CSCI566 FinalReport

OpacifyDepth: Reliable Depth for Transparent Objects

Yumeng He
University of Southern California
heyumeng@usc.edu

Xiaoying Wang
University of Southern California
xwang648@usc.edu

Jingkai Shi
University of Southern California
shikings@usc.edu

Figure 1: We propose OpacifyDepth, a two-stage appearance-normalization pipeline that converts transparent objects into opaque surrogates, enabling stable monocular depth estimation without retraining depth models.

Abstract

Monocular depth estimation remains fundamentally unreliable for transparent and reflective objects, where light transmission and specular reflection violate the appearance assumptions baked into modern depth networks. State-of-the-art monocular estimators consistently fail to produce stable or meaningful depth predictions in regions containing glass, mirrors, or glossy materials. In this work, we present a two-stage solution that converts transparent-object depth estimation into a standard opaque-surface inference problem. First, we perform appearance normalization by detecting transparent and reflective regions and replacing their textures with physically plausible opaque counterparts. Second, we apply an off-the-shelf monocular depth estimator directly to the modified image without requiring any model retraining or architectural changes. We implemented multiple texture–replacement strategies, including semantic-guided matte synthesis and material-consistent inpainting, and found that they substantially restore depth stability and boundary sharpness across diverse scenes. Experiments on a hybrid

Preprint. Under review.

rendered–real dataset containing glass, plastic, and metallic objects show consistent improvements over baseline depth estimators, particularly in regions where transparency previously caused severe distortions. Qualitative evaluations further demonstrate smoother geometry, reduced boundary artifacts, and significantly improved alignment with ground-truth depth. Our results indicated that appearance normalization is an effective and generalizable front-end for handling transparent and reflective objects within existing monocular depth pipelines.

1 Introduction

Monocular depth estimation is a core problem underpinning applications in 3D reconstruction, robotic manipulation, navigation, scene understanding, and visual content generation [1, 2]. In settings involving transparent objects, accurate depth inference is particularly important for grasping, collision avoidance, and precise placement, as well as for stable real-to-simulation pipelines and automated hyperparameter tuning in robotic workflows [3–5]. While recent monocular depth models such as MiDaS [6], DPT [7], ZoeDepth [8], Depth Anything V2 [9], and Marigold [10] achieve strong performance on standard benchmarks, they continue to exhibit systematic failures when applied to scenes containing transparent or reflective materials.

These failure modes arise from fundamental optical effects. Refraction and light transmission distort the mapping between observed appearance and underlying geometry, causing background leakage and spatially inconsistent depth cues [11]. Transparent regions often lack reliable texture and shading gradients, rendering their surfaces ambiguous under photometric inference. Specular highlights, reflections, and mixed illumination further confound boundary localization, leading to fragmented geometry and unstable depth predictions [12–14]. Together, these effects violate the assumptions under which most monocular depth estimators are trained, making transparent objects a persistent and well-documented failure case.

Rather than attempting to infer depth directly through transparent media, we reframe the task as an appearance normalization problem. Our approach decomposes depth estimation into two stages: first, transparent or reflective regions are identified and their appearance is converted into geometrically consistent opaque surfaces; second, any off-the-shelf monocular depth estimator is applied to the resulting image. This decoupling bypasses the physical complexity of light transport in transparent materials and allows existing depth models to operate under their intended photometric assumptions without architectural modification or retraining.

To localize regions requiring normalization, we rely on vision–language-guided segmentation to obtain object-level masks for transparent materials. The subsequent appearance-conversion stage replaces refractive observations with plausible diffuse textures while preserving object boundaries and shading continuity. The resulting image maintains the original scene geometry but removes appearance artifacts that are incompatible with monocular depth inference.

A central challenge in studying transparent-object depth is the lack of suitable training and evaluation data. Existing datasets are often limited in scale, restricted to a narrow set of object categories, or lack reliable ground-truth depth. To address these limitations, we construct a synthetic dataset consisting of three complementary components: paired transparent–opaque object renders with ground-truth depth, additional multi-object Blender-rendered scenes designed to test generalization beyond the training distribution, and real photographs of transparent objects captured under natural conditions without depth supervision. Quantitative evaluation is conducted on a held-out synthetic test split with ground-truth depth, while qualitative evaluation focuses on both unseen synthetic scenes and real-world images to assess robustness and in-the-wild behavior.

Across all settings, the proposed preprocessing step consistently improves the depth predictions produced by existing monocular estimators. On synthetic data, the method yields more coherent geometry and cleaner object boundaries, while on real images it suppresses background interference and produces more stable depth within transparent regions.

In summary, our contributions are as follows:

- We introduce a dedicated pipeline for transparent-object depth estimation that reframes the problem through appearance normalization rather than direct inference on refractive observations.

- We propose a material-conversion module that transforms transparent regions into geometry-preserving opaque surrogates, enabling depth models to operate under well-defined photometric assumptions.
- We demonstrate that the edited images are compatible with a wide range of existing monocular depth estimators without requiring architectural changes or retraining.
- We construct paired transparent–opaque synthetic data and evaluate the method using quantitative results on a held-out synthetic test set together with qualitative case studies on unseen synthetic scenes and real photographs.

2 Related Work

2.1 Monocular Depth Estimation

Monocular depth estimation has seen rapid progress with the adoption of large-scale pretraining and global reasoning mechanisms. Early CNN-based approaches [15] first demonstrated direct depth regression from a single image, followed by transformer architectures such as DPT [16] and DINOv2-derived backbones [17], which improved long-range feature aggregation and depth consistency. More recent models including ZoeDepth [18], Marigold [19], and Depth Anything V2 [20] have further incorporated metric-relative fusion, diffusion priors, and large-scale pseudo-labeling to obtain strong performance across diverse benchmarks.

Despite these advancements, existing monocular depth models consistently fail on transparent objects. Multiple recent evaluations [21] provide explicit evidence that state-of-the-art foundation models such as DPT, ZoeDepth, and Depth Anything V2. These models produce highly unstable, distorted, or background-leaking depth in transparent regions due to refraction-induced appearance–geometry mismatch. This study demonstrates that transparent surfaces violate the photometric assumptions under which these models are trained, leading to depth collapse, discontinuities, and incorrect foreground–background separation.

These findings align with earlier observations in transparent-object geometry literature (e.g., ClearGrasp [11]), where refraction and texture scarcity are shown to remove the shading cues necessary for monocular inference. Together, these results establish transparent materials as a reproducible and well-documented failure mode for modern monocular depth estimation, motivating approaches that explicitly modify or normalize appearance before applying depth prediction.‘

2.2 Segmentation

Segmentation plays a central role across a wide range of computer vision tasks, enabling models to reason about scene structure, object boundaries, and region-specific appearance. As modern systems increasingly require object-level understanding, access to reliable segmentation masks has become fundamental. Advances in both semantic segmentation and instance segmentation have made it possible to obtain accurate masks even for objects with complex geometry or visual ambiguity.

