ZeroShape: Regression-based Zero-shot Shape Reconstruction

Zixuan Huang[1*] Stefan Stojanov[2*] Anh Thai[2] Varun Jampani[3] James M. Rehg[1]
[1]University of Illinois at Urbana-Champaign,
[2]Georgia Institute of Technology, [3]Stability AI

Abstract

We study the problem of single-image zero-shot 3D shape reconstruction. Recent works learn zero-shot shape reconstruction through generative modeling of 3D assets, but these models are computationally expensive at train and inference time. In contrast, the traditional approach to this problem is regression-based, where deterministic models are trained to directly regress the object shape. Such regression methods possess much higher computational efficiency than generative methods. This raises a natural question: is generative modeling necessary for high performance, or conversely, are regression-based approaches still competitive? To answer this, we design a strong regression-based model, called ZeroShape, based on the converging findings in this field and a novel insight. We also curate a large real-world evaluation benchmark, with objects from three different real-world 3D datasets. This evaluation benchmark is more diverse and an order of magnitude larger than what prior works use to quantitatively evaluate their models, aiming at reducing the evaluation variance in our field. We show that ZeroShape not only achieves superior performance over state-of-the-art methods, but also demonstrates significantly higher computational and data efficiency.[1]

1. Introduction

Inferring the properties of individual objects such as their category or 3D shape is a fundamental task in computer vision. The ultimate goal is to do this accurately for any object, generally referred to as zero-shot generalization. For machine learning methods, this means high accuracy on data distributions that may be significantly different from the training set, such as images of novel types of objects like machine parts or images from uncommon visual contexts like underwater imagery. An object representation capable of zero-shot generalization, therefore, needs to accurately capture the visual properties that are shared across all objects in the world—an extremely ambitious goal.

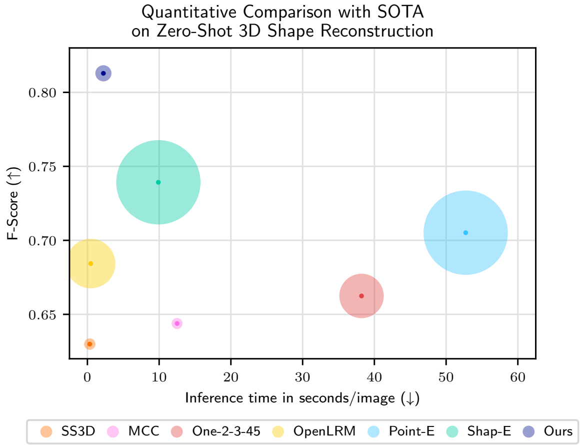


Figure 1. We outperform SOTA methods for zero-shot 3D shape reconstruction, while having faster inference time and less training data. Circle size indicates the number of 3D assets used for training, with biggest being 3M[2]. F-Score with threshold 0.05 is averaged over Octroc3D [51], Pix3D [52] and OmniObject3D [66].

Recent work in computer vision has taken the broader challenge of zero-shot generalization head-on, with impressive developments for 2D vision tasks like segmentation [26, 42], visual question answering [3, 27], image generation [1, 48, 49], and in training general vision representations that can be easily adapted for any vision task [38, 43]. This progress has largely been enabled by increasing model size and scaling training dataset size to the order of tens to hundreds of millions of images.

These developments have inspired efforts which aim at zero-shot generalization for single image 3D object shape reconstruction [22, 30, 31, 37]. This is a classical and famously ill-posed problem, with important applications like virtual object placement in scenes in AR and object manipulation in robotics. These works aim to learn a "zero-shot 3D shape prior" by relying on generative diffusion models for 3D point clouds [37], NeRFs [22], or for 2D images fine-

[*] Both authors contributed equally to this work.

[1] Project website at: https://zixuanh.com/projects/zeroshape.html

[2] We use 3M as a reference value. Point-E [37] and Shape-E [22] state a dataset size of "several million".

