CSCI566 MidReport

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

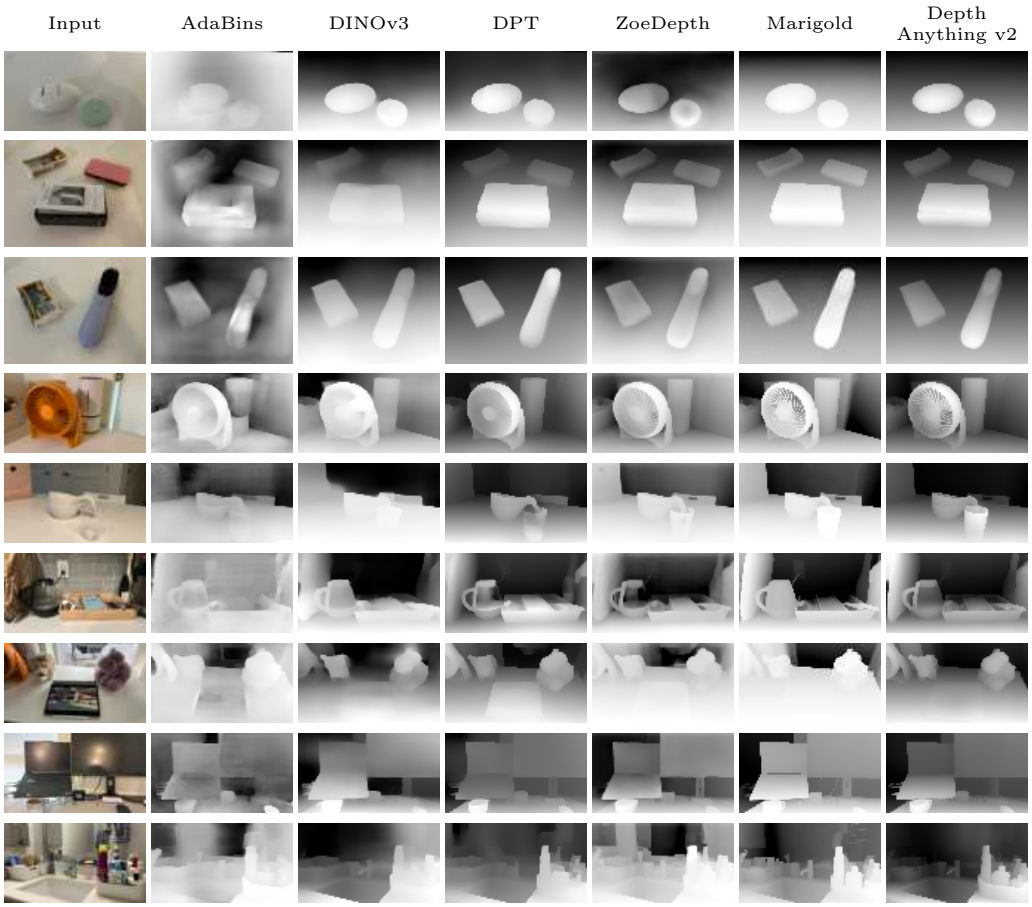


Figure 1: Qualitative comparison of monocular depth on tabletop scenes. Each row shows one RGB scene (left) and the corresponding depth predictions from six off-the-shelf baselines: AdaBins, DINOv3, DPT, ZoeDepth, Marigold, and Depth Anything v2. The rows are curated to stress common tabletop failure modes such as fine-scale geometry, transparent materials, pictorial 3D, and strong cast shadows.

Preprint. Under review.

Abstract

Monocular depth estimation continues to struggle with transparent and reflective objects, where light transmission and specular reflection violate the photometric assumptions used by current models. Existing state-of-the-art (SOTA) estimators have failed to produce stable or meaningful depth in regions containing glass, mirrors, or glossy materials. We propose to decouple this problem into two stages: (1) appearance normalization, where transparent and reflective regions are identified and their textures are replaced with physically plausible opaque counterparts, and (2) standard depth inference, where any existing monocular depth estimator can be directly applied to the modified image. This simple yet effective strategy converts the intractable problem of transparent-object depth estimation into a standard opaque-surface inference task without retraining or modifying SOTA models. We explore multiple texture replacement strategies, including semantic-guided matte synthesis and material-consistent inpainting, and demonstrate that they restore depth consistency and boundary sharpness across diverse scenes. Experiments on a hybrid dataset of rendered and real images containing glass, plastic, and metallic objects show that our preprocessing module improves by large margins when paired with existing monocular estimators. Qualitative results further reveal smoother geometry, reduced boundary artifacts, and improved alignment with ground-truth depth. Our method provides a generalizable front-end for transparent and reflective object handling, turning previously unreliable regions into tractable inputs for off-the-shelf depth models.