Within the domain of transparent objects, segmentation has received increasing attention. Trans2Seg [22] proposes a Transformer-based architecture specifically designed to distinguish transparent materials and introduces the Trans10K dataset for benchmarking. Trans4Trans [23] extends this direction with a dual-head Transformer tailored for real-world navigation and robotic manipulation. Beyond segmentation alone, several works incorporate geometric cues: RFTrans [24] estimates refractive flow to recover surface normals, whereas sensor-based techniques [25] combine refraction modeling with photometric constraints for transparent-surface reconstruction.

These studies highlight the critical role of segmentation and material-aware processing in perceiving transparent objects. However, existing approaches often require specialized training, dataset-specific supervision, or strong geometric priors. Our framework takes a different direction by leveraging general-purpose vision–language segmentation models. Because our method does not depend on transparent-object datasets for training, we use them solely for evaluation to assess cross-dataset generalization.

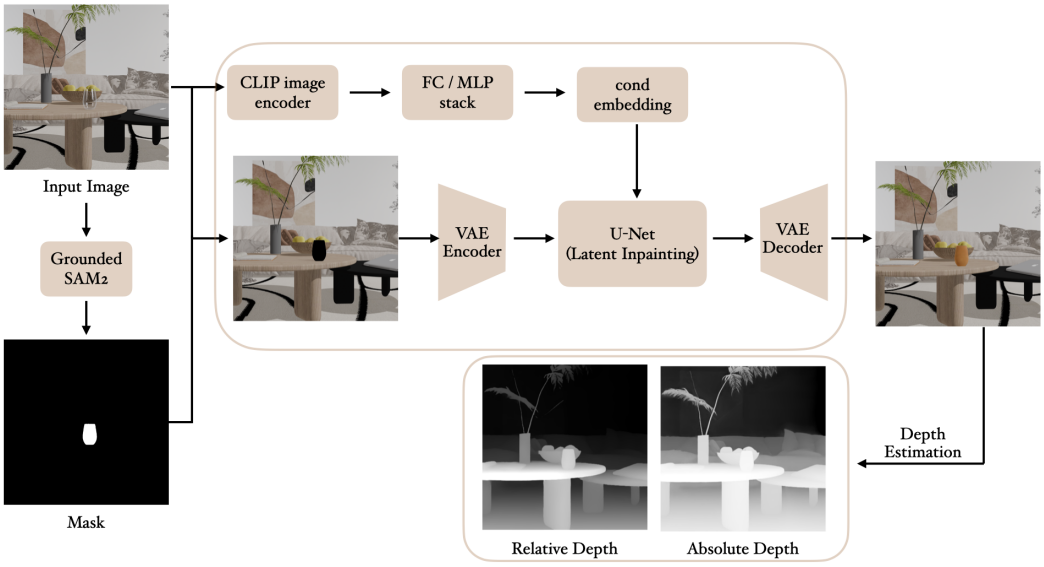


Figure 2: Overview of the OpacifyDepth framework. The image editing network converts transparent regions into opaque surfaces, enabling monocular depth estimators to produce accurate depth in transparent-object scenes.

2.3 Image Inpainting and Material Editing

Image inpainting, intrinsic decomposition, and material editing are closely related to our goal of transforming transparent regions into opaque, geometrically consistent surfaces. Prior work in material manipulation seeks to disentangle or adjust surface appearance so that images better reflect physically plausible properties. Careaga et al. [26] decompose images into intrinsic albedo and shading, reducing ambiguity introduced by complex illumination. Materialist [27] investigates learned transformations between transparent and opaque materials, enabling controllable edits under different lighting and viewpoint conditions. More recent diffusion-based approaches such as Alchemist [28] allow parametric modification of reflectance, roughness, and translucency while maintaining visual coherence.

Although these studies do not specifically address ambiguity introduced by transparent materials, they nevertheless motivate the idea that modifying surface appearance can help reduce certain forms of visual uncertainty. Inspired by this broader perspective, our method applies a geometry-preserving opaque substitution targeted to transparent regions, producing edited images that better conform to the visual assumptions underlying standard monocular depth estimators.

3 Method

3.1 Overview

Our method, **OpacifyDepth**, reformulates the challenge of depth estimation for transparent objects as an image editing problem. Formally, denote the transparent input image as $I^{tr} \in \mathbb{R}^{H\times W\times 3}$, where H and W are the height and width. The edit region is represented by a binary mask $M^{tr} \in \{0,1\}^{H\times W}$ where value 1 specifies the transparent object pixels. Our goal is to synthesize an edited image I'^{op} from $\{I^{tr}, M^{tr}\}$, so that the masked region exhibits opaque appearance while the background remains unchanged. The edited image is then processed by a frozen monocular depth estimator to obtain depth predictions.

3.2 Data Construction

Training Dataset Generation. We render paired transparent and opaque images that share identical geometry, camera pose, and illumination, as illustrated in Fig. 3. Rendering is performed in Blender

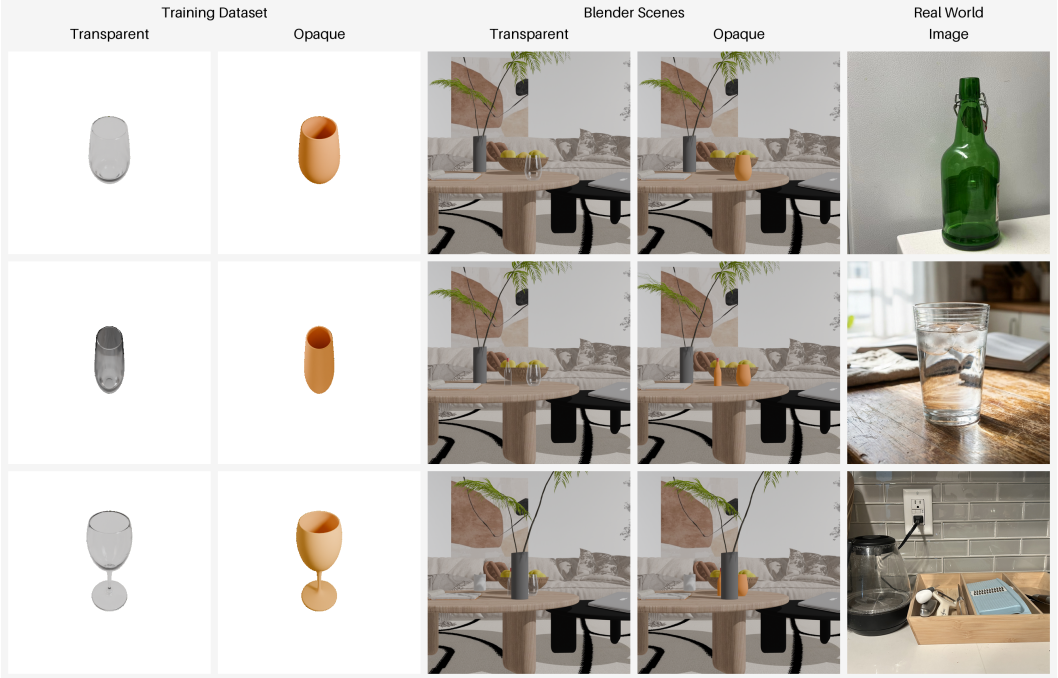


Figure 3: Overview of the datasets used in this work. We include a synthetic training set with paired transparent and opaque objects, Blender-rendered scenes under controlled conditions, and real-world images of transparent objects captured in the wild.

