Fitness of General-Purpose Monocular Depth Estimation Architectures for Transparent Structures

T. Wirth[1], A. Jamili[1], M. von Buelow[1], V. Knauthe[1] and S. Guthe[1,2]

[1]Technical University of Darmstadt, Germany [2]Fraunhofer IGD, Germany

Abstract

Due to material properties, monocular depth estimation of transparent structures is inherently challenging. Recent advances leverage additional knowledge that is not available in all contexts, i.e., known shape or depth information from a sensor. General-purpose machine learning models, that do not utilize such additional knowledge, have not yet been explicitly evaluated regarding their performance on transparent structures. In this work, we show that these models show poor performance on the depth estimation of transparent structures. However, fine-tuning on suitable data sets, such as ClearGrasp, increases their estimation performance on the task at hand. Our evaluations show that high performance on general-purpose benchmarks translates well into performance on transparent objects after fine-tuning. Furthermore, our analysis suggests, that state-of-the-art high-performing models are not able to capture a high grade of detail from both the image foreground and background at the same time. This finding shows the demand for a combination of existing models to further enhance depth estimation quality.

CCS Concepts

• ***Computing methodologies*** → ***Computer vision; Shape inference;***

1. Introduction

Monocular Depth Estimation (MDE) is closely linked to many computer vision tasks like 3D Reconstruction, semantic segmentation or object detection. With the rising popularity of Artificial Neural Networks (ANN), multiple data sets and architectures have been proposed to address this task. Transparent objects, due to their reflective and refractive nature, pose a challenge in computer vision in general and in MDE in particular.

Current research on transparent MDE leverages additional depth information, like known shape, structured light or depth estimates from an additional sensor. Since this information is not readily available in most uncontrolled environments, we examine existing techniques that do not require such additional information.

The main contribution of this work is a qualitative comparison of the performance of state-of-the-art MDE models on small transparent objects like bottles and introducing the problematic that fine-tuned models are unable to preserve details in the image foreground and background at the same time on images that are captured in the wild.

2. Related Work

Existing methods for MDE of transparent objects rely on additional information, such as an additional depth channel [SMP*20, ARS13, ZMX*21], structured light [HWL15, QGY16], or known object shape [KCB11, LR13, PLD16]. In contrast, multiple ANN architectures for MDE use image data exclusively. These architectures focus on general-purpose MDE, i.e., opaque objects only. Most of these approaches [AW18, GXSL18, CW19] leverage the local feature processing of Convolutional Neural Networks (CNN). Approaches using encoder-decoder architecture [AW18, FBAW21, RBK21, AW18] have been shown to have good performance in this context as well. Recent research [FBAW21, RBK21, YTD*21, CZHP21] applies model attention in terms of Vision Transformers (ViT) and achieves superior results.

3. Methodology

Architectures. We examine architectures with high performance on both the KITTI [GLSU13] and the NYUv2 [SHKF12] benchmark, i.e., LAPDepth [SLK21] and the attention-based architectures AdaBins [FBAW21] and DPTHybird [RBK21]. DenseDepth [AW18] is included as a reasonable baseline for our analysis being an early-stage encoder-decoder architecture with comparably high performance.

Data Sets. In this work, we employ the ClearGrasp [SMP*20] data set containing synthetic and real images depicting simple transparent and opaque objects with ground truth depth information. Their synthetic data set is split into a training set (*Syn-Train*) containing 49,500 annotated images of 2-5 objects and a validation set (*Syn-known*) with 500 images. Additionally they provide a synthetic test set (*Syn-novel*) that consists of 407 images of 4 transparent objects,

DOI: 10.2312/egs.20221020

that have not been used in the other data sets.
They provide a real-world test set containing 113 images (*Real-novel*). These images depict transparent objects which have not been used in training and validation sets. Example images of the test sets can be seen in Fig. 1 and Fig. 2.
In addition to the ClearGrasp data set we captured 30 images of simple transparent and opaque objects, i.e., mugs and glassware, on similar surfaces, i.e., tables. We refer to this data set as *Real-wild*.