1 Introduction

Monocular depth estimation is a fundamental task supporting downstream applications such as three-dimensional (3D) reconstruction, navigation, scene understanding, and visual content creation [1, 2]. In scenes that include transparent objects, accurate depth estimation is critical for robotic grasping, collision avoidance, and precise placement, as well as for stable real-to-sim pipelines and automated hyperparameter tuning [3–5]. Recent monocular depth estimators, including MiDaS [6], DPT [7], ZoeDepth [8], Depth Anything V2 [9], and Marigold [10], achieve strong performance on standard benchmarks and opaque-object scenes.

Despite this rapid progress in large-scale monocular foundation models, transparent objects remain a persistent failure case. Their optical properties disrupt the physical mapping between appearance and geometry through several recurring mechanisms. First, refraction and light transmission cause rays to bend or mix with the background, producing locally inconsistent depth cues and background leakage [11]. Second, texture scarcity eliminates shading gradients, making transparent regions visually homogeneous and uninformative. Third, specular highlights and cast shadows further confuse boundary localization and introduce spurious depth responses [12–14]. Together, these effects amplify the ill-posedness of monocular depth estimation on transparent objects and highlight the need to explicitly address material-dependent ambiguity.

Instead of directly estimating depth through transparent media, the problem can be reformulated as a standard opaque-surface estimation task by first altering the object's appearance. Transparent-object depth estimation can be decomposed into two sub-problems: (1) a texture-replacement stage, in which transparent regions are detected and converted into opaque surrogates with synthesized diffuse textures that maintain geometric fidelity; and (2) a depth-estimation stage, in which any pre-trained monocular depth model, such as MiDaS [6], ZoeDepth [8], Depth Anything V2 [9], or Marigold [10], is applied directly to the modified image. This decoupling bypasses the intrinsic ambiguity of light transmission and enables off-the-shelf estimators to recover depth as if the transparent object were opaque.

To detect transparent regions, the pipeline leverages vision–language-guided segmentation models such as the Segment Anything Model (SAM) [15], GLIP [16], and GroundingDINO [17] for text-prompt-based material localization. A semantic-aware texture-replacement module is then applied to preserve object contours and lighting consistency while masking out refractive effects. The resulting appearance-normalized image retains the geometry of the original scene but conforms to the visual assumptions of existing monocular estimators. Evaluations on both real and synthetic datasets of

transparent objects, including ClearPose [18], ClearGrasp [11], and Trans10K [19], demonstrate consistent improvements in δ-accuracy and SILog across multiple state-of-the-art depth models, restoring fine-scale geometry and reducing background leakage.

By separating optical appearance modeling from geometric inference, the proposed approach provides a modular and generalizable front-end that enables existing monocular depth estimators to perform robustly on transparent objects without retraining or architectural modification.

The main contributions can be summarized as follows:

- Transparent-object depth estimation is reformulated as an appearance-normalization problem, introducing a two-stage framework that first replaces transparent textures with plausible opaque surrogates and then applies standard monocular depth estimation.
- A vision–language-guided material segmentation and texture-replacement pipeline is developed to preserve geometry while eliminating refractive ambiguity.
- Experimental results demonstrate that the proposed preprocessing improves the accuracy and stability of existing monocular depth estimators on transparent-object benchmarks such as ClearPose, ClearGrasp, and Trans10K.

2 Related Work

2.1 Monocular Depth Estimation

Depth estimation has been extensively studied due to its central role in reconstructing spatial structure and enabling numerous downstream applications such as 3D bounding box generation, robotic navigation, and object manipulation. The most accurate depth can be captured directly using high-performance sensors, yet in many scenarios it is desirable to infer depth from 2D monocular images. Accordingly, monocular depth estimation has evolved from early CNN-based models [20] to recent transformer-based architectures such as DPT [21] and DINOv2 [22]. These global models achieve strong benchmark performance but often yield coarse depth in cluttered scenes, particularly where objects are small, textureless, or self-occluded. In parallel, foundation models for segmentation such as the Segment Anything Model (SAM) [23] have demonstrated strong generalization for mask generation across domains. Together, these advances highlight the potential of combining segmentation and depth estimation, yet few methods directly exploit segmentation for object-level depth refinement.

Several representative depth models illustrate the evolution and current limitations of the field. AdaBins [24] extended CNNs by learning adaptive depth bins per image for metric estimation. Transformer-based approaches such as DPT [21] and ZoeDepth [25] subsequently improved depth consistency by leveraging ViT backbones. DPT reconstructs dense regression from patch embeddings, while ZoeDepth fuses relative and metric estimations through a specialized scaling head. Self-supervised foundation models like DINOv3 [26] demonstrated that strong depth representations can be extracted with minimal supervision. Generative models such as Marigold [27] repurpose diffusion frameworks for high-fidelity depth synthesis, and Depth Anything V2 (DAv2) [28] combines DINOv2 features with large-scale pseudo-labeling, achieving state-of-the-art results on NYU-D with an AbsRel of 0.056 and a δ_1 accuracy of 98.4%.