with randomized HDR environment maps to increase visual diversity. The transparent view I^{tr} is generated by varying material parameters including transmission, index of refraction, thickness, and micro-roughness. The opaque target I^{op} replaces the transparent material with a uniform metallic BRDF chosen to preserve clear geometric shading cues. These paired renders provide well defined supervision for material translation, while also serving as controlled inputs for evaluating depth inference under matched geometric conditions.

Normalization of 2D Image Crops. To reduce intra-dataset variation and eliminate scale-related bias, we normalize all object crops to a unified canonical resolution. For each rendered object, we compute its axis-aligned bounding box and take the longer side as the reference scale. The cropped region is then resized so that every object has the same normalized extent regardless of its original physical size or camera distance. This ensures that the editing module receives consistent spatial statistics across training samples, preventing the network from learning spurious correlations tied to object scale or framing rather than transparency-specific appearance cues.

Additional Non-Supervised Images. Beyond the paired supervisory dataset, we collect an additional set of Blender-rendered multi-object scenes (xxx images) and a set of real photographs captured in indoor environments (xxx images). These images do not contain ground truth depth but are used to evaluate robustness and generalization. The multi-object Blender scenes allow testing in cluttered compositions with complex occlusion patterns, while the real images demonstrate that the appearance-normalization module produces stable edits on in-the-wild inputs, even when lighting, textures, or background statistics differ significantly from the training distribution.

3.3 Image Editing Network

The core of our framework is an image editing network designed for opacification. This network performs a guided inpainting task, leveraging a conditional diffusion model adapted from Paint by Example [29]. The model operates in the latent space of a pretrained VAE, consisting of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. The architecture is designed to integrate two primary sources of information: spatial context from the background and semantic guidance from the transparent object itself.

Architectural Components. The network's central component is a U-Net denoiser. Its design incorporates the following key elements:

- **Multi-Channel Latent Input:** To provide explicit spatial context and define the synthesis region, the U-Net takes a 9-channel latent input. This input concatenates the noisy opaque latent z_t, the encoded background from the masked transparent input z_{bg}, and the downsampled mask $\bar{M}^{tr}$:
$$z_{input} = [z_t, z_{bg}, \bar{M}^{tr}]. \tag{1}$$
- **Semantic Conditioning:** To ensure the synthesized object is semantically consistent with the original, the network is conditioned on high-level features from the transparent input I^{tr}. A frozen CLIP image encoder extracts a class token, which is then projected by an MLP to obtain a conditioning vector c. This vector is injected into the U-Net via cross-attention layers to guide the generation process.

Training. The network is trained end-to-end using the standard DDPM objective [30]. The objective is to train the U-Net, ϵ_θ, to predict the noise added to the clean opaque latent z, conditioned on the multi-channel input and semantic vector c. During inference, this process is reversed: starting from a random Gaussian noise latent z_T, the model iteratively denoises it to produce the final opacified latent z_0, which is then decoded to the edited image $I'^{op} = \mathcal{D}(z_0)$.

3.4 Inference Pipeline

The end-to-end inference pipeline consists of three main stages: pre-processing, core model inference, and post-processing. First, in the pre-processing stage, inputs are prepared for the editing network. For a given test image, Grounded-SAM2 generates a segmentation mask for the target transparent object. A square crop containing the object is then extracted and resized to 512×512. Next, the prepared crop and mask are fed into the trained image editing network for the core inference step, which produces an opacified crop. This model inference process is illustrated in Fig. 2. Finally, in the post-processing stage, the generated crop is seamlessly integrated back into the full image to form the final output. Mask-based blending is applied to preserve background fidelity:
$$I'^{op}_{crop,blended} = M^{tr} \odot I'^{op}_{crop} + (1 - M^{tr}) \odot I^{tr}_{crop}. \tag{2}$$
The blended crop is placed back at its original location, and the resulting image I'^{op} is fed to a frozen Depth Anything 3 model for final depth prediction.

4 Experimental Evaluation

This section evaluates the proposed image editing strategy for improving monocular depth estimation on transparent objects. The experimental setup and evaluation protocol are described first, followed by quantitative analysis on the rendered test split and qualitative comparisons. Additional controlled studies and real-world editing attempts are provided in Appendix 7.2.

4.1 Experimental Setup

Rendered dataset. Evaluation is conducted on a Blender-rendered dataset comprising 1530 images, partitioned into 1500 training samples and 30 test samples. Each test sample provides: (i) a transparent input image I^{tr} containing transparent objects, (ii) a paired opaque rendering I^{op} with identical geometry and camera parameters, (iii) a binary object mask M defining the target region, and (iv) a reference depth map D^{gt} obtained from Blender's Z-pass rendering.

Compared settings. Depth Anything 3 (DA3; `depth-anything/DA3NESTED-GIANT-LARGE-1.1`) serves as the depth estimation backbone for settings (i), (ii), and (iv). Four configurations are compared:

- **Transparent baseline**: DA3 applied to the transparent input I^{tr}.
- **Color-filling baseline**: Masked region filled with solid color prior to DA3 inference.
- **Depth4ToM** [31]: A specialized method for transparent-object depth estimation.
- **OpacifyDepth**: DA3 applied to the edited image $I'^{op} = T_\theta(I^{tr}, C)$.

