O^2-Recon: Completing 3D Reconstruction of Occluded Objects in the Scene with a Pre-trained 2D Diffusion Model

Yubin Hu[1], Sheng Ye[1], Wang Zhao[1], Matthieu Lin[1], Yuze He[1], Yu-Hui Wen[3], Ying He[2], Yong-Jin Liu[1]

[1]Tsinghua University
[2]Nanyang Technological University
[3]Beijing Jiaotong University

Abstract

Occlusion is a common issue in 3D reconstruction from RGB-D videos, often blocking the complete reconstruction of objects and presenting an ongoing problem. In this paper, we propose a novel framework, empowered by a 2D diffusion-based in-painting model, to reconstruct complete surfaces for the hidden parts of objects. Specifically, we utilize a pre-trained diffusion model to fill in the hidden areas of 2D images. Then we use these in-painted images to optimize a neural implicit surface representation for each instance for 3D reconstruction. Since creating the in-painting masks needed for this process is tricky, we adopt a human-in-the-loop strategy that involves very little human engagement to generate high-quality masks. Moreover, some parts of objects can be totally hidden because the videos are usually shot from limited perspectives. To ensure recovering these invisible areas, we develop a cascaded network architecture for predicting signed distance field, making use of different frequency bands of positional encoding and maintaining overall smoothness. Besides the commonly used rendering loss, Eikonal loss, and silhouette loss, we adopt a CLIP-based semantic consistency loss to guide the surface from unseen camera angles. Experiments on ScanNet scenes show that our proposed framework achieves state-of-the-art accuracy and completeness in object-level reconstruction from scene-level RGB-D videos. Code: https://github.com/THU-LYJ-Lab/O2-Recon.

1 Introduction

The task of reconstructing 3D objects within a scene has been a longstanding challenge in computer vision. Unlike scene-level reconstruction techniques (Azinovic et al. 2022; Wang, Bleja, and Agapito 2022), object-level 3D reconstruction focuses on creating individual representations for each instance within a scene. This technique is crucial for applications in computer vision, robotics, and mixed reality that require fined-grained scene modeling and understanding.

Many works approach object-level 3D reconstruction as a task of estimating an object's pose and shape code, using a categorical generative model (Rünz et al. 2020; Shan et al. 2021). While these methods create complete shapes, they are limited to reconstructing objects from specific categories, like tables or chairs. Even within these categories, the generated shape codes often struggle to accurately match the actual object surfaces. There are also a few approaches focusing retrieving suitable CAD models from a database and estimating their 9 degrees of freedom poses (Avetisyan et al. 2019). These methods also face similar issues, such as limited scalability and low accuracy in reconstruction.

Figure 1: Occlusion presents significant hurdles for object-level reconstruction. For the occluded armchair, highlighted in blue, existing methods can only yield a partial reconstruction where a significant portion of the geometry is missing.

Benefiting from the emerging technology of neural radiance fields (NeRF) (Mildenhall et al. 2020; Sucar et al. 2021), vMap (Kong et al. 2023) is able to reconstruct a wider variety of objects, moving beyond just categorical instances. However, it does not address the issue of occlusion in scene-level videos, which results in incomplete observations of objects and reduced reconstruction quality. As illustrated in Figure 1, the camera paths in 3D indoor scenes often limit the coverage of scene-level videos. As a result, objects close to walls or to each other are frequently only partially recorded. The lack of complete visuals, especially the absence of information for the occluded regions, makes these images inadequate for neural rendering-based reconstruction methods (Wang et al. 2021; Yariv et al. 2021).

Inspired by the recent success of diffusion-based image in-painting (Wang et al. 2023b), we explore the application of a pre-trained diffusion model to in-paint the occluded regions in the input video frames. While the latent diffusion model (Rombach et al. 2022) is adept at in-painting missing regions in images, it may produce drastically incorrect content without precise in-painting masks that identify the missing parts. In this paper, we address this challenge by in-

troducing affordable human interaction into our framework, thereby ensuring both the accuracy of the masks and the overall quality of the in-painting process.

Provided with an RGB-D video sequence accompanied by object masks, our system requires a user to choose between 1 to 3 frames containing occlusion. The user is then guided to sketch the in-painting masks for these frames, utilizing their experience and judgement. These sketched masks are subsequently re-projected to all other views, utilizing depth information in-painted by the diffusion model, and then merged to create the in-painting masks for the remaining frames. By incorporating cost-effective human engagement, our proposed approach ensures the generation of high-quality in-painting masks. These masks maintain robust geometric consistency across various views, thereby guiding the 2D diffusion model to create convincing and coherent in-paintings for the occluded regions. As for the reconstruction stage, we utilize the neural implicit surface representation like NeuS (Wang et al. 2021) and optimize it with rendering loss. Given the possible visual inconsistency across the in-painted images, the implicit representation can filter the inconsistency during the multi-view rendering-based optimization and reconstruct reasonable underlying surfaces.

To mitigate the reconstructed artifacts in areas that are entirely unseen, our system enhances the rendering-based reconstruction from two perspectives: first, by adopting semantic supervision over the unseen regions; second, by applying a smoothness prior of the neural implicit surface. In the case of semantic supervision, we guide the reconstruction by supervising the CLIP (Radford et al. 2021) features of renderings from novel views within both the image and text domains. For smoothness, we introduce a cascaded architecture for predicting signed distance field (SDF), which is specially designed to prevent noisy artifacts in the unseen regions. To achieve this, we utilize a shallow MLP equipped with low-frequency positional encodings (PEs), ensuring overall smoothness of the surface. Concurrently, we adopt a deeper auxiliary branch, armed with high-frequency PEs, to predict residuals of SDF. This dual approach is effective in maintaining superior expressiveness of visible regions while ensuring a balanced and coherent reconstruction.

To sum up, the main contributions of our work include:

1) A 3D reconstruction framework for occluded objects in the scene, termed O^2-Recon, that addresses the occlusion problem by employing diffusion-based in-painting within the 2D image domain. 2) A human-in-the-loop strategy for in-painting mask generation, enabling the production of high-quality masks with minimal human engagement, which are used to guide the diffusion-based 2D in-painting process. 3) The creation of a novel cascaded SDF prediction network, coupled with semantic consistency supervision using CLIP, to enhance the surface quality of completely unseen regions in occluded objects.

We conduct extensive experiments on ScanNet scenes, demonstrating that our proposed framework achieves state-of-the-art reconstruction accuracy and completeness for occluded objects in the scene. With the complete objects reconstructed by our method, we enable further object-level manipulations with highly free translations and rotations.

2 Related Works

Object-Level 3D Scene Reconstruction. Generating an independent 3D representation for individual objects within a scene is an active research area. Many methods seek to address this problem through the joint optimization of an object's shape code and pose. For instance, FroDO (Rünz et al. 2020) utilizes a pre-trained encoder-decoder network inspired by DeepSDF (Park et al. 2019) to map RGB images to a sparse point cloud and a dense SDF field using a latent shape code as a proxy. ELLIPSDF (Shan et al. 2021) introduces a bi-level object model that captures both the coarse-level scale and the fine-level shape details, enhancing the joint optimization process for object pose and shape code. To enable real-time reconstruction, MOLTR (Li, Rezatofighi, and Reid 2021) removes the backward optimization and focuses on predicting the shape code by multi-view image encodings. It leverages another pre-trained 3D detector to predict the objects' 9-DoF poses. Departing from the typical two-stage pipeline, CenterSnap (Irshad et al. 2022) and RayTran (Tyszkiewicz et al. 2022) unifies pose and shape estimation into a single-stage network. Instead of predicting shape codes, RayTran directly predicts the SDF volume. Despite their ability for reconstructing complete shapes for individual objects, these methods are typically constrained to specific categories such as tables or chairs. Additionally, models that are pre-trained on synthetic datasets like ShapeNet (Chang et al. 2015) often struggle when applied to real-world scenarios, since the surfaces decoded from shape codes might not accurately represent actual objects.

Another class of methods leverages CAD databases instead of the generative models, retrieving suitable models (Avetisyan et al. 2019) and applying deformations (Ishimtsev et al. 2020) to align with actual objects. However, the inherent limitation of deformation operation implies that these methods lack the flexibility to accurately represent real-world objects and the ability for high-fidelity reconstruction.

There are also approaches that utilize NeRF (Mildenhall et al. 2020) for object-level reconstruction of arbitrary 3D objects. For example, vMap (Kong et al. 2023) represents each object with an independent NeRF, and optimizes it through photometric loss. While effective in certain scenarios, this approach fails to handle occlusion, often resulting in incomplete and degenerated surfaces when parts of objects are not visible. Object-NeRF (Yang et al. 2021) addresses the misleading supervision of incomplete instance masks with the use of a 3D guard mask, however it still relies on the intrinsic smoothness bias of NeRF (similar to vMap) to mitigate the occluded regions. RICO (Li et al. 2023) regularizes the unseen areas through object-background relationship, but it falls short in providing effective supervision for the occluded parts, leaving room for further improvement.

Our approach differs from the above methods in that it explicitly supervises the occluded regions using in-paintings generated by a pre-trained 2D diffusion model. It offers two unique features. Firstly, it reconstructs accurate surfaces for *arbitrary* objects by relying on a neural implicit surface representation. Secondly, the application of diffusion-based in-painting model enables our method to reconstruct *complete* shapes of objects, even when they are partially occluded.

