

TRAN-D: 2D Gaussian Splatting-based Sparse-view Transparent Object Depth Reconstruction via Physics Simulation for Scene Update

Supplementary Material

A. Implementation Details

Segmentation Generic prompts like “glass” or “transparent” often misclassify background regions or merge overlapping objects. To avoid these failures, we create an intentionally non-dictionary, category-specific text prompt “786dvpteg” to unambiguously denote “transparent object,” while all other objects use the generic prompt “object.” We did not include the “object” prompt during training. We fine-tuned Grounded DINO for 1 epoch with a batch size of 12, conducted on a single NVIDIA A6000 GPU. The training was performed on a synthetic TRAnsPose [16] dataset generated using BlenderProc [3]. Only the image backbone portion was fine-tuned, while the text backbone BERT [4] layers remained frozen.

Gaussian Optimization The Gaussian Splatting model optimization was performed on a single NVIDIA RTX 2080 Ti GPU. The model parameters are mostly identical to those used in the 2DGS method. We began the optimization process with random points. The learning rate for the object index one-hot vector was initialized at 0.1 and decayed to 0.0025 over 1000 iterations. Following object removal, the scene was refined over the course of 100 iterations.

Physics Simulation In our physics-based scene update, we employ the Material Point Method (MPM), which represents bodies as material points carrying mass, momentum, and material properties, and simulates their interactions under gravity and collisions with other objects to compute forces and update positions.

To generate a suitable particle distribution, we first convert our optimized 2D Gaussian splatting surface into a mesh via surface reconstruction of the rendered depth map. We then sample particles at approximately uniform spacing over this mesh. Each particle is assigned material properties—Young’s modulus 5×10^4 Pa and Poisson’s ratio 0.4—values selected after evaluating several combinations because this pair consistently yielded the fastest simulation times without any observable loss in dynamic fidelity or visual quality. Other parameter settings provided comparable physical accuracy but incurred higher computational cost.

We run the MPM simulation for 100 timesteps. At each step, particle masses and velocities are used to compute stresses and body forces including gravity. Collisions are resolved against static geometry, with the ground height set to 0 to account for contact and collisions with the floor, as well as neighboring particles. Particle positions are updated accordingly. After simulation, we perform a brief 100 it-

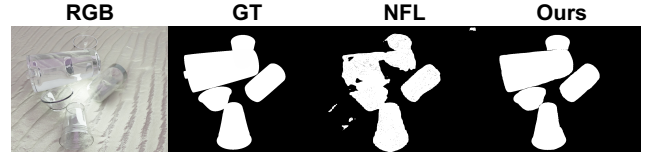


Figure 9. Even in boundary-ambiguous regions, our method delivers accurate object segmentation, whereas the mask-based NFL produces noisy or incomplete masks.

erations of Gaussian re-optimization, omitting the object-aware 3D loss since surface consistency was already enforced during the initial optimization.

B. Segmentation Results

We compare the segmentation performance of NFL for transparent-object reconstruction (Fig. 9). Despite targeting transparent objects, these methods struggle when boundaries become ambiguous due to lighting variations, background clutter, or occlusions.

Moreover, despite ambiguous boundaries, cluttered backgrounds, and varied viewpoints, our model consistently produces accurate masks for transparent objects across all scenarios (Fig. 10).

C. Additional Qualitative Results

In this supplementary material, we present additional qualitative results highlighting the robustness of TRAN-D.

As illustrated in Fig. 12, TRAN-D consistently produces accurate and stable depth reconstructions across a variety of object types and background textures. Notably, TRAN-D maintains high performance even in challenging scenarios, such as scenes featuring complex, marble-like backgrounds where transparent objects are visually difficult to distinguish, and cases with significant object overlap or occlusion.

Moreover, Fig. 13 presents additional real-world results demonstrating the practicality and effectiveness of TRAN-D in real-world applications. Unlike synthetic scenarios, real-world conditions introduce complexities such as closely placed or interdependent objects, significantly complicating the reconstruction task. Despite these challenges, TRAN-D achieves reliable depth reconstructions, highlighting its potential for use in complex, real-world environments. For comprehensive visualization of results across more scenes, please refer to the supplementary video.

