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

Anonymous submission

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.

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; Xie et al. 2023), 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 introducing 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:

- 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.
- 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.
- 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.

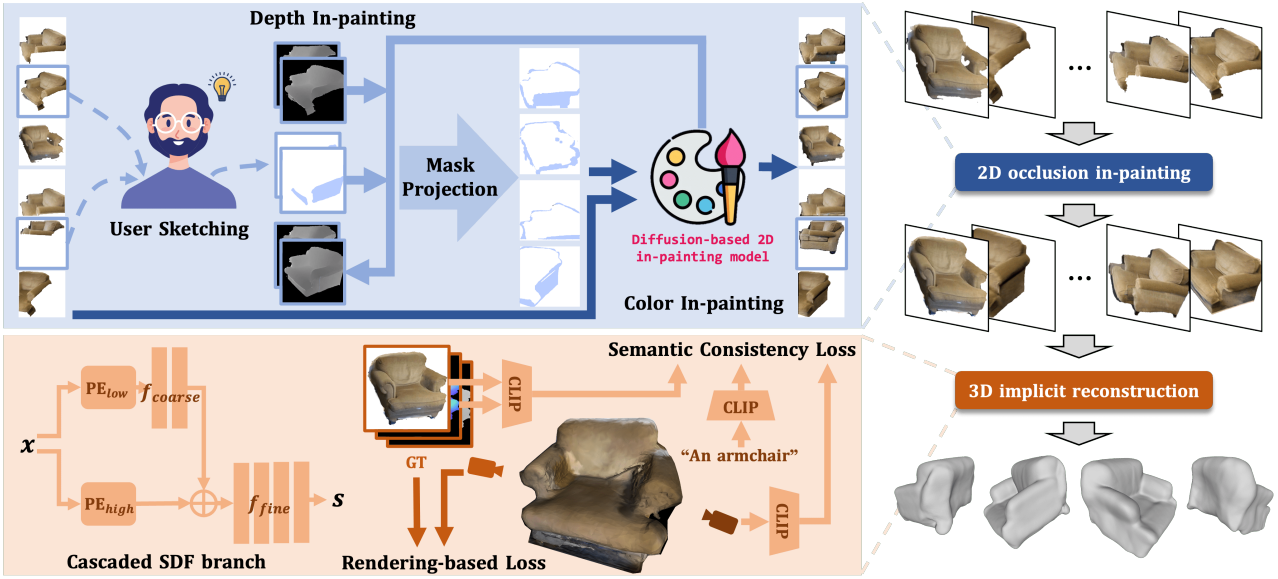


Figure 2: The proposed O^2-Recon framework. We utilize the Stable Diffusion in-painting model in our implementation.

NeRFs Empowered by 2D Diffusion Models. Motivated by the success of diffusion models in 2D image generation and editing (Saharia et al. 2022; Croitoru et al. 2023; Ruiz et al. 2023; Kawar et al. 2023), some methods have incorporated the pre-trained 2D diffusion models into NeRF representations. NerDi (Deng et al. 2022) provides arbitrary-view supervisions using a pre-trained diffusion model conditioned on semantic features, enabling single-view NeRF synthesis. To achieve NeRF-level editing, Haque et al. (Haque et al. 2023) iteratively updates the multi-view image dataset with edited images produced by the pre-trained InstructPix2Pix model (Brooks, Holynski, and Efros 2023). And Kamata et al. (Kamata et al. 2023) supervises the rendered images with the pre-trained InstructPix2Pix model via an SDS strategy proposed in DreamFusion (Poole et al. 2022). Both methods achieve style transfer of NeRF scenes based on the text instructions. Under the supervision of pre-trained Stable Diffusion model, the recently proposed RePaint-NeRF (Zhou et al. 2023) achieves local editing of the selected areas in the NeRF scenes. Moreover, a series of works generates 3D NeRFs from text prompts with the help of 2D diffusion models. Researchers propose to train a NeRF representation using different strategies, including score distillation sampling (Poole et al. 2022; Lin et al. 2023), score Jacobian chaining (Wang et al. 2023a), and variational score distillation (Wang et al. 2023c), etc.

