

Gaussian Object Carver: Object-Compositional Gaussian Splatting with Surfaces Completion

Liu Liu*, Xinjie Wang*, Jiaxiong Qiu, Tianwei Lin, Xiaolin Zhou, Zhizhong Su
 Horizon Robotics, Beijing, China
 * Equal contribution

Abstract

3D scene reconstruction is a foundational problem in computer vision. Despite recent advancements in Neural Implicit Representations (NIR), existing methods often lack editability and compositional flexibility, limiting their use in scenarios requiring high interactivity and object-level manipulation. In this paper, we introduce the Gaussian Object Carver (GOC), a novel, efficient, and scalable framework for object-compositional 3D scene reconstruction. GOC leverages 3D Gaussian Splatting (GS), enriched with monocular geometry priors and multi-view geometry regularization, to achieve high-quality and flexible reconstruction. Furthermore, we propose a zero-shot Object Surface Completion (OSC) model, which uses 3D priors from 3d object data to reconstruct unobserved surfaces, ensuring object completeness even in occluded areas. Experimental results demonstrate that GOC improves reconstruction efficiency and geometric fidelity. It holds promise for advancing the practical application of digital twins in embodied AI, AR/VR, and interactive simulation environments. The code will be available at <https://github.com/liuliu3dv/GOC>.

1. Introduction

In embodied AI, collecting data from real-world environments is prohibitively expensive, making simulators a more efficient alternative. However, traditional simulators that rely on graphical assets face two main challenges: limited diversity and a domain gap between synthetic assets and real-world scenes. Recent advancements in Neural Implicit Representations (NIR) enable scalable digital twins of the real world from captured data. This capability has driven numerous downstream applications in embodied AI. For instance, in autonomous driving, neural simulators like UniSim and NeuRAD [35, 51] enable safe, modifiable environments for effective closed-loop testing. In the field of robotics, building digital twins using NIR, often referred to

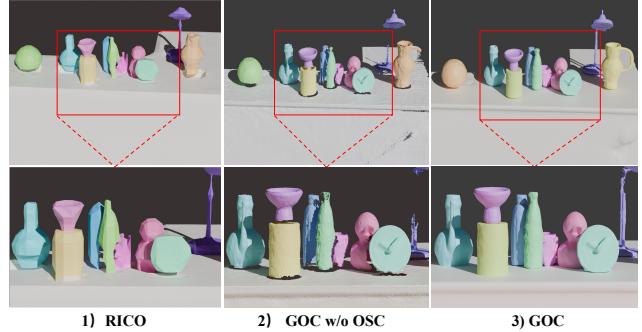


Figure 1. Invisible Surface Completion: We introduce a novel, efficient, and scalable framework for object-compositional 3D scene reconstruction, specifically designed to complete object surfaces in occluded regions. Compared to RICO [20], GOC without OSC has better detail but suffers from surface holes. With the incorporation of OSC, our method generates watertight, separable object meshes, even in the presence of occlusions.

as Real2Sim, shows promising potential for data collection and closed-loop training [36]. However, robotics scenarios bring additional challenges, including higher-frequency physical interactions, complex occlusions, and open-set object categories.

Our goal is to design an efficient, scalable, and high-quality object-compositional scene reconstruction framework that enhances editability and interactivity, thereby expanding the applicability of NIR in embodied AI. Recent methods [20, 28, 44, 45] achieve object-compositional reconstruction by jointly optimizing scene geometry and segmentation. However, the training of SDF-based approaches is computationally intensive, which limits their scalability. Furthermore, in indoor scenes, complex occlusions and constrained viewpoints are common, which severely affect the quality and usability of object reconstruction.

The motivation behind this work is to combine scene observations with data-driven priors to create digital twins from real-world log data. By leveraging observations, we can capture as much information as possible from the scene, while data-driven priors enable robust reconstruction even in cases with insufficient observations. In this paper,

we introduce Gaussian Object Carver (GOC), an efficient object-compositional reconstruction framework, which is presented in 7. We are the first to apply 3D Gaussian Splatting (3D GS) to object-compositional reconstruction, significantly improving efficiency with fast differentiable rasterization. To achieve object-separable and accurate surface reconstructions, we integrate monocular semantic and geometric priors with multi-view geometric regularization. Additionally, to address unobserved surface reconstruction, we propose a novel generative Object Surface Completion (OSC) module that completes missing regions using 3D object priors. As shown in 1, GOC offers better detail compared to existing methods, and can generate watertight object meshes, even in the presence of occlusions.

The contributions are summarized as follows:

1. We propose Gaussian Object Carver, a novel and efficient framework that combines 3D Gaussian Splatting with a generalizable object completion model. Compared to existing methods, our approach achieves more than 10 times efficiency, and generates watertight, separable object meshes, even in scenarios involving occlusion.
2. We develop a 3D GS-based object-compositional reconstruction method incorporating monocular geometry priors and multi-view geometry regularization to improve geometric accuracy and scalability.
3. We introduce a zero-shot 3D Object Surface Completion (OSC) model, trained on a large-scale dataset, demonstrating generalizability for unseen surface completion at the object level.

2. Related Work

2.1. 3D representations and Surface Reconstruction

Neural Radiance Fields (NeRF) [25] utilize volume rendering to create photorealistic scene representations through stable optimization. However, NeRF alone struggles with precise geometric reconstruction, leading to the development of methods that integrate geometry-based representations, such as iso-surfaces (e.g., occupancy fields [24] and Signed Distance Functions (SDF) [39, 55]), as well as volume density, to improve surface reconstruction fidelity. To further enhance the quality and robustness of surface reconstructions, recent approaches like MonoSDF and NeuRIS [38, 55] incorporate geometric regularization from monocular models, adding constraints that help to capture fine details. Additionally, GeoNeuS[9] introduces geometric consistency from the multi-view stereo, addressing issues of scale ambiguity and improving cross-view alignment for higher-fidelity reconstructions. These advancements have collectively enhanced the reliability of NIR for high-quality surface reconstruction.

Despite progress with NeRF and SDF approaches, optimization remains time-intensive. Recently, 3D Gaussian

Splatting (3D GS) [52] has redefined efficiency in 3D reconstruction, offering high-quality, fast rendering through differentiable rasterization of 3D Gaussians. Achieving geometric accuracy and meshable surfaces from Gaussian primitives is increasingly critical for 3D scene understanding. Recent works [4, 11, 14] aim to refine geometric precision and mesh generation from these representations. Likewise, DN-Splatter [37] demonstrates that depth and normal priors can significantly enhance the training of 3D Gaussian splatters, leading to higher fidelity in reconstruction and meshing. A crucial aspect of compositional scene reconstruction is high-quality 3D segmentation. Recent approaches [32, 47, 53, 61] combine 2D scene understanding with 3D Gaussians, enabling real-time, editable 3D scene representations that address the computational inefficiencies of NeRF-based methods. By using consistent 2D masks across views, Gaussian Grouping achieves enhanced segmentation quality and computational efficiency compared to NeRF-based techniques.

