# SeeClear: Reliable Transparent Object Depth Estimation via Generative Opacification

Xiaoying Wang[1⋆], Yumeng He[1,2*], Jingkai Shi[1,2*], Jiayin Lu[1], Yin Yang[3], Ying Jiang[1], and Chenfanfu Jiang[1]

[1] University of California, Los Angeles, USA
[2] University of Southern California, USA
[3] University of Utah, USA

**Fig. 1: SeeClear** is a novel framework that converts transparent objects into diffusion-generated opaque images, predicting stable and accurate depth for transparent objects.

**Abstract.** Monocular depth estimation remains challenging for transparent objects, where refraction and transmission are difficult to model and break the appearance assumptions used by depth networks. As a result, state-of-the-art estimators often produce unstable or incorrect depth predictions for transparent materials. We propose **SeeClear**, a novel framework that converts transparent objects into diffusion-generated opaque images, enabling stable monocular depth estimation for transparent objects. Given an input image, we first localize transparent regions and transform their refractive appearance into geometrically consistent opaque appearances using a diffusion-based generative opacification module. The processed image is then fed into an off-the-shelf monocular depth estimator without retraining or architectural changes. To train the opacification model, we construct **SeeClear-396k**, a synthetic dataset containing 396k rendered images across 66k paired transparent-opaque configurations. Experiments on both synthetic and real-world datasets show that **SeeClear** significantly improves depth estimation for transparent objects.

**Keywords:** Monocular Depth Estimation · Transparent Objects

⋆ Equal contribution. xwang648@usc.edu, heyumeng@usc.edu, shikings@usc.edu, jiayin_lu@math.ucla.edu, yangzzzy@gmail.com, anajymua@gmail.com, cffjiang@ucla.edu

## 1 Introduction

Monocular depth estimation underpins 3D reconstruction [15, 31], navigation [17], and robotic manipulation [23, 42, 44], and is a key enabler for grasping [30, 33], collision avoidance [12, 13], and real-to-simulation pipelines [34, 43]. Though recent monocular depth models achieve strong performance [46, 47], they often fail in real-world scenes containing objects with complex appearance properties. In particular, transparent objects pose a severe challenge due to refractive and transmissive effects that violate common photometric assumptions [28], leading to erroneous depth predictions that can propagate to downstream simulation and manipulation tasks. This discrepancy is particularly problematic in physical scene understanding. For example, inaccurate depth estimation of transparent cabinets can lead to incorrect reconstruction of collision regions.

Prior work addresses transparent depth estimation by fine-tuning depth models specifically for transparent objects [9,49,55], achieving strong performance on transparent scenes. However, these approaches rely on large-scale paired transparent RGB–depth data and introduce distribution shifts. Moreover, as state-of-the-art depth models continue to evolve rapidly [58], every updated backbone requires additional fine-tuning. Although modern foundation depth models remain challenged by transparent depth estimation, they demonstrate strong generalization and high accuracy for opaque objects. This naturally raises the question: can we leverage transparent-object data without modifying or retraining the depth model, while remaining compatible with foundation depth models, to enable accurate depth estimation for transparent objects?

As illustrated in Fig. 1, we propose **SeeClear**, a plug-and-play front-end framework for accurate and stable monocular transparent-object depth estimation. The method first localizes transparent regions and converts them into geometrically consistent opaque appearances. Specifically, a segmentation model provides object masks as structural guidance, while a diffusion model generates diffuse textures to replace refractive appearances while preserving object boundaries and shading. The opacified image is then composited with the original background using a mask refinement module to form the final input for the depth estimator. To train the generative opacification module, we construct **SeeClear-396k**, a large-scale synthetic dataset containing 396k rendered images across 66k paired transparent–opaque configurations with aligned depth, normals, and object masks. Extensive experiments on synthetic and real-world datasets demonstrate the effectiveness of the proposed method.

In summary, our contributions are as follows: (a). We propose a novel framework **SeeClear** that first transforms transparent regions into geometry-consistent opaque images, enabling further depth prediction for transparent objects using monocular depth models. (b). We introduce a plug-and-play front-end adapter with a Mask Refinement Module to transform transparent regions into geometry-consistent opaque representations. To support training, we construct a large-scale high-quality paired transparent–opaque synthetic dataset, **SeeClear-396k**. (c). We conduct extensive experiments on both synthetic and real-world datasets,

demonstrating consistent improvements over baselines and validating the effectiveness and generality of the proposed method.

## 2 Related Work

### 2.1 Monocular Depth Estimation

Monocular depth estimation infers per-pixel scene depth from an image. Early CNN-based methods [10, 22] demonstrated direct depth regression, and later transformer architectures such as DPT [37] and DINO-derived backbones [32,48] improved long-range aggregation and depth coherence. Recent depth estimation models [2, 20, 38], including ZoeDepth [1], Marigold [19], and Depth Anything [24, 57, 58], further combine relative-depth pretraining with metric-depth supervision, exploit diffusion priors, and scale training with large-scale pseudo-labeling. These advances improve zero-shot transfer across domains, strengthen metric depth prediction when absolute scale is required, and produce finer and more robust depth maps, making monocular depth practical for many real-world applications. Despite these advances, depth estimation for transparent objects remains challenging, as optical effects such as refraction and reflection violate the standard appearance–geometry correspondence and often lead to unreliable depth predictions [28]. To address this, prior work trains depth models specifically for transparent objects [9, 49, 55]. However, these approaches require large-scale paired transparent RGB–depth data for supervision, which is difficult to obtain in reality [8, 11, 41]. Moreover, these training strategies are model-dependent, requiring retraining for each new depth backbone. Instead, we propose a plug-and-play approach that converts transparent appearances into geometry-consistent opaque images, followed by a monocular depth estimator. This reframes transparent depth estimation as an opacification problem, enabling compatibility with existing depth models without retraining.

### 2.2 Transparent Object Segmentation

Image segmentation [21, 27, 39, 59] provides object-level structural priors that are critical for geometry reasoning, as accurate boundaries help constrain depth inference and prevent cross-region ambiguity. For transparent objects in particular, segmentation has attracted increasing attention due to the prevalence of glass and specular materials in real-world environments [41]. However, segmenting transparent objects remains challenging due to complex optical effects such as refraction, reflection, and background distortion [18]. Trans2Seg [53] proposes a transformer model to distinguish transparent materials and introduces the Trans10K dataset as a large-scale benchmark that also supports supervised model training. Trans4Trans [60] extends this direction with a dual-head transformer for transparent object segmentation to support real-world navigation. However, such methods are designed as task-specific segmentation models, relying on dedicated annotations and optimizing mask quality rather than depth adaptation. Instead, our framework uses general-purpose vision-language segmentation to obtain coarse object-level masks that localize regions requiring appearance opacification before depth estimation.

### 2.3 Material Editing

Broad image editing methods provide useful context but are not designed for material-aware appearance conversion. InstructPix2Pix [4] performs instruction-guided image editing, while AnyDoor [7] focuses on object-level image customization. Although these methods can alter object appearance, they do not explicitly model material properties or target transparent-to-opaque conversion. In contrast, we focus on material-level editing to convert transparent objects into opaque appearances. To address image-based material editing, early works such as [25] adopt physically based inverse rendering to estimate intrinsic properties (e.g., albedo, roughness, metallic) and enable edits through PBR re-rendering, offering interpretability and physical consistency. Careaga et al. [5] use diffusion models to decompose images into albedo and shading, reducing illumination ambiguity without explicit physical modeling. Recent work Materialist [50] combines learned material prediction with differentiable rendering for physically consistent editing, including transparency manipulation. Alchemist [45] instead uses diffusion models for parametric control over reflectance attributes, but lacks explicit physical constraints. Despite their success, these methods mainly target photorealistic editing or controllable material manipulation rather than edits for downstream depth estimation. In contrast, our method uses paired transparent–opaque guidance to learn appearance transformation that preserves surface geometry and boundaries, producing more reliable inputs for depth prediction.