In this paper, we utilize a diffusion-based 2D in-painting model to aid the 3D reconstruction of occluded objects. The pre-trained diffusion model is used to in-paint the occluded regions in the 2D images. And the in-painted images are then utilized 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[1] (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:

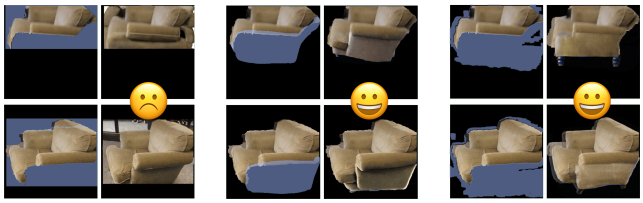
$$M_n^i = \mathbb{1}_{A_i}(x, y), \quad A_i = \{(x, y) | S_n^I(x, y) = i\}. \tag{1}$$

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

[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.

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 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 $\{\tilde{M}n^i\}^N_{n=1}$, 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. S 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. More details about the novel view generation are presented in supplementary material.

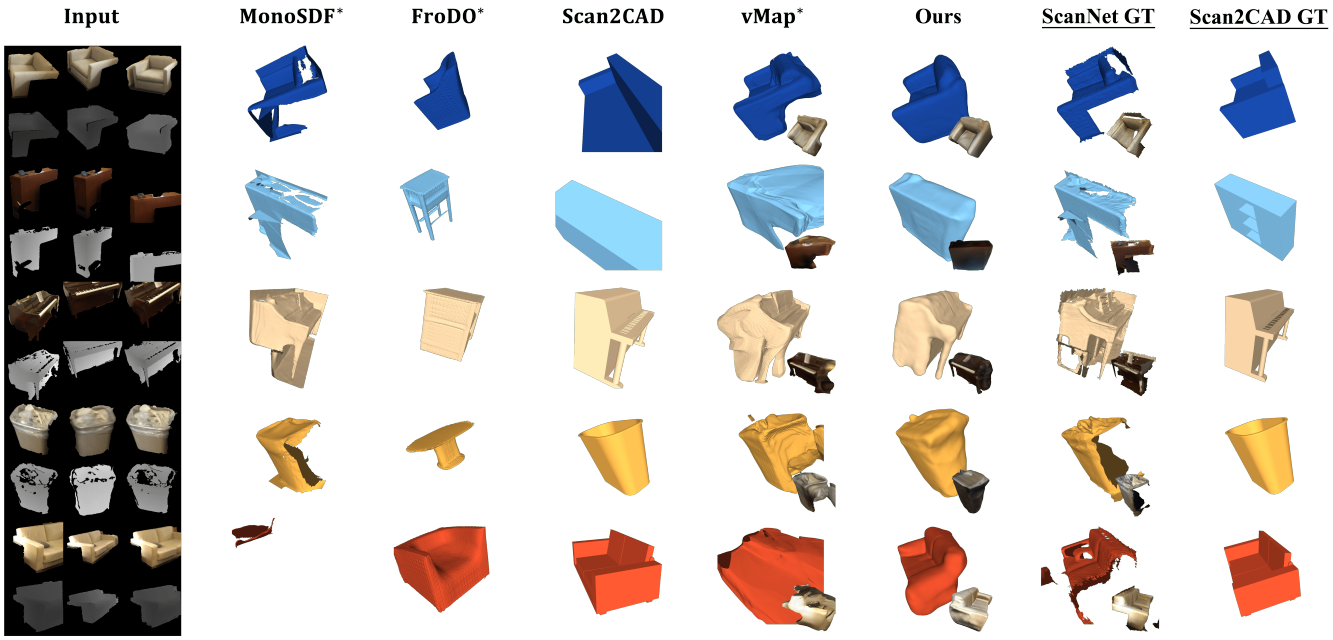


Figure 4: Visualization of occluded objects reconstructed by different methods. The occlusion conditions can be visualized in the column of Input and ScanNet GT, the missing parts indicate occlusion in the corresponding regions.