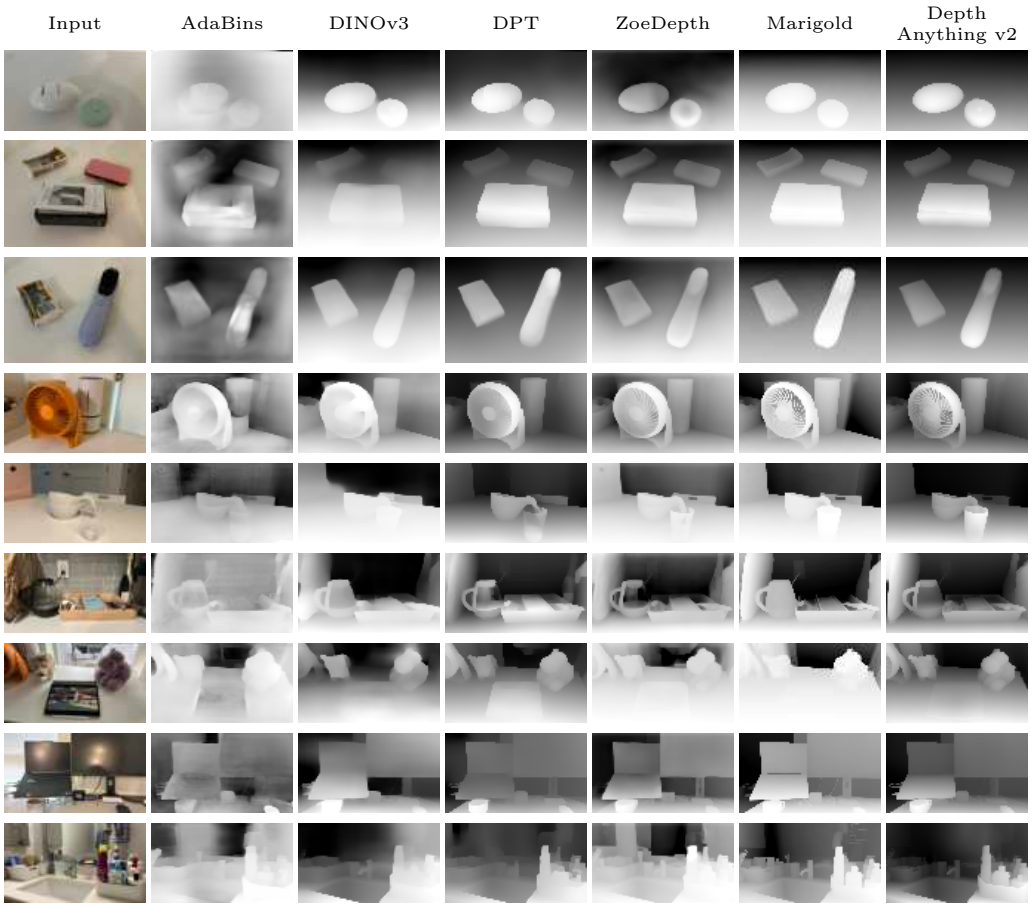


Figure 4: Qualitative comparison of monocular depth on tabletop scenes. Each row shows one RGB scene (left) and the corresponding depth predictions from six off-the-shelf baselines: AdaBins, DINOv3, DPT, ZoeDepth, Marigold, and Depth Anything V2. The rows are curated to stress common tabletop failure modes such as fine-scale geometry, transparent materials, pictorial 3D, and strong cast shadows.

Setting	**RMSE** ↓	**MAE** ↓	**AbsRel** ↓
Transparent baseline	0.8992	0.8398	0.5155
Color-filling baseline	0.9756	0.8989	0.5798
Depth4ToM [31]	1.0880	0.9721	0.8911
OpacifyDepth	**0.8499**	**0.7823**	**0.3960**

Table 1: Masked-region depth evaluation on the 30-image rendered test set. Lower values indicate better performance (↓).

4.2 Quantitative Analysis

Evaluation protocol. To quantitatively evaluate performance, three standard depth estimation metrics are computed exclusively within the target region specified by the binary mask M: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Absolute Relative Error (AbsRel). To enable fair comparison with Depth4ToM [31], which produces affine-invariant predictions, a single global scale-and-shift alignment (s, t) is estimated via least squares over all masked pixels. The metrics for Depth4ToM are then computed using the aligned prediction.

Table 1 reports masked-region depth errors on the 30-image test split. OpacifyDepth achieves the lowest error across all three metrics and improves upon the transparent baseline. The color-filling baseline underperforms the transparent baseline on this test set. Depth4ToM yields higher errors under the global affine alignment protocol.

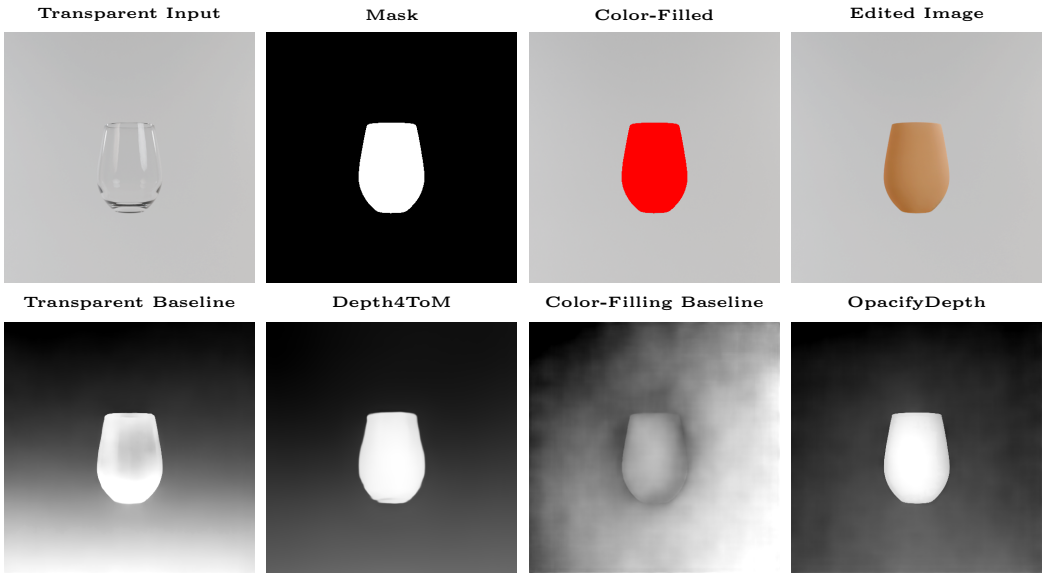


Figure 5: Qualitative comparison across evaluated settings. Top: transparent input, object mask, color-filled input, and edited image. Bottom: depth predictions from Transparent Baseline, Depth4ToM, Color-Filling Baseline, and OpacifyDepth.

4.3 Qualitative Analysis

Off-the-shelf Baseline Comparison. To evaluate how existing models handle transparency, the performance of six off-the-shelf monocular depth baselines is evaluated on complex, real-world tabletop scenes. The results are presented in Figure 4. The visualization reveals a common failure mode: most estimators fail on transparent surfaces, resulting in "see-through" artifacts. A notable exception is Marigold, which shows some resilience to transparency but introduces geometric inaccuracies in opaque regions of the scene. This finding indicates that transparency remains a significant challenge for current state-of-the-art models.

Comparison across compared settings. A direct comparison of the four **compared settings** is conducted on samples from the rendered test set, as shown in Figure 5. The transparent baseline exhibits a prominent see-through artifact, where the depth estimator fails to perceive the glass as a solid surface and instead assigns it the depth of the background. The color-filling baseline produces a noticeably noisier depth map in this example, which may be attributed to the sparse surrounding context and a potential domain mismatch. While the Depth4ToM method successfully mitigates the see-through issue, its predicted object boundaries do not align perfectly with the object's silhouette. In contrast, OpacifyDepth not only resolves the see-through problem but also yields object boundaries that are more accurately aligned. This demonstrates that the proposed appearance normalization is a valid strategy, enabling a frozen depth estimator to achieve leading performance without any fine-tuning.