## 3 Method

Given an input image $I^{tr} \in \mathbb{R}^{H\times W\times 3}$ containing a transparent object, our goal is to estimate its depth map $I^{d} \in \mathbb{R}^{H\times W}$. To achieve this, we introduce **SeeClear** (Fig. 2), a two-stage framework that integrates diffusion-based generative opacification, mask-aware compositing, and off-the-shelf monocular depth estimation. Starting from the input image $I^{tr}$, we first apply a segmentation model to obtain the object mask $M^{seg} \in \{0,1\}^{H\times W}$. Conditioned on both $I^{tr}$ and $M^{seg}$, a latent diffusion model transforms refractive appearances into a geometry-consistent opaque image $I^{pred}$. To integrate the generated opaque region with the original background, we introduce a mask refinement module that predicts a blending mask $\hat{M}_{ref} \in \{0,1\}^{H\times W}$. The composited image $I^{blend}$ is obtained by binary compositing using the refined mask $\hat{M}_{ref}$. The blended image is then fed into an off-the-shelf foundation depth model to produce the final depth map $I^{d}$.

### 3.1 Generative Opacification

Given the input image $I^{tr}$ and transparent object mask $M^{seg}$, this stage generates the opaque object image using a diffusion model.

**Conditional Latent Diffusion Formulation.** We adopt a conditional latent diffusion model to learn the mapping from transparent observations to their opaque counterparts. The ground-truth opaque image $I^{op}$ is first encoded into the latent space of a pre-trained Variational Autoencoder (VAE) [40], yielding

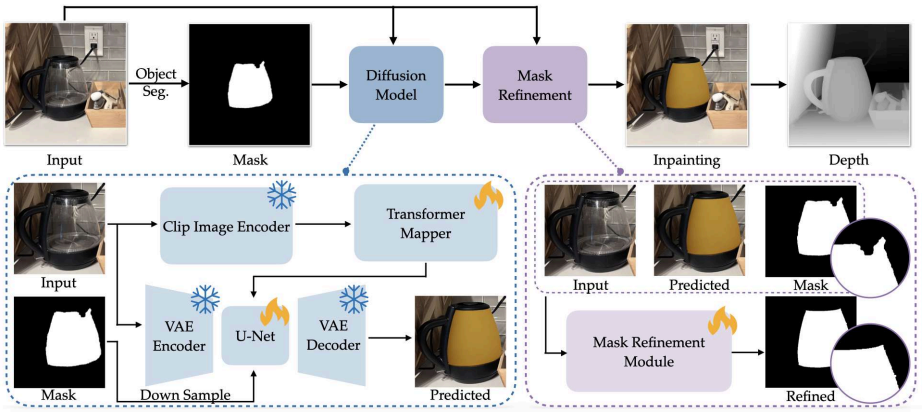


**Fig. 2: Pipeline Overview.** Starting from an image, we first apply a segmentation model to obtain the transparent object mask. Guided by the mask and the image, a latent diffusion model generates an opacified image of the transparent object. A mask refinement module then predicts a blending mask, binarized at inference, to hard-composite the generated opaque region with the original background, producing the composited image, which is fed into a depth model to estimate depth.

$z_0 = \mathcal{E}(I^{op})$, where $\mathcal{E}$ denotes the VAE encoder. During the forward diffusion process, Gaussian noise is progressively added to $z_0$ to obtain a noisy latent $z_t$. To condition generation on the transparent observation, the denoising network $\epsilon_\theta$ takes as input a concatenated latent tensor

$$z_{input} = \text{Concat}\big(z_t,\ \mathcal{E}(I^{tr}),\ \text{Down}(M^{seg})\big)\,, \tag{1}$$

where $I^{tr}$ is the input image containing transparent objects and $M^{seg}$ denotes the transparent object mask. This results in a concatenated feature consisting of the noisy latent $z_t$, the encoded transparent image $\mathcal{E}(I^{tr})$, and a downsampled mask $\text{Down}(M^{seg})$. The mask provides explicit spatial guidance, encouraging the model to preserve object geometry while removing refractive distortions.

**CLIP-based Semantic Conditioning.** To guide the generation toward an opaque result that preserves the original object shape, an image-level condition encoding object identity is required. Explicit geometric conditions, such as depth or normal maps, would be desirable, but they are unreliable for transparent objects. We therefore use the original transparent-object image $I^{tr}$ as the condition source. Since a transparent object and its diffuse counterpart typically belong to the same semantic category even when their appearances differ significantly, we adopt a CLIP [35] image encoder, whose image-text contrastive training yields semantically oriented representations. We retain only the class token as the condition $c$, since patch tokens may encode local appearance details contaminated by background leakage. The class token is then fed into a lightweight Transformer Mapper to project it into the cross-attention conditioning space of the denoising network $\epsilon_\theta$, yielding the final semantic condition $c$.

**Training Objectives.** Following the standard DDPM objective [16], we train the denoising network $\epsilon_\theta$ to predict the injected noise $\epsilon$ at timestep $t$:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{t,z_0,\epsilon} \left\| \epsilon - \epsilon_\theta(z_{input},\, c,\, t) \right\|_2^2 . \tag{2}$$

While $\mathcal{L}_{\text{LDM}}$ supervises the denoising process in the compressed latent space, it does not directly constrain the perceptual quality of the decoded output: the VAE's lossy compression discards high-frequency details, and element-wise $\ell_2$ supervision is insensitive to perceptually significant structures such as sharp edges and fine-grained textures [61]. We therefore introduce a complementary perceptual loss computed directly in pixel space using multi-level VGG features:

$$\mathcal{L}_{\text{LPIPS}} = \text{LPIPS}(\mathcal{D}(z_{pred}) \odot M^{\text{seg}},\; I^{op} \odot M^{\text{seg}}) , \tag{3}$$

where $z_{pred} = (z_t - \sqrt{1-\bar{\alpha}_t}\, \epsilon_\theta(z_{input}, c, t))/\sqrt{\bar{\alpha}_t}$ is the denoised latent, $\mathcal{D}(z_{pred})$ is the decoded prediction in pixel space, $\mathcal{D}(\cdot)$ denotes the VAE decoder, $\odot$ denotes element-wise multiplication, and $M^{\text{seg}}$ confines the perceptual supervision to the transparent object region, focusing texture fidelity on the opacified content. Following [29], $\mathcal{L}_{\text{LPIPS}}$ is applied only at low-noise timesteps, since at high-noise timesteps the decoded prediction deviates significantly from the clean image, and perceptual supervision risks reducing sample diversity. The overall training objective is $\mathcal{L} = \mathcal{L}_{\text{LDM}} + \lambda\, \mathcal{L}_{\text{LPIPS}}$, where $\lambda$ is a weighting coefficient.

**Mask augmentation.** Since the segmentation mask $M^{seg}$ at inference may contain boundary inaccuracies, the training mask is randomly perturbed by adding or subtracting small geometric shapes (e.g., circles and rectangles) near object boundaries. This augmentation improves robustness to imperfect masks from the automatic segmentation stage.