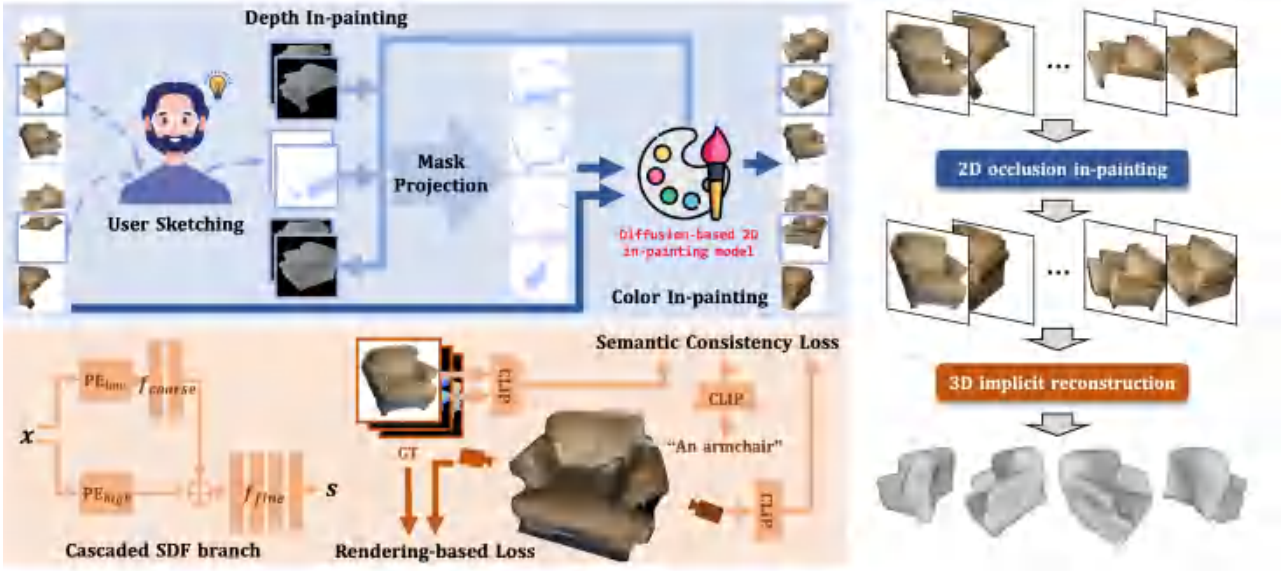


Figure 2: The proposed O^2-Recon framework. We utilize the Stable Diffusion in-painting model[1] in our implementation.

NeRFs Empowered by 2D Diffusion Models. The success of diffusion models in 2D image generation and editing (Saharia et al. 2022; Kawar et al. 2023) has motivated interest in combining these advanced 2D models with NeRF representations. To achieve NeRF-level editing guided by text instructions, instruct-NeRF2NeRF (Haque et al. 2023) iteratively updates the multi-view image dataset with the edited images, which are rendered by the pre-trained InstructPix2Pix model (Brooks, Holynski, and Efros 2023). Guided by the pre-trained Stable Diffusion model, the recently proposed RePaint-NeRF (Zhou et al. 2023) facilitates local editing within selected areas in NeRF scenes. There are also works for generating NeRFs from text prompts with the aid of 2D diffusion models by different approaches, such as score distillation sampling (Poole et al. 2023; Lin et al. 2023), score Jacobian chaining (Wang et al. 2023a), and variational score distillation (Wang et al. 2023c).

In this paper, we utilize a diffusion-based 2D in-painting model to aid 3D reconstruction of occluded objects. The pre-trained diffusion model is used to in-paint the occluded regions in the 2D images. These enhanced images subsequently serve as the foundation to reconstruct complete shapes for occluded objects.

3 Method

Given an RGB-D video clip composed of N image frames $\{I_n\}_{n=1}^N$ and depth frames $\{D_n\}_{n=1}^N$, we assume that high-quality instance and semantic segmentation results $\{S_n^I\}_{n=1}^N$ and $\{S_n^S\}_{n=1}^N$ are already obtained by existing methods such as (Kong et al. 2023; Xie et al. 2021). In this paper, we aim to reconstruct the complete shapes of occluded objects. As shown in Figure 2, our method begins by in-painting the occluded regions in images, utilizing a pre-trained diffusion model, which in our implementation is the Stable Diffusion in-painting model (Rombach et al. 2022). We then reconstruct the 3D object using a neural implicit surface representation that compensates for the entirely unseen regions (an example is provided in the rightmost part of Figure 2).

In Section 3.1, we elaborate on our proposed 2D in-painting process for occluded objects in images, employing the pre-trained diffusion model with minimal human engagement. In the subsequent neural implicit surface based reconstruction, we design a cascaded network architecture for the SDF branch, effectively preventing degenerated high-frequency artifacts in the unseen areas (see Section 3.2). Finally, we discuss the loss functions utilized in the entire optimization process in Section 3.3.

3.1 Diffusion-based 2D Occlusion In-painting

Utilizing the instance segmentation, we first extract the object mask $\{M_n^i\}_{n=1}^N$ for each object with the identifier i:

$$M_n^i = \mathbb{1}_{A_i}(x, y), \quad A_i = \{(x, y) | S_n^I(x, y) = i\}. \tag{1}$$

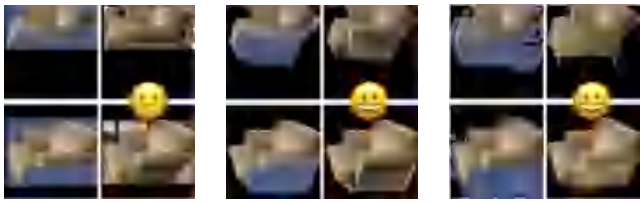
Subsequently, we apply the Hadamard product to the extracted masks and RGB-D frames. This process yields the masked RGB images $\{I_n^i\}_{n=1}^N$ and depths $\{D_n^i\}_{n=1}^N$ for each object i, as defined by

$$I_n^i = I_n \circ M_n^i, \quad D_n^i = D_n \circ M_n^i. \tag{2}$$

These masked data, typically incomplete for occluded objects, can present incorrect boundaries that may disrupt downstream rendering-based geometry optimization. We address this challenge by completing the occluded objects in images using a pre-trained diffusion model.

Note that to achieve satisfactory in-painting results, text prompts and high-quality mask prompts must be provided to the 2D diffusion model. However, generating accurate in-panting masks for occluded objects is a highly non-trivial task. Simply utilizing the 2D bounding box of the visible region as an in-painting mask may lead to background

[1]https://huggingface.co/runwayml/stable-diffusion-inpainting



(a) (b) (c)

Figure 3: Illustration of results produced by different in-painting masks: (a) bounding box masks, (b) user-sketched masks, and (c) masks projected from selected views.

contents appearing inside the object region. Sometimes, the bounding box only encompasses part of the whole object, resulting in incomplete in-painting, as shown in Figure 3a. Moreover, since the occluded areas may vary significantly between different views, predicting geometrically consistent in-painting masks through automated algorithms poses a technical challenge. To overcome the challenge, we propose a human-in-the-loop mask generation strategy that requires minimal human engagement.

User Sketching. As shown in the upper part of Figure 2, we enlist the assistance of a user to sketch the in-painting mask on 1 to 3 representative images. This process does not necessitate specialized expertise from the participants and can be completed in just 1 to 2 minutes for each object.

Depth In-painting. Building upon the user-sketched 2D masks, we aim to re-project these masks to generate the in-painting masks for all other frames. However, this operation is complicated by the absence of depth information in the sketched area due to occlusion. To project in-painting masks from the sketched images to the correct regions of other images, we utilize the pre-trained diffusion model to predict pseudo depth for the sketched area. By treating the depth map as a grayscale image, we feed both the masked depth frame and the sketched in-painting masks into the diffusion model, which yields a predicted completed depth map.

Mask Projection. We formulate the mask projection as:

$$\tilde{M}^i_m = Merge(\{\tilde{M}^i_{n\to m} | n \in \text{selected views}\}), \quad (3)$$

$$\tilde{M}^i_{n\to m} = Proj(\tilde{M}^i_n, P_n, P_m, K_m), \quad (4)$$

where n is the source view (i.e., the view containing user sketches) and m denotes the target view. $\tilde{M}^i_m$ represents the in-painting mask for object i in frame m. P_m and K_m are the extrinsic and intrinsic matrix of the depth frame m. $Proj(\cdot)$ and $Merge(\cdot)$ denote the mask projection and merging process, respectively.

Color In-painting. We take the in-painting masks and incomplete color images as inputs and feed them into the diffusion model for in-painting. Thanks to the human-in-the-loop strategy, we are able to generate high-quality in-painting masks $\{\tilde{M}^i_n\}^N_{n=1}$, which effectively guide and enhance the filling process. Though these projected masks might not be precisely accurate, they serve as valuable indicators of the visible contours for the occluded objects. This helps to prevent the intrusion of background contents into the object region and effectively directs the creation of plausible object shapes, as illustrated in Figure 3c.

Mask Refining. Once the in-painting has been completed, we further refine the object masks according to the in-painted RGB images. The updated masks for object i are denoted by $\{\hat{M}^i_n\}^N_{n=1}$. For both depth and color in-painting processes, we adopt text prompts that reflect the semantic class identified in $\{S^S_n\}^N_{n=1}$ by *"A/an ${CLASS}."* We refer readers to the supplementary material for more details about mask projection and refining.

3.2 Cascaded SDF Prediction

In scene-level videos, the limited number of camera views available for each individual object falls short in guiding rendering-based optimization, creating a unique challenge in sparse-view object reconstruction. As highlighted in recent works (Mu et al. 2023; Long et al. 2022; Ren et al. 2023), the SDF network involved in neural implicit surface reconstruction tends to overfit to the color appearance, rather than accurately learning the surface geometry. This overfitting often leads to artifacts, such as degeneration in the unseen regions.

To tackle this issue, we propose a cascaded network architecture to enhance the smoothness priors in the SDF branch. As shown in the bottom-left section of Figure 2, our architecture adopts a two-part structure, differing from popular neural implicit surfaces, such as NeuS (Wang et al. 2021) that typically uses a single large MLP. The first part is a coarse prediction block with low-frequency positional encodings PE_{low} and shallow MLP layers f_{coarse}. The second part is a refinement block, employing high-frequency positional encodings PE_{high} and deep MLP layers f_{fine}.