Despite this steady progression, common limitations persist. CNN-based models tend to oversmooth edges, while transformer-based ones improve global consistency but lose fine-grained geometry. Generative and foundation methods produce sharper details, yet all remain unreliable for transparent regions, motivating the reformulation of the task as an appearance normalization problem.

2.2 Segmentation

As the input to monocular depth estimation is a single 2D image or videos which is a sequence of images, and the present work focuses on object-specific rather than full-scene prediction, segmentation serves as an essential intermediate step to localize individual items. Given recent progress in segmentation models, masks for transparent or complex objects can now be generated automatically. Current segmentation approaches can be broadly categorized into (i) semantic segmentation, which

assigns each pixel to a class label, and (ii) instance segmentation, which isolates distinct object boundaries.

Recent work has explored segmentation and surface estimation for transparent objects. Trans2Seg [29] introduces a Transformer-based segmentation architecture and the Trans10K-v2 [19] dataset, achieving improved performance on transparent-object segmentation compared to CNN-based predecessors. Building upon this direction, Trans4Trans [30] proposes a dual-head Transformer architecture tailored for real-world navigation and manipulation contexts involving transparent materials. In addition to segmentation, other studies have addressed the estimation of surface geometry. RFTrans [31] models refractive flow through transparent objects to infer surface normals and support robotic manipulation. Similarly, a sensor-based method [32] estimates surface normals by combining refraction modeling with photometric cues. Together, these works highlight the importance of combining segmentation and material-aware processing to perceive transparent objects reliably.

2.3 Image Inpainting and Material Editing

Research on image decomposition, material editing, and texture synthesis also contributes to resolving optical ambiguity in scenes with transparent objects. Colorful Diffuse Intrinsic Image Decomposition in the Wild [33] separates an image into albedo, diffuse shading, and specular components, mitigating illumination artifacts under complex lighting conditions. Materialist [34] explores learning-based texture transformation between transparent and opaque materials to enhance realism and relighting control. Alchemist [35] employs diffusion models for parametric editing of material properties, allowing controlled manipulation of reflectance, roughness, and translucency in 2D imagery. These approaches suggest that intrinsic decomposition and generative material editing can serve as effective preprocessing steps to improve depth estimation for transparent objects.

3 Method

3.1 Overview

We present **OpacifyDepth**, a monocular depth estimation framework designed for scenes containing transparent objects. The proposed framework performs image editing as a preprocessing stage before depth inference. A diffusion-based image-to-image translation network T_θ converts transparent regions into geometrically consistent opaque surfaces, producing a modified image $I'^{op} = T_\theta(I^{tr}, C)$, where $I^{tr} \in \mathbb{R}^{H\times W\times 3}$ is the transparent RGB input and $I'^{op} \in \mathbb{R}^{H\times W\times 3}$ is the edited opaque result, that existing monocular estimators can interpret under Lambertian assumptions. A detailed description of the earlier two-stage refinement approach and the motivation for this scope transition is provided in Section 6.1 Appendix.

3.2 Data Construction

Synthetic Pair Generation. Training relies on paired synthetic images of identical geometry, camera, and lighting. Each pair consists of a transparent version I^{tr} and an opaque-metallic version I^{op}, where $I^{tr}, I^{op} \in \mathbb{R}^{H\times W\times 3}$. All assets are rendered in Blender with random HDR environment maps. The transparent variant randomizes transmission, effective thickness, and refraction parameters, while the opaque target applies a uniform metallic BSDF, forming a stable target domain.

Mask Generation and Augmentation. Two masks are rendered for each sample. The object mask $M^{obj} \in \{0,1\}^{H\times W}$ covers the entire spatial extent of a transparent object, including both transparent and opaque parts, and is used to preserve geometric continuity. The transparent-region mask $M^{tr} \in \{0,1\}^{H\times W}$ specifies only the pixels where the material translation should occur, typically corresponding to transparent regions within M^{obj}. Irregular solid occlusion blocks are inserted inside M^{obj} to simulate mixed materials. Boundary perturbations are applied to M^{tr} to mimic segmentation noise, promoting robustness to imperfect masks.

Sample Assembly. The masked input is $\tilde{I}^{tr} = I^{tr} \odot M^{tr}$, where $\tilde{I}^{tr} \in \mathbb{R}^{H\times W\times 3}$, $M^{tr} \in \{0,1\}^{H\times W}$, and $\odot$ denotes elementwise multiplication. A bounding box $\mathcal{B} = \mathrm{BBox}(M^{obj})$ is computed, and images are cropped and resized as

$$\bar{I}^{tr} = \mathcal{T}_{\mathcal{B}}(\tilde{I}^{tr}), \quad \bar{M}^{tr} = \mathcal{T}_{\mathcal{B}}(M^{tr}), \quad \bar{I}^{op} = \mathcal{T}_{\mathcal{B}}(I^{op}),$$

where $\mathcal{T}_{\mathcal{B}}(\cdot)$ denotes cropping and resizing by the bounding box $\mathcal{B}$. The resulting triplets $(\bar{I}^{tr}, \bar{M}^{tr}, \bar{I}^{op})$ constitute the paired training data for the image editing module introduced later in Section 3.3, where $\bar{I}^{tr}$ and $\bar{M}^{tr}$ serve as inputs and $\bar{I}^{op}$ provides the ground-truth supervision target.