### 3.2 Mask Refinement Module

Directly using the decoded image $\mathcal{D}(z_{pred})$ for depth estimation is suboptimal, since VAE encoding-decoding introduces blurring and color drift in non-object regions. To preserve the original scene context, the generated object region should be composited back into the transparent-object image $I^{tr}$. However, latent-space inpainting can introduce boundary halos and color shifts due to VAE nonlinearity [3], while hard pasting in pixel space with $M^{seg}$ often produces seam artifacts due to boundary inaccuracies in the segmentation mask. We therefore introduce a lightweight Mask Refinement Module (MRM) that outputs a soft blending mask $M_{ref} \in [0,1]^{H\times W}$, which is further binarized at inference to obtain $\hat{M}_{ref} \in \{0,1\}^{H\times W}$ for pixel-level compositing of the generated region with the original image. The MRM is implemented as a fully convolutional network consisting of three 3×3 convolutional layers followed by a 1×1 projection layer, with GroupNorm and SiLU activations. During training, the module takes as input the concatenation of the opaque-object image $I^{op}$, the transparent-object image $I^{tr}$, and an augmented version of the ground-truth object mask using the same mask augmentation scheme described in Sec. 3.1. The module is supervised

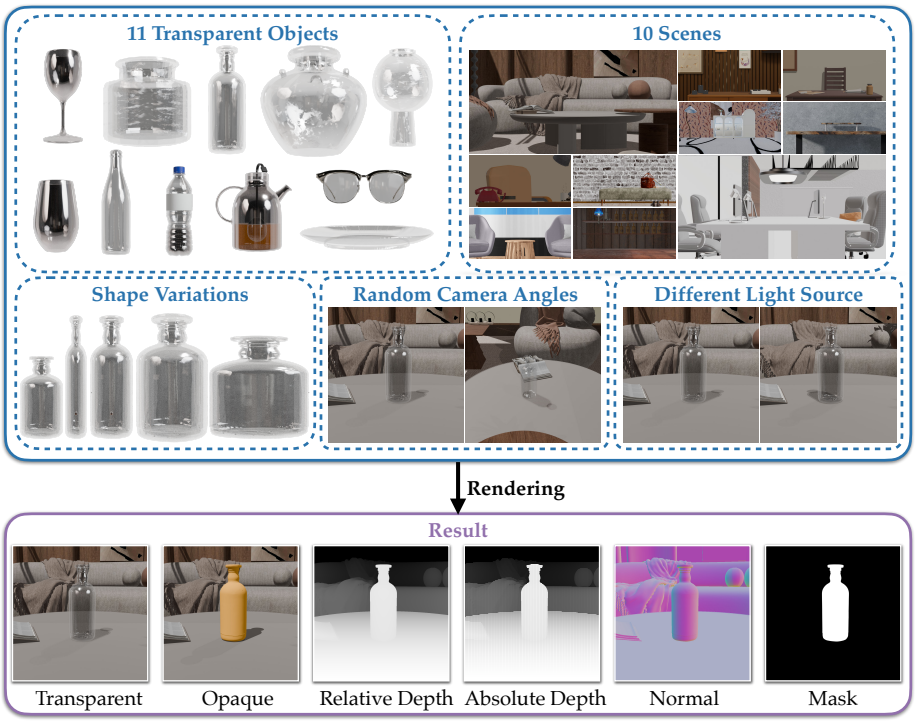


**Fig. 3: Rendering Pipeline.** We build the **SeeClear-396k** dataset for transparent-object depth estimation. For each object–scene configuration, Blender renders paired transparent ($I^{tr}$) and opaque ($I^{op}$) images with identical geometry, camera pose, and illumination. Viewpoint, lighting, and anisotropic shape variations are systematically sampled to produce diverse training data with aligned depth, normals, and masks.

with a binary cross-entropy loss together with an additional mid-value penalty:

$$\mathcal{L}_{refine} = \mathcal{L}_{BCE} + \mu \sum M_{ref}(1 - M_{ref}) \tag{4}$$

where $\mu$ is a weighting coefficient and the second term penalizes mid-range mask values, encouraging confident separation between object and background regions. At inference, the opaque-object image and augmented training mask are replaced by the diffusion-generated image $I^{pred}$ and the segmentation mask $M^{seg}$, respectively. The final composited image is then obtained by binary compositing:

$$I^{blend} = \hat{M}_{ref} \odot I^{pred} + \left(1 - \hat{M}_{ref}\right) \odot I^{tr}. \tag{5}$$

### 3.3 Data Curation

Our training dataset is designed to improve generalization to in-the-wild transparent objects. Following recent transparent-depth benchmarks, we adopt a

*failure-mode-driven* construction strategy inspired by Booster [36], which organizes dataset coverage around settings where depth prediction breaks down. We curate objects and scenes to capture three key challenges: non-Lambertian cue distortion (e.g., refraction and textureless glass), depth-definition ambiguity (e.g., object surfaces versus background visible through glass), and complex scene structure (e.g., thin boundaries, multi-layer layouts, and occlusions between transparent and opaque objects). To provide explicit supervision for generative opacification, we render paired transparent and opaque images for each sampled configuration that share identical geometry, camera pose, and illumination (Fig. 3). The transparent image $I^{tr}$ uses a glass-like material model, while the corresponding opaque image $I^{op}$ replaces the transparent shader with a diffuse Lambertian material of fixed color (e.g., `#E7A034FF`), isolating appearance changes caused by transparency while keeping all geometric factors fixed.

To ensure realistic diversity, we curate 11 everyday transparent objects and furniture placed in 10 indoor scenes, forming 110 object–scene units. Each unit is expanded through deterministic sampling of viewpoint, illumination, and shape variation. We sample 10 camera viewpoints within a predefined angular range, use two lighting setups (canonical overhead and randomized placement), and apply anisotropic scaling under six deformation modes $\{x, y, z, xy, yz, xz\}$ with five magnitudes to diversify object shapes. All parameters are sampled with a fixed random seed for reproducibility. In total, the dataset contains 110 units with 600 image sets each, including transparent images $I^{tr}$, opaque images $I^{op}$, absolute depth (OpenEXR), relative depth (PNG), surface normals (PNG), and object masks (PNG), yielding 396,000 rendered images.

## 4 Experimental Evaluation

**Implementation Details** The opacification network is implemented as a latent diffusion model [40] initialized from Paint-by-Example [56] and fine-tuned on the **SeeClear**-396k dataset described in Sec. 3.3. For transparent object localization, we adopt a coarse-to-fine segmentation strategy. Trans4Trans [60] first provides an initial localization prior, which serves as a spatial prompt for SAM 3 [6] to produce a refined transparent-object mask $M^{seg}$. Training uses a base learning rate of $1 \times 10^{-5}$ with a linear warm-up schedule and a batch size of 4, and runs for 328,339 iterations on four NVIDIA H100 NVL GPUs. At inference, the model is applied per instance: for each transparent object, a local patch centered on its bounding box is cropped from the input image, resized to $512 \times 512$, processed by the diffusion model, and composited back into the original image using the Mask Refinement Module. By mapping to this canonical space, the model can ignore global spatial layout and focus on object-level opacification regardless of where or how large the object appears in the scene. When multiple transparent objects are present, their patches are processed together as a single batch, minimizing computational overhead. The diffusion process is solved using the UniPC sampler [62] with 10 steps. For depth estimation, we explore two off-the-shelf monocular depth models, Depth Anything V3 [24] and MoGe-2 [51].