In this work, to improve training efficiency, we introduce 3D Gaussian Splatting into object-compositional scene reconstruction. These enable our framework to achieve state-of-the-art quality and efficiency in object-compositional reconstruction.

2.2. Compositional Scene Reconstruction

Recent methods [44, 45] achieve object-compositional reconstruction by disentangling objects within a scene. Building upon ObjectSDF++, additional work [46] addresses the dependency on annotations, while [28] introduces physically differentiable constraints, collectively enhancing the practicality of SDF-based object-compositional reconstruction. However, in indoor scenes, complex occlusions and restricted viewpoints are common, severely affecting the quality and usability of object reconstruction. Methods like [12, 20] leverage scene geometry priors, such as background smoothness, object compactness, and object-background relationships, through the designed SDF regularization. Relying on manually designed regularization terms has limitations, as it cannot address challenging scenes with complex occlusions and restricted viewpoints.

Our approach combines scene observations with data-driven priors to address incomplete object observations. First, we leverage a 3D GS-based object-compositional reconstruction to extract observed geometry. Then, we complete the unobserved surfaces using 3D object priors, enabling more accurate and realistic scene reconstructions.

2.3. 3D Object Completion

For 3D shape completion, methods like PCN [56], TopNet [34], and GRNet [48] address the task by transforming partial point clouds into complete ones. However, to obtain a watertight mesh, these approaches require additional sur-

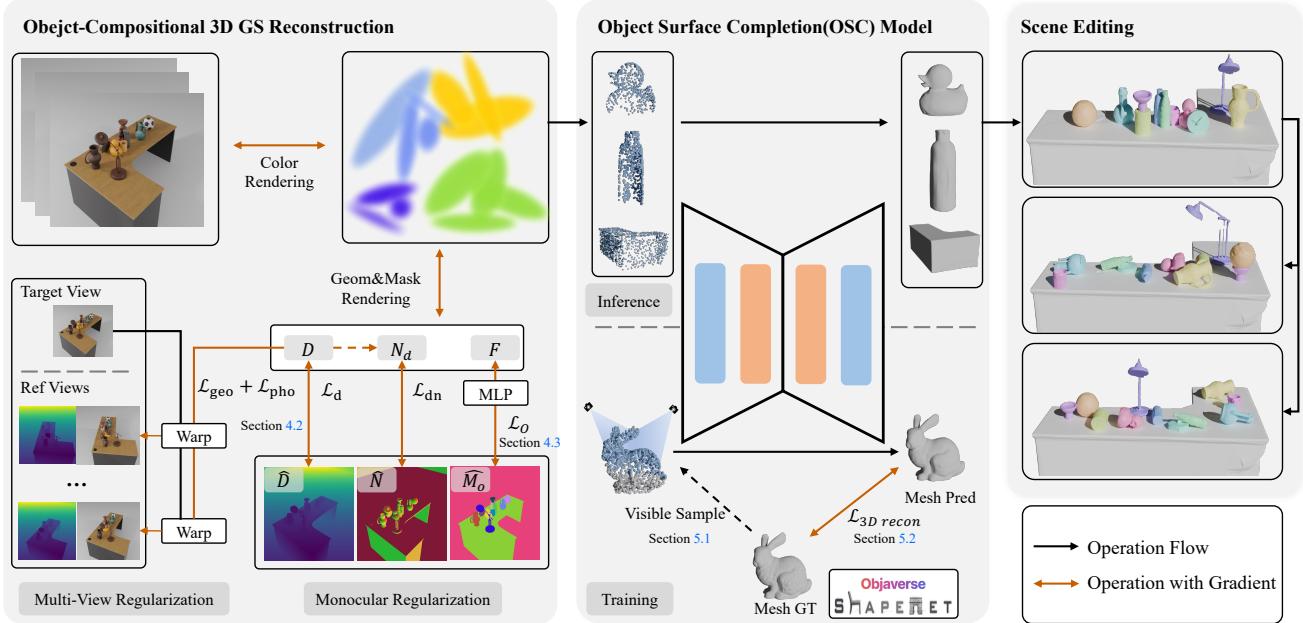


Figure 2. Overview of GOC: Given multi-view images of a scene, we optimize 3D Gaussian Splatting (3D GS) to generate scene geometry and segmentation, applying regularization from both multi-view geometry and monocular priors. Next, incomplete objects from partially observed inputs are fed into the Object Completion Model (OSC), which performs zero-shot completion to fill in missing geometry and produce complete 3D shapes. Finally, this process yields watertight and separable object meshes, enabling flexible scene rearrangement and object-level manipulation.

face reconstruction algorithms (e.g. traditional [16, 27] or neural kernel-based surface reconstruction [15, 41, 42]) to post-process the completed point clouds. Newer methods, such as PatchComplete [33] and DiffComplete [6], focus on directly completing missing signed distance fields (SDFs), resulting in a complete SDFs for a watertight shape. However, acquiring partial SDFs from captured or reconstructed point clouds is challenging, which limits their practicality in real-world applications where SDFs may not be readily available. Transformer-based models like ShapeFormer [50] enhance global feature learning by taking partial points as input and decoding them into a local deep implicit function, from which a mesh can be extracted via methods such as Marching Cubes [40]. Nevertheless, these models often rely on small, category-specific datasets, restricting their ability to generalize to unseen objects and complex scenes.

Methods like [21, 22, 43, 49] use a single-view image to generate the object mesh and have trained on a large amount of data. Additionally, works like OccNet [24], 3dshape2vecset [58], GEM3D [30] and CLAY [59] leverage additional modalities, such as point clouds or text prompt as conditions for diffusion models. However, these methods typically focus on individual object generation and struggle with object separation from the whole scene.

Our approach introduces a unified framework that combines scene reconstruction, instance segmentation, and object completion. This framework not only reconstructs and completes individual watertight meshes for each object but

also maintains the original geometric structure, achieving great generalization across diverse shape collections.

3. Overall Framework

As shown in Fig. 7, Gaussian Object Carver (GOC) is 3D GS-based object-compositional reconstruction framework. Given multi-view images of a scene as input, GOC efficiently generates separable object meshes, enabling flexible scene editing and object-level manipulation. To address the challenge of compositional reconstruction, the framework consists of two primary modules. The first is a 3D GS-based object-compositional reconstruction method, which is detailed in Sec. 4. In this section, we describe our approach and the design of regularization techniques for optimization. The second module is an general Object Completion Model (OSC), which is specifically designed to handle incomplete or occluded objects. Leveraging object priors, OSC generates complete geometric reconstructions of partially observed objects. Further details of the OSC are provided in Sec. 5, where we describe its architecture and functionality in greater depth.

4. 3D GS-based Compositional Reconstruction

This section is organized into four parts. Firstly, we review the preliminary concepts of 3D GS in Section 4.1. Then, in Section 4.2, we present the regularization of geometry. Next, we discuss the rendering and regularization of seg-

mentation in Section 4.3. Finally, in Section 4.4, we describe the optimization procedure. More implementation details are explained in supplementary materials.