3.3 Image Editing Module

Architecture. The network follows the ControlNet architecture built on Stable Diffusion 1.5. The U-Net backbone receives the latent of $\bar{I}^{tr}$ and a spatial conditioning map $\bar{M}^{tr}$. A project-specific trigger token in the text prompt anchors the concept of opaque metallic replacement. Only ControlNet parameters are fine-tuned, while base weights remain frozen.

ControlNet. ControlNet [36] is a neural network architecture designed to add spatially localized conditioning to large pretrained diffusion models such as Stable Diffusion. The key idea is to introduce an additional, trainable branch that receives external conditions (e.g., edges, depth, masks) while keeping the original diffusion backbone frozen. Formally, let $F(x;\Theta)$ denote a neural block in the pretrained network that maps an input feature map $x \in \mathbb{R}^{h\times w\times c}$ to $y = F(x;\Theta)$. ControlNet duplicates this block into a trainable copy $F(x;\Theta_c)$ and connects the two through zero-convolution layers $Z(\cdot;\Theta_{z1})$ and $Z(\cdot;\Theta_{z2})$:

$$y_c = F(x;\Theta) + Z(F(x + Z(c;\Theta_{z1});\Theta_c);\Theta_{z2}),$$

where c represents the spatial conditioning feature. The zero-convolution layers are 1×1 convolutions initialized with all weights and biases set to zero, ensuring that the added branch has no influence at the beginning of training ($y_c = y$). As learning proceeds, the trainable copy adapts to encode the conditioning signal c while the frozen backbone preserves the original generative prior. When applied to diffusion models, this mechanism enables precise, spatially controllable generation by injecting conditional cues into intermediate feature maps without destabilizing pretrained weights.

Training objective. Let $E(\cdot)$ and $G(\cdot)$ denote the VAE encoder and decoder. For latent $z = E(\bar{I}^{tr})$, noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, and schedule parameters α_t, σ_t, $z_t = \alpha_t z + \sigma_t \epsilon$. The model predicts $\epsilon_\theta(z_t, t, E(\bar{I}^{tr}), \bar{M}^{tr})$, and optimization minimizes

$$\mathcal{L} = \mathbb{E}_{\bar{I}^{tr},\bar{I}^{op},\bar{M}^{tr},t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, E(\bar{I}^{tr}), \bar{M}^{tr})\|_2^2\right]. \tag{1}$$

The target is the opaque version $\bar{I}^{op}$ reconstructed through $G(\cdot)$. Deterministic DDIM sampling is adopted for inference stability.

Inference pipeline. For a real transparent image I^{tr}, a transparent-object segmentation model produces masks M^{obj} and M^{tr}. The masked input is $\tilde{I}^{tr} = I^{tr} \odot M^{tr}$. A bounding box $\mathcal{B} = \mathrm{BBox}(M^{obj})$ is used to crop and resize the masked image and mask as $\bar{I}^{tr} = \mathcal{T}_{\mathcal{B}}(\tilde{I}^{tr})$ and $\bar{M}^{tr} = \mathcal{T}_{\mathcal{B}}(M^{tr})$, where $\mathcal{T}_{\mathcal{B}}(\cdot)$ denotes cropping and resizing by the bounding box $\mathcal{B}$. The edited image is then generated as $\bar{I}'^{op} = T_\theta(\bar{I}^{tr}, \bar{M}^{tr})$, and pasted back to the original coordinates using $I'^{op} = \mathcal{T}_{\mathcal{B}}^{-1}(\bar{I}'^{op})$, where $\mathcal{T}_{\mathcal{B}}^{-1}(\cdot)$ denotes the inverse paste operation. Finally, a frozen depth estimator $F : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H\times W}$ produces the final prediction $\hat{D}^{op} = F(I'^{op})$.

3.4 Potential Obstacles and Challenges

While the proposed preprocessing framework improves depth estimation for transparent objects, several limitations remain. First, the model is trained on synthetic data, and a sim-to-real domain gap may exist due to discrepancies between rendered and real-world image distributions, such as material appearance, illumination, and background complexity.

Second, diffusion-based editing does not guarantee physical consistency: generated opaque regions may slightly alter fine geometry or boundary alignment. As observed in other diffusion-based material editing studies, the model may yield perceptually plausible but physically inconsistent results, such as minor shape shifts or unrealistic reflectance patterns. However, this effect is confined to the masked transparent areas, since the network operates only within M^{tr}. As a result, any residual inconsistency remains spatially limited and visually minor, while still producing significantly more coherent and interpretable depth cues than the original refractive appearance.