We train the cascaded network using a two-stage strategy that separates low-frequency geometry and high-frequency fine details. In the first stage, we focus on training the coarse SDF prediction block for generating a smooth surface. This initial stage is pivotal as it guards against rapid overfitting to the restricted camera views, thereby providing a stable initialization. Recognizing surfaces obtained in the first stage may be over-smoothed and lack fine details, the second stage comes into play. In this phase, we activate the refinement block, working in conjunction with the coarse block, to predict SDF residuals. In the cascaded network, the SDF value s at a certain 3D point x is formulated as

$$s = f_{fine}([f_{coarse}(PE_{low}(x)), PE_{high}(x)]). \quad (5)$$

Experiments show that our two-stage training strategy, separating the learning of low-frequency and high-frequency signals, stabilizes the training process while enhancing the network's ability of capturing fine-grained geometry of visible areas. This approach strikes a balance between overall smoothness and intricate detail. Furthermore, by employing this cascaded architecture, we enhance the reconstruction quality of entirely unseen surfaces, enabling the manipulation of reconstructed objects even under large rotations.

3.3 Loss Functions

Given the in-painted images, the updated object masks, and the original depth information, we optimize the implicit representation with the sum of following loss functions.

Rendering-based Loss. Using the volume rendering equation in MonoSDF (Yu et al. 2022), we can render the

expected color, normal and depth values of ray $\mathbf{r}$ and supervise them with the ground truth values. The ground truth surface normal maps are predicted from the in-painted RGB images by SNU (Bae, Budvytis, and Cipolla 2021) following the practice in NeuRIS (Wang et al. 2022). For the color and normal values, we sample $\mathbf{r}$ from valid rays $\hat{\mathcal{R}}$ in the updated object masks $\{\hat{M}_n^i\}_{n=1}^N$ after in-painting. For the depth values, we sample $\mathbf{r}$ from valid rays $\mathcal{R}$ in the original incomplete object masks $\{M_n^i\}_{n=1}^N$, because we empirically find neither the in-painted nor the predicted depth values are reliable. The rendering-based loss can be summarized as

$$\begin{aligned}\mathcal{L}_r =& \lambda_{\mathcal{C}} \mathbb{E}_{\mathbf{r}\in\hat{\mathcal{R}}}(\|\hat{\mathcal{C}}(\mathbf{r}) - \mathcal{C}(\mathbf{r})\|_1) \\ &+ \lambda_{\mathcal{N}} \mathbb{E}_{\mathbf{r}\in\hat{\mathcal{R}}}(\|1 - \hat{\mathcal{N}}(\mathbf{r})^T \mathcal{N}(\mathbf{r})\|_1) \\ &+ \lambda_{\mathcal{D}} \mathbb{E}_{\mathbf{r}\in\mathcal{R}}(\|\hat{\mathcal{D}}(\mathbf{r}) - \mathcal{D}(\mathbf{r})\|_1),\end{aligned} \tag{6}$$

where $\hat{\mathcal{C}}(\mathbf{r})$, $\hat{\mathcal{N}}(\mathbf{r})$ and $\hat{\mathcal{D}}(\mathbf{r})$ denote the rendered color, normal and depth values of ray $\mathbf{r}$, respectively.

Eikonal Loss. For all sampled points along the ray, we add an Eikonal term (Gropp et al. 2020) following the common practice to regularize SDF values in the 3D space

$$\mathcal{L}_{eik} = \lambda_e \mathbb{E}_{x\in\mathcal{X}}(\|1 - \nabla_x s(x)\|_1), \tag{7}$$

where $\mathcal{X}$ represents the set of sampled points.

Silhouette Loss. Inspired by methods like GET3D (Gao et al. 2022), we incorporate a binary cross entropy loss for the summed weights along the ray to supervise the 3D shape from the 2D silhouette projection, which is formulated as

$$\mathcal{L}_{si} = \lambda_{si} \mathcal{L}_{CE}(w(\mathbf{r}), \hat{M}(\mathbf{r})), \tag{8}$$

where $w(\mathbf{r})$ denotes the summation of weights along $\mathbf{r}$ and $\hat{M}(\mathbf{r})$ denotes the binary value in the updated object mask.

Semantic Consistency Loss. To improve the supervision of totally unseen areas, we apply a semantic consistency loss to the rendered color and normal images from novel views. Using the pre-trained CLIP (Radford et al. 2021) as encoder, we align the rendered images from novel views to the semantic space represented by the categorical text prompt $\mathcal{T}$ for in-painting and the color image $\mathcal{C}_r$ and normal image $\mathcal{N}_r$ from the reference view. The reference view is selected by the CLIP similarity of color images and the text prompt. Our proposed semantic consistency loss can be formulated as

$$\begin{aligned}\mathcal{L}_{se} = \lambda_{se}(&\|2 - \phi(\mathcal{C}_n)^T \phi(\mathcal{T}) - \phi(\mathcal{N}_n)^T \phi(\mathcal{T})\|_1 \\ &+ \|1 - \phi(\mathcal{C}_n)^T \phi(\mathcal{C}_r)\| + \|1 - \phi(\mathcal{N}_n)^T \phi(\mathcal{N}_r)\|),\end{aligned} \tag{9}$$

where $\phi(\cdot)$ denotes the CLIP encoder, $\mathcal{C}_n$ and $\mathcal{N}_n$ denote the color and normal image rendered from the novel views.

Since the semantic features need to be calculated from the whole rendered images, we render $\mathcal{C}_n$ and $\mathcal{N}_n$ at a low resolution for efficiency. The semantic consistency loss is applied every several iterations to the randomly generated novel views. We refer readers to supplementary material for more details about the novel view generation.

3.4 Implementation Details

We train our model on an NVIDIA GeForce RTX 3090 GPU for total 50k iterations using Adam optimizer, including 20k iterations for the coarse block alone and 30k iterations for both the coarse and fine blocks. The semantic consistency loss is turned on after 10k iterations and applied every 5 iterations. We set the initial learning rate to 2e-4 and decrease it by 0.5× every 20k iterations. The overall training process for one object occupies around 3GB GPU memory and takes about 4 hours. The loss weights are set to $\lambda_{\mathcal{C}} = \lambda_{\mathcal{N}} = \lambda_{\mathcal{D}} = \lambda_{se} = 1.0$, $\lambda_e = 0.1$, and $\lambda_{si} = 5.0$.

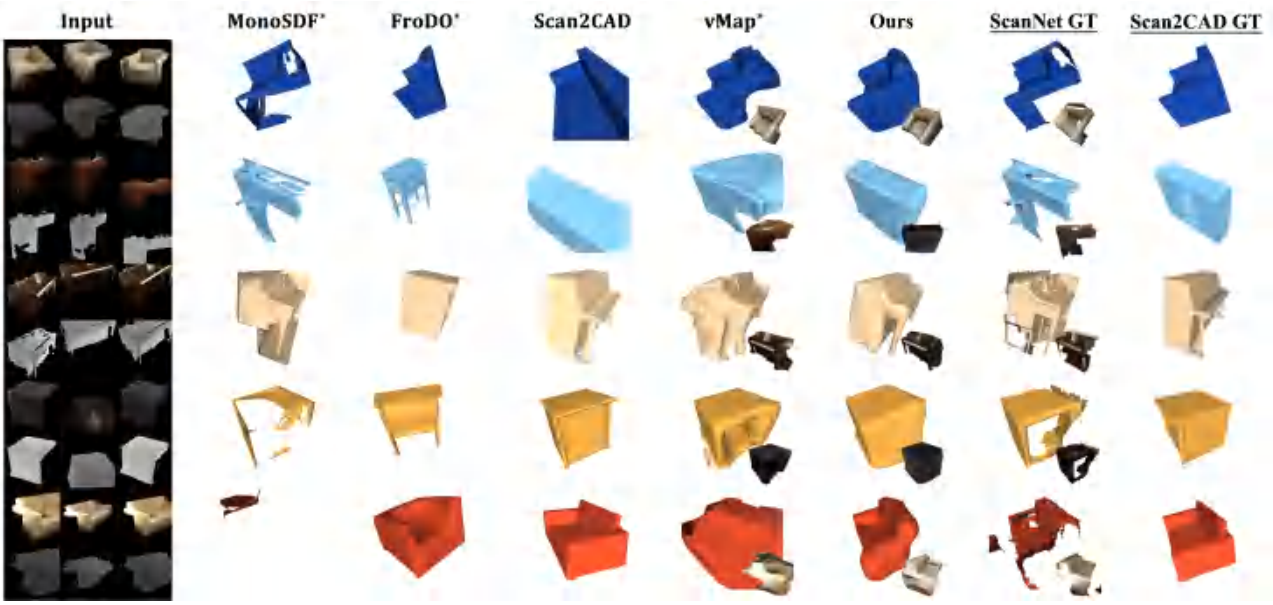


Figure 4: Visualization of occluded objects reconstructed by different methods. The occlusion conditions can be visualized in the column of Input and ScanNet GT, the missing parts indicate occlusion in the corresponding regions.

4 Experiments

4.1 Datasets, Baselines and Evaluation Metrics

Datasets. We evaluate O^2-Recon on 6 scenes from ScanNet (Dai et al. 2017), encompassing 77 objects. A significant portion of these objects are occluded, presenting difficulties in achieving complete reconstructions. We follow the practice in vMap (Kong et al. 2023) to get the segmentation results of the scene-level videos. Since the 3D ground truth of occluded regions are missing in the ScanNet, we leverage the aligned CAD models in Scan2CAD (Avetisyan et al. 2019) to evaluate the reconstruction accuracy of unseen regions. In our experiments, 59% of the occluded objects require 1 user-sketched mask, 29% require 2 user-sketched masks, and 12% require 3 user-sketched masks.

Baselines. We compare our method with the following state of the arts. (1) The scene-level reconstruction method MonoSDF (Yu et al. 2022). We leverage the ground truth depth in its optimization and denote it as MonoSDF*. (2) The re-implemented shape-code-based method FroDO (Rünz et al. 2020), denoted by FroDO*. (3) The Scan2CAD method (Avetisyan et al. 2019) based on CAD model retrieval. (4) The general object-level reconstruction method vMap (Kong et al. 2023). We optimize it with more iterations for fair comparison and denote it as vMap*. We refer readers to SM for more details about the baseline methods.