4.1. Preliminary

Our work builds on 3D Gaussian Splatting [17], and the scene is explicitly represented by numerous differentiable 3D Gaussian primitives G . Each primitive is parameterized by a mean $\mu \in \mathbb{R}^3$, a covariance matrix $\Sigma \in \mathbb{R}^{3 \times 3}$, which is decomposed into a scaling vector $s \in \mathbb{R}^3$ and a rotation quaternion $q \in \mathbb{R}^4$, along with opacity $o \in \mathbb{R}$ and color $c \in \mathbb{R}^3$, the latter represented using spherical harmonics.

With patchwise parallelization, 3DGS achieves efficient alpha-blending for rendering and training. For each camera view, after 3D Gaussian primitives are projected as 2D space and sorted by z-buffer. Then the color C of a pixel could be computed by volumetric rendering[25] using front-to-back depth order. The composite pixel-wise color C and alpha A are given by:

$$\begin{aligned} C &= \sum_{i \in \mathcal{N}} c_i \alpha_i T_i, \\ A &= \sum_{i \in \mathcal{N}} \alpha_i T_i, \end{aligned} \quad (1)$$

Where $T_i = \prod_{j=1}^{i-1} (1 - \alpha_j)$, \mathcal{N} is the set of sorted Gaussians on the ray of rendered pixel, and T_i is the transmittance, defined as the product of opacity values of previous Gaussians overlapping the same pixel.

4.2. Geometry Regularization

For depth rendering, to eliminate the transparency impact on depth rendering, the rendered depth D needs to be alpha-normalized based on pixel alpha and is computed as:

$$D = \sum_{i \in \mathcal{N}} d_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) / A \quad (2)$$

where d is the distance between the 3D Gaussian center and the camera center.

Monocular Geometry Regularization Surface reconstruction in complex indoor scenes is inherently challenging due to the lack of texture. To address this, we draw inspiration from [37, 55] and propose incorporating monocular priors, specifically normal \hat{N} and depth \hat{D} , into the reconstruction process.

First, the rendered depth D can be directly constrained through depth prior \hat{D} by:

$$\mathcal{L}_d = \sum_{i,j} |D - \hat{D}| \quad (3)$$

By deriving the depth result, we can compute the normal from rendered depth N_d . We then leverage \hat{N} to regularize

this result. To mitigate the impact of transparency, we use α to weight the normal loss:

$$\mathcal{L}_{dn} = \sum_{i,j} \alpha (1 - N_d^T \hat{N}) \quad (4)$$

Multi-View Geometry Regularization Monocular depth estimation often suffers from ambiguities, leading to inconsistencies across views and degrading surface reconstruction quality. To address this, we integrate multi-view geometry regularization into optimization, inspired by prior work in multi-view geometry ([9, 10]).

We introduce a photometric reprojection loss inspired by self-supervised depth estimation [10]. Unlike methods requiring optical flow priors or multi-plane projections, our approach is computationally efficient and needs no preprocessing. With monocular geometry priors providing reasonably accurate depth estimates, this loss also avoids local optima:

$$\mathcal{L}_{pho} = \frac{1}{N} \sum_{i,j} \lambda \frac{1 - \text{SSIM}(C_{ij}, \tilde{C}_{ij})}{2} + (1-\lambda) |C_{ij} - \tilde{C}_{ij}|, \quad (5)$$

Where $\lambda = 0.85$, \tilde{C}_{ij} are the colors from the reference frame projected using rendered depth D , and C_{ij} are the colors in the target frame.

Photometric reprojection constraints may struggle in low-texture or overexposed regions. To enhance 3D consistency, we adopt the geometry reprojection consistency loss \mathcal{L}_{geom} from [4]. This term enforces depth alignment across viewpoints by computing the circular projection error between depth maps from the reference and target frames.

4.3. Segmentation Regularization

For object segmentation, inspired by [53], we first give 3D Gaussian a group of learnable features f encoded semantic information and render the feature for each pixel through alpha-blending. Similar to color rendering, the rendered semantic features F can be produced by:

$$F = \sum_{i \in \mathcal{N}} f_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \quad (6)$$

where f_i is the semantic feature of each Gaussian primitive.

Then, we use a multilayer perceptron network (MLP) and softmax to get the classification from the rendered semantic feature F , thus obtaining an instance mask for each pixel M_o . We use a cross entropy loss \mathcal{L}_o between instance mask GT \hat{M}_o and M_o .

4.4. Optimization

We adopt the photometric loss \mathcal{L}_c from vanilla 3DGS [17]. All loss functions are simultaneously optimized by training

from scratch. The total loss function \mathcal{L} can be defined as:

$$\mathcal{L} = \mathcal{L}_c + \lambda_d \mathcal{L}_d + \lambda_{dn} \mathcal{L}_{dn} + \lambda_{geo} \mathcal{L}_{geo} + \lambda_{pho} \mathcal{L}_{pho} + \lambda_o \mathcal{L}_o \quad (7)$$

In our experiments, $\lambda_d = 0.3$, $\lambda_n = 0.1$, $\lambda_{pho} = 0.3$, $\lambda_{geo} = 0.3$, $\lambda_o = 0.1$. And we use the training strategy in [18], because we observed that the 3DGS strategy [17] is sensitive to initialization and hyperparameter settings.

5. Object Surfaces Completion Model

We propose a general Object Surface Completion (OSC) model, designed to recover complete, watertight meshes from sparse or partial point clouds. The objective of OSC is to reconstruct meshes that closely resemble the original geometric structure of the object, rather than the diversity in generated meshes. To achieve this, we adopt a lightweight VAE [19] framework. OSC demonstrates strong generalization capability, enabling surface completion and reconstruction of objects with arbitrary geometries without fine-tuning in different domains. It supports inputs from incomplete point clouds collected by real sensors, depth recovery, or 3DGS [17] and NeRF [25] based reconstructions.

This section is organized into three parts. First, Sections 5.1 and 5.2 introduce the model details and the loss design of the OSC model. Next, Section 5.3 provides detailed insights into the training process of the OSC model.

5.1. Model Details

The OSC model encodes the geometric structure of the input point cloud into an implicit latent space through the Surface Points Encoder. Using grid query points and the embedding of the encoded point cloud, the surface completion decoder then outputs the occupancy probability, the likelihood that each point lies within the object’s surface, for each query point. Finally, the Marching Cubes algorithm [40] is applied to extract the surface mesh.

Masking To enable the model to recover complete watertight surfaces from incomplete point clouds, we extend the MAE [13] masking strategy into 3D space. Specifically, during model training, we uniformly sample points from the mesh surface and project the points into the camera’s pixel coordinates using the intrinsic and extrinsic parameters of the virtual camera. Points that do not fall onto the imaging plane are masked out. Considering depth occlusions between points, non-visible points are also masked. The retained point set is denoted as \mathbf{P}_s . Gaussian noise is added to \mathbf{P}_s to augment data with slight random perturbations.