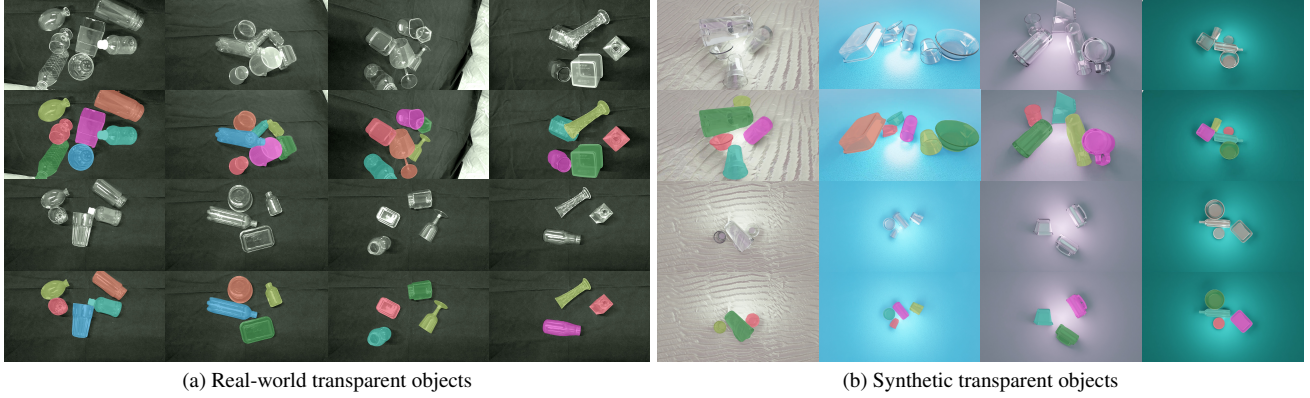


Figure 10. Segmentation result visualization for transparent objects.

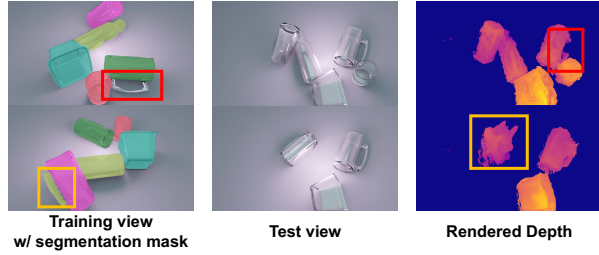


Figure 11. Failure cases due to segmentation inaccuracies. In the first case (yellow box), incorrect labeling in the segmentation process results in poor reconstruction of the pink object, which causes the physics simulation to fail at $t = 1$. In the second case (red box), part of the object is not segmented, leading to incomplete reconstruction at $t = 0$.

D. Qualitative Analysis on the Number of Views

In this section, we analyze the qualitative performance of TRAN-D with different numbers of training views. Fig. 14 illustrates that increasing the number of training views generally leads to better depth reconstruction, with only a minimal difference observed between 6 and 12 views. This indicates that while additional views offer extra information, the model’s performance stabilizes after a certain number of views (around 6 in our case).

TRAN-D performs well even with just 6 views, offering reliable depth reconstructions with minimal artifacts. However, increasing the number of views beyond 6 does not result in a substantial improvement, which implies diminishing returns in terms of reconstruction accuracy as the number of views increases.

E. Failure Cases and Discussion

Our Grounded SAM, trained on transparent objects, provides object-level segmentation masks that distinguish ob-

jects from the background and enable object-aware loss computation. This offers a significant advantage to our model; however, it is also highly sensitive to segmentation quality.

In sparse-view scenarios, large viewpoint shifts complicate consistent object tracking, and physical properties of transparent objects, such as reflection and refraction under intense lighting, introduce ambiguity in object boundaries. These factors, combined with inadequate segmentation coverage in occluded areas, further degrade reconstruction quality. As shown in Fig. 11, when parts of an object are not grouped together or are incorrectly segmented as different objects, the reconstruction suffers and the physics simulation may fail.

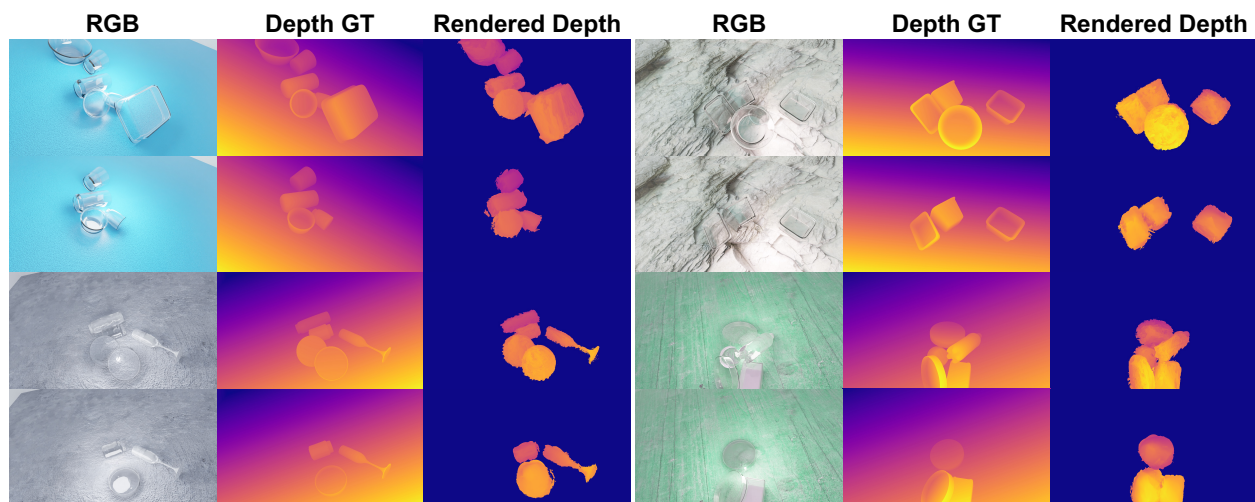


Figure 12. Depth reconstruction results of synthetic sequences.