Training. We initialize pre-existing models in the state they achieved their highest results on their respective benchmark. We further refer to them as baseline models. These models are fine-tuned using the *Syn-train* training set, further refered to as fine-tuned models.
The architectures and the training and evaluation process are implemented in PyTorch. For training we employ a learning rate according to 1-cycle [ST19] with a maximum learning rate of 10^{-5}. As a loss function we use the pixel-wise Scale-Invariant loss (SI) [EF15] with parameters chosen according to Bhat et al. [FBAW21] for all models – partially deviating from the loss functions the models are proposed with – in order to ensure a fair comparison.
The training process is terminated when the respective model's loss converges on the validation set consisting of *Syn-known* and *Real-known*.

4. Evaluation

Baseline Model Performance. The baseline models differ in the depth range they are trained to estimate from the ClearGrasp data set. Therefore, a quantitative analysis of depth estimation quality on the ClearGrasp test set is impractical. We employ a qualitative comparison instead.
Figure 1 illustrates some exemplary depth estimations of the baseline models on the ClearGrasp test sets. Note that the baseline architectures have a different depth range and our experiment only examines their near-field performance. Our investigations show, that especially DenseDepth and LAPDepth, struggle with the depth estimation of the transparent objects. In some cases, the baseline models show reasonable results (Fig. 1a). In most cases however, the objects are either not registered at all (Fig. 1a) or only the object contours are reasonably estimated (Fig. 1b).
Overall, the presented results question the suitability of the baseline models for the task at hand and therefore indicate that further fine-tuning on a dedicated data set, i.e., ClearGrasp, might be beneficial for their performance estimating the depth of transparent structures.

Fine-Tuned Model Performance. We compare the performance of our models on the ClearGrasp *Real-novel* test set based on the standard metrics used in previous work [EPF14]: percentages of pixels with predicted depth within the intervals 1.05 (δ_1), 1.10 (δ_2) and 1.25 (δ_3), the median error relative to the depth (REL), root mean squared error (RMS) and the logarithmic root mean squared error(log_{10}). Thereby, we distinguish between the performance on the complete image (Table 1) and the performance on the regions containing transparent objects (Table 2).
Our results indicate that the fine-tuned models' performance shows a similar tendency as their respective baseline models on the NYUv2 and KITTI benchmark in both evaluated scenarios. This indicates that performance on the general-purpose task is a good predictor for the performance of the fine-tuned model. The superiority of DPTHybrid and AdaBins suggest that ViT yield a relevant performance overhead for MDE of transparent objects as well.
While the results are considerably worse on transparent regions compared to the complete image, the comparative performance of the evaluated models is similar in both scenarios. A qualitative analysis (Fig. 2) further underlines the superiority of the fine-tuned DPTHybrid model, which is the only model that preserves details at the edges of transparent objects.

Real-novel	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL$\downarrow$	RMS$\downarrow$	$log_{10} \downarrow$
DenseDepth [AW18]	**0.661**	0.852	0.946	0.260	0.189	0.094
LapDepth [SLK21]	0.575	0.918	0.989	0.203	0.163	0.096
AdaBins [FBAW21]	0.384	0.828	0.954	0.254	0.189	0.128
DPTHybrid [RBK21]	0.591	**0.950**	**0.998**	**0.187**	**0.142**	**0.091**

Table 1: *Fine-tuned model performance on the ClearGrasp test set for complete image.*

Real-novel	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL$\downarrow$	RMS$\downarrow$	$log_{10} \downarrow$
DenseDepth [AW18]	**0.596**	0.691	0.767	-	0.303	1.154
LapDepth [SLK21]	0.424	0.773	0.808	-	0.231	1.147
AdaBins [FBAW21]	0.296	0.690	0.777	-	0.229	1.162
DPTHybrid [RBK21]	0.416	**0.781**	**0.811**	-	**0.219**	**1.143**

Table 2: *Fine-tuned model performance on the ClearGrasp test set for transparent image areas.*