Surface points Encoder We apply Farthest Point Sampling (FPS) [31] to extract \mathbf{P}_s core structure into the dimensions $\mathbb{R}^{M \times 3}$, which is encoded into the latent space through

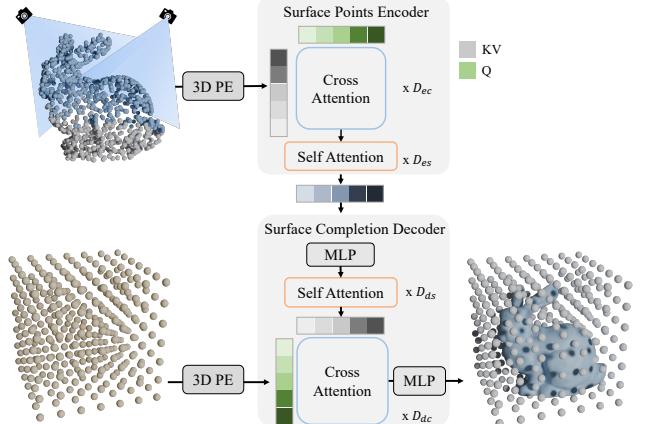


Figure 3. Illustration of OSC for surface reconstruction from sparse or incomplete point clouds. During training, points are uniformly sampled from the object’s surface mesh, filtered by specified virtual cameras, and encoded into embedding in latent space via an encoder. The decoder predicts the occupancy probability of each query point in a predefined 3D grid with the embedding. The reconstructed complete mesh is obtained by extracting the isosurface from the occupancy field.

Fourier Positional Encodings (FPE) [25] to serve as learnable queries. The point set \mathbf{P}_s is also encoded with FPE as Key-Value (KV). After passing through cross-attention layers with depth D_{ec} , followed by self-attention layers with depth D_{es} , we obtain the embedding representation $\mathbf{E} \in \mathbb{R}^{M \times D}$ of \mathbf{P}_s in the latent space, where $M = 2048$, $D_{ec} = 10$, $D_{es} = 10$, and $D = 16$.

$$\mathbf{E} = \text{SelfAttns}(\text{CrossAttns}(\text{FPE}(\text{FPS}(\mathbf{P}_s)), \text{FPE}(\mathbf{P}_s))) \quad (8)$$

By applying KL regularization, we constrain \mathbf{E} to a Gaussian distribution, ensuring that similar data have similar positions in the latent space, facilitating the generation of continuous data as in [19]. Specifically, two MLP layers are used to learn the Gaussian distribution’s mean μ_i and variance σ_i^2 from $\mathbf{E} \in \mathbb{R}^{M \times D}$. The regularization loss \mathcal{L}_{KL} is applied to enforce this constraint:

$$\mathcal{L}_{KL} = \frac{1}{2} \sum_{i=1}^D (\sigma_i^2 + \mu_i^2 - 1 - \log \sigma_i^2) \quad (9)$$

A sampled latent $\mathbf{E}_s \in \mathbb{R}^{M \times D}$ is then drawn from the distribution and used as input to the subsequent decoder.

Surface Completion Decoder Unlike MAE [13], which compresses and reconstructs images, our surface completion decoder takes as input a set of initialized query points, $\mathbf{Q}_g \in \mathbb{R}^{k \times 3}$, along with the latent space encoding \mathbf{E}_s of the surface point cloud. The decoder outputs an occupancy probability $\hat{\mathcal{O}}_s(\mathbf{Q}_g) \in \mathbb{R}^k$ for each query point similar to methods in [24, 58]:

$$\hat{\mathcal{O}}_s(\mathbf{Q}_g) = \text{MLP}(\text{CrossAttns}(\text{FPE}(\mathbf{Q}_g), \text{SelfAttns}(\mathbf{E}_s))) \quad (10)$$

5.2. Optimization

Directly supervising the decoder with the ground truth $\mathcal{O}(\mathbf{Q}_g) \in \{0, 1\}$ often results in noticeable artifacts, such as streaks and voxel-like patterns on the reconstructed object surface. Inspired by label smoothing [26], we set the values of $\mathcal{O}(\mathbf{Q}_g)$ near the object's surface by utilizing a signed distance field (SDF) and a threshold T_{iso} to map values within the range of 0 to 1. This approach enhances reconstruction precision and smoothness around the surface as shown in supplementary materials. The threshold T_{iso} is set to 1/128:

$$\text{smooth}(\mathcal{O}(\mathbf{Q}_g)) = 0.5 \cdot \mathbf{1} - 0.5 \times \frac{\text{SDF}(\mathbf{Q}_g)}{T_{\text{iso}}} \quad (11)$$

$$\mathcal{O}_s(\mathbf{Q}_g) = \begin{cases} 0, & \text{if } \text{SDF}(\mathbf{Q}_g) > T_{\text{iso}} \\ \text{smooth}(\mathcal{O}(\mathbf{Q}_g)), & \text{if } -T_{\text{iso}} \leq \text{SDF}(\mathbf{Q}_g) \leq T_{\text{iso}} \\ 1, & \text{otherwise} \end{cases} \quad (12)$$

We optimize the model by using a binary cross-entropy loss to minimize the distribution difference between the predicted occupancy probability and the ground truth, similar to [24, 58], θ represents the learnable parameters of the model f_θ :

$$\mathcal{L}_{\text{BCE}}(\theta) = \mathbb{E}[\text{BCE}(f_\theta(\mathbf{P}_s, \mathbf{Q}_g), \mathcal{O}_s(\mathbf{Q}_g))] \quad (13)$$

The model f_θ outputs the occupancy probability for each query point, and the final mesh is formed by applying a pre-defined isosurface threshold T_b using the Marching Cubes algorithm [40]. However, optimizing solely with \mathcal{L}_{BCE} does not guarantee a clear surface boundary under threshold T_b , which may lead to incomplete mesh generation as shown in supplementary materials. To address this, we introduce \mathcal{L}_{IoU} , which converts each query point's occupancy probability to an actual occupancy state based on the threshold T_b . The Intersection over Union (IoU) is then computed with the original ground truth $\mathcal{O}(\mathbf{Q}_g) \in \{0, 1\}$ to achieve clear boundaries. T_b is set to 0.3 during both training and inference:

$$\hat{\mathcal{O}}(\mathbf{Q}_g) = \begin{cases} 1, & \text{if } \hat{\mathcal{O}}_s(\mathbf{Q}_g) > T_b \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

$$\mathcal{L}_{\text{IoU}}(\theta) = \mathbb{E}\left[1 - \text{IoU}\left(\hat{\mathcal{O}}(\mathbf{Q}_g), \mathcal{O}(\mathbf{Q}_g)\right)\right] \quad (15)$$

The total loss can be written as below, where λ_{BCE} , λ_{IoU} , and λ_{KL} are set to 1.0, 0.01, and 0.0001 respectively:

$$\mathcal{L}(\theta) = \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{IoU}}\mathcal{L}_{\text{IoU}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} \quad (16)$$

5.3. Implementation Details

The OSC training dataset consists of meshes tagged as train-set from the ShapeNet Core v2 dataset [3] and a diverse selection of meshes from Objaverse [8]. Meshes from Objaverse were filtered to exclude those with very few faces or vertices and any that contained multiple disconnected objects. Additionally, low-quality meshes, as indicated by the annotations from [62], were excluded. Each mesh was converted to a watertight form using TSDF [27]. We then used Open3D [60] to sample surface points, query points, and calculate SDF values for each query point as ground truth for training. After discarding meshes for which TSDF or SDF calculations failed, the resulting dataset included a curated set of around 400,000 high-quality diverse meshes.