Generalization to Complex Scenes. The generalization of OpacifyDepth is evaluated on complex rendered scenes with multiple transparent objects, which are not part of the test set. Figure 6 visualizes the results. The transparent baseline exhibits see-through artifacts within the masked regions (red boxes), assigning background depth to transparent surfaces. The normalized masked-region depth reveals the depth inconsistencies. OpacifyDepth resolves the see-through artifacts and produces consistent depth estimates within the masked regions, indicating that the method can generalize to more complex scenes and still outperform the baseline.

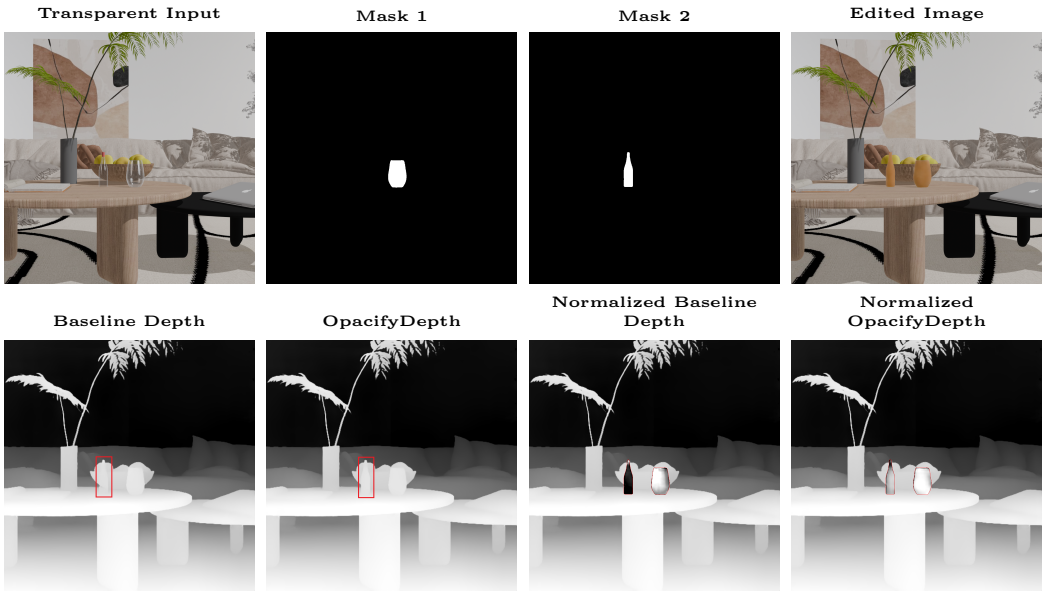


Figure 6: From left to right - Top: transparent input I^{tr}, object masks (Mask 1 and Mask 2), and edited image I'^{op} generated by the image editing network. Bottom: depth by Depth Anything 3 on transparent input, depth by Depth Anything 3 on edited image, normalized masked-region depth from transparent input, and normalized masked-region depth from edited image. Red boxes highlight the left transparent object where the transparent baseline exhibits see-through artifacts.

5 Conclusion

This study examined how current state-of-the-art monocular depth estimators behave in the presence of transparent objects. Through preliminary experiments and an analysis of existing datasets and models, we identified a consistent failure pattern: transparent regions cause depth collapse, unstable boundary predictions, and background leakage across multiple architectures. These observations indicated that the dominant source of error is not geometric complexity or clutter, but the optical ambiguities introduced by refraction and transmission.

The experimental findings further showed that even under controlled synthetic conditions, where geometry, illumination, and noise are eliminated, transparent objects remain a systematic failure mode. Combined with an assessment of transparent-object benchmarks, we found that many available datasets rely heavily on synthetic repetition and limited shape diversity, constraining generalization to real-world materials and optical properties.

Overall, the study supports two key conclusions. First, transparent objects represent a reproducible and architecture-independent failure case for modern monocular depth estimation methods, consistently degrading depth quality across different model families. Second, these failures are closely tied to biased data distributions and a fundamental mismatch between appearance and geometry in transparent regions, which together limit generalization and motivate a shift away from direct depth regression toward appearance-normalization-based solutions.

6 Discussion

This project was initially motivated by the importance of depth estimation across a wide range of computer vision and robotics applications, including 3D reconstruction, manipulation, and scene understanding. Over the project, we deliberately chose to explore a perception problem beyond large language models in order to engage with visual reasoning and geometry-driven challenges. At the outset of the project, the primary difficulty was expected to be depth estimation in cluttered tabletop scenes with complex object interactions. Instead, preliminary experiments revealed that state-of-the-art monocular depth estimators handled clutter and occlusion reliably, while transparent

objects consistently produced the most severe errors. This contrast reshaped the research direction from one centered on geometric complexity to one dominated by optical ambiguity. Appendix 7.1 provides a detailed account of this transition and the initial framework that motivated the shift.

Building on this revised understanding, our analysis examined the persistent challenges that transparent objects pose for modern monocular depth estimators. Across baseline models, transparent materials consistently disrupted the appearance and geometry relationship assumed by monocular prediction, producing depth collapse, boundary distortion, and background leakage. The consistency of these failure modes across synthetic renders, real images, and multiple architectures suggests that the limitation is fundamental rather than specific to a particular model design or training strategy.

Early attempts to address these issues focused on post-hoc depth refinement by integrating semantic reasoning and geometric priors. Although this initial framework provided a principled way to incorporate boundary constraints, occlusion handling, and confidence modeling, it encompassed several research-level subproblems that each required dedicated solutions. In practice, preliminary experiments showed that simple gradient-based edge alignment was unreliable, as image edges frequently diverged from true geometric discontinuities. Developing a unified model capable of robustly resolving boundary refinement, occlusion recovery, and confidence estimation proved computationally expensive and theoretically unstable within the scope of this project.

As a result, the research scope was narrowed to a more tractable but still challenging domain: transparent objects. These objects represent a dominant failure mode for monocular depth estimation due to refraction and transmission ambiguities. Rather than continuing to correct depth predictions after inference, the problem was reframed as one of input preprocessing, emphasizing that the root cause of failure lies in RGB appearance rather than in the depth regression process itself. This shift led to the current appearance-normalization and material-replacement framework described in Section 3, which translates transparent appearances into opaque counterparts prior to depth inference.

Motivated by these observations, the proposed framework replaces refractive transparent regions with geometrically consistent opaque surrogates, producing inputs that better conform to the inductive biases of existing monocular depth estimators. By keeping the downstream depth models frozen, the approach remains modular and compatible with a broad class of architectures, allowing improvements without retraining or architectural modification.