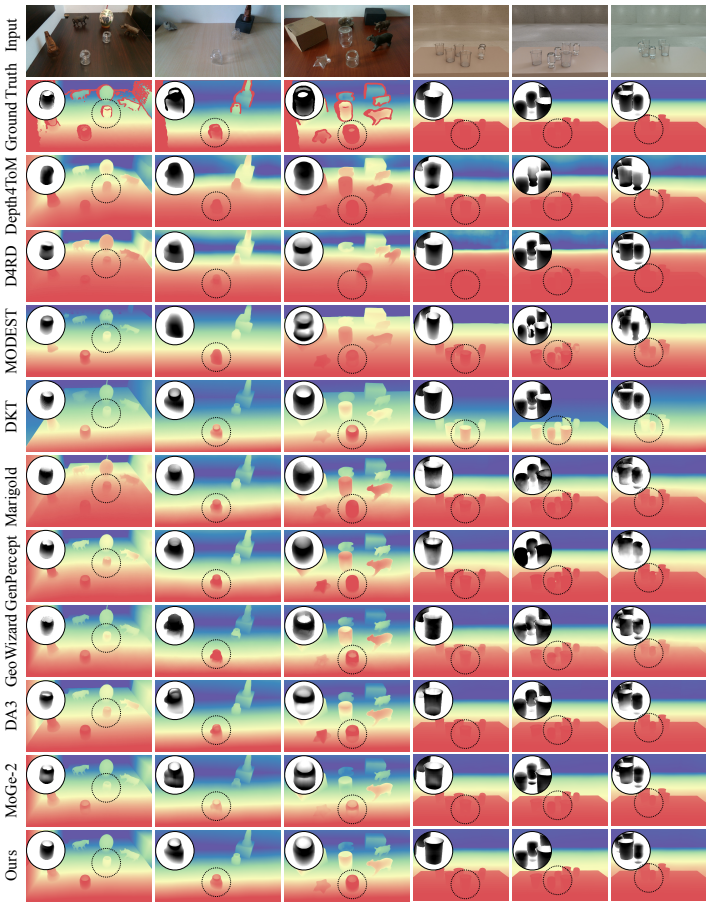


**Fig. 4: Qualitative Comparison.** We evaluate transparent-object depth estimation on ClearGrasp [41] (Columns 1–3) and TransPhy3D [55] datasets (Columns 4-6). Compared with baselines, **SeeClear** produces accurate transparent-object depth. Depth maps in the circle are normalized in grayscale.

Neither backbone is fine-tuned, and all evaluations are conducted in a zero-shot setting. On a single H100 GPU, processing one image takes approximately 7.65 s.

**Baselines** We compare **SeeClear** with several state-of-the-art methods for transparent-object depth estimation, including Depth4ToM [9], Diffusion4RobustDepth (D4RD) [49], MODEST [26], and DKT [55]. In addition, we compare with general-purpose monocular depth estimators. These include diffusion-based depth models such as Marigold [19], GenPercept [54], and GeoWizard [14], which leverage pre-trained generative diffusion priors. We also evaluate discriminative depth models, including Depth Anything V3 (DA3) [24] and MoGe-2 [51], which are feed-forward ViT-based estimators trained on large-scale depth datasets. Note that DKT is fine-tuned on both evaluation datasets and therefore does not operate in a zero-shot setting. Other methods, including the proposed method, are

evaluated without training on the evaluation datasets. To compare with DKT in a zero-shot setting, we also evaluate on additional datasets where DKT is not trained, with results reported in the Appendix.

### 4.1 Quantitative Evaluation

We evaluate transparent-object depth estimation on both real-world and synthetic benchmarks, including *ClearGrasp* [41], an RGB-D dataset of transparent objects, and *TransPhy3D* [55], a synthetic dataset with physically rendered depth for transparent objects. Following prior work [54], affine-invariant and disparity-output models are aligned to metric ground-truth depth via least-squares regression before computing evaluation metrics. For disparity models, the regression is performed in the disparity domain and converted back to depth. Evaluation on *ClearGrasp* [41] uses three disjoint regions: the transparent-object region (**ToM**), the full image (**All**), and the non-transparent region (**Other**). For *TransPhy3D* [55], evaluation is performed only on the full image, as no ground-truth object masks are provided for region-specific evaluation. We report depth-estimation metrics including AbsRel (↓), SiLog (↓), RMSE (↓, mm), iRMSE (↓, $\text{mm}^{-1}$), MAE (↓, mm), and threshold accuracies $\delta$<1.025, $\delta$<1.05, and $\delta$<1.10 (↑), measuring relative depth error (AbsRel), scale-invariant logarithmic error (SiLog), root-mean-square error sensitive to large depth deviations (RMSE), inverse-depth error emphasizing near-range accuracy (iRMSE), mean absolute error (MAE), and the percentage of pixels whose predicted depth falls within a multiplicative threshold of the ground truth ($\delta$ metrics).

**Quantitative Results on ClearGrasp.** As shown in Tab. 1, **SeeClear** with MoGe-2 achieves the best performance on the **All** and **Other** regions across all eight metrics (iRMSE tied in Other), while also surpassing most baselines on the **ToM** region. Although DKT achieves higher scores on $\delta_{1.025}$ and $\delta_{1.05}$ in the **ToM** region owing to fine-tuning on the evaluation dataset, **SeeClear** with DA3 outperforms DKT on the remaining six metrics in a zero-shot setting. **SeeClear** also outperforms DKT on the **All** and **Other** regions, where DKT's performance is limited by a tendency to overestimate depth in background areas. For a direct zero-shot comparison with DKT on an unseen dataset, we additionally evaluate on TDoF20 [52] (see Appendix), where **SeeClear** outperforms DKT on seven of the eight metrics, with iRMSE tied. For both DA3 and MoGe-2, generative opacification consistently improves performance. With DA3, RMSE decreases from 24.03 mm to 21.45 mm (10.7%). The improvement is more pronounced for MoGe-2, where RMSE is reduced from 56.83 mm to 26.62 mm and $\delta_{1.025}$ increases from 5.63% to 42.75%. These results suggest that replacing the transparent appearance with a generated opaque counterpart allows depth models to better estimate depth on transparent objects.

Compared with specialized transparent-object depth methods not fine-tuned on the evaluation data, **SeeClear** with DA3 achieves lower RMSE than Depth4ToM (24.83 mm), D4RD (39.95 mm), and MODEST (39.31 mm). It also outperforms diffusion-based depth models including GenPercept (24.09 mm), Marigold (28.10 mm), and GeoWizard (31.54 mm), despite keeping the backbone frozen. Beyond RMSE,

**Table 1: ClearGrasp.** Comparison on the real-world test set across three evaluation regions. **ToM**: transparent-object region. **All**: full image. **Other**: non-transparent region. †: fine-tuned on this dataset. : best result.