The training was conducted with the AdamW optimizer, using a learning rate of 1e-4 and a batch size of 8, utilizing 32 NVIDIA 4090 GPUs over 4 days.

6. Experiments

In this section, we first introduce our experimental setup. Then, in Section 6.2 and Section 6.3, we compare our framework with state-of-the-art methods to evaluate surface reconstruction on both synthetic and real datasets. Finally, we present ablation studies in Section 6.4.

6.1. Settings

Datasets In our experiments, we used two public datasets, ShapeNet [3] and ScanNet [7], as well as a custom synthetic dataset. ShapeNet, a large-scale and richly annotated shape repository represented by 3D CAD models, was employed to evaluate the surface reconstruction quality of the OSC model with complete point cloud inputs. However, our primary focus was to assess the geometric quality of each object after reconstructing the scene and completing the segmentation of all objects. Since ShapeNet consists of near-perfect individual CAD models, it is unsuitable for evaluating surface reconstruction and completion on incomplete point clouds. Similarly, ScanNet features incomplete ground truth meshes, with missing object surfaces in unobserved views, making it unsuitable for evaluating the quality of full-object completion. Consequently, a custom synthetic dataset was essential for our experiments. We created five synthetic indoor scenes, each containing approximately ten fully detailed 3D assets from BlenderKit [2]. We manually configured camera paths around each scene, rendering 170 RGB-D images along with instance masks using Blender [1] to create a **full observation** dataset. Additionally, a **sparse observation** dataset was generated by sampling 30% of the viewpoints, resulting in 50 images, to simulate a more challenging scenario where handheld data capture provides only partial views of object surfaces.

Table 1. Comparison of Methods on Object and Scene Reconstruction under **Full Observation**. GOC achieves the best geometric accuracy and completeness while being highly efficient, requiring only 5% of the time consumed by current state-of-the-art methods. The top-performing metrics are **highlighted**.

Method	Time ↓	Object Recon				Scene Recon			
		Accuracy ↓	Completion ↓	CD ↓	F-score ↑	Accuracy ↓	Completion ↓	CD ↓	F-score ↑
ObjectSDF++(MLP) [45]	21h 26min	0.0232	0.0511	0.0371	0.8741	0.0203	0.0252	0.0240	0.9164
RICO [20]	17h 59min	0.0203	0.0629	0.0416	0.8429	0.0248	0.0354	0.0330	0.8642
GOC w/o OSC	1h 7min	0.0045	0.0543	0.0294	0.9124	0.0136	0.0271	0.0211	0.9570
GOC w ShapeFormer [50]	1h 11min	0.0239	0.0689	0.0464	0.7875	-	-	-	-
GOC	1h 9min	0.0062	0.0501	0.0282	0.9228	-	-	-	-

Table 2. Comparison of Methods on Object and Scene Reconstruction under **Sparse Observation**. The best metrics are **highlighted**.

Method	Accuracy↓	Completion↓	CD↓	F-score↑
ObjectSDF++(MLP) [45]	0.0140	0.0654	0.0397	0.8749
RICO [20]	0.0177	0.0635	0.0406	0.8354
GOC w/o OSC	0.0038	0.0677	0.0357	0.8682
GOC w ShapeFormer [50]	0.0201	0.0802	0.0502	0.7835
GOC	0.0073	0.0575	0.0324	0.9033

Metrics For scene reconstruction performance, we report Chamfer Distance(CD), F-score, and normal consistency(NC) for evaluation on ScanNet. For synthetic scenes, we separate the metrics into two aspects: scene reconstruction and object completion. For object reconstruction evaluation on ShapeNet, we use Intersection over Union (IoU) as an additional metric.

Baselines For the object reconstruction, segmentation, and completion task, we selected ObjSDF++[45] and RICO[20] as baseline methods. We report the performance metrics separately for both the reconstruction phase and the object completion phase. For object completion, we also compare our method with ShapeFormer[50], which serves as the completion network for partial point cloud inputs. Additionally, we evaluated our OSC model on the ShapeNet test set for surface reconstruction quality with complete point cloud inputs, comparing its performance to state-of-the-art methods such as 3D2VS [58] and IF-Net [5].

6.2. Reconstruction in Synthetic Scenes

To align with the settings of ObjectSDF++ [45] and RICO [20], all images were downsampled to a resolution of 384x384. Both ObjectSDF++ and RICO were trained for up to 3000 epochs until convergence, while our GOC’s 3D GS reconstruction was trained for 30,000 steps. For testing, sampled segmented point clouds of each reconstructed object were used as inputs to the OSC model. Pre-trained on ShapeNet [3] and Objaverse [8], OSC required no additional fine-tuning to generalize to the domain of reconstructed point clouds, and was used only for inference.

As shown in Table 1, our method (GOC w/o OSC, where “w/o OSC” refers to reconstruction without completion) achieves state-of-the-art performance in scene reconstruc-

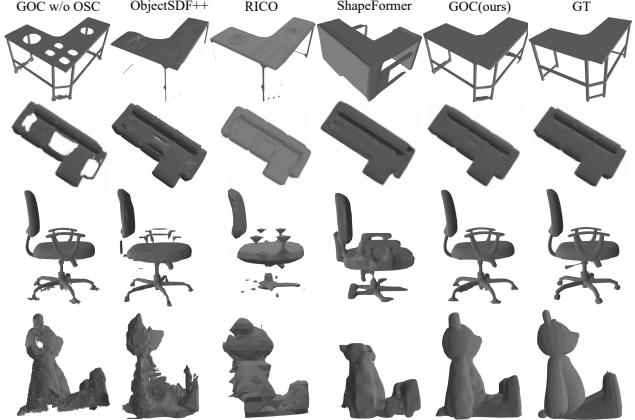


Figure 4. Qualitative comparison of surface reconstruction and completion quality across different methods on objects from the synthetic dataset. Zoom in for details.

tion. Compared to recent approaches like ObjectSDF++ and RICO, GOC achieves higher accuracy. However, the Completion metric reveals that our reconstruction completeness is slightly lower than ObjectSDF++. We compare the object reconstruction metrics after integrating the OSC model, which completes the segmented and sampled point clouds of reconstructed objects, we achieve a marked improvement in the completion metric, resulting in the best overall performance. When comparing our completed point cloud reconstruction with ShapeFormer, another surface completion method based on point cloud input, GOC demonstrates significant advantages in both CD and F-score, see the qualitative comparison across different methods in Figure 4. Notably, GOC is highly efficient, consuming only 5% of the time required by the current state-of-the-art methods. On average, it takes just 1 hour and 9 minutes to reconstruct an entire scene and complete all objects within it.