Metrics. We follow the previous work (Kong et al. 2023) to evaluate the reconstruction accuracy with the F-score within 5cm and the Chamfer distance (cm). We also report the accuracy and completion terms for detailed analysis.

4.2 Comparisons

Qualitative Evaluation. Figure 4 compares the reconstruction results of different methods on objects that suffer from various occlusion conditions. Notably, the scene-level method MonoSDF* cannot reconstruct complete surfaces for occluded regions, and sometimes fails on certain cases, e.g., the last row. As for the FroDO* method based on shape code, although complete meshes can be derived from the latent space, it cannot match the actual surface very well, and cannot reconstruct 3D objects of arbitrary categories, e.g., the piano and the trash bin. The Scan2CAD method can retrieve proper CAD models from the database, but the optimized scale and pose parameters are often unsatisfactory. The NeRF-based method vMap* generates accurate surfaces for visible areas of arbitrary objects, but produces holes or degenerated artifacts in the unseen areas. In contrast, our proposed system O^2-Recon simultaneously reconstructs accurate surfaces for visible regions and plausible surfaces for invisible regions. We ensure the accuracy of visible surfaces by rendering-based SDF optimization and the plausibility of unseen surfaces by the 2D in-painting.

Quantitative Evaluation. We use the ground truth in ScanNet and Scan2CAD datasets to evaluate the accuracy of the object-level reconstructions. The ScanNet ground truth contains accurate surfaces fused from depth inputs, which can be utilized to measure the accuracy of visible regions. As for the occluded areas, we measure the accuracy and plausibility using ground truth in the Scan2CAD dataset, which contains annotated CAD models for objects in the scenes that are complete and roughly match the actual objects.

Method	ScanNet GT				Scan2CAD GT			
	F-score ↑	Acc. ↓	Comp. ↓	Chamfer Dist. ↓	F-score ↑	Acc. ↓	Comp. ↓	Chamfer Dist. ↓
MonoSDF*	0.627	8.60	11.04	9.82	0.217	8.18	14.25	11.22
FroDO*	0.357	11.00	11.44	11.22	0.387	8.92	11.20	10.05
Scan2CAD	0.219	8.05	20.61	14.33	0.328	9.05	19.90	14.45
vMap*	0.636	17.47	**3.33**	10.40	0.471	21.17	**5.28**	13.23
O^2-Recon **(Ours)**	**0.715**	**4.32**	4.57	**4.45**	**0.568**	**5.96**	6.34	**6.15**

Table 1: Evaluation of object reconstruction on the ScanNet and Scan2CAD datasets.

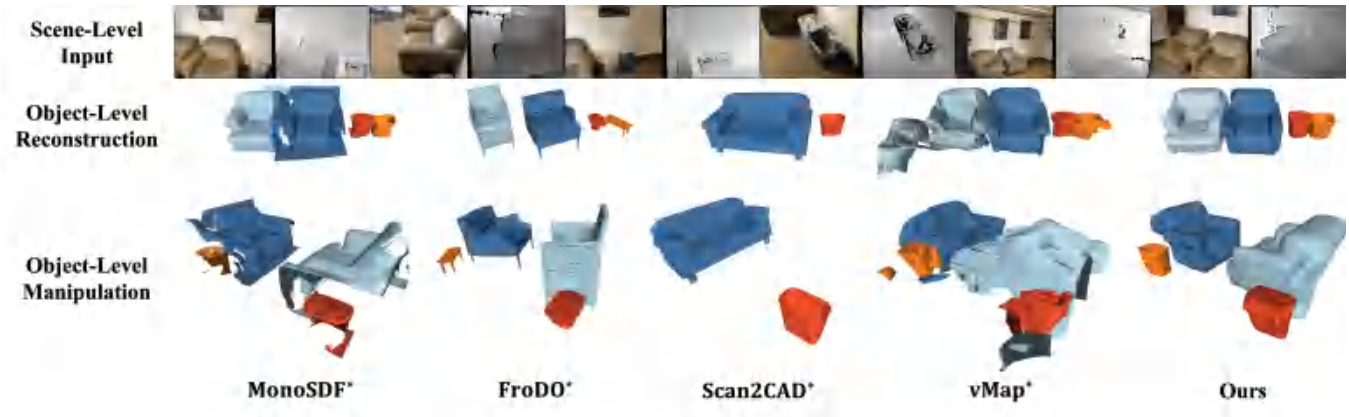


Figure 5: Comparison of object-level manipulation results.

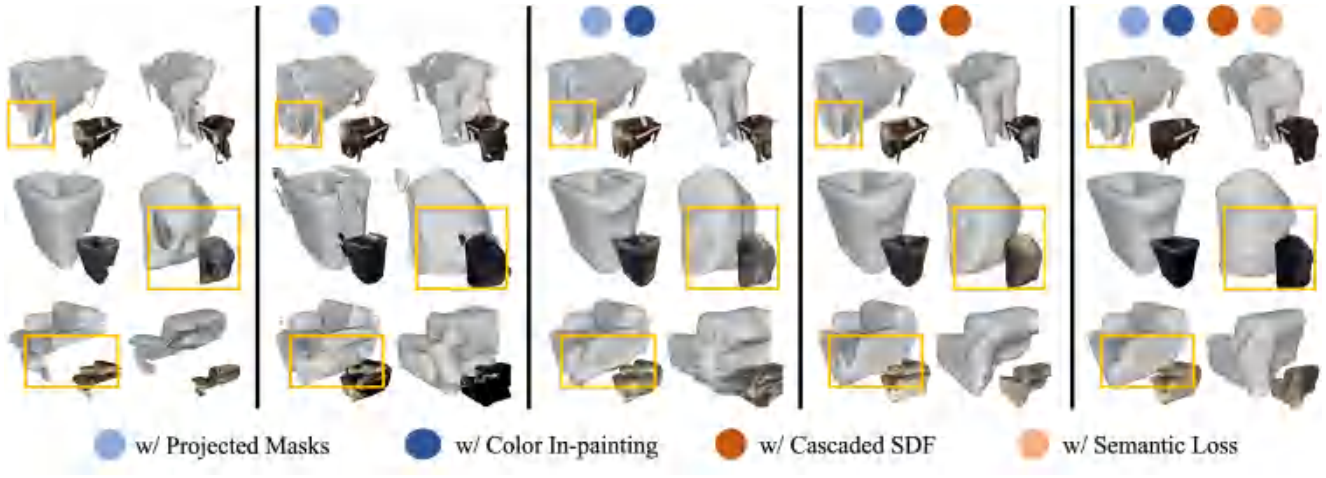


Figure 6: The ablation studies demonstrate how our carefully designed components progressively fill in the occluded regions.

	V1	V2	V3	V4	Full
w/ Projected Masks		✓	✓	✓	✓
w/ Color In-painting			✓	✓	✓
w/ Cascaded SDF				✓	✓
w/ Semantic Loss					✓
ScanNet F-score ↑	0.718	0.697	**0.720**	0.713	0.715
ScanNet C.D. ↓	4.70	5.17	4.65	4.53	**4.45**
Scan2CAD F-score ↑	0.502	0.511	0.553	0.560	**0.568**
Scan2CAD C.D. ↓	9.73	8.72	6.80	6.51	**6.15**

Table 2: Quantitative results for the ablation studies.

As shown in Table 1, our method outperforms all baseline methods in terms of the overall F-score and Chamfer distance. Compared to the baseline methods, our method reduces the Chamfer distance by around 50% and improves the F-score by more than 10%. We also notice that vMap performs better in the completion term but receives the largest error in the accuracy term, since it reconstructs a lot of surfaces in the empty space, as shown in Figure 4. These quantitative results are consistent with our qualitative analysis, and demonstrate the superiority of our proposed method.

Object-Level Manipulation. Based on the independent reconstructed objects, we can achieve object-level manipulation with few artifacts due to the high accuracy and completeness of O^2-Recon. As shown in Figure 5, 3D reconstructions generated by O^2-Recon maintain a good visualization effect after large-scale manipulation. While the 3D manipulation results based on other methods contain artifacts like missing or floating parts and inaccurate geometry.

4.3 Ablation Study

We evaluate the impact of different components on achieving complete 3D reconstruction by comparing our full method to four other variants as shown in Table 2.

Color In-painting. The 2D color in-painting is a crucial component in our method. While we can already reconstruct some occluded areas using the projected masks without color in-painting, these masks are not precise enough. This lack of precision leads to surfaces that can look distorted or noisy, as seen in the second column of Figure 6. By adopting color in-painting followed by mask refinement, we produce more accurate shapes for obscured areas. This is evident in the third column of Figure 6. The quantitative results reported in Table 2 also confirms the effectiveness of the color in-painting step for improving both the F-score and Chamfer distance measures.

Cascaded SDF Architecture and Semantic Loss. To enhance the reconstruction of completely hidden areas, we introduce a cascaded SDF architecture coupled with a semantic consistency loss. Both strategies, as illustrated in the last two columns of Figure 6 and Table 2, contribute to a smoother and more precise reconstructed surface in these unseen regions. In particular, the supervision provided by the semantic consistency loss significantly boosts the color consistency of the resulting surface.

5 Conclusion

In this paper, we introduce O^2-Recon for reconstructing complete 3D geometry of occluded objects in a scene using a pre-trained 2D diffusion model. We utilize the diffusion model to in-paint the occluded parts in multi-view 2D images, and then reconstruct 3D objects using neural implicit surface from the in-painted images. To prevent inconsistency in mask generation, we adopt a human-in-the-loop strategy that can effectively guide the 2D in-painting process with only a few human interaction. During the optimization process of neural implicit surfaces, we design a cascaded SDF architecture to guarantee smoothness, and also leverage the pre-trained CLIP model to supervise novel views with semantic consistency loss. Our experiments on the ScanNet scenes show that O^2-Recon is able to reconstruct accurate and complete 3D surfaces for occluded objects from any category. The reconstructed 3D objects can be utilized in further manipulation like large rotations and translations.