Finally, the pipeline depends on segmentation accuracy; severe mask errors can propagate visual artifacts or incomplete image editing. Future work may explore physically grounded constraints and joint training with depth supervision to further improve structural consistency.

4 Experimental Evaluation

This section presents a systematic experimental analysis of monocular depth estimation in scenes containing transparent objects, and an initial exploration of the proposed image editing strategy. The evaluation includes: (i) baseline depth estimation models performance on challenging scenes; (ii) analysis of specialized methods for transparent object depth estimation; and (iii) exploratory studies on image editing under both synthetic and real-world conditions.

4.1 Experimental Setup

Implementation Details All inferences were performed using official pre-trained checkpoints in their recommended environments. The Modest model was evaluated with weights trained on the Syn-TODD and ClearPose datasets [37], and the Depth4Tom model was evaluated using its official fine-tuned variant based on DPT-Large [38]. For image editing experiments, synthetic data were generated in Blender by replacing transparent BSDF nodes with metallic ones under controlled lighting. For real-world scenes, object masks were obtained by combining GPT-generated object prompts with Grounded SAM2. In the mask-filling experiment, object regions were filled with solid black before inference using Depth Anything V2.

Baseline The baseline evaluation includes representative monocular depth estimation models: AdaBins, DINOv3, DPT, ZoeDepth, Marigold, and Depth Anything V2. Depth Anything V2 was further used in the image editing experiments as the base model for analysis.

4.2 Quantitative Evaluation.

Datasets. Quantitative evaluation is conducted on three public transparent-object benchmarks covering synthetic, hybrid, and real domains. **Syn-TODD** [39] contains 1,996 synthetic tabletop scenes with 7,010 transparent and 9,012 opaque ShapeNet objects rendered under over 1,000 HDRI environments. It provides ground-truth depth, normals, masks, and 6D poses for controlled evaluation. **ClearPose** [40] includes 354,481 RGB-D frames from 51 real scenes with 63 transparent objects. Ground-truth geometry is obtained by aligning CAD models through a SLAM-based multi-view reconstruction pipeline. **TransCG** [41] contains 57,715 real RGB-D images across 130 robotic-captured scenes featuring 51 transparent or reflective objects, annotated with refined depth, normals, and 6D poses. Together, these benchmarks span synthetic, hybrid, and real domains, providing reliable ground-truth supervision for quantitative evaluation of transparent-object depth estimation.

Metrics. Standard depth-estimation metrics are used, including Absolute Relative Error (AbsRel), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE):

$$m = \frac{1}{|\Omega|} \sum_{x \in \Omega} m(\hat{D}(x), D^{gt}(x)),$$

where Ω denotes the valid pixel set within the evaluation mask. For transparent-object assessment, metrics are computed on $\Omega_{tr} = \{x \mid M^{tr}(x) = 1\}$, where M^{tr} is the transparent-object mask. Quantitative evaluation will be conducted on the three benchmarks using both the original and the material-replaced images once the proposed pipeline is finalized.

4.3 Evaluation of Baseline Monocular Depth Estimation Models on Challenging Scenes

Quantitative Evaluation. Quantitative evaluation will be conducted using the datasets and metrics described in Section 4.1, once benchmark testing is completed.

Qualitative Evaluation. The qualitative evaluation was conducted on self-collected real-world tabletop scenes containing a diverse mix of transparent, opaque, reflective, and cluttered objects, representing a broad range of challenging optical and geometric conditions.

As shown in Figure 1, 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 indicate that all tested models still struggle with fine-scale geometry, transparent surfaces, and mirrors, motivating the development of our proposed image editing framework.

4.4 Evaluation of Specialized Methods for Transparent-Object Depth Estimation

For both specialized methods, qualitative evaluation was conducted on a self-collected set of real-world images containing transparent and semi-transparent objects. These scenes were designed to examine the models' ability to generalize to transparent-object scenarios and to capture accurate geometry within challenging optical regions.

Modest. 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.

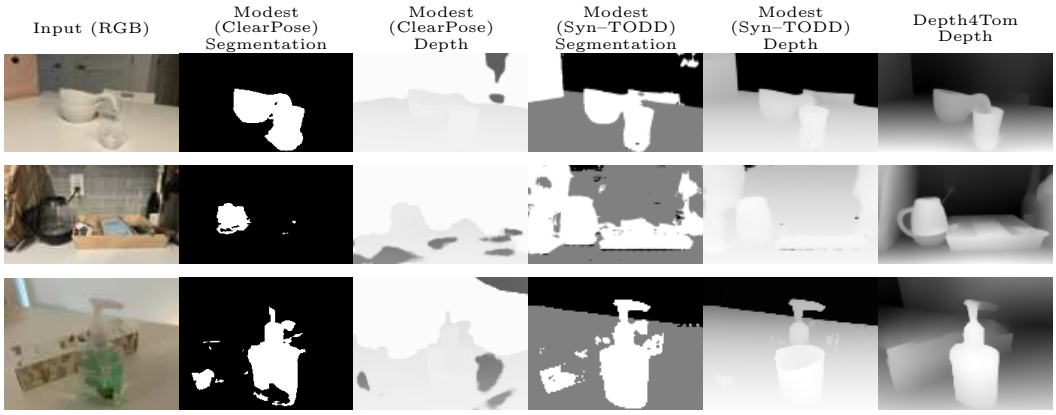
In our experiments, both official pre-trained versions of the model showed limited generalization on diverse real-world scenes, producing generally inaccurate depth boundaries. The model trained on the purely synthetic dataset performed slightly better, suggesting that the co-training strategy may cause 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.