In the Wild. Our *Real-wild* data set does not contain corresponding depth information. Therefore, we conducted a qualitative analysis of the fine-tuned models. Exemplary images with their respective depth estimates are illustrated in Fig. 3.
Our analysis shows that DenseDepth and LAPDepth tend to generate noisy depth estimates in general. The fine-tuned DPTHybrid model shows sharp estimates especially in the area of the objects' edges. These results coincide with our previous findings.
We additionally observe, that fine-tuned DPTHybrid models are not able to preserve details of the image background. In contrast to that, our fine-tuned AdaBins model shows a high level of detail in the image background, while its depiction of the objects in focus is imprecise. While this finding is not supported by our analysis on the ClearGrasp dataset, this could be caused by the relatively controlled setup in which the ClearGrasp dataset is captured. Recent research has shown the existence of a trade-off between consistent scene structure and high-level details depending on the depth map resolution [MDM*21]. Our results indicate that this trade-off differs between different MDE architectures.

5. Conclusion

In our work, we evaluated the fitness of state-of-the-art MDE architectures for images including transparent objects. Models trained on the KITTI and NYUv2 benchmark are not well suited for the this task. Nevertheless, fine-tuning these models on data sets that

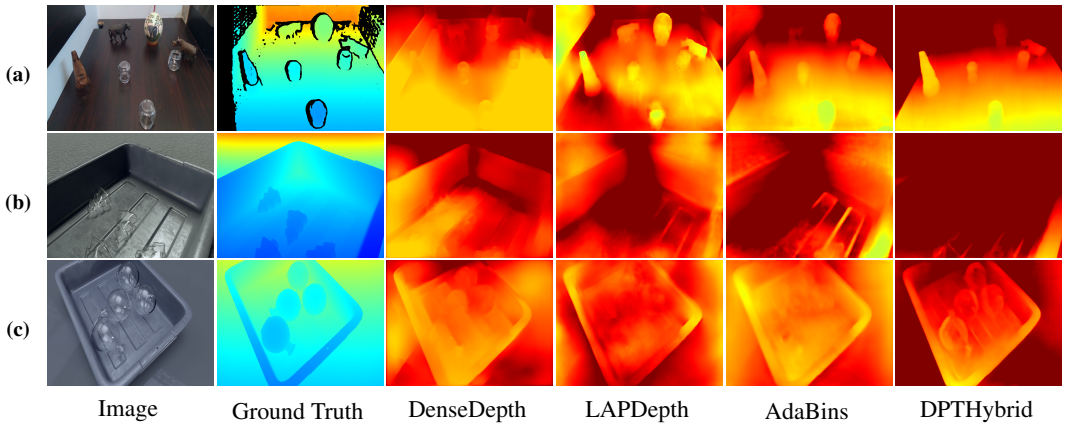

Figure 1: *Baseline models' depth estimates on exemplary images of the ClearGrasp test sets.*

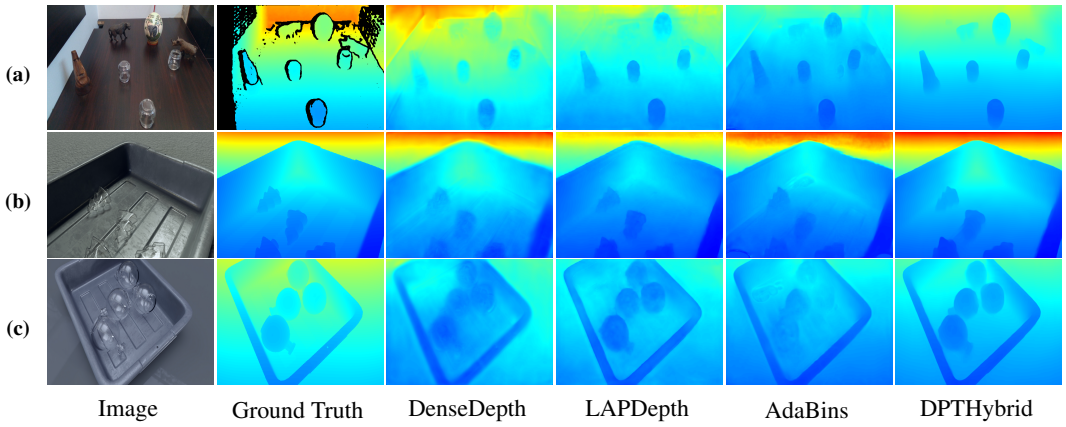

Figure 2: *Fine-tuned models' depth estimates on exemplary images of the ClearGrasp test sets.*

