACKNOWLEDGMENTS

These investigations by Mr. Simon were carried out as part of a Federal Volunteer Service (FJN) program at the Technische Universität Ilmenau.

References

[1] Kaixin Bai et al. *ClearDepth: Enhanced Stereo Perception of Transparent Objects for Robotic Manipulation*. https://arxiv.org/abs/2409.08926. 2024.

[2] Xianghui Fan et al. "TDCNet: Transparent Objects Depth Completion with CNN-Transformer Dual-Branch Parallel Network". In: *arXiv: 2412.14961* (2024).

[3] Jiaqi Jiang et al. "Robotic Perception of Transparent Objects: A Review". In: *IEEE Transactions on Artificial Intelligence* (2023), pp. 1–21. DOI: 10.1109/TAI.2023.3326120.

[4] Shreeyak Sajjan et al. "Clear Grasp: 3D Shape Estimation of Transparent Objects for Manipulation". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 2020, pp. 3634–3642.

[5] Hongda Quan, Wenqi Shi, and Lingbao Kong. "Non-destructive optical measurement of transparent objects: a review". In: *Light: Advanced Manufacturing* 6.LAM2024030022 (2025), p. 333. DOI: 10.37188/lam.2025.022.

[6] Zhoujie Wu et al. "Dynamic 3D shape reconstruction under complex reflection and transmission conditions using multi-scale parallel single-pixel imaging". In: *Light: Advanced Manufacturing* 5.LAM2024020017 (2024), p. 373. DOI: 10.37188/lam.2024.034.

[7] Martin Landmann et al. "High-resolution sequential thermal fringe projection technique for fast and accurate 3D shape measurement of transparent objects". In: *Appl. Opt.* 60.8 (Mar. 2021), pp. 2362–2371. DOI: 10.1364/AO.419492.

[8] Martin Brenner et al. "RGB-D and Thermal Sensor Fusion: A Systematic Literature Review". In: *IEEE Access* 11 (2023), pp. 82410–82442.

[9] Pierluigi Zama Ramirez et al. "Open Challenges in Deep Stereo: the Booster Dataset". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. CVPR. 2022.

[10] Zhiyuan Wu et al. "Transparent Objects: A Corner Case in Stereo Matching". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 2023, pp. 12353–12359.

[11] Ju He et al. "Semi-synthesis: A fast way to produce effective datasets for stereo matching". In: *arXiv:2101.10811* (2021).

[12] Christina Junger et al. "TranSpec3D: A Novel Measurement Principle to Generate A Non-Synthetic Data Set of Transparent and Specular Surfaces without Object Preparation". In: *Sensors* 23.20 (2023). DOI: 10.3390/s23208567.

[13] Lihe Yang et al. "Depth Anything V2". In: *arXiv:2406.09414* (2024).

[14] Luca Bartolomei et al. "Stereo Anywhere: Robust Zero-Shot Deep Stereo Matching Even Where Either Stereo or Mono Fail". In: *arXiv: 2412.04472* (2025).

[15] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *arXiv:2103.00020* (2021). URL: https://arxiv.org/abs/2103.00020.

[16] Alexander Kirillov et al. "Segment Anything". In: *arXiv:2304.02643* (2023).

[17] Nikhila Ravi et al. "SAM 2: Segment Anything in Images and Videos". In: *arXiv:2408.00714* (2024).

[18] Lihe Yang et al. "Depth Anything V2". In: *arXiv:2406.09414* (2024).

[19] Aleksei Bochkovskii et al. "Depth Pro: Sharp Monocular Metric Depth in Less Than a Second". In: *arXiv:2410.02073* (2024).

[20] Bingxin Ke et al. "Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation". In: *Proceedings of the IEEE/CVF and CVPR*. 2024.

[21] Xiao Fu et al. "GeoWizard: Unleashing the Diffusion Priors for 3D Geometry Estimation from a Single Image". In: *arXiv:2403.12013* (2024).

[22] Reiner Birkl, Diana Wofk, and Matthias Müller. "MiDaS v3.1 – A Model Zoo for Robust Monocular Relative Depth Estimation". In: *arXiv: 2307.14460* (2023).

[23] Mu Hu et al. "Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-shot Metric Depth and Surface Normal Estimation". In: *arXiv:2404.15506* (2024).

[24] Xianda Guo et al. "Stereo Anything: Unifying Stereo Matching with Large-Scale Mixed Data". In: *arXiv:2411.14053* (2024).