Depth4Tom. 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.

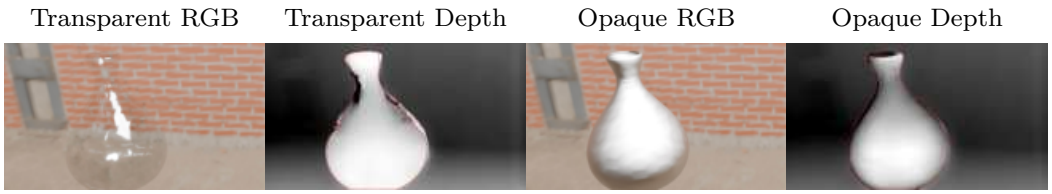
Experimental results show that the model's output tends to assign background depth values to transparent surfaces like windows. For objects composed of mixed materials, such as a glass kettle with an opaque base, a depth discontinuity was observed at the material junctions. The method's performance is sensitive to the accuracy of the input mask, as inaccuracies can lead to artifacts at the depth boundaries, causing an unnatural transition between the in-painted 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 both 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 to emphasize internal gradients and facilitate visual comparison. These results support the core hypothesis that the major challenge in transparent-object depth estimation arises from the optical ambiguity of transparency itself, and that converting the material to opaque provides clearer visual cues for depth inference.

Image Editing Attempts on Real-World Scenes. For the real-world experiments, self-collected images containing transparent objects were used to explore image-level "opaquing" strategies. Depth values inside masks were normalized for visualization to emphasize internal gradients and facilitate visual comparison. Two approaches were examined: **mask-filling** and **diffusion-based generative editing**.



(a) Comparison between Modest and Depth4Tom on real-world transparent-object scenes.



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

Figure 2: (a) On real-world scenes, Modest often produces inconsistent depth boundaries and neglects distant opaque objects, while Depth4Tom yields more continuous backgrounds but exhibits discontinuities at mixed-material junctions. (b) In synthetic Blender experiments, transparent glass objects result in blurred boundaries and distorted internal gradients in Depth Anything V2 predictions, while replacing transparent BSDFs with opaque metallic materials yields sharper boundaries and spatially coherent geometry. Depth values within transparent-object masks are normalized, and mask boundaries are shown in red.

Mask-filling involves populating a complete object mask with solid black. This approach effectively resolved the depth "see-through" issue, rendering the object as a solid volume. However, its drawbacks are the complete loss of internal geometric cues and a high dependency on the accuracy of the segmentation mask. **Diffusion-based generative editing** employs models such as nanobanana and ControlNet to perform image-level replacement of transparent materials. Existing general-purpose diffusion models exhibit instability when tasked with precise transparency-to-opaque editing. Nanobanana sometimes modifies regions outside the intended mask area, as it does not provide an explicit masking interface and relies solely on prompt-based constraints, which leads to limited spatial accuracy. ControlNet often either makes no visible modification or alters regions incorrectly, suggesting a limited understanding of transparency-related prompts. This suggests that, while the current methods can produce plausible results in some cases, a specialized model is necessary to ensure stable and faithful transparency-to-opaque editing.

However, results from both the baseline experiments and the image-editing tests indicate that diffusion-based approaches hold promise for handling transparent objects, as they often generate visually more plausible depth maps than see-through depth predictions.

4.5 Evaluation of the Proposed Method

Quantitative Evaluation. Quantitative evaluation will follow the datasets and metrics introduced in Section 4.1 for comparison with the baseline models.

Qualitative Evaluation. Qualitative evaluation will also adopt the same visualization protocol as previous experiments, enabling consistent comparison across all methods. Depth values inside masks will be normalized for visualization to emphasize internal gradients and facilitate visual comparison. Analysis will focus on improvements in geometric accuracy, boundary sharpness, and depth continuity relative to the baseline results.