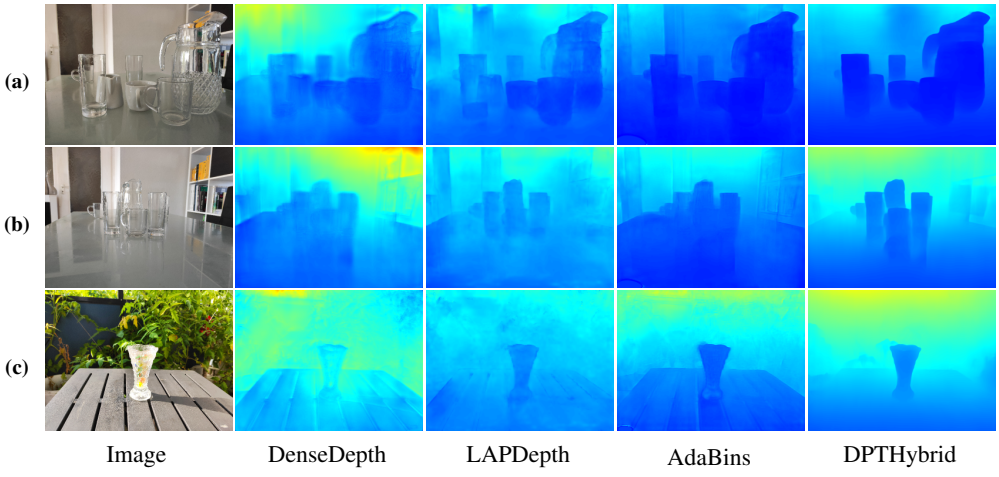

Figure 3: *Fine-tuned models depth estimates for images from our Real-wild dataset*

explicitly contain transparent structures leads to significant performance improvement. We found that a high performance on general purpose benchmarks is a good predictor for high performance of the performance of fine-tuned models for MDE of transparent structures. Out of the examined architectures, a fine-tuned version of DenseDepth showed highly superior performance.
Furthermore, our evaluation based on images captured in the wild suggests, that the examined attention-based models (AdaBins, DPTHybrid) are not able to generate depth estimates with a high grade of detail in the image foreground as well as in the image background. We therefore suggest a combination of these architectures in order to further increase the performance for scenes that include transparent structures.

Limitations and Future Work. Our examination is restricted by the limited availability of data sets that offer reliable depth information for transparent objects. Our results are restricted to a limited set of transparent objects, that are only captured in a relatively small near-field distance range, due to the use of the ClearGrasp dataset. Future research could investigate the merit of a combination of methods with high foreground and high background detail preservation for the task at hand. Furthermore, the suitability of state-of-the-art MDE models for special material properties such as reflective or planar surfaces could be investigated in the future.

Acknowledgements. The project this publication is based on has been sponsored by the German Federal Ministry of Education and Research under contract number 01IS17050. The authors are responsible for its content. Part of the research in this paper was funded by DFG project 407714161. We thank the anonymous reviewers whose comments helped improve this manuscript.

References

[ARS13] ALT N., RIVES P., STEINBACH E.: Reconstruction of transparent objects in unstructured scenes with a depth camera. In *2013 IEEE International Conference on Image Processing* (2013), IEEE, pp. 4131–4135. doi:10.1109/ICIP.2013.6738851. 1

[AW18] ALHASHIM I., WONKA P.: High quality monocular depth estimation via transfer learning. arXiv:1812.11941. 1, 2

[CW19] CHANG J., WETZSTEIN G.: Deep optics for monocular depth estimation and 3d object detection. 10192–10201. doi:10.1109/ICCV.2019.01029. 1

[CZHP21] CHEN Y., ZHAO H., HU Z., PENG J.: Attention-based context aggregation network for monocular depth estimation. *International Journal of Machine Learning and Cybernetics 12*, 6 (2021), 1583–1596. doi:10.1007/s13042-020-01251-y. 1