We assessed the geometric accuracy and completeness of object completion and reconstruction under the challenging Sparse Observation setting, as presented in Table 2. Although our accuracy decreased slightly after applying completion (GOC w/o OSC), from 0.0073 to 0.0038, the completeness of object reconstruction improved significantly, from 0.0677 to 0.0575. Compared to other methods, our approach achieves superior performance in both Chamfer Distance and F-score.

6.3. Reconstruction in Real-world Scenes

For indoor surface reconstruction, we conduct comparisons on widely used real-world datasets [7]. The results are reported in Tab. 3. Since the ground truth data also exhibits missing occluded regions in real-world datasets, we evaluate reconstruction performance without including the Object Surface Completion (OSC) model for comparison in this experiment.

Table 3. Comparison of Methods on Object and Scene Reconstruction on Scannet[7]. The top-performing metrics are **highlighted** and second-best performing metrics are highlighted.

Method	Image size	Object Recon			Scene Recon		
		CD ↓	F-score ↑	NC ↑	CD ↓	F-score ↑	NC ↑
MonoSDF [55]	384x384	-	-	-	0.0897	0.6030	0.844
RICO [20]	384x384	0.0929	0.7310	0.7944	0.0892	0.6144	0.8458
ObjectSDF++ [45]	384x384	<u>0.0921</u>	0.7482	0.8105	0.0886	0.6168	0.8520
PHYRECON [28]	384x384	0.0792	0.7554	0.8254	0.0834	0.6301	0.8657
GOC w/o OSC	384x384	0.1177	<u>0.7831</u>	0.7979	0.0530	<u>0.7331</u>	0.8324
	640x480	0.1444	0.7956	<u>0.8134</u>	0.0556	0.8243	<u>0.8591</u>

Our method outperforms competing approaches in scene reconstruction metrics, achieving the lowest CD and the highest F-score, demonstrating superior geometric accuracy and completeness. In object reconstruction, Our method achieves the highest F-score, indicating strong object-level reconstruction quality. Overall, our method demonstrates strong performance in both object-compositional and full scene reconstruction.

6.4. Ablation Study

Geometric Regularizations To quantitatively analyze the effectiveness of the proposed regularizations, we evaluate real scenes by comparing our full method to four variants in Tab. 4. The ablation study demonstrates the significance of each regularization in our framework. Depth regularization \mathcal{L}_d has the most substantial impact, with its removal causing severe performance drops, particularly in F-score, indicating its critical role in accurate geometry capture. Photo-consistency loss \mathcal{L}_{pho} aids multi-view alignment, reducing inconsistencies across views, while denoising loss \mathcal{L}_{dn} contributes to structural integrity and smoothness, as shown by its positive effect on F-score and NC. Geometric regularization \mathcal{L}_{geo} , though less impactful, helps fine-tune surface details. Together, these components enable robust, high-quality object-compositional reconstruction.

Table 4. Ablation Study of Scene Recon on Scannet [7]

Method	CD ↓	F-score ↑	NC ↑
Full	0.0556	0.8243	0.8591
w/o \mathcal{L}_{dn}	0.0534	0.8430	0.8201
w/o \mathcal{L}_d	0.1330	0.3483	0.7651
w/o \mathcal{L}_{pho}	0.0568	0.8087	0.8591
w/o \mathcal{L}_{geo}	0.0569	0.8246	0.8507

OSC Model Object Reconstruction Quality We evaluated the OSC model on the ShapeNet [3] test set for surface reconstruction quality under complete point cloud inputs. Compared to state-of-the-art methods such as 3D2VS [58] and IF-Net [5], OSC achieved the best reconstruction quality as shown in supplementary materials. This demonstrates that OSC is robust to different forms of point cloud inputs.

OSC Model Structure We experimented with different model architectures and evaluated the geometric accuracy of surface reconstruction on the ShapeNet test set, as shown in Table 5. Similar to MAE[13], we made the decoder lighter and concentrated more challenging learning tasks in the encoder, allowing for better adaptation to various potential downstream tasks. Ultimately, we selected the medium-sized model, OSC-M, with 101 million parameters, as it provides the optimal trade-off between performance and efficiency.

Table 5. Ablation study of the OSC model structure. Metrics were evaluated on the ShapeNet test set. “enc. Depth” refers to the number of self-attention and cross-attention layers in the encoder of the OSC model, while “dec.” refers to the decoder.

Model	enc. Depth	dec. Depth	Params.	IoU↑	CD↓	F-score↑
OSC-S	6	4	63 M	0.969	0.018	0.983
OSC-M	10	6	101 M	0.975	0.018	0.987
OSC-L	14	8	164 M	0.976	0.017	0.990

OSC Model Training Data Augmentation In the OSC model training, we experimented with three different data preprocessing approaches: using complete point clouds, applying a random 50% dropout to the point clouds, and simulating occlusions by filtering point clouds based on camera visibility. Results in Table 6 from the synthetic dataset under Sparse Observation showed that simulating camera occlusions produced the best completion and reconstruction quality. Combining the second and third masking strategies did not yield further performance improvements. As a result, we adopted the occlusion-based data augmentation strategy for training.

Table 6. The impact of different masking strategies on completion quality. Metrics were evaluated on the synthetic dataset under Sparse Observation.

Method	Accuracy↓	Completion↓	CD↓	F-score↑
w/o mask	0.0115	0.0748	0.0432	0.8380
Random drop	0.0152	0.0567	0.0360	0.8822
Visible mask	0.0073	0.0575	0.0324	0.9033

7. Conclusion

In this work, we present Gaussian Object Carver (GOC), a novel and efficient framework for object-compositional

scene reconstruction. Compared to existing methods, GOC achieves more than 10 times efficiency, and generates watertight, separable object meshes, even in scenarios involving occlusion. We introduce the zero-shot 3D Object Surface Completion (OSC) model, trained on a large-scale dataset, demonstrating generalizability for unseen surface completion at the object level.