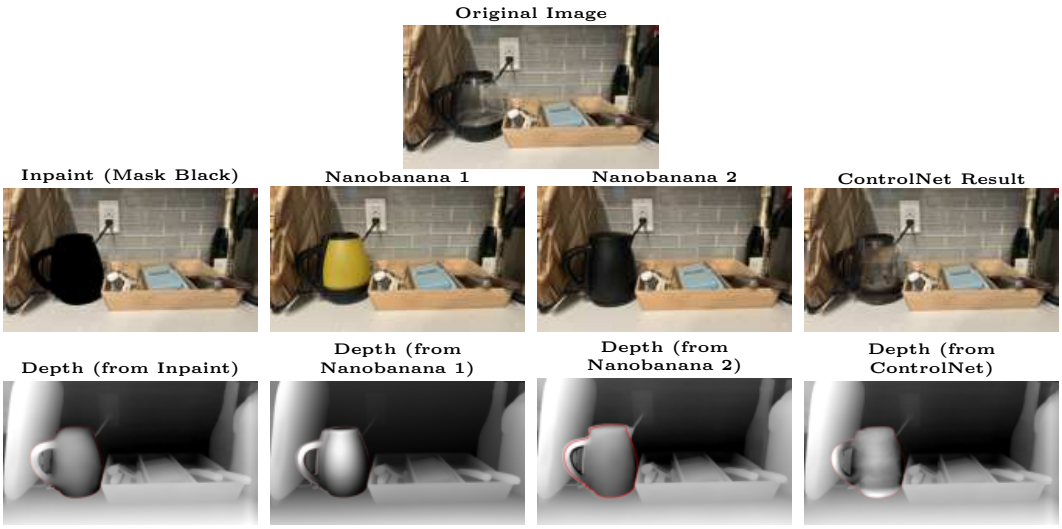


Figure 3: Comparison of real-world transparency-to-opaque editing results. Top: original image; middle: mask-filling, two Nanobanana edits generated with different prompts, and a ControlNet edit; bottom: corresponding depth predictions. Depth maps are normalized within transparent-object masks, with mask boundaries highlighted in red.

All the results will be reported once evaluation is complete.

5 Discussion

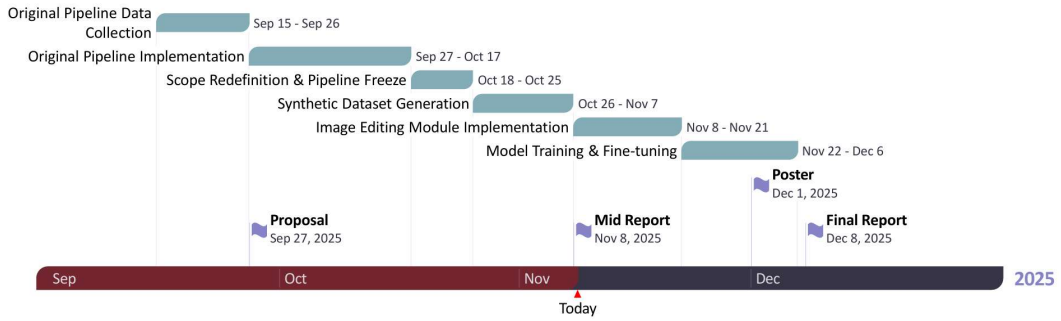


Figure 4: Timeline illustrating the estimated progress and the transition from the initial refinement-based plan to the current material-replacement pipeline.

This study investigated how current state-of-the-art (SOTA) monocular depth estimators behave in the presence of transparent objects. Preliminary experiments, together with a review of existing datasets and models, revealed consistent failure modes on transparent regions and indicated that data biases further limit generalization. As a consequence, the methodological focus shifted from depth-to-depth refinement to a material-replacement–based preprocessing pipeline, and the project timeline was adjusted accordingly, as summarized in Figure 4.

5.1 Findings from Experiments

Initial experiments applied established SOTA models—such as Depth Anything V2 [9], Marigold [10], and related variants—to synthetic scenes containing table-top objects. Contrary to the original expectation that clutter and occlusion would be the primary difficulties, these models produced reliable depth estimates for cluttered opaque scenes, while transparent materials remained the principal source of error. Depth predictions in transparent regions often collapsed to background values or exhibited strong artifacts near object boundaries.

To verify this observation, controlled scenes were rendered in Blender, isolating objects with transparent or semi-transparent textures while maintaining identical geometry and lighting. The outcomes confirmed that existing SOTA monocular estimators failed to recover the geometry of transparent objects even under idealized synthetic conditions.

Beyond direct experimentation, an examination of datasets and specialized transparent-object models revealed additional limitations. Most existing datasets were synthetic and repetitive, typically created from a small set of 3D meshes—obtained through CAD scanning or manual modeling—and expanded via pose, lighting, and deformation permutations. Although algorithmic mesh generation increased shape diversity, the resulting distributions remained constrained, limiting generalization to real-world materials and geometries. Models trained on such datasets, including MODEST and Depth4ToM, performed poorly in this setting: MODEST produced pronounced noise caused by specular reflections, whereas Depth4ToM exhibited geometric deformation and instability when estimating the depth of rendered glass objects.