Acknowledgments

This work was partially supported by the Fundamental Research Funds for the Central Universities (2023XKRC045), the Natural Science Foundation of China (Project Number U2336214) and the Key Laboratory of Pervasive Computing, Ministry of Education, China.

References

Avetisyan, A.; Dahnert, M.; Dai, A.; Savva, M.; Chang, A. X.; and Nießner, M. 2019. Scan2CAD: Learning CAD Model Alignment in RGB-D Scans. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2614–2623. Computer Vision Foundation / IEEE.

Azinovic, D.; Martin-Brualla, R.; Goldman, D. B.; Nießner, M.; and Thies, J. 2022. Neural RGB-D Surface Reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 6280–6291. IEEE.

Bae, G.; Budvytis, I.; and Cipolla, R. 2021. Estimating and Exploiting the Aleatoric Uncertainty in Surface Normal Estimation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 13117–13126. IEEE.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 18392–18402. IEEE.

Chang, A. X.; Funkhouser, T. A.; Guibas, L. J.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; Xiao, J.; Yi, L.; and Yu, F. 2015. ShapeNet: An Information-Rich 3D Model Repository. *CoRR*, abs/1512.03012.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T. A.; and Nießner, M. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2432–2443. IEEE Computer Society.

Gao, J.; Shen, T.; Wang, Z.; Chen, W.; Yin, K.; Li, D.; Litany, O.; Gojcic, Z.; and Fidler, S. 2022. GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images. In *NeurIPS*.

Gropp, A.; Yariv, L.; Haim, N.; Atzmon, M.; and Lipman, Y. 2020. Implicit Geometric Regularization for Learning Shapes. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 3789–3799. PMLR.

Haque, A.; Tancik, M.; Efros, A. A.; Holynski, A.; and Kanazawa, A. 2023. Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 19740–19750.

Irshad, M. Z.; Kollar, T.; Laskey, M.; Stone, K.; and Kira, Z. 2022. CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and Categorical 6D Pose and Size Estimation. In *2022 International Conference on Robotics and Automation, ICRA 2022, Philadelphia, PA, USA, May 23-27, 2022*, 10632–10640. IEEE.

Ishimtsev, V.; Bokhovkin, A.; Artemov, A.; Ignatyev, S.; Niessner, M.; Zorin, D.; and Burnaev, E. 2020. Cad-deform: Deformable fitting of cad models to 3d scans. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, 599–628. Springer.

Kawamura, R. 2017. RectLabel. https://rectlabel.com/about.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-Based Real Image Editing With Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6007–6017.

Kong, X.; Liu, S.; Taher, M.; and Davison, A. J. 2023. vMAP: Vectorised Object Mapping for Neural Field SLAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 952–961.

Li, K.; Rezatofighi, H.; and Reid, I. 2021. MOLTR: Multiple Object Localization, Tracking and Reconstruction From Monocular RGB Videos. *IEEE Robotics Autom. Lett.*, 6(2): 3341–3348.

Li, Z.; Lyu, X.; Ding, Y.; Wang, M.; Liao, Y.; and Liu, Y. 2023. RICO: Regularizing the Unobservable for Indoor Compositional Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 17761–17771.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023. Magic3D: High-Resolution Text-to-3D Content Creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 300–309.

Long, X.; Lin, C.; Wang, P.; Komura, T.; and Wang, W. 2022. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, 210–227. Springer.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision*, 405–421. Springer.

Mu, T.-J.; Chen, H.-X.; Cai, J.-X.; and Guo, N. 2023. Neural 3D reconstruction from sparse views using geometric priors. *Computational Visual Media*, 1–11.

Park, J. J.; Florence, P. R.; Straub, J.; Newcombe, R. A.; and Lovegrove, S. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 165–174. Computer Vision Foundation / IEEE.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Ren, Y.; Zhang, T.; Pollefeys, M.; Süsstrunk, S.; and Wang, F. 2023. Volrecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16685–16695.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 10674–10685. IEEE.

Rünz, M.; Li, K.; Tang, M.; Ma, L.; Kong, C.; Schmidt, T.; Reid, I. D.; Agapito, L.; Straub, J.; Lovegrove, S.; and Newcombe, R. A. 2020. FroDO: From Detections to 3D Objects. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 14708–14717. Computer Vision Foundation / IEEE.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, S. K. S.; Lopes, R. G.; Ayan, B. K.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*.

Shan, M.; Feng, Q.; Jau, Y.; and Atanasov, N. 2021. ELLIPSDF: Joint Object Pose and Shape Optimization with a Bi-level Ellipsoid and Signed Distance Function Description. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 5926–5935. IEEE.

Sucar, E.; Liu, S.; Ortiz, J.; and Davison, A. J. 2021. iMAP: Implicit Mapping and Positioning in Real-Time. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 6209–6218. IEEE.

Tyszkiewicz, M. J.; Maninis, K.-K.; Popov, S.; and Ferrari, V. 2022. RayTran: 3D pose estimation and shape reconstruction of multiple objects from videos with ray-traced transformers. In *European Conference on Computer Vision*, 211–228. Springer.

Wang, H.; Du, X.; Li, J.; Yeh, R. A.; and Shakhnarovich, G. 2023a. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12619–12629.

Wang, J.; Bleja, T.; and Agapito, L. 2022. GO-Surf: Neural Feature Grid Optimization for Fast, High-Fidelity RGB-D Surface Reconstruction. In *International Conference on 3D Vision, 3DV 2022, Prague, Czech Republic, September 12-16, 2022*, 433–442. IEEE.

Wang, J.; Wang, P.; Long, X.; Theobalt, C.; Komura, T.; Liu, L.; and Wang, W. 2022. Neuris: Neural reconstruction of indoor scenes using normal priors. In *European Conference on Computer Vision*, 139–155. Springer.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 27171–27183.

Wang, S.; Saharia, C.; Montgomery, C.; Pont-Tuset, J.; Noy, S.; Pellegrini, S.; Onoe, Y.; Laszlo, S.; Fleet, D. J.; Soricut, R.; Baldridge, J.; Norouzi, M.; Anderson, P.; and Chan, W. 2023b. Imagen Editor and EditBench: Advancing and Evaluating Text-Guided Image Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18359–18369.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023c. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *CoRR*, abs/2305.16213.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 12077–12090.

Yang, B.; Zhang, Y.; Xu, Y.; Li, Y.; Zhou, H.; Bao, H.; Zhang, G.; and Cui, Z. 2021. Learning Object-Compositional Neural Radiance Field for Editable Scene Rendering. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 13759–13768. IEEE.

Yariv, L.; Gu, J.; Kasten, Y.; and Lipman, Y. 2021. Volume Rendering of Neural Implicit Surfaces. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 4805–4815.

Yu, Z.; Peng, S.; Niemeyer, M.; Sattler, T.; and Geiger, A. 2022. MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction. In *NeurIPS*.

Zhou, X.; He, Y.; Yu, F. R.; Li, J.; and Li, Y. 2023. RePaint-NeRF: NeRF Editting via Semantic Masks and Diffusion Models. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, 1813–1821. ijcai.org.

O^2-Recon: Completing 3D Reconstruction of Occluded Objects in the Scene with a Pre-trained 2D Diffusion Model

Supplementary Material

	Mean	Std	Max	Min
F-score [<5cm %] ↑	0.573	0.016	0.588	0.550
Acc. Dist. [cm] ↓	5.79	0.36	6.30	5.37
Comp. Dist. [cm] ↓	6.37	0.32	0.68	0.60
Chamfer Dist. [cm] ↓	6.08	0.11	6.20	5.91

Table S1: Reconstruction accuracy statistics of O^2-Recon with different user engagement.

S1 The Tool for User Sketching

We use the RectLabel software (Kawamura 2017) as a tool to facilitate the user sketching step. The interface, as depicted in Figure S1, enables users to conveniently navigate through all the RGB images associated with an object. By utilizing the brush tool provided by RectLabel, users can perform segmentation annotation, sketching the areas that require in-painting based on their expertise and experience.

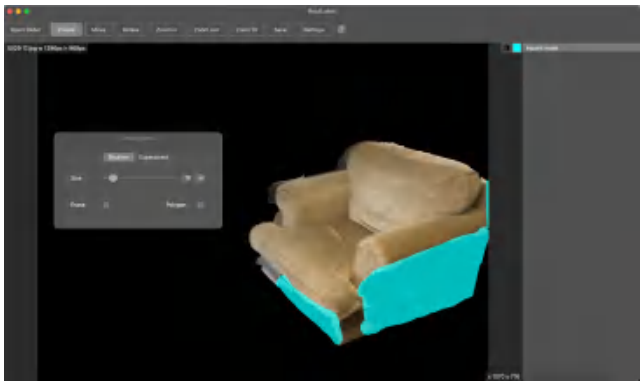


Figure S1: A screenshot of the user sketching GUI.

S2 Masks Sketched by Different Users

To study the influence of masks sketched by different users, we ask 5 people to select frames and draw in-painting masks for all the 77 objects. We evaluate the reconstruction results on Scan2CAD annotations and report the statistics in Table S1. The results show that our system consistently reconstructs 3D meshes with good quality given different sketched masks of different selected frames.

S3 Comparison with All User-Sketched Masks

To validate the effectiveness of our proposed human-engaged mask generation process, we conducted a comparison with high-quality all human-sketched in-painting masks. The results of this comparison are depicted in Figure S2. In the figure, we can observe that both the quality of the in-painted images and the reconstructed 3D mesh are comparable to those obtained using all-user-sketched masks. This indicates that our semi-automatic in-painting process yields satisfactory results with significantly less cost of labor and time.