4 Experiments

4.1 Datasets, Baselines and Evaluation Metrics

Datasets. We select 6 ScanNet (Dai et al. 2017) scenes for evaluation, which contain 77 objects, including many occluded objects that are difficult to completely reconstruct. We follow the practice in vMap (Kong et al. 2023) to get the instance and semantic 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. (a) The scene-level reconstruction method MonoSDF (Yu et al. 2022). We leverage the ground truth depth in its optimization and denote it as MonoSDF*. (b) The re-implemented shape-code-based method FroDO (Rünz et al. 2020), denoted by FroDO*. (c) The Scan2CAD method (Avetisyan et al. 2019) based on CAD model retrieval. (d) 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*.

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 (measured in cm). We also report the accuracy and completion terms of Chamfer distance for detailed analysis.

4.2 Comparisons

Qualitative Evaluation. As shown in Figure 4, we compare the reconstruction results of different methods on objects that suffer from various occlusion conditions. 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.

As a comparison, our proposed system O^2-Recon simultaneously reconstructs accurate surfaces for visible regions and plausible surfaces for invisible regions. We refer readers to the supplementary material for qualitative comparisons from more camera views.

Quantitative Evaluation. We leverage 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 ground truth depth maps, which can be utilized to measure the reconstruction accuracy for 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

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

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**

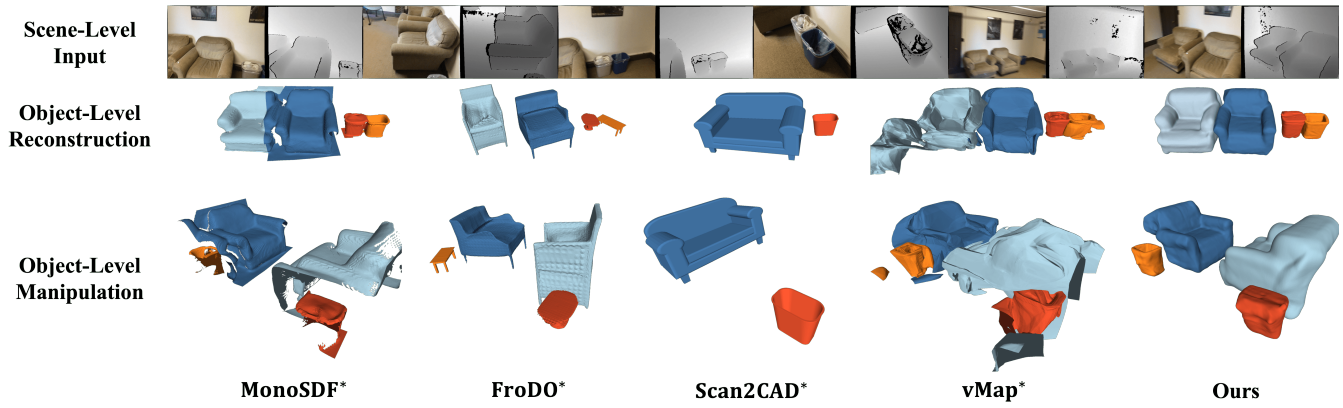


Figure 5: Comparison of object-level manipulation results.