| Region | Method | AbsRel↓ | SiLog↓ | RMSE↓ | iRMSE↓ | MAE↓ | $\delta_{1.025}$↑ | $\delta_{1.05}$↑ | $\delta_{1.10}$↑ |
|---|---|---|---|---|---|---|---|---|---|
| **ToM** | Depth4ToM [9] | 0.040 | 0.034 | 24.83 | 0.009 | 20.94 | 41.41 | 70.48 | 92.56 |
| | MODEST [26] | 0.062 | 0.041 | 39.31 | 0.013 | 33.09 | 29.22 | 54.01 | 79.60 |
| | D4RD [49] | 0.069 | 0.031 | 39.95 | 0.014 | 36.18 | 16.21 | 42.70 | 78.25 |
| | DKT† [55] | 0.035 | 0.024 | 22.43 | 0.008 | 18.59 | **52.69** | **81.29** | 94.04 |
| | Marigold [19] | 0.047 | 0.036 | 28.10 | 0.011 | 24.31 | 38.58 | 64.49 | 88.10 |
| | GenPercept [54] | 0.039 | 0.028 | 24.09 | 0.009 | 20.30 | 47.79 | 75.21 | 92.71 |
| | GeoWizard [14] | 0.052 | 0.029 | 31.54 | 0.012 | 27.54 | 34.86 | 62.25 | 87.87 |
| | DA3 [24] | 0.038 | 0.028 | 24.03 | 0.008 | 19.95 | 46.89 | 75.82 | 92.24 |
| | MoGe-2 [51] | 0.101 | 0.035 | 56.83 | 0.019 | 52.96 | 5.63 | 17.81 | 60.01 |
| | Ours (DA3) | **0.033** | **0.023** | **21.45** | **0.007** | **17.94** | 51.27 | 78.13 | **95.63** |
| | Ours (MoGe-2) | 0.043 | 0.024 | 26.62 | 0.009 | 22.94 | 42.75 | 71.43 | 91.55 |
| **All** | Depth4ToM [9] | 0.044 | 0.055 | 44.94 | 0.009 | 29.74 | 51.14 | 71.87 | 85.68 |
| | MODEST [26] | 0.087 | 0.123 | 111.43 | 0.019 | 58.05 | 59.23 | 72.61 | 79.44 |
| | D4RD [49] | 0.026 | 0.037 | 27.71 | 0.007 | 16.65 | 74.08 | 86.20 | 93.27 |
| | DKT† [55] | 0.111 | 0.144 | 137.48 | 0.021 | 70.21 | 61.05 | 74.63 | 81.00 |
| | Marigold [19] | 0.049 | 0.062 | 48.64 | 0.011 | 31.76 | 49.05 | 74.36 | 87.07 |
| | GenPercept [54] | 0.034 | 0.042 | 32.89 | 0.008 | 22.10 | 59.75 | 83.34 | 92.58 |
| | GeoWizard [14] | 0.027 | 0.036 | 26.80 | 0.008 | 17.33 | 75.65 | 86.41 | 93.93 |
| | DA3 [24] | 0.027 | 0.046 | 38.88 | 0.007 | 18.19 | 79.53 | 88.96 | 96.21 |
| | MoGe-2 [51] | 0.020 | 0.031 | 22.30 | 0.006 | 13.31 | 77.51 | 89.31 | 96.82 |
| | Ours (DA3) | 0.025 | 0.041 | 34.85 | 0.006 | 16.87 | 79.93 | 89.28 | 96.58 |
| | Ours (MoGe-2) | **0.016** | **0.022** | **17.87** | **0.004** | **11.19** | **80.63** | **92.77** | **99.05** |
| **Other** | Depth4ToM [9] | 0.044 | 0.055 | 45.56 | 0.009 | 30.09 | 51.74 | 72.23 | 85.55 |
| | MODEST [26] | 0.087 | 0.124 | 113.28 | 0.019 | 59.09 | 60.50 | 73.52 | 79.56 |
| | D4RD [49] | 0.023 | 0.033 | 25.71 | 0.006 | 15.54 | 76.76 | 88.53 | 94.44 |
| | DKT† [55] | 0.113 | 0.146 | 139.87 | 0.021 | 72.04 | 61.39 | 74.35 | 80.49 |
| | Marigold [19] | 0.049 | 0.062 | 49.09 | 0.011 | 32.04 | 49.60 | 74.97 | 87.25 |
| | GenPercept [54] | 0.034 | 0.042 | 33.07 | 0.008 | 22.17 | 60.27 | 83.79 | 92.69 |
| | GeoWizard [14] | 0.025 | 0.034 | 26.28 | 0.007 | 16.91 | 77.40 | 87.45 | 94.27 |
| | DA3 [24] | 0.026 | 0.045 | 38.78 | 0.006 | 18.04 | 81.19 | 89.94 | 96.58 |
| | MoGe-2 [51] | 0.016 | 0.021 | 17.97 | **0.003** | 11.24 | 80.61 | 92.55 | 99.03 |
| | Ours (DA3) | 0.024 | 0.041 | 34.93 | 0.006 | 16.87 | 81.07 | 89.73 | 96.71 |
| | Ours (MoGe-2) | **0.015** | **0.020** | **17.12** | **0.003** | **10.71** | **82.17** | **93.64** | **99.44** |

our method improves error metrics such as AbsRel and SiLog while maintaining competitive threshold accuracies. On the **All** and **Other** regions, **SeeClear** with MoGe-2 achieves the best performance across all metrics, indicating that improvements on transparent regions do not degrade predictions elsewhere. **SeeClear** with DA3 improves over the baseline on most metrics, with the largest gain in RMSE (38.88 mm to 34.85 mm).

**Quantitative Results on TransPhy3D.** As shown in Tab. 2, **SeeClear** achieves consistent improvements on the synthetic TransPhy3D benchmark. In particular, **SeeClear** with MoGe-2 attains the best performance on six of the eight metrics and improves over the MoGe-2 baseline across seven of eight metrics (tied on iRMSE). The largest gains compared to MoGe-2 are observed in AbsRel, which decreases from 0.016 to 0.012 (25%), and in $\delta_{1.025}$, which increases from 89.58% to 94.43%, reflecting more pixels whose predicted depth

| Method | AbsRel↓ | SiLog↓ | RMSE↓ | iRMSE↓ | MAE↓ | $\delta_{1.025}$↑ | $\delta_{1.05}$↑ | $\delta_{1.10}$↑ |
|---|---|---|---|---|---|---|---|---|
| Depth4ToM [9] | 0.027 | 0.051 | 60.13 | 0.006 | 30.27 | 72.39 | 88.79 | 96.29 |
| MODEST [26] | 0.035 | 0.061 | 67.54 | 0.007 | 37.40 | 66.97 | 85.26 | 93.81 |
| D4RD [49] | 0.017 | 0.039 | 51.76 | 0.004 | 23.67 | 90.94 | 96.40 | 98.20 |
| DKT† [55] | 0.014 | 0.029 | **33.20** | **0.003** | **13.88** | 89.20 | 96.33 | 98.57 |
| Marigold [19] | 0.027 | 0.044 | 50.72 | 0.005 | 27.88 | 66.37 | 88.76 | 97.36 |
| GenPercept [54] | 0.037 | 0.059 | 58.19 | 0.008 | 33.31 | 53.56 | 78.34 | 92.78 |
| GeoWizard [14] | 0.025 | 0.045 | 55.88 | 0.005 | 29.63 | 74.64 | 92.74 | 97.91 |
| DA3 [24] | 0.023 | 0.043 | 53.25 | 0.005 | 25.27 | 76.00 | 93.91 | 97.99 |
| MoGe-2 [51] | 0.016 | 0.030 | 39.78 | **0.003** | 19.21 | 89.58 | 96.17 | 98.62 |
| Ours (DA3) | 0.017 | 0.037 | 47.46 | 0.004 | 21.11 | 88.01 | 96.42 | 98.29 |
| Ours (MoGe-2) | **0.012** | **0.028** | 38.91 | **0.003** | 16.81 | **94.43** | **97.50** | **98.67** |

**Table 2: TransPhy3D.** Comparison on the synthetic test set. : best result. †: fine-tuned on this dataset.

| Region | Method | AbsRel↓ | SiLog↓ | RMSE↓ | iRMSE↓ | MAE↓ | $\delta_{1.025}$ ↑ | $\delta_{1.05}$ ↑ | $\delta_{1.10}$ ↑ |
|---|---|---|---|---|---|---|---|---|---|
| ToM | DA3 [24] | 0.031 | 0.036 | 28.53 | 7.5e-5 | 18.94 | 69.17 | 86.88 | 94.95 |
| | **Ours (DA3)** | **0.024** | **0.032** | **20.71** | **6.7e-5** | **12.85** | **73.67** | **88.67** | **95.92** |
| All | DA3 [24] | 0.020 | 0.031 | 21.03 | 6.4e-5 | 13.65 | 77.62 | 91.04 | 97.17 |
| | **Ours (DA3)** | **0.017** | **0.027** | **16.22** | **5.2e-5** | **10.26** | **80.96** | **92.99** | **98.04** |