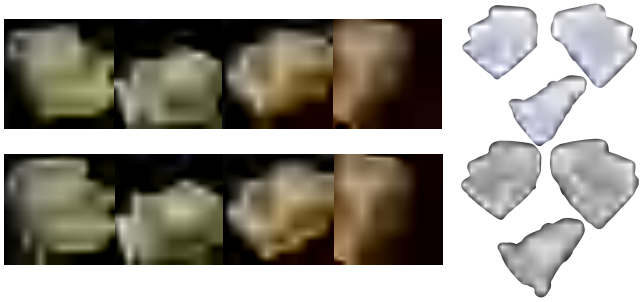


Figure S2: Comparison of surface reconstruction from all user-sketched in-painting masks (top) and inpainting masks generated from our pipeline (bottom).

S4 Depth In-painting and Mask Projection

In the mask projection step, our goal is to propagate the in-painting masks from the selected views to all other views.

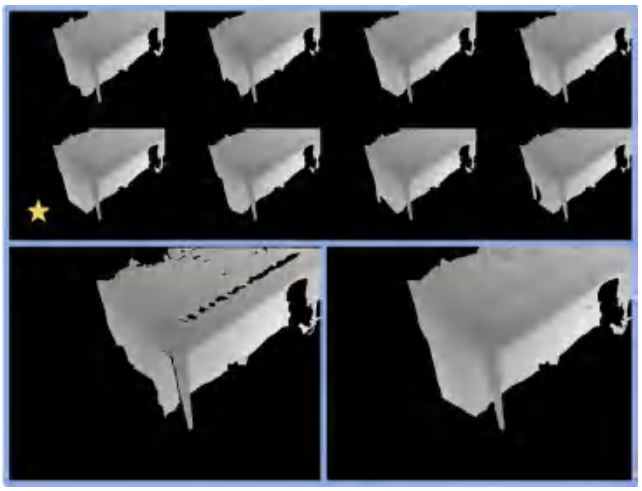


Figure S3: Middle and final results of the depth in-painting step. The upper part shows 8 in-painting results, and the lower part shows the incomplete input depth and the produced complete depth result.

To propagate the mask information through re-projection, we utilize the diffusion model to in-paint depth values within the sketched area. We leverage the Stable Diffusion in-

painting model (Rombach et al. 2022) as the diffusion model in our implementation. The depth maps are firstly normalized from $[0, D_{max}]$ to $[0, 255]$, where D_{max} denotes the maximum depth value. Treating the normalized depth maps as grayscale images, we input them into the diffusion-based in-painting model. After undergoing the in-painting process, the in-painted gray images are then scaled back to $[0, D_{max}]$ and form the in-painted depth maps. To resist the randomness of the diffusion model and ensure the quality of in-painted depth values, we in-paint each selected view 8 times and select the output with the largest valid area. Since we have only 1-3 selected views per instance, the time cost consumed by the repeated in-painting process is acceptable. As shown in Figure S3, our repeat-and-select process is effective and can produce good in-painted depth maps.

Using the depth information predicted by the diffusion model, we can perform a back-projection of the mask pixels from the 2D images to the corresponding 3D space. This process generates a point cloud that encompasses the mask pixels from all the selected views. Subsequently, these 3D points are projected back onto the 2D images of all other views. However, it is worth noting that the point cloud projection often leads to sparsity in the resulting 2D images. Moreover, the presence of inaccurate depth predictions can lead to incorrect projections, resulting in isolated points that fall outside the subject area. To obtain good in-painting masks, we apply morphological processing operations after the mask projection step. These operations facilitate refinement and improvement of the masks by filling in sparse regions and eliminating isolated points that fall outside the subject area.

Specifically, after the projection of 3D points, we firstly discard the pixels that have less than 5 positive neighbors in their 3x3 neighborhood. Such operation filters the noisy isolated points outside the subject area. To connect the sparse components within the subject area, we then apply a dilation operation to the filtered mask image. This dilation process helps bridge gaps and fills in missing regions, resulting in solid and connected regions. These regions can then be interpreted as reliable masks for further processing. To avoid modifications to the visible regions, we then subtract the original object mask from the dilated areas. To address potential pixel-level error in the instance masks, we perform an additional dilation operation after the subtraction step, and leverage the results as in-painting masks for the diffusion model.

With the generated in-painting masks, the pre-trained diffusion model is able to plausibly in-paint the occluded regions of objects. However, due to variations in neighborhood information across different pixels, certain regions within the in-painting masks may be filled with *black pixels* by the diffusion model. It is important to note that these *black pixels* cannot be considered as valid areas and should be excluded from the instance masks of the in-painted images. Thus we update the instance masks after in-painting by excluding all-zero pixels. This post-inpainting mask update not only eliminates the *black pixels* but also helps filter out any remaining noisy regions that may have resulted from the mask projection step.

Figure S4 illustrates the evolving process of object masks throughout the entire in-painting procedure. The final instance masks exhibit significant improvements compared to the original occluded masks, offering more comprehensive guidance for the geometry optimization process. By applying various operations such as dilation, subtraction, and post-inpainting mask updating, we refine the object masks to ensure better alignment with the completed regions. These refined instance masks play a crucial role in providing accurate and detailed supervision during the optimization of the geometry, ultimately leading to more visually pleasing and faithful reconstructions.

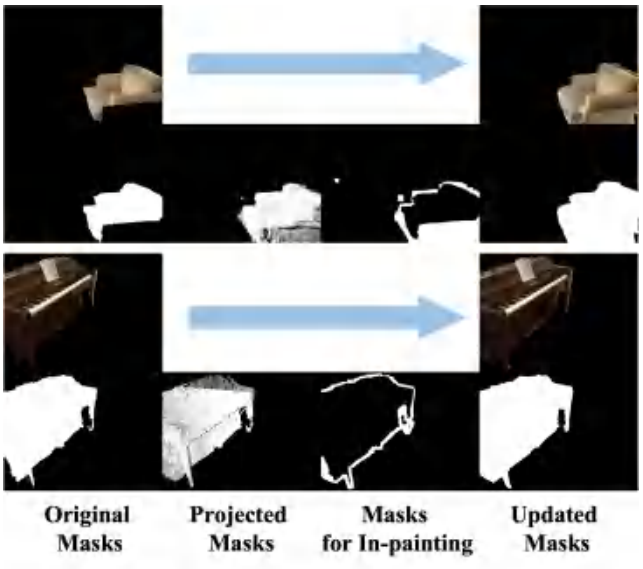


Figure S4: The middle results of in-painting masks for views that are not selected.

S5 Novel View Generation

To effectively supervise the implicit representation from novel views, we first randomly select novel camera poses based on the captured views, and then render the low-resolution outputs from that viewpoint for efficiency. In this section, we describe our novel view selection strategy and the techniques for fast rendering during the training process.

For iterations that requires the semantic consistency supervision, we randomly choose a reference camera pose from the captured viewpoints. However, since most reference poses are not oriented towards the center of the object, we perform a correction step to ensure that a significant portion of the object falls within the center region of the novel view. To correct the view direction of the reference pose, we set it as a vector pointing towards the center of the object from the reference pose's position. The center point coordinates can be obtained from the point cloud of the object, which is generated by fusing the masked depth frames. Next, we generate a novel camera pose near the corrected reference pose. Specifically, we parameterize the camera pose using three parameters: yaw angle, pitch angle, and radius with

respect to the center point of the object. The angle parameters are sampled from normal distributions with predefined standard deviations, where the mean values are set to the corresponding parameters of the corrected reference pose. The radius of the novel pose is sampled from a uniform distribution, with the maximum and minimum values determined as follows:

$$r_{min} = \min(\frac{D_{bbox}}{2}, \{d_i\}_{\min}), \tag{S1}$$

$$r_{max} = \max(\frac{D_{bbox}}{2}, \{d_i\}_{\max}) \times 0.9, \tag{S2}$$

where D_{bbox} denotes the diagonal length of the object point cloud's 3D bounding box and d_i denotes the distance between the object center and the position of viewpoint i. With the sampled parameters, we calculate the camera pose of novel viewpoint, and leverage it to render the RGB images and surface normal maps.

To ensure efficient training with semantic consistency supervision, we employ two techniques to accelerate the rendering process: region of interest localization and rendering resolution reduction. Both techniques effectively reduce the number of pixels to be rendered, improving rendering efficiency and training speed. In particular, we firstly project corner points of the object point cloud's 3D bounding box onto the novel view to approximate its location in the 2D image. From these projected points, we extract an axis-aligned 2D bounding box that represents the region of interest. By focusing on this smaller region, we only need to render a cropped patch instead of the entire image, significantly reducing computation. Moreover, we also down-sample the rendering resolution to further reduce the rendering burden and improve the encoding speed of CLIP. In Figure S5, we provide a visualization of the RGB image at the reference pose and the renderings obtained from the randomly sampled novel view using the aforementioned techniques. This demonstrates how these techniques contribute to efficient rendering and training in our approach.

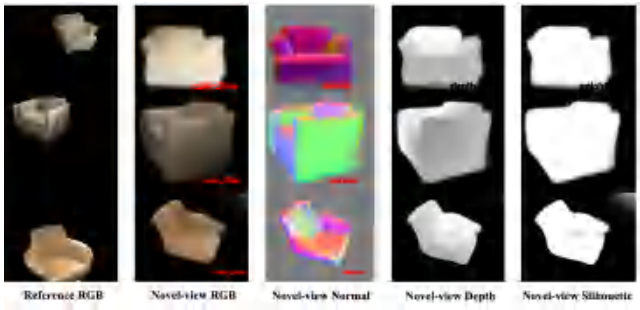


Figure S5: The reference image and renderings from novel views.

S6 Implementation Details

Our proposed model is experimented on an NVIDIA GeForce RTX 3090 GPU. For the color branch, we adopt a 2-layer MLP with 64 hidden units, which is much smaller than the scene-level configuration in NeuRIS (Wang et al. 2022). As for the cascaded SDF branch, we set the frequency of positional encodings for the coarse and fine blocks as $0.5\times$ and $1.0\times$ of that in NeuS (Wang et al. 2021), respectively. We use a four-layer MLP for the coarse block and another two-layer MLP for deeper layers of the fine block, with the same 64 hidden units as the color branch.