6.1 Future Work

Future work will focus on completing quantitative evaluation on public transparent-object benchmarks and validating the approach under real-world lighting conditions, material variation, and scene complexity. An important direction is to analyze how optical factors such as refractive index, object thickness, and internal reflections influence the effectiveness of appearance normalization. Another priority is strengthening the material replacement model to ensure faithful and geometry-preserving transformations across a broader range of transparent objects.

Longer-term extensions include integrating the preprocessing module into robotic perception and 3D reconstruction pipelines, evaluating cross-model consistency among different monocular depth estimators, and exploring transparency-aware data augmentation strategies. Together, these directions aim to clarify whether the observed limitations arise primarily from architectural constraints or from biased training distributions that dominate current monocular depth pipelines.

References

[1] Richard A. Newcombe, Shahram Izadi, et al. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE Int. Symp. on Mixed and Augmented Reality (ISMAR)*. doi: 10.1109/ISMAR.2011.6092378. URL https://doi.org/10.1109/ISMAR.2011.6092378.

[2] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*. URL https://openaccess.thecvf.com/content_cvpr_2016/papers/Schonberger_Structure-From-Motion_Revisited_CVPR_2016_paper.pdf.

[3] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, et al. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. In *Robotics: Science and Systems (RSS)*. URL https://arxiv.org/abs/1703.09312.

[4] Andreas ten Pas, Marcus Gualtieri, Kate Saenko, and Robert Platt. Grasp pose detection in point clouds. URL https://arxiv.org/abs/1706.09911.

[5] Tomas Hodan et al. Bop challenge 2020 on 6d object localization. In *ECCV Workshops*. URL https://arxiv.org/abs/2009.07378.

[6] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. . URL https://arxiv.org/abs/1907.01341.

[7] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, . URL https://arxiv.org/abs/2103.13413.

[8] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Zoedepth: Zero-shot transfer by combining relative and metric depth. URL https://arxiv.org/abs/2302.12288.

[9] Lintu Yang et al. Depth anything v2. URL https://arxiv.org/abs/2406.09414.

[10] Bingxin Ke, Alexander Becker, Konrad Schindler, and Anton Obukhov. Repurposing diffusion-based image generators for monocular depth estimation (marigold). In *CVPR*. URL https://arxiv.org/abs/2312.02145.

[11] Shreeyak S. Sajjan et al. Cleargrasp: 3d shape estimation of transparent objects for manipulation. In *ICRA*. URL https://arxiv.org/abs/1910.02550.

[12] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, . URL https://openaccess.thecvf.com/content_cvpr_2017/papers/Godard_Unsupervised_Monocular_Depth_CVPR_2017_paper.pdf.

[13] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, . URL https://openaccess.thecvf.com/content_ICCV_2019/papers/Godard_Digging_Into_Self-Supervised_Monocular_Depth_Estimation_ICCV_2019_paper.pdf.

[14] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *ECCV*. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123640562.pdf.

[15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014. URL https://arxiv.org/abs/1406.2283.

[16] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction, 2021. URL https://arxiv.org/abs/2103.13413.

[17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

[18] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. URL https://arxiv.org/abs/2302.12288.

[19] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation, 2024. URL https://arxiv.org/abs/2312.02145.

[20] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024. URL https://arxiv.org/abs/2406.09414.

[21] Ke Ma, Yizhou Fang, Jean-Baptiste Weibel, Shuai Tan, Xinggang Wang, Yang Xiao, Yi Fang, and Tian Xia. Phys-liquid: A physics-informed dataset for estimating 3d geometry and volume of transparent deformable liquids, 2025. URL https://arxiv.org/abs/2511.11077.

[22] Enze Xie, Wenjia Wang, Wenhai Wang, Peize Sun, Hang Xu, Ding Liang, and Ping Luo. Segmenting transparent object in the wild with transformer, 2021. URL https://arxiv.org/abs/2101.08461.

[23] Jiaming Zhang, Kailun Yang, Angela Constantinescu, Kunyu Peng, Karin Müller, and Rainer Stiefelhagen. Trans4trans: Efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world, 2021. URL https://arxiv.org/abs/2107.03172.

[24] Tutian Tang, Jiyu Liu, Jieyi Zhang, Haoyuan Fu, Wenqiang Xu, and Cewu Lu. Rftrans: Leveraging refractive flow of transparent objects for surface normal estimation and manipulation. *IEEE Robotics and Automation Letters*, 9(4):3735–3742, April 2024. ISSN 2377-3774. doi: 10.1109/lra.2024.3364837. URL http://dx.doi.org/10.1109/LRA.2024.3364837.

[25] Chi Xu, Jiale Chen, Mengyang Yao, Jun Zhou, Lijun Zhang, and Yi Liu. 6dof pose estimation of transparent object from a single rgb-d image. *Sensors*, 20(23), 2020. ISSN 1424-8220. doi: 10.3390/s20236790. URL https://www.mdpi.com/1424-8220/20/23/6790.

[26] Chris Careaga and Yağız Aksoy. Colorful diffuse intrinsic image decomposition in the wild. *ACM Transactions on Graphics*, 43(6):1–12, November 2024. ISSN 1557-7368. doi: 10.1145/3687984. URL http://dx.doi.org/10.1145/3687984.

[27] Lezhong Wang, Duc Minh Tran, Ruiqi Cui, Thomson TG, Anders Bjorholm Dahl, Siavash Arjomand Bigdeli, Jeppe Revall Frisvad, and Manmohan Chandraker. Materialist: Physically based editing using single-image inverse rendering, 2025. URL https://arxiv.org/abs/2501.03717.

[28] Prafull Sharma, Varun Jampani, Yuanzhen Li, Xuhui Jia, Dmitry Lagun, Fredo Durand, William T. Freeman, and Mark Matthews. Alchemist: Parametric control of material properties with diffusion models, 2023. URL https://arxiv.org/abs/2312.02970.

[29] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18381–18391, 2023.

[30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[31] Alex Costanzino, Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Learning depth estimation for transparent and mirror surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9244–9255, 2023.

7 Appendix

7.1 Initial Framework and Scope Transition

7.1.1 Overview of the Initial Framework

The early investigation formulated tabletop depth refinement as a single-image, depth-to-depth learning problem. Given an RGB image $I \in \mathbb{R}^{H\times W\times 3}$ without known intrinsics, the objective was to predict a refined depth map $\hat{D} \in \mathbb{R}^{H\times W}$ that corrects structural artifacts and stabilizes scale. The pipeline was designed in two stages: (i) pseudo ground-truth construction and (ii) transformer-based depth refinement.