References

- [1] Blender team. *Blender: 3D modelling and rendering package*. Available at <https://www.blender.org>. 6
- [2] BlenderKit Team. Blenderkit. Online; accessed 14 November 2024. Available at <https://www.blenderkit.com>.
- [3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 6, 7, 8, 1
- [4] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *arXiv preprint arXiv:2406.06521*, 2024. 2, 4
- [5] Julian Chibane, Thiendo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6970–6981, 2020. 7, 8, 1
- [6] Ruihang Chu, Enze Xie, Shentong Mo, Zhenguo Li, Matthias Nießner, Chi-Wing Fu, and Jiaya Jia. Diffcomplete: Diffusion-based generative 3d shape completion. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6, 8
- [8] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 6, 7
- [9] Qiancheng Fu, Qingshan Xu, Yew-Soo Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 4
- [10] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 4
- [11] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5354–5363, 2024. 2
- [12] Gemmechu Hassena, Jonathan Moon, Ryan Fujii, Andrew Yuen, Noah Snavely, Steve Marschner, and Bharath Hariharan. Objectcarver: Semi-automatic segmentation, reconstruction and separation of 3d objects. *arXiv preprint arXiv:2407.19108*, 2024. 2
- [13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 5, 8
- [14] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2
- [15] Jiahui Huang, Zan Gojcic, Matan Atzmon, Or Litany, Sanja Fidler, and Francis Williams. Neural kernel surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4369–4379, 2023. 3
- [16] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006. 3
- [17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 4, 5, 1
- [18] Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Jeff Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 3d gaussian splatting as markov chain monte carlo. *arXiv preprint arXiv:2404.09591*, 2024. 5, 1
- [19] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. 5
- [20] Zizhang Li, Xiaoyang Lyu, Yuanyuan Ding, Mengmeng Wang, Yiyi Liao, and Yong Liu. Rico: Regularizing the unobservable for indoor compositional reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17761–17771, 2023. 1, 2, 7, 8
- [21] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10072–10083, 2024. 3
- [22] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [23] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 1
- [24] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings*

- of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2, 3, 5, 6, 1
- [25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 4, 5
- [26] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019. 6
- [27] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. Ieee, 2011. 3, 6, 1
- [28] Junfeng Ni, Yixin Chen, Bohan Jing, Nan Jiang, Bin Wang, Bo Dai, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Phyrecon: Physically plausible neural scene reconstruction. *arXiv preprint arXiv:2404.16666*, 2024. 1, 2, 8
- [29] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020. 1
- [30] Dmitry Petrov, Pradyumn Goyal, Vikas Thamizharasan, Vladimir Kim, Matheus Gadelha, Melinos Averkiou, Siddhartha Chaudhuri, and Evangelos Kalogerakis. Gem3d: Generative medial abstractions for 3d shape synthesis. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3
- [31] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 5
- [32] Yansong Qu, Shaohui Dai, Xinyang Li, Jianghang Lin, Liu-juan Cao, Shengchuan Zhang, and Rongrong Ji. Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5328–5337, 2024. 2
- [33] Yuchen Rao, Yinyu Nie, and Angela Dai. Patchcomplete: Learning multi-resolution patch priors for 3d shape completion on unseen categories. *Advances in Neural Information Processing Systems*, 35:34436–34450, 2022. 3
- [34] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 3d point cloud generative adversarial network based on tree structured graph convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3859–3868, 2019. 2
- [35] Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson. Neurad: Neural rendering for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14895–14904, 2024. 1
- [36] Marcel Torne, Anthony Simeonov, Zechu Li, April Chan, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. *arXiv preprint arXiv:2403.03949*, 2024. 1
- [37] Matias Turkulainen, Xuqian Ren, Iaroslav Melekhov, Otto Seiskari, Esa Rahtu, and Juho Kannala. Dn-splat: Depth and normal priors for gaussian splatting and meshing. *arXiv preprint arXiv:2403.17822*, 2024. 2, 4
- [38] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. In *European Conference on Computer Vision*, pages 139–155. Springer, 2022. 2
- [39] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2
- [40] LORENSEN WE. Marching cubes: A high resolution 3d surface construction algorithm. *Computer graphics*, 21(1): 7–12, 1987. 3, 5, 6
- [41] Francis Williams, Matthew Trager, Joan Bruna, and Denis Zorin. Neural splines: Fitting 3d surfaces with infinitely-wide neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9949–9958, 2021. 3
- [42] Francis Williams, Zan Gojcic, Sameh Khamis, Denis Zorin, Joan Bruna, Sanja Fidler, and Or Litany. Neural fields as learnable kernels for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18500–18510, 2022. 3
- [43] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343*, 2024. 3
- [44] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In *European Conference on Computer Vision*, pages 197–213. Springer, 2022. 1, 2
- [45] Qianyi Wu, Kaisiyuan Wang, Kejie Li, Jianmin Zheng, and Jianfei Cai. Objectsdf++: Improved object-compositional neural implicit surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21764–21774, 2023. 1, 2, 7, 8
- [46] Tianhao Wu, Chuanxia Zheng, Qianyi Wu, and Tat-Jen Cham. Clusteringsdf: Self-organized neural implicit surfaces for 3d decomposition. In *European Conference on Computer Vision*, pages 255–272. Springer, 2025. 2
- [47] Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, et al. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. *arXiv preprint arXiv:2406.02058*, 2024. 2
- [48] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Jiageng Mao, Shengping Zhang, and Wenxiu Sun. Grnet: Gridding residual network for dense point cloud completion. In *European conference on computer vision*, pages 365–381. Springer, 2020. 2

- [49] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 3
- [50] Xinguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6239–6249, 2022. 3, 7
- [51] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023. 1
- [52] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. *arXiv preprint arXiv:2312.00732*, 2023. 2
- [53] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *ECCV*, 2024. 2, 4
- [54] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for Gaussian splatting. *arXiv preprint arXiv:2409.06765*, 2024. 1
- [55] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Satler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 2, 4, 8
- [56] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 international conference on 3D vision (3DV)*, pages 728–737. IEEE, 2018. 2
- [57] Biao Zhang, Matthias Nießner, and Peter Wonka. 3dilg: Irregular latent grids for 3d generative modeling. *Advances in Neural Information Processing Systems*, 35:21871–21885, 2022. 1
- [58] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023. 3, 5, 6, 7, 8, 1
- [59] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 3
- [60] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. 6
- [61] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024. 2
- [62] Qi Zuo, Xiaodong Gu, Yuan Dong, Zhengyi Zhao, Weihao Yuan, Lingteng Qiu, Liefeng Bo, and Zilong Dong. High-fidelity 3d textured shapes generation by sparse encoding and adversarial decoding. In *European Conference on Computer Vision*, 2024. 6

Gaussian Object Carver: Object-Compositional Gaussian Splatting with Surfaces Completion

Supplementary Material

1. OSC Reconstruction Quality on ShapeNet

We evaluated the OSC model on the ShapeNet [3] test set to assess surface reconstruction quality using complete point-cloud inputs. Compared to state-of-the-art methods such as 3D2VS [58] and IF-Net [5], OSC demonstrated superior performance across all metrics.

Table 7. Comparison of Single Object Surface Reconstruction Quality on the ShapeNet Test Set.