We implement model in Pytorch and train it using Adam optimizer. The loss weights are set to $\lambda_{\mathcal{C}} = \lambda_{\mathcal{N}} = \lambda_{\mathcal{D}} = \lambda_{se} = 1.0$, $\lambda_e = 0.1$, and $\lambda_{si} = 5.0$. The total training process takes 50k iterations. Specifically, we train the coarse block alone for 20k iterations and then together with the fine block for another 30k iterations. The semantic consistency loss is turned on after 10k iterations and applied every 5 iterations. And the initial learning rate is set to 2e-4 and decreases by 0.5 every 20k iterations. We randomly select 512 rays for each iterations and sample 64 points on each ray. The overall training process for one object occupies around 3GB GPU memory and takes about 4 hours.

S7 Details of the Baseline Methods

We compare O^2-Recon with four baselines, each representing a type of methods that can be utilized in object-level 3D reconstruction. The baseline methods are described as follows.

- The scene-level reconstruction method MonoSDF (Yu et al. 2022). MonoSDF is based on implicit surface in the form of MLP layers and utilizes geometry priors to improve the reconstruction accuracy. Instead of the depth information estimated from monocular images, we leverage the groundtruth depth to optimize the MLP for fair comparison and denote it as MonoSDF*.
- The most representative shape code based method FroDO (Rünz et al. 2020). Since most of the shape code based methods, like (Li, Rezatofighi, and Reid 2021; Irshad et al. 2022; Shan et al. 2021), are not open-source, we re-implement the representative FroDO method and denote it as FroDO*. We integrate the FroDO models of two categories, tables and chairs, to output the final results.
- The Scan2CAD method (Avetisyan et al. 2019) based on CAD databased retrieval. We follow the instructions in the officially released codebase and evaluate this method.
- The general object-level reconstruction method vMap (Kong et al. 2023). Since vMap is proposed for real-time reconstruction and is trained with limited iterations, we increase its training iteration to the same as O^2-Recon for fair comparison and denote it as vMap*.

S8 Comparison with the 3D Completion Method

We utilize a recent 3D completion method, SDFusion (Cheng et al., CVPR 2023), to complete the 3D surface reconstructed by MonoSDF conditioned on image and text prompts. The 3D completion mask is set to $|SDF(x)| > 1/32$ for the 64^3 SDF volume. The results presented in Table S2 and Figure S6 clearly show that this method cannot produce reasonable geometry for the occluded parts and

even compromises MonoSDF's performance. The core issue stems from an overreliance on highly idealized training datasets such as ShapeNet, which fails to capture the complexity of real-world scenarios.

Method	ScanNet GT		Scan2CAD GT	
	F-score ↑	C.D. ↓	F-score ↑	C.D. ↓
MonoSDF	0.627	9.82	0.217	12.22
MonoSDF + SDFusion	0.509	12.02	0.169	12.05
Ours w/o $\mathcal{T}$	0.715	4.50	0.562	6.32
Ours	**0.715**	**4.45**	**0.568**	**6.15**

Table S2: Additional results of geometry evaluation.

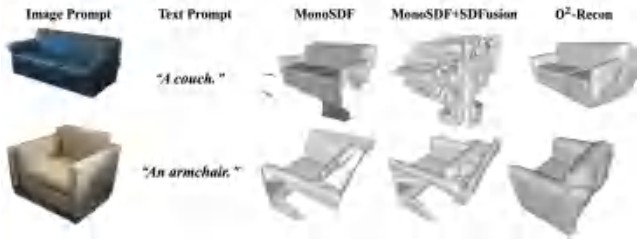


Figure S6: 3D completion results of SDFusion conditioned on an in-painted image and the categorical text prompt.

S9 More Visualizations of the Reconstruction Results

Due to the space limitation, we only show some selected views for each object in the main paper. In Figure S7, S8, and S9, we compare our method with baseline methods from more perspectives and show the visualization results. We also provide the visualizations of object-level manipulation results from more perspectives in Figure S10. We can observe that O^2-Recon produces much better surfaces in the invisible areas of occluded objects.

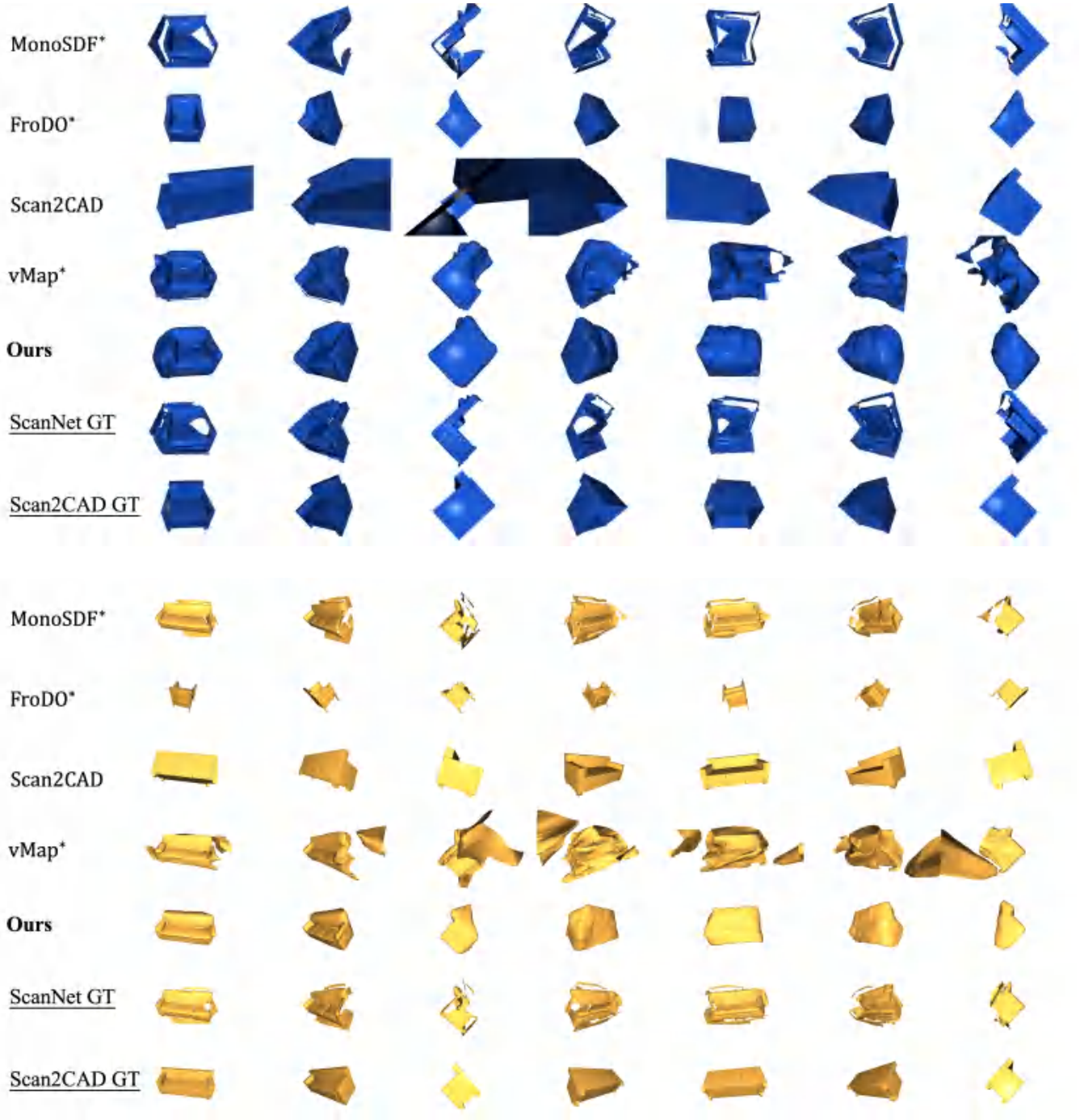


Figure S7: More visualizations of the qualitative comparisons (1-2).