**Table 3: Booster.** DA3 vs. **SeeClear**+DA3. **ToM**: transparent region; **All**: full image. : ours.

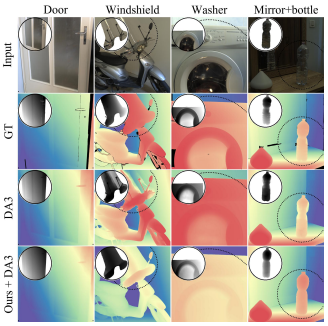


**Fig. 5: Booster Qualitative. SeeClear**+DA3 recovers more accurate depth than DA3 alone across diverse unseen categories.

closely match the ground truth. Similarly, **SeeClear** with DA3 consistently improves over DA3 across all eight metrics, reducing $\delta_{1.025}$ from 76.00% to 88.01% and RMSE from 53.25 mm to 47.46 mm. These results further demonstrate that appearance opacification enables general-purpose depth models to better handle transparent objects without requiring dataset-specific fine-tuning, offering stable and accurate depth estimation for complex transparent objects. Compared with specialized transparent-object depth methods, **SeeClear** with MoGe-2 achieves comparable or better performance to DKT (fine-tuned on this dataset) across most metrics without any dataset-specific training—surpassing DKT on AbsRel (0.012 vs. 0.014) while remaining competitive on RMSE and threshold accuracies. It also outperforms D4RD (AbsRel 0.017), Depth4ToM (0.027), and MODEST (0.035), as well as general diffusion-based depth models including Marigold (0.027), GeoWizard (0.025), and GenPercept (0.037).

**Quantitative Results on Booster.** We further evaluate **SeeClear** on the Booster Monocular Benchmark [36], which provides high-quality depth ground-truth and covers a wide variety of transparent-object categories. As shown in Tab. 3 and Fig. 5, **SeeClear** with DA3 consistently improves over DA3 on all eight metrics in both the transparent-object region (**ToM**) and the full image (**All**). The largest gain is on ToM MAE, which drops from 18.94 mm to 12.85 mm (a relative reduction of 32.2%). Importantly, Booster includes categories outside our training data, such as doors, mirrors, windshields, and washers, and **SeeClear** yields consistent improvements across all of them, demonstrating generalization to unseen transparent-object categories.

### 4.2 Qualitative Evaluation

Fig. 4 presents qualitative comparisons with the baseline in the same test set in Sec. 4.1. General depth models such as DA3 and MoGe-2 produce accurate depth in non-transparent regions but fail on transparent surfaces, which are often pre-

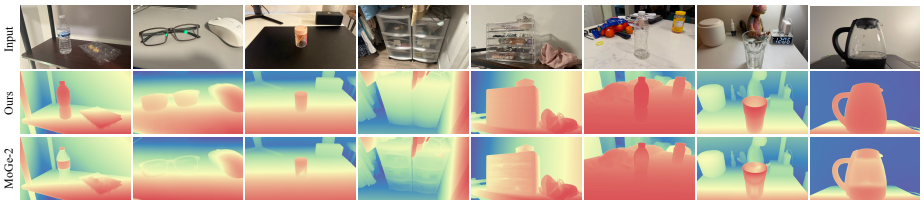


**Fig. 6: Qualitative comparison on in-the-wild images.** MoGe-2 exhibits depth leakage through transparent objects (e.g., a water bottle), while our method produces more accurate and consistent depth predictions.

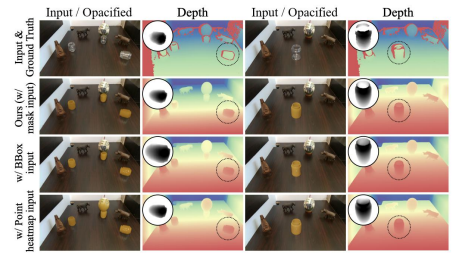


**Fig. 7: Transparent Object Localization Ablation.** The mask $M^{seg}$ provides precise spatial guidance, while bounding boxes and point heatmaps cause inaccurate opaque images and depths. Depth maps in the circle are normalized.

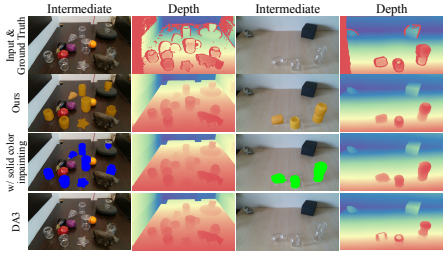


**Fig. 8: Opacified Image Ablation.** Flat-color opaque images lack shading cues, causing large depth errors, while our generative opacification restores plausible appearance for depth estimation and predicts accurate depth maps.

dicted as translucent due to the lack of reliable visual cues. Depth4ToM tends to produce solid but overly smooth depth predictions, resulting in vague object boundaries and reduced geometric detail. Our zero-shot method produces visually comparable transparent-object depth to DKT, which requires fine-tuning on this dataset. MODEST and D4RD often generate vague depth estimates, with interior regions appearing overly bright or translucent. GenPercept struggles with boundary accuracy and complex transparent structures, frequently producing translucent predictions and distorted depth near object edges. GeoWizard tends to generate piecewise or discretized depth with abrupt changes, leading to non-smooth geometry. Similarly, Marigold often produces discontinuous depth transitions and translucent artifacts. In contrast, our method produces stable and accurate depth predictions for complex transparent objects by converting transparent appearances into geometry-consistent opaque representations. For instance, for an open jam bottle, our method accurately predicts the normalized depth map (circled in Fig. 4), where the opening appears lighter due to the hollow region, while the front bottle body appears darker because it is closer to the camera. We also present qualitative comparisons on in-the-wild images in Fig. 6, demonstrating the robustness of our method on real-world transparent objects across diverse categories.

### 4.3 Ablation Study

We conduct an ablation study on the key components of the proposed framework, with quantitative results summarized in Tab. 4.

| Method | AbsRel↓ | SiLog↓ | RMSE↓ | iRMSE↓ | MAE↓ | $\delta_{1.025}$↑ | $\delta_{1.05}$↑ | $\delta_{1.10}$↑ |
|---|---|---|---|---|---|---|---|---|
| w/ BBox | 0.035 | 0.025 | 22.96 | 0.008 | 18.99 | 50.90 | 76.05 | 93.77 |
| w/ Point | 0.034 | 0.024 | 22.38 | 0.008 | 18.68 | 49.87 | 77.32 | 94.47 |
| w/ DINOv2 CLS | 0.037 | 0.025 | 23.85 | 0.008 | 20.08 | 49.78 | 75.84 | 92.73 |
| w/ DINOv2 full | 0.034 | 0.024 | 22.35 | 0.008 | 18.67 | 51.53 | 78.80 | 94.32 |
| w/ CLIP full | 0.034 | 0.024 | 22.30 | 0.008 | 18.58 | 48.91 | 77.99 | 95.08 |
| w/o MRM | **0.033** | 0.025 | 21.73 | 0.008 | 18.00 | 50.48 | **80.66** | 94.58 |
| w/o blending | 0.037 | 0.026 | 23.60 | 0.008 | 19.83 | 45.45 | 76.19 | 93.46 |
| DA3 (baseline) | 0.038 | 0.028 | 24.03 | 0.008 | 19.95 | 46.89 | 75.82 | 92.24 |
| Solid-color | 0.053 | 0.035 | 33.09 | 0.012 | 28.11 | 34.00 | 61.52 | 84.64 |
| **Ours** | **0.033** | **0.023** | **21.45** | **0.007** | **17.94** | **51.27** | 78.13 | **95.63** |

**Table 4: Ablation study** on ClearGrasp Real (ToM region). All variants use DA3 as the depth backbone. : best result.

**Transparent Object Localization.** To localize transparent objects, we use the mask $M^{seg}$ as spatial conditioning for generating opaque images. We compare two alternatives: bounding box, and point heatmap inputs. As shown in Tab. 4, the mask $M^{seg}$ consistently outperforms coarser cues. Using a bounding box increases RMSE from 21.45 mm to 22.96 mm, while a point heatmap yields 22.38 mm, since imprecise spatial guidance leads to over- or under-opacification near object boundaries. As illustrated in Fig. 7, bounding-box inputs provide correct localization but lack precise boundary delineation, resulting in inaccurate object contours; point heatmaps additionally suffer from ambiguous spatial extent, leading to unintended edits in surrounding regions.