Model	IoU↑	CD↓	F-score↑
OccNet [24]	0.825	0.072	0.858
ConvOccNet [29]	0.888	0.052	0.933
IF-Net [5]	0.934	0.041	0.967
3DILG [57]	0.953	0.040	0.970
3D2VS [58]	0.965	0.038	0.967
OSC	0.975	0.018	0.987

Table 7 provides a detailed comparison of single-object surface reconstruction methods on the ShapeNet test set. The OSC model outperformed prior approaches, achieving the highest Intersection over Union (IoU) at 0.975, the lowest Chamfer Distance (CD) at 0.018, and the highest F-score at 0.987. These results highlight OSC’s robustness and effectiveness in reconstructing precise geometric surfaces.

The significant improvements in IoU, CD, and F-score metrics underline the model’s ability to capture fine-grained geometric details and achieve accurate surface reconstructions. This establishes OSC as a leading approach for robust surface reconstruction, particularly when using complete point-cloud data.

2. Additional Ablation Results for OSC Model

To evaluate the effectiveness of key components in the OSC model, we performed an ablation study with additional experiments focusing on \mathcal{L}_{IoU} loss and label smoothing. The results, presented in Figures 5 and 6, reveal the significant impact of these components on reconstruction quality. Excluding \mathcal{L}_{IoU} results in poorly defined mesh boundaries, highlighting its role in establishing accurate isosurface thresholds during inference. In contrast, including \mathcal{L}_{IoU} ensures sharp and precise boundary delineation. Similarly, the absence of label smoothing leads to voxel-like artifacts that degrade surface quality, whereas its inclusion enhances smoothness and detail, producing refined and artifact-free meshes. These additional results confirm the critical contributions of \mathcal{L}_{IoU} loss and label smoothing to the overall performance and robustness of the OSC model.

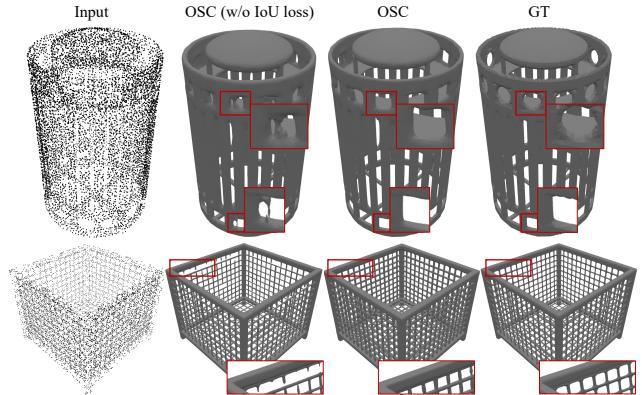


Figure 5. Qualitative comparison of reconstruction quality with (second column) and without \mathcal{L}_{IoU} (third column). Using \mathcal{L}_{IoU} aids OSC in establishing a well-defined isosurface threshold during the inference stage, resulting in clear and sharp mesh boundaries. Zoom in for details.

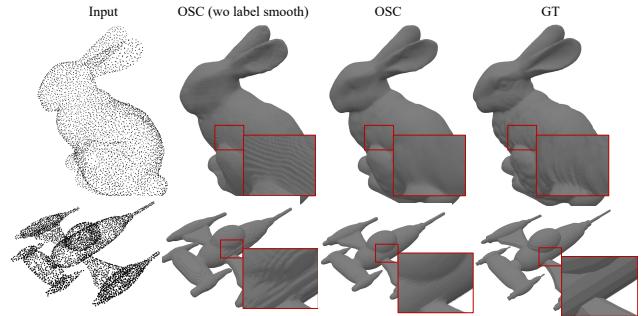


Figure 6. Qualitative comparison of reconstruction quality with (second column) and without label smoothing (third column). Label smoothing enhances the precision and smoothness of the reconstructed mesh surface, effectively reducing voxel-like artifacts on the surface. Zoom in for details.

3. Additional 3D GS Implementation Details

Implementation Details Our code is built based on gsplat [54] and training strategy are consistent with [18], because we observed that 3DGS [17] strategy is sensitive with initialization and hyperparameter settings. The training iterations for all scenes are set to 30,000. All experiments in this paper are conducted on Nvidia RTX 4090 GPU.

Mesh Exaction We start by rendering the depth for each training view and then apply the TSDF Fusion algorithm [27] to construct the corresponding TSDF field. From this field, the mesh is subsequently extracted using the Marching Cubes algorithm [23].

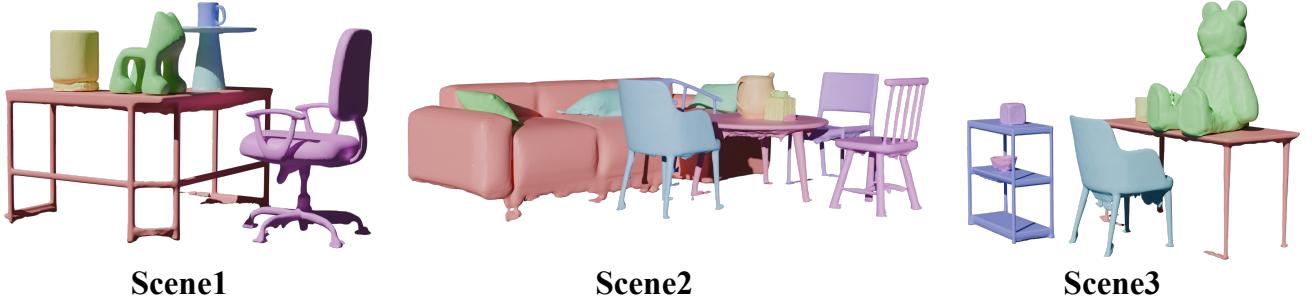


Figure 7. Semantic mesh results of GOC on Synthetic Scenes

Depth Regularization For datasets with sensor-provided depth at scene scale, L1 loss can be directly used for supervision. However, in the absence of such sensor data, a monocular depth model is employed to generate a prior, albeit without actual depth measurements. The inherent scale ambiguity in monocular depth estimates must be addressed to align them with true scene geometry. To achieve this, we employ least squares optimization to refine both the scaling parameter k and the offset parameter b for each image. This ensures that the monocular depth estimates are consistent with the rendered depth in terms of scale:

$$\hat{k}, \hat{b} = \arg \min_{k,b} \sum_{i,j} \left| \left(k \cdot \hat{D}_{i,j} + b \right) - D_{i,j} \right|_2^2, \quad (17)$$

where $\hat{D}_{i,j}$ and $D_{i,j}$ are the per-pixel depth values of the predicted and rendered depth maps, respectively. Once aligned, we apply the same loss function as used for sensor depth regularization.

4. Additional results

Per-scene quantitative results of GOC on the Synthetic Scenes are reported in Fig. 7. This process yields watertight and separable object meshes while preserving highly detailed features, enabling flexible scene rearrangement and object-level manipulation

5. Limitation

Currently, our approach supports geometry completion based solely on reconstructed point cloud data. It's simple and efficient but may struggle with complex object models due to ambiguity. In future work, we aim to integrate additional observations into the 3D model input, such as multiview CLIP features and texture information, leveraging multimodal data to achieve more accurate 3D completion and generation. This enhancement will enable a more robust integration of scene observations with data-driven priors.