7.1.2 Stage 1: Pseudo Ground-Truth Construction

Stage 1 aimed to synthesize a high-quality pseudo ground truth $D^\star$ and a per-pixel confidence map $C \in [0,1]^{H\times W}$ for supervision, together with instance masks $\{M_k\}_{k=1}^K$. Semantic guidance was provided by a vision–language model (VLM). The VLM produced a structured scene description including object categories, materials, coarse locations, and semantic flags (e.g., mirrors, screens, transparent surfaces). This semantic pass identified unreliable regions for geometric supervision and suggested geometry priors such as planar, piecewise planar, cylindrical, spherical, or thin-plate surfaces to regularize local fitting.

Instance proposals were generated by a text-conditioned detector and refined into clean masks. Low-confidence or semantically inconsistent proposals were filtered using text–image similarity checks. Each accepted mask was morphologically cleaned, and a signed-distance transform defined inner and outer boundary bands for localized boundary operations.

Two complementary monocular predictors provided initial depth estimates. After log-depth normalization to $[0,1]$, a fused depth $D^{(0)}$ was formed by per-pixel weighting that preferred sharper edges where predictors disagreed, while a variance map V captured local disagreement. Boundary fattening and occlusion ambiguities were mitigated using band-limited, edge-aware filtering guided by image gradients. Across mask boundaries, plausible near–far depth transitions were enforced, preferring the lower-variance hypothesis.

Within each instance M_k, the semantically suggested geometry was fitted by robust estimation (e.g., RANSAC). Residuals $R_k = D^{(0)} - D_k^{\text{model}}$ were denoised using instance-restricted total variation or anisotropic diffusion to preserve thin ridges. Small instances were temporarily upsampled to prevent oversmoothing, and missing regions were filled by in-instance interpolation or Poisson/Laplacian inpainting with ridge preservation. At the scene level, instance reconstructions were fused, resolving overlaps via residual and variance comparisons. When a tabletop plane was detected, planar support and non-penetration constraints were applied at contact bands, followed by edge-aware harmonization near seams.

The confidence map combined geometric residuals, boundary proximity, estimator variance, hole-fill ratios, and semantic uncertainty. For pixel x,

$$C(x) = \sigma(-\alpha r(x) + \beta d(x) - \gamma v(x) - \delta h(x) - \eta s(x)), \tag{3}$$

and $C(x) = 0$ in ignored regions. Quality checks on boundary thickness, contact consistency, and abnormal residual variance triggered local fallbacks when inconsistencies were detected. The outputs of Stage 1 were $(D^\star, C, \{M_k\})$ and corresponding semantics.

7.1.3 Stage 2: Transformer-Based Depth Refinement

Stage 2 trained a compact transformer f_ϕ that refined noisy depth maps from arbitrary monocular estimators. The mapping $f_\phi : \mathbb{R}^{H\times W} \to \mathbb{R}^{H\times W}$ produced $\hat{D} = f_\phi(D)$ to approximate $D^\star$. The architecture was an encoder–decoder transformer operating on non-overlapping $p \times p$ patches in log-depth space. Each patch was flattened, linearly projected into a token, and augmented with 2D sine positional encodings. The backbone consisted of L layers of windowed multi-head self-attention with periodic global tokens to propagate long-range context for consistent scale and symmetry. An optional instance raster and distance-to-edge map could be concatenated to the embeddings. A convolutional head predicted a residual R in log-depth, yielding $\hat{D} = D + R$, which preserved coarse geometry while emphasizing boundary corrections.

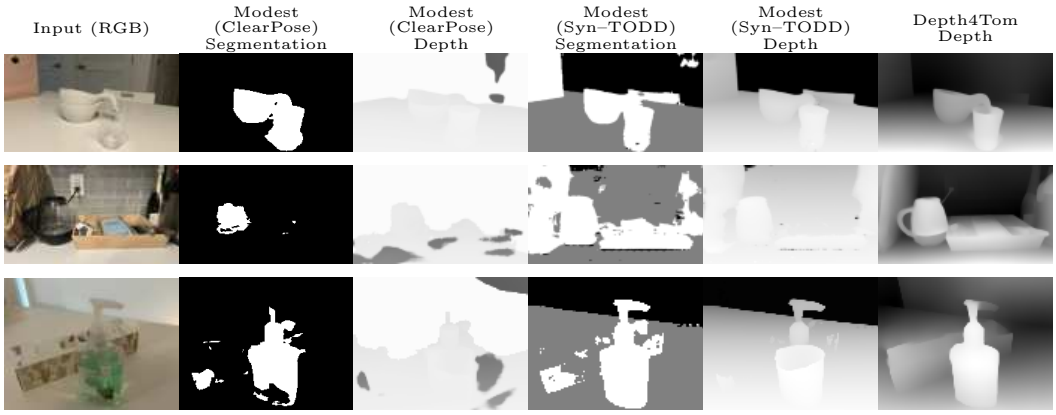
The loss function balanced scale-invariant depth fidelity, edge sharpness, and surface normal alignment:

$$\ell(x) = \lambda_{\text{si}}\ell_{\text{SILog}}\big(\hat{D}(x), D^\star(x)\big) + \lambda_\nabla \|\nabla\hat{D}(x) - \nabla D^\star(x)\|_1 + \lambda_{\text{n}}(1 - \langle \hat{n}(x), n^\star(x)\rangle), \tag{4}$$

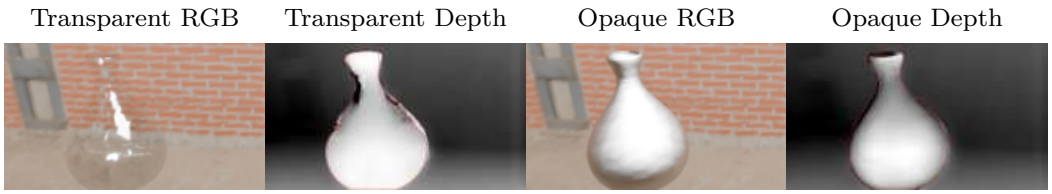
and the global objective was

$$\mathcal{L} = \sum_x w(x)\,\ell(x), \quad w(x) = C(x)\,b(x)\,s_{k(x)}. \tag{5}$$

Here $b(x)$ emphasized boundary bands to enforce crisp discontinuities, while $s_k = 1/\sqrt{\text{area}(M_k)}$ compensated for bias against small objects. Standard photometric and geometric augmentations were



(a) Comparison between Modest and Depth4ToM on real-world transparent-object scenes.



(b) Synthetic Blender experiment demonstrating material replacement effects on depth prediction.

Figure 7: (a) Modest often produces inconsistent depth boundaries and can neglect distant opaque objects, while Depth4ToM yields more continuous backgrounds but may exhibit discontinuities at mixed-material junctions. (b) In controlled Blender experiments, transparent glass objects result in blurred boundaries and distorted internal gradients in depth predictions, while replacing transparent materials with opaque metallic materials yields sharper boundaries and more spatially coherent geometry. Mask boundaries are shown in red.

applied during pseudo-label generation, and mild perturbations in depth space improved robustness to upstream estimator noise.