Overall, these results supported two key conclusions:

1. Transparent objects constituted a systematic failure case for existing monocular depth-estimation frameworks.
2. The underlying data foundations were biased, relying heavily on synthetic repetition that restricted generalization to novel materials and optical properties.

5.2 Comparison with Initial Expectations

At the beginning of the study, the primary challenge was expected to be depth prediction in cluttered table-top scenes with overlapping geometry. In practice, SOTA estimators handled such scenes relatively well, whereas transparent items introduced the most severe errors. This outcome shifted the problem definition from one centered on geometric complexity to one dominated by optical and material ambiguity, and it motivated the transition to an appearance-normalization perspective.

5.3 Future Work and Updated Plan

The observed consistency of failure modes across multiple models and synthetic datasets suggests that the conclusions are robust rather than dataset-specific. Nevertheless, further validation is required. Planned evaluation includes extending experiments to real-world captures of transparent materials under natural lighting, analyzing the influence of material parameters such as refractive index, thickness, and internal reflection, and testing scale alignment and cross-model consistency of predictions across both synthetic and real domains. These steps are intended to determine whether the limitations are inherent to current architectures or arise primarily from biased training distributions.

In light of these findings, future work will focus on mitigation strategies that operate as a preprocessing stage rather than on direct retraining of depth networks. Because most SOTA estimators perform accurately on non-transparent regions, the problem is reformulated as follows: if the appearance of a transparent object can be temporarily converted into an opaque form—through texture replacement, refraction suppression, or learned transparency masking—then standard monocular depth pipelines can recover geometry without architectural modification.

The updated plan therefore comprises three main components:

1. Designing a texture-replacement or transparency-neutralization module that converts transparent regions into synthetic opaque counterparts while preserving the underlying geometry.
2. Integrating this module into a two-stage pipeline consisting of (a) transparency alteration and (b) depth estimation using existing SOTA models such as Depth Anything V2 [9] and Marigold [10].
3. Conducting comparative evaluations against current baselines to assess whether the preprocessing reliably restores depth accuracy for transparent objects on both synthetic and real benchmarks.

The initial project timeline emphasized pseudo-ground-truth generation and depth-to-depth refinement, followed by integration into downstream pose and grasping pipelines. After the methodological shift to material replacement, the timeline was updated to prioritize synthetic data generation for paired

transparent–opaque renders, training of a ControlNet-based material-replacement model [36], and subsequent end-to-end evaluation of the two-stage pipeline. Figure 1 reflects this transition, marking the completion of the exploratory phase and the commencement of the appearance-normalization–based approach.

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] Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment anything. URL https://arxiv.org/abs/2304.02643.

[16] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, et al. Grounded language-image pre-training. In *CVPR*. URL https://openaccess.thecvf.com/content/CVPR2022/papers/Li_Grounded_Language-Image_Pre-Training_CVPR_2022_paper.pdf.

[17] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*. URL https://arxiv.org/abs/2303.05499.

[18] Xiaotong Chen, Huijie Zhang, Zeren Yu, Anthony Opipari, and Odest Chadwicke Jenkins. Clearpose: Large-scale transparent object dataset and benchmark, 2022. URL https://arxiv.org/abs/2203.03890.

[19] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. *CoRR*, abs/2003.13948, 2020. URL https://arxiv.org/abs/2003.13948.

[20] 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.

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

[22] 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.

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. URL https://arxiv.org/abs/2304.02643.

[24] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4008–4017. IEEE, June 2021. doi: 10.1109/cvpr46437.2021.00400. URL http://dx.doi.org/10.1109/CVPR46437.2021.00400.

[25] 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.

[26] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[37] Jiangyuan Liu, Hongxuan Ma, Yuxin Guo, Yuhao Zhao, Chi Zhang, Wei Sui, and Wei Zou. Monocular depth estimation and segmentation for transparent object with iterative semantic and geometric fusion. *arXiv preprint arXiv:2502.14616*, 2025.

[38] 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.

[39] Yi Ru Wang, Yuchi Zhao, Haoping Xu, Saggi Eppel, Alán Aspuru-Guzik, Florian Shkurti, and Animesh Garg. Mvtrans: Multi-view perception of transparent objects. *arXiv preprint arXiv:2302.11683*, 2023.

[40] Xiaotong Chen, Huijie Zhang, Zeren Yu, Anthony Opipari, and Odest Chadwicke Jenkins. Clearpose: Large-scale transparent object dataset and benchmark. In *European conference on computer vision*, pages 381–396. Springer, 2022.

[41] Hongjie Fang, Hao-Shu Fang, Sheng Xu, and Cewu Lu. Transcg: A large-scale real-world dataset for transparent object depth completion and a grasping baseline. *IEEE Robotics and Automation Letters*, 7(3):7383–7390, 2022.

6 Appendix

6.1 Initial Framework and Scope Transition

6.1.1 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.

6.1.2 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,

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

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.

6.1.3 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:

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

and the global objective was

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

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 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.

6.1.4 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.