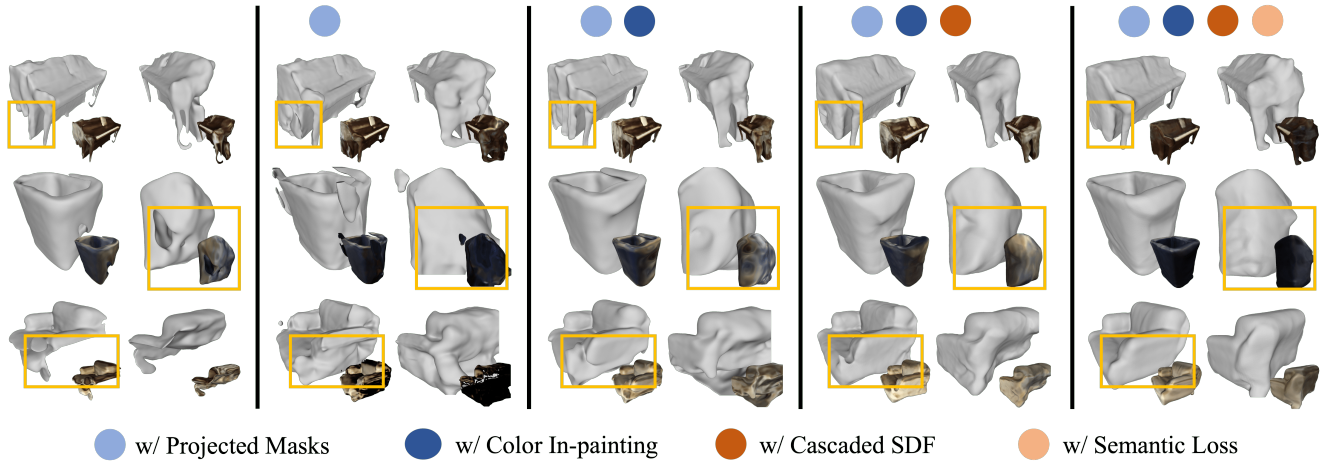


Figure 6: The ablation studies. Note how our carefully designed components progressively complete the occluded regions.

Table 2: Quantitative results for ablation study. C.D. is short for Chamfer distance.

	#1	#2	#3	#4	#5
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**

roughly match the actual objects.

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

Color In-painting. 2D color in-painting is one of the most important steps in our method. Although we can already reconstruct some occluded parts of the object according to the projected masks without color in-painting, the projected masks are not accurate enough, which indicate unreasonable geometry and lead to noisy reconstructed surfaces containing black areas due to the lack of color information, as shown in the second column of Figure 6. With the color in-painting and the following mask refinement steps, we generate cleaner and more reasonable shapes for the occluded area, as visualized in the third column of Figure 6. From the quantitative results reported in Table 2, the color in-painting step brings a large improvement in terms of F-score and Chamfer distance.

Cascaded SDF architecture and Semantic Loss. To compensate for the totally unseen areas, we propose a cascaded SDF architecture and a semantic consistency loss. As shown in the last two columns of Figure 6 and Table 2, both techniques improve the smoothness and accuracy of reconstructed surface for the totally unseen areas. In particular, the supervision of semantic consistency loss greatly improves the color consistency of the reconstructed mesh results.

5 Conclusion

In this paper, we propose O^2-Recon, a framework that reconstructs complete 3D surfaces of occluded objects in the scene empowered by the pre-trained 2D diffusion model. We utilize the diffusion model to in-paint the occluded parts in the multi-view 2D images, and then reconstruct the 3D mesh with neural implicit surface from the in-painted images. Our proposed human-in-the-loop mask generation strategy can effectively guide the 2D in-painting process with 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 of arbitrary categories. The reconstructed 3D objects can be utilized in further manipulation like large rotations and translations.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18392–18402.

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.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion Models in Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20.

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.

Deng, C.; Jiang, C. M.; Qi, C. R.; Yan, X.; Zhou, Y.; Guibas, L. J.; and Anguelov, D. 2022. NeRDi: Single-View NeRF Synthesis with Language-Guided Diffusion as General Image Priors. *CoRR*, abs/2212.03267.

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. *CoRR*, abs/2303.12789.

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 Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIII*, volume 12358 of *Lecture Notes in Computer Science*, 599–628. Springer.

Kamata, H.; Sakuma, Y.; Hayakawa, A.; Ishii, M.; and Narihira, T. 2023. Instruct 3D-to-3D: Text Instruction Guided 3D-to-3D conversion. *CoRR*, abs/2303.15780.

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. *CoRR*, abs/2303.08605.

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 Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, 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. 2022. DreamFusion: Text-to-3D using 2D Diffusion. *CoRR*, abs/2209.14988.

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.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22500–22510.

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.; Popov, S.; and Ferrari, V. 2022. RayTran: 3D Pose Estimation and Shape Reconstruction of Multiple Objects from Videos with Ray-Traced Transformers. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part X*, volume 13670 of *Lecture Notes in Computer Science*, 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 Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXII*, volume 13692 of *Lecture Notes in Computer Science*, 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.

Xie, S.; Zhang, Z.; Lin, Z.; Hinz, T.; and Zhang, K. 2023. SmartBrush: Text and Shape Guided Object Inpainting With Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22428–22437.

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. *CoRR*, abs/2306.05668.