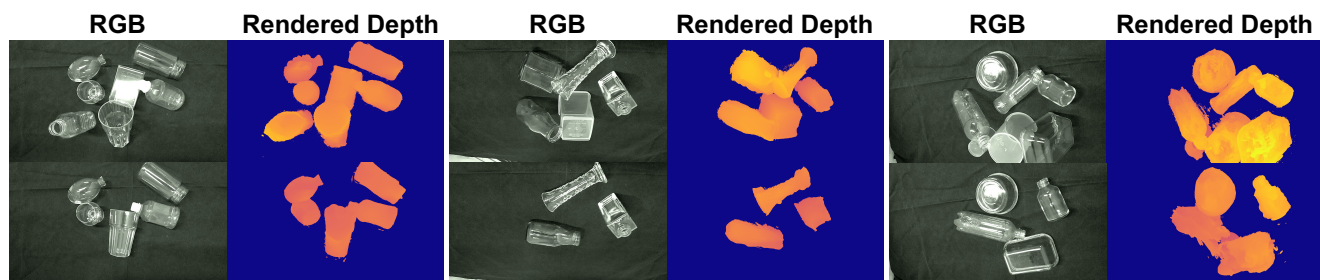


Figure 13. Depth reconstruction results of real-world sequences.

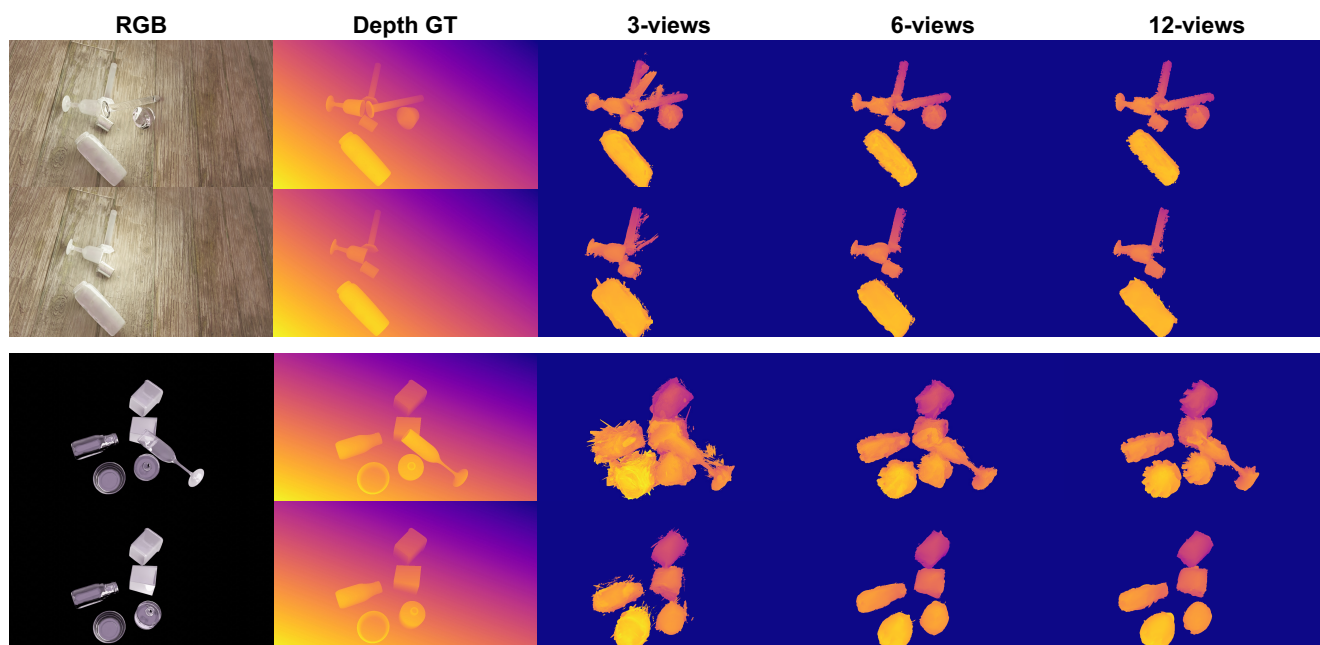


Figure 14. Depth reconstruction results of 3, 6, 12 views in our model.