**Image Encoder.** We ablate four variants: CLIP CLS tokens, CLIP full tokens, DINOv2 CLS tokens, and DINOv2 full tokens. As shown in Tab. 4, the CLIP class token performs best (RMSE 21.45 mm), while all alternatives degrade performance. As illustrated in Fig. 9, weaker conditioning tokens lead to incomplete opacification, where regions near the object boundary or interior remain translucent, causing the depth model to produce translucent depth artifacts in those areas.

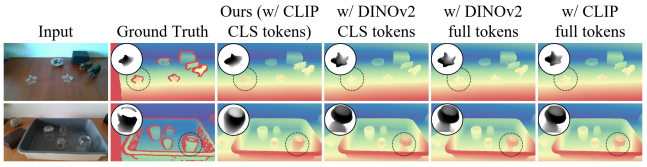


**Fig. 9: Image Encoder Ablation.** CLIP class tokens provide the most stable signal and achieve the best depth accuracy over DINOv2 variants.

**Image Compositing.** We explore a mask refinement module with binary compositing to integrate the opacified region from the generated image $I^{pred}$ with the background $I^{tr}$, preserving scene context. Without MRM, segmentation errors cause incomplete opacification near edges, increasing iRMSE from 0.007 to 0.008. Without blending, compositing is determined by patch boundaries; in multi-object scenes, a region shared by multiple patches may be overwritten by a later patch, restoring the

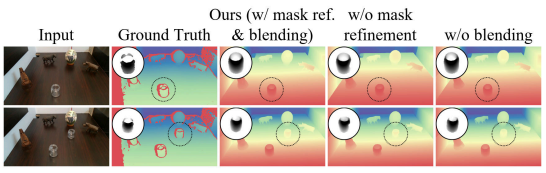


**Fig. 10: Image Compositing Ablation.** Removing MRM and binary compositing causes blurry boundaries and incomplete opacification.

transparent appearance. As shown in Fig. 10, this causes translucent artifacts in the region. This raises RMSE to 23.60 mm and reduces $\delta$<1.025 to 45.45%.

**Opacified Image.** Our method employs a generative opacification module to convert transparent regions into shading- and geometry-consistent opaque representations. To evaluate its effect, we replace the generated opaque image with a flat-filled baseline and feed the original input directly to the depth model. Following the solid-color inpainting strategy of Depth4ToM [9], we fill the transparent region with red, green, and blue colors respectively and take the pixel-wise median of the three depth predictions, using the same $M^{seg}$. This baseline yields the largest degradation across all ablations (RMSE 33.09 mm), falling below even the unmodified DA3 baseline (24.03 mm): flat fills carry no structural cues, and mask inaccuracies produce incorrect object boundaries that mislead the depth backbone. As shown in Fig. 8, while DA3 itself produces translucent depth artifacts on transparent regions, our diffusion-generated opaque images restore physically plausible surface appearance, enabling accurate depth estimation.

## 5 Limitations

**Failure Cases.** A typical failure mode occurs on large, nearly invisible glass panes where the surrounding frame is only partially visible: with limited contextual cues, **SeeClear** may fail to fully opacify the transparent region and depth can leak through, as illustrated in Fig. 11. Another failure mode occurs when the input mask severely mislocalizes the transparent region, causing the opacified content to be composited onto the wrong area.

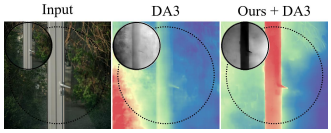


**Fig. 11: Failure Case. SeeClear** may fail to opacify large, nearly invisible glass panes where contextual cues are limited.

## 6 Conclusion

We presented **SeeClear**, a two-stage framework for depth estimation of transparent objects from a single image. By integrating a plug-and-play generative opacification module with mask-aware refinement and off-the-shelf monocular depth estimation, **SeeClear** decouples appearance opacification and depth inference. By converting transparent appearances into geometry-consistent opaque appearances, the framework removes refractive and transmissive effects that lie outside the training distribution of foundation depth models, enabling foundation depth models to produce accurate predictions without retraining. We plan to extend **SeeClear** to video-based depth estimation and more complex non-Lambertian materials, including specular, translucent, and multi-layer transparent objects. We aim to incorporate temporal consistency and spatial coherence constraints to ensure stable predictions and improved robustness under occlusions.

## References


1. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth (2023), https://arxiv.org/abs/2302.12288
2. Birkl, R., Wofk, D., Müller, M.: Midas v3. 1–a model zoo for robust monocular relative depth estimation. arXiv preprint arXiv:2307.14460 (2023)
3. Bradbury, R., Zhong, D.: Your latent mask is wrong: Pixel-equivalent latent compositing for diffusion models (2025), https://arxiv.org/abs/2512.05198
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18392–18402 (2023)
5. Careaga, C., Aksoy, Y.: Colorful diffuse intrinsic image decomposition in the wild. ACM Transactions on Graphics **43**(6), 1–12 (Nov 2024). https://doi.org/10.1145/3687984, http://dx.doi.org/10.1145/3687984
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts (2025), https://arxiv.org/abs/2511.16719
7. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6593–6602 (2024)
8. Chen, X., Zhang, H., Yu, Z., Opipari, A., Chadwicke Jenkins, O.: Clearpose: Large-scale transparent object dataset and benchmark. In: Eur. Conf. Comput. Vis. pp. 381–396. Springer (2022)
9. Costanzino, A., Ramirez, P.Z., Poggi, M., Tosi, F., Mattoccia, S., Di Stefano, L.: Learning depth estimation for transparent and mirror surfaces. In: Int. Conf. Comput. Vis. pp. 9244–9255 (2023)
10. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Adv. Neural Inform. Process. Syst. (2014)
11. Fang, H., Fang, H.S., Xu, S., Lu, C.: 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). https://doi.org/10.1109/LRA.2022.3183256
12. Flacco, F., Kroeger, T., De Luca, A., Khatib, O.: A depth space approach for evaluating distance to objects: with application to human-robot collision avoidance. Journal of Intelligent & Robotic Systems **80**(Suppl 1), 7–22 (2015)
13. Flacco, F., Kröger, T., De Luca, A., Khatib, O.: A depth space approach to human-robot collision avoidance. In: IEEE Int. Conf. Robot. Autom. pp. 338–345. IEEE (2012)
14. Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., Long, X.: Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In: Eur. Conf. Comput. Vis. pp. 241–258 (2024)
15. Guo, H., Zhu, H., Peng, S., Lin, H., Yan, Y., Xie, T., Wang, W., Zhou, X., Bao, H.: Multi-view reconstruction via sfm-guided monocular depth estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5272–5282 (2025)