[EF15] EIGEN D., FERGUS R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 2650–2658. doi:10.1109/ICCV.2015.304. 2

[EPF14] EIGEN D., PUHRSCH C., FERGUS R.: Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems 27* (2014). arXiv:1406.2283. 2

[FBAW21] FAROOQ BHAT S., ALHASHIM I., WONKA P.: Adabins: Depth estimation using adaptive bins. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 4008–4017. doi:10.1109/CVPR46437.2021.00400. 1, 2

[GLSU13] GEIGER A., LENZ P., STILLER C., URTASUN R.: Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research 32*, 11 (2013), 1231–1237. doi:10.1177/0278364913491297. 1

[GXSL18] GAN Y., XU X., SUN W., LIN L.: Monocular depth estimation with affinity, vertical pooling, and label enhancement. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 224–239. doi:10.1007/978-3-030-01219-9_14. 1

[HWL15] HAN K., WONG K.-Y. K., LIU M.: A fixed viewpoint approach for dense reconstruction of transparent objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 4001–4008. doi:10.1109/CVPR.2015.7299026. 1

[KCB11] KLANK U., CARTON D., BEETZ M.: Transparent object detection and reconstruction on a mobile platform. In *2011 IEEE International Conference on Robotics and Automation* (2011), IEEE, pp. 5971–5978. doi:10.1109/ICRA.2011.5979793. 1

[LR13] LYSENKOV I., RABAUD V.: Pose estimation of rigid transparent objects in transparent clutter. In *2013 IEEE International Conference on Robotics and Automation* (2013), IEEE, pp. 162–169. doi:10.1109/ICRA.2013.6630571. 1

[MDM*21] MIANGOLEH S. M. H., DILLE S., MAI L., PARIS S., AKSOY Y.: Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 9680–9689. doi:10.1109/CVPR46437.2021.00956. 2

[PLD16] PHILLIPS C. J., LECCE M., DANIILIDIS K.: Seeing glassware: from edge detection to pose estimation and shape recovery. In *Robotics: Science and Systems* (2016), vol. 3, p. 3. doi:10.15607/RSS.2016.XII.021. 1

[QGY16] QIAN Y., GONG M., YANG Y. H.: 3d reconstruction of transparent objects with position-normal consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 4369–4377. doi:10.1109/CVPR.2016.473. 1

[RBK21] RANFTL R., BOCHKOVSKIY A., KOLTUN V.: Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 12159–12168. doi:10.1109/ICCV48922.2021.01196. 1, 2

[SHKF12] SILBERMAN N., HOIEM D., KOHLI P., FERGUS R.: Indoor segmentation and support inference from rgbd images. In *European conference on computer vision* (2012), Springer, pp. 746–760. doi:10.1007/978-3-642-33715-4_54. 1

[SLK21] SONG M., LIM S., KIM W.: Monocular depth estimation using laplacian pyramid-based depth residuals. *IEEE Transactions on Circuits and Systems for Video Technology* (2021). doi:10.1109/TCSVT.2021.3049869. 1, 2

[SMP*20] SAJJAN S., MOORE M., PAN M., NAGARAJA G., LEE J., ZENG A., SONG S.: Clear grasp: 3d shape estimation of transparent objects for manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)* (2020), IEEE, pp. 3634–3642. doi:10.1109/ICRA40945.2020.9197518. 1

[ST19] SMITH L. N., TOPIN N.: Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications* (2019), vol. 11006, International Society for Optics and Photonics, p. 1100612. doi:10.1117/12.2520589. 2

[YTD*21] YANG G., TANG H., DING M., SEBE N., RICCI E.: Transformer-based attention networks for continuous pixel-wise prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 16249–16259. doi:10.1109/ICCV48922.2021.01596. 1

[ZMX*21] ZHU L., MOUSAVIAN A., XIANG Y., MAZHAR H., EENBERGEN J. V., DEBNATH S., FOX D.: Rgb-d local implicit function for depth completion of transparent objects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 4647–4656. doi:10.1109/CVPR46437.2021.00462. 1