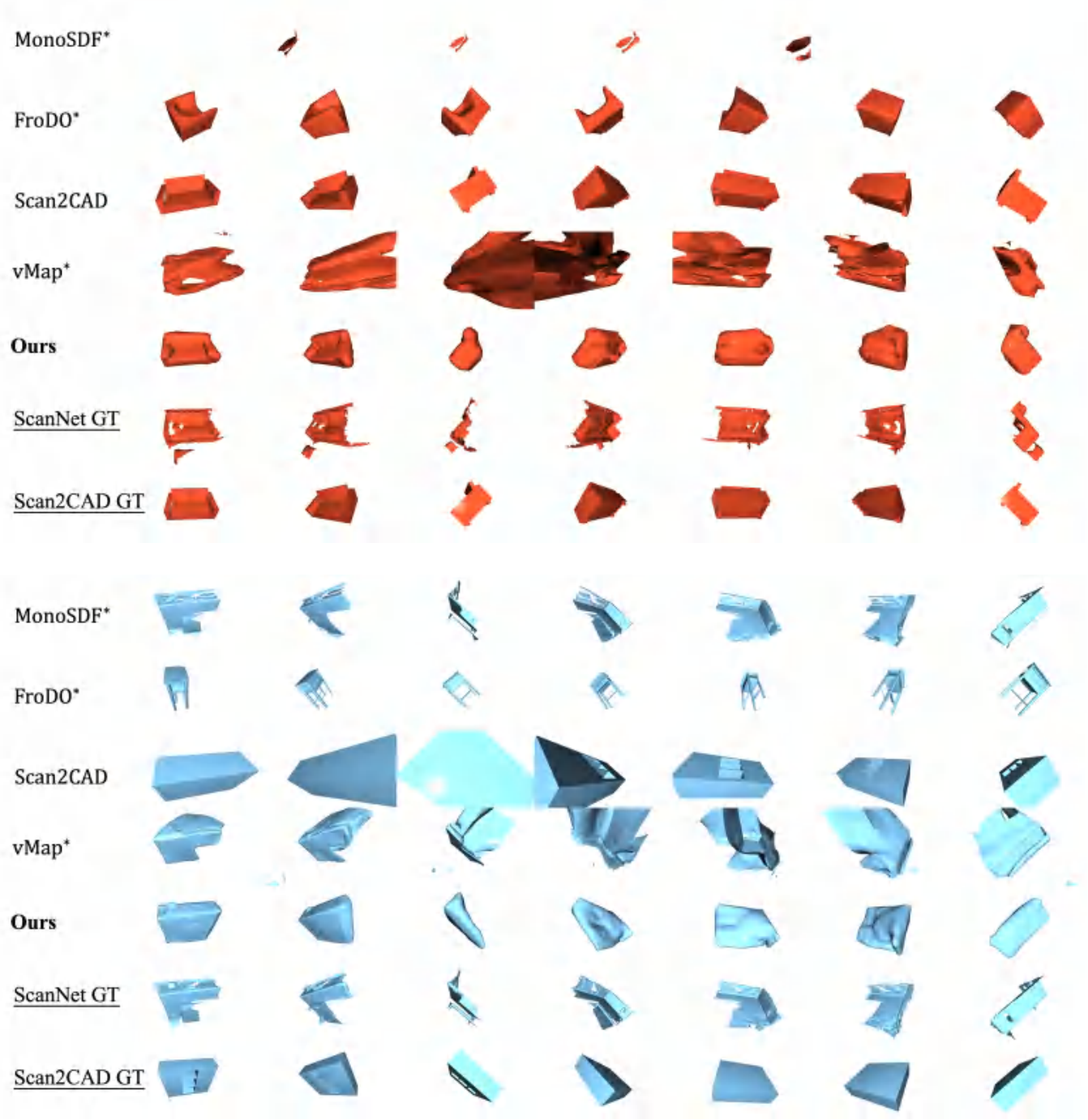


Figure S8: More visualizations of the qualitative comparisons (3-4).

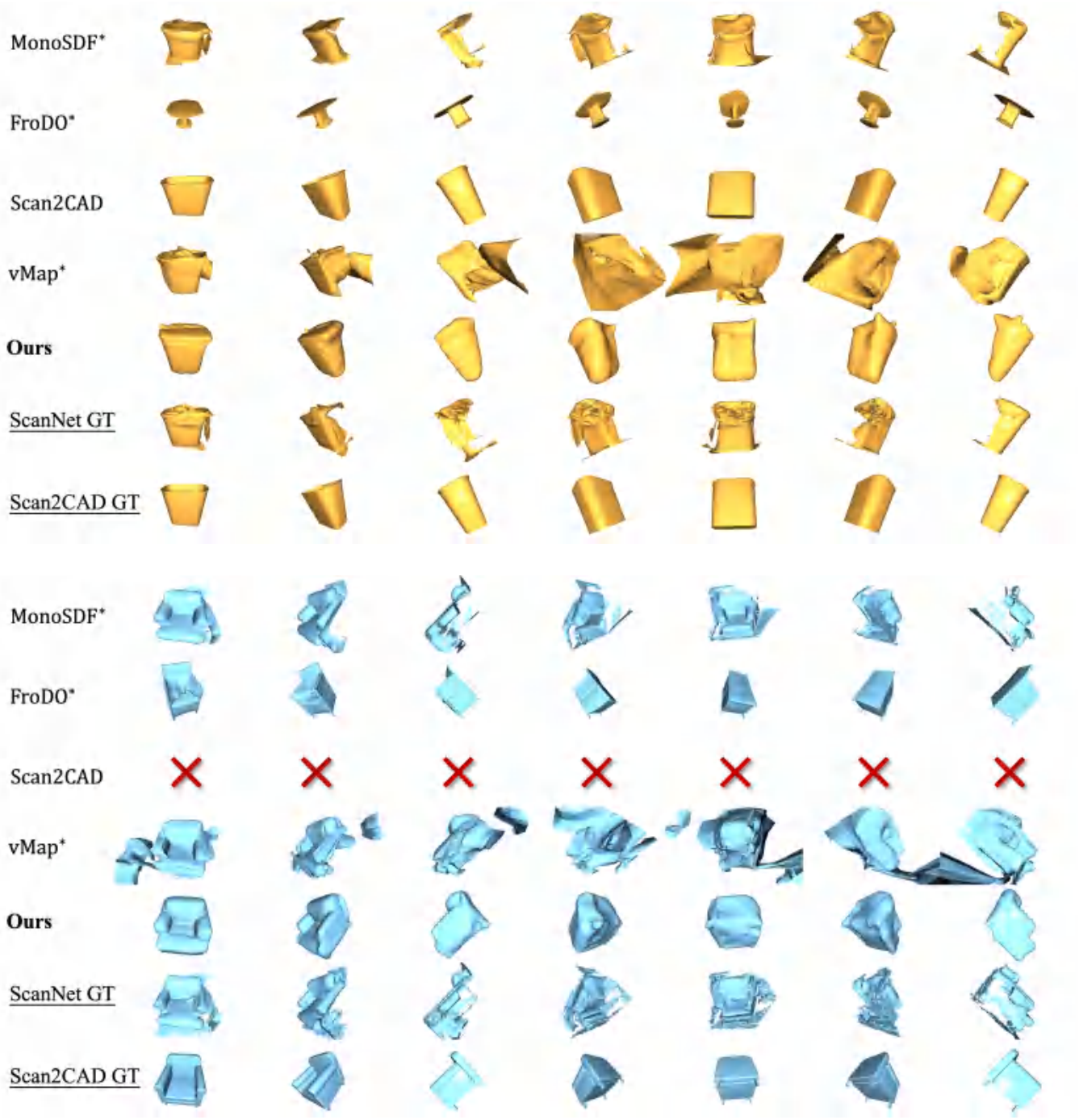


Figure S9: More visualizations of the qualitative comparisons (5-6).

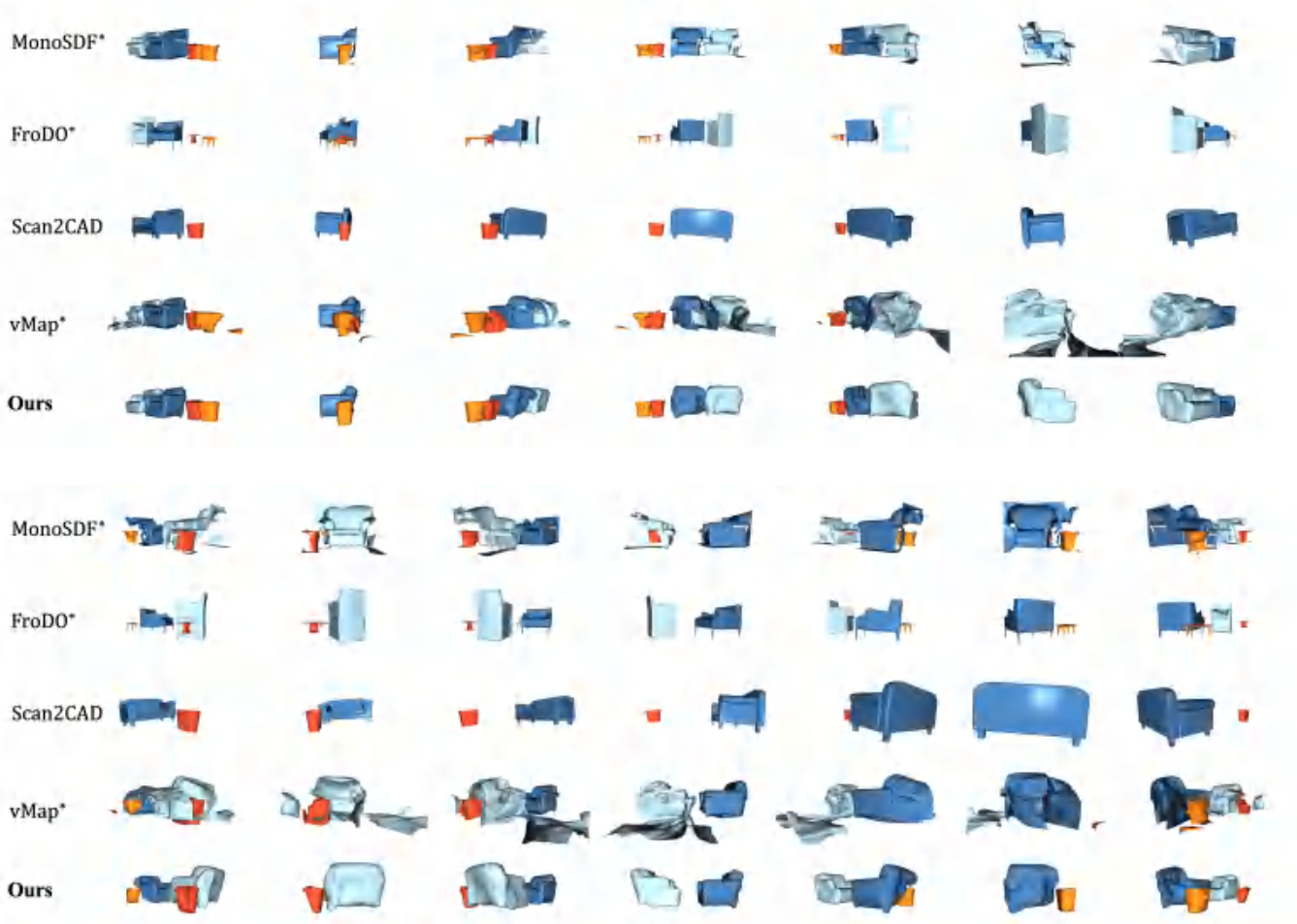


Figure S10: More visualizations of the object-level manipulations.