At inference, the transformer accepted a single noisy depth map and produced a refined map in one forward pass, requiring no intrinsics, semantics, or masks. This design yielded sharper boundaries, stable global scale, and accurate small-object reconstruction, functioning as a plug-and-play refinement module for generic monocular depth predictors.

7.1.4 Scope Transition and Motivation for the Current Work

Although the above framework effectively integrated semantic reasoning and geometric priors, it encompassed several research-level subproblems—boundary refinement, occlusion recovery, and confidence modeling as each requiring dedicated solutions. Preliminary experiments showed that simple gradient-based edge alignment was unreliable, as image edges often diverge from true geometric discontinuities. Developing a unified model to robustly handle all these cases was computationally expensive and theoretically unstable.

The research scope was therefore narrowed to a more tractable but still challenging domain: transparent objects. These represent a dominant failure mode for monocular estimators due to refraction and transmission ambiguities. The problem was reframed from post-hoc depth correction to input preprocessing, emphasizing that the root cause lies in RGB appearance rather than in depth regression itself. This shift led to the current material-replacement framework described in Section 3, which translates transparent appearances into opaque counterparts before depth inference.

7.2 Additional Analyses

Qualitative Evaluation of Baseline Monocular Depth Estimation Models on Challenging Scenes
The qualitative evaluation was conducted on self-collected real-world tabletop scenes containing a

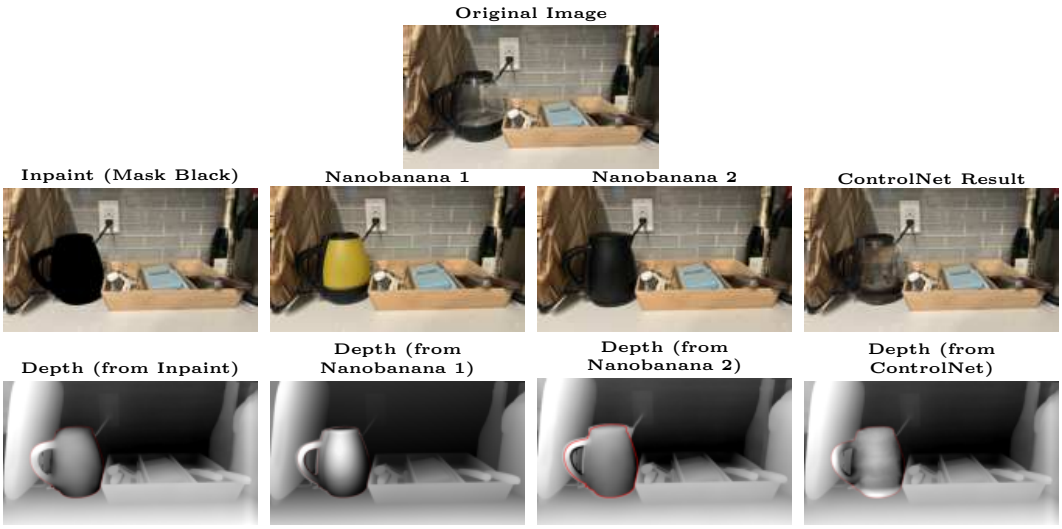


Figure 8: Comparison of real-world transparency-to-opaque editing results. Top: original image; middle: mask-filling, two Nanobanana edits generated with different prompts, and a ControlNet edit; bottom: corresponding depth predictions. Depth maps are normalized within transparent-object masks, with mask boundaries highlighted in red.

diverse mix of transparent, opaque, reflective, and cluttered objects, representing a broad range of challenging optical and geometric conditions.

As shown in Figure 4, the qualitative assessment reveals a clear performance hierarchy and common failure modes. The CNN-based AdaBins performs the weakest, producing severely blurred contours. Foundation models such as DINOv3 yield smooth reconstructions but tend to be overly flat. The Transformer-based DPT and ZoeDepth produce more coherent results, with DPT excelling at local details while ZoeDepth offers better global consistency. The Stable Diffusion-based generative model, Marigold, provides highly detailed depth maps and handles transparent regions more reasonably than other models, but can introduce geometric hallucinations such as disproportionate object scaling or missing small objects. Depth Anything V2 consistently delivers state-of-the-art performance in terms of detail and sharpness on non-transparent objects but sometimes fails on transparent or reflective regions.

These findings indicated that all tested models still struggle with fine-scale geometry, and transparent surfaces, motivating the development of our proposed image editing framework.

Evaluation of Specialized Methods for Transparent-Object Depth Estimation The **Modest** model exemplifies a co-training approach, jointly predicting depth and segmentation in an end-to-end framework. Its architecture uses a visual encoder and an iterative fusion decoder, where semantic and geometric features are progressively refined through shared-weight gates to allow the two tasks to mutually inform each other.

In our experiments, both official pre-trained versions of the model showed limited generalization on diverse real-world scenes, producing generally inaccurate depth boundaries. The model trained on the purely synthetic dataset performed slightly better, suggesting that the co-training strategy may cause the model to over-focus on transparent objects and their immediate surroundings, at the expense of ignoring non-transparent objects located further away in the scene.

The **Depth4Tom** model utilizes a monocular distillation pipeline. It first in-paints transparent regions, identified by a mask, with random uniform colors. These augmented images are then processed by a pre-trained depth network to generate pseudo-labeled "virtual depths," which are used to fine-tune the network itself.

Experimental results show that the model's output tends to assign background depth values to transparent surfaces such as windows. For objects composed of mixed materials, such as a glass kettle with an opaque base, a depth discontinuity was observed at material junctions. The method's performance

is sensitive to mask accuracy; segmentation errors lead to artifacts at the depth boundaries, producing unnatural transitions between the edited region and its surroundings.

Analysis of Image Editing on Synthetic Data. Controlled experiments were conducted in a Blender synthetic environment under a fixed HDR environment map and identical camera settings. With the transparent glass material, Depth Anything V2 produced inaccurate depth predictions, exhibiting blurred boundaries and distorted internal gradients. After replacing the material with an opaque metallic one, the predicted depth maps became substantially more accurate and spatially coherent across the entire object region. Depth values inside masks were normalized for visualization. These results support the hypothesis that optical ambiguity is the primary challenge in transparent-object depth estimation, and that converting the material to opaque yields clearer depth cues.

Image Editing Attempts on Real-World Scenes. Self-collected images containing transparent objects were used to explore image-level material replacement. Two approaches were examined: mask-filling and diffusion-based generative editing.

Mask-filling resolves the "see-through" depth issue but eliminates internal geometric cues, and its success is tightly coupled to mask accuracy. Diffusion-based editing (nanobanana, ControlNet) demonstrates inconsistent behavior: sometimes modifying regions outside the mask or failing to alter the transparent region at all. This suggests that general-purpose diffusion models lack a strong notion of transparency, motivating the development of a specialized editing module.