16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
17. Hodan, T., et al.: Bop challenge 2020 on 6d object localization. In: ECCV Workshops (2020)
18. Kalra, A., Taamazyan, V., Rao, S.K., Venkataraman, K., Raskar, R., Kadambi, A.: Deep polarization cues for transparent object segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (June 2020)
19. Ke, B., Qu, K., Wang, T., Metzger, N., Huang, S., Li, B., Obukhov, A., Schindler, K.: Marigold: Affordable adaptation of diffusion-based image generators for image analysis. IEEE Trans. Pattern Anal. Mach. Intell. (2025). https://doi.org/10.1109/TPAMI.2025.3591076
20. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025)
21. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: Int. Conf. Comput. Vis. (2023), https://arxiv.org/abs/2304.02643
22. Lee, B.U., Lee, K., Oh, J., Kweon, I.S.: Cnn-based simultaneous dehazing and depth estimation. In: IEEE Int. Conf. Robot. Autom. pp. 9722–9728. IEEE (2020)
23. Li, J., Wang, W., Peng, Y., Shen, C., Zhu, Y., Xu, Z.: Visual robotic manipulation with depth-aware pretraining. In: IEEE Int. Conf. Robot. Biomim. pp. 843–850. IEEE (2024)
24. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
25. Liu, G., Ceylan, D., Yumer, E., Yang, J., Lien, J.M.: Material editing using a physically based rendering network. In: Int. Conf. Comput. Vis. pp. 2261–2269 (2017)
26. Liu, J., Ma, H., Guo, Y., Zhao, Y., Zhang, C., Sui, W., Zou, W.: Monocular depth estimation and segmentation for transparent object with iterative semantic and geometric fusion. arXiv preprint arXiv:2502.14616 (2025)
27. Ma, B., Ma, M., Li, R., Zheng, J., Li, D.: Tosq: Transparent object segmentation via query-based dictionary lookup with transformers. Sensors **25**(15), 4700 (2025)
28. Ma, K., Fang, Y., Weibel, J.B., Tan, S., Wang, X., Xiao, Y., Fang, Y., Xia, T.: Phys-liquid: A physics-informed dataset for estimating 3d geometry and volume of transparent deformable liquids. In: AAAI. pp. 7782–7790 (2026)
29. Ma, Z., Xu, R., Zhang, S.: Pixelgen: Improving pixel diffusion with perceptual supervision. arXiv preprint arXiv:2602.02493 (2026)
30. Mahler, J., Liang, J., Niyaz, S., Laskey, M., Doan, R., Liu, X., Ojea, J.A., Goldberg, K.: Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics (2017), https://arxiv.org/abs/1703.09312
31. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking. In: IEEE Int. Symp. on Mixed and Augmented Reality (ISMAR) (2011). https://doi.org/10.1109/ISMAR.2011.6092378
32. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2024), https://arxiv.org/abs/2304.07193

33. ten Pas, A., Gualtieri, M., Saenko, K., Platt, R.: Grasp pose detection in point clouds. The International Journal of Robotics Research (2017)
34. Qiu, T., Du, R., Spine, N., Cheng, L., Jiang, Y.: Joint 3d point cloud segmentation using real-sim loop: From panels to trees and branches (2025), https://arxiv.org/abs/2503.05630
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. pp. 8748–8763 (2021)
36. Ramirez, P.Z., Costanzino, A., Tosi, F., Poggi, M., Salti, S., Mattoccia, S., Stefano, L.D.: Booster: A benchmark for depth from images of specular and transparent surfaces. IEEE Trans. Pattern Anal. Mach. Intell. **46**(1), 85–102 (2024). https://doi.org/10.1109/TPAMI.2023.3323858
37. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Int. Conf. Comput. Vis. (2021)
38. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Trans. Pattern Anal. Mach. Intell. **44**(3), 1623–1637 (2020)
39. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded sam: Assembling open-world models for diverse visual tasks (2024), https://arxiv.org/abs/2401.14159
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10684–10695 (2022)
41. Sajjan, S., Moore, M., Pan, M., Nagaraja, G., Lee, J., Zeng, A., Song, S.: ClearGrasp: 3D shape estimation of transparent objects for manipulation. In: IEEE Int. Conf. Robot. Autom. pp. 3634–3642 (2020)
42. Schmidt, T., Hertkorn, K., Newcombe, R., Marton, Z., Suppa, M., Fox, D.: Depth-based tracking with physical constraints for robot manipulation. In: IEEE Int. Conf. Robot. Autom. pp. 119–126. IEEE (2015)
43. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: IEEE Conf. Comput. Vis. Pattern Recog. (2016)
44. Shankar, K., Tjersland, M., Ma, J., Stone, K., Bajracharya, M.: A learned stereo depth system for robotic manipulation in homes. IEEE Robotics and Automation Letters **7**(2), 2305–2312 (2022)
45. Sharma, P., Jampani, V., Li, Y., Jia, X., Lagun, D., Durand, F., Freeman, W.T., Matthews, M.: Alchemist: Parametric control of material properties with diffusion models (2023), https://arxiv.org/abs/2312.02970
46. Silberman, N., Fergus, R.: Indoor scene segmentation using a structured light sensor. In: Proceedings of the International Conference on Computer Vision - Workshop on 3D Representation and Recognition (2011)
47. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Eur. Conf. Comput. Vis. (2012)
48. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
49. Tosi, F., Zama Ramirez, P., Poggi, M.: Diffusion models for monocular depth estimation: Overcoming challenging conditions. In: Eur. Conf. Comput. Vis. (2024)
50. Wang, L., Tran, D.M., Cui, R., TG, T., Dahl, A.B., Bigdeli, S.A., Frisvad, J.R., Chandraker, M.: Materialist: Physically based editing using single-image inverse rendering (2025), https://arxiv.org/abs/2501.03717

51. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. arXiv preprint arXiv:2507.02546 (2025)
52. Wang, Y., Wu, T., Zou, Q.: 6dof pose estimation of transparent objects: Dataset and method. Sensors **26**(3), 898 (2026). https://doi.org/10.3390/s26030898
53. Xie, E., Wang, W., Wang, W., Sun, P., Xu, H., Liang, D., Luo, P.: Segmenting transparent object in the wild with transformer (2021), https://arxiv.org/abs/2101.08461
54. Xu, G., Ge, Y., Liu, M., Fan, C., Xie, K., Zhao, Z., Chen, H., Shen, C.: What matters when repurposing diffusion models for general dense perception tasks? arXiv preprint arXiv:2403.06090 (2024)
55. Xu, S., Wei, S., Wei, Q., Geng, Z., Li, H., Shen, L., Sun, Q., Han, S., Ma, B., Li, B., et al.: Diffusion knows transparency: Repurposing video diffusion for transparent object depth and normal estimation. arXiv preprint arXiv:2512.23705 (2025)
56. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18381–18391 (2023)
57. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10371–10381 (2024)
58. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything V2. In: Adv. Neural Inform. Process. Syst. (2024)
59. Zhang, B., Wang, Z., Ling, Y., Guan, Y., Zhang, S., Li, W., Wei, L., Zhang, C.: Shuffletrans: Patch-wise weight shuffle for transparent object segmentation. Neural Networks **167**, 199–212 (2023)
60. Zhang, J., Yang, K., Constantinescu, A., Peng, K., Müller, K., Stiefelhagen, R.: Trans4trans: Efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world. In: ICCV Workshops. pp. 1760–1770 (2021)
61. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 586–595 (2018)
62. Zhao, W., Bai, L., Rao, Y., Zhou, J., Lu, J.: UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In: Adv. Neural Inform. Process. Syst. vol. 36, pp. 49842–49869 (2023)