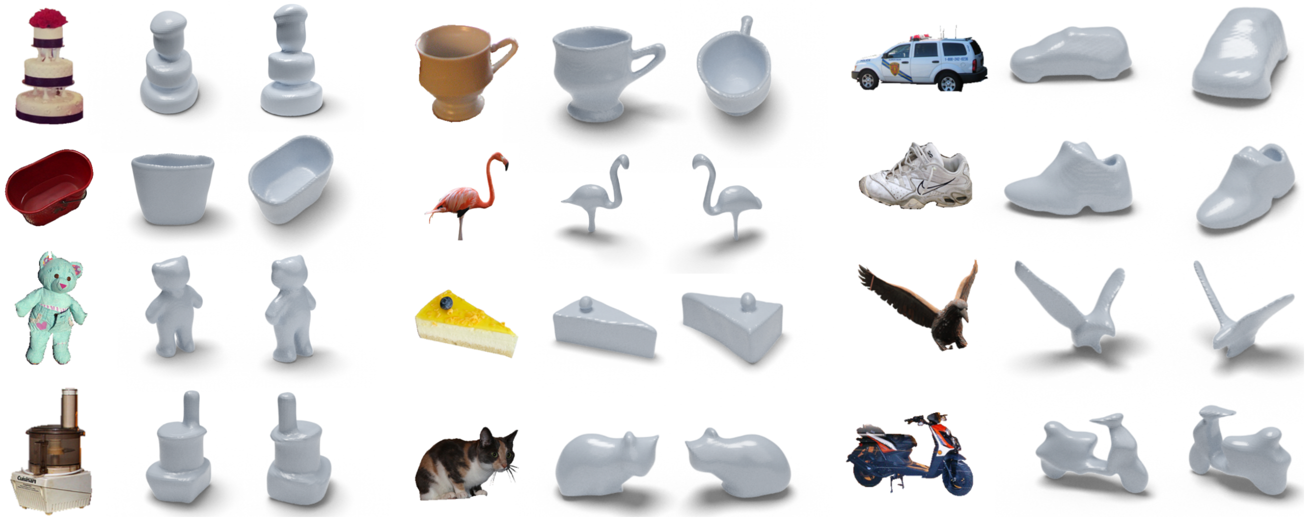


Figure 2. **ZeroShape reconstructions from in-the-wild images.** Our method produces detailed and accurate object reconstructions from single-view images on a diverse set of objects.

tuned for novel-view synthesis [30, 31], enabled by million-scale 3D data curation efforts such as Objaverse [9, 10]. While these methods show impressive zero-shot generalization ability, it comes at a great compute cost due to large model parameter counts and the inference-time sampling required by diffusion models. Using expensive generative modeling for zero-shot 3D shape from single images diverges from the approach of early deep learning-based works on this task [50, 56, 63, 67, 70]. These works define the task as a 3D occupancy or signed distance regression problem and predict the shape of objects in a single forward pass. This raises a natural question: is generative modeling necessary for high performance at learning zero-shot 3D shape prior, or conversely, can a regression-based approach still be competitive?

In this work, we find that regression approaches are indeed competitive if designed carefully, and computationally more efficient by a large margin compared to the generative counterparts. We propose ZeroShape: a regression-based 3D shape reconstruction approach that achieves state-of-the-art zero-shot generalization, trained entirely on synthetic data, while requiring a fraction of the compute and data budget of prior work (see Fig. 1). We build our model upon key ingredients that facilitate generalization based on prior works: 1) usage of intermediate geometric representation (e.g. depth) [33, 56, 64, 67, 70], 2) explicit reasoning with local features [5, 63, 68]. Specifically, we decompose the reconstruction into estimating the shape of the visible portion of the object, and then predicting the complete 3D object shape based on this initial prediction. The accurate estimation of the visible 3D surface is enabled by a joint modeling of camera intrinsics and depth, which we find to be essential for high accuracy.

Another thrust of our work is a large benchmark for evaluating zero-shot reconstruction performance. The 3D vision community is working on developing a zero-shot 3D shape prior, but what is the correct way to evaluate our progress? Currently we lack a well-defined benchmark, which has lead to well-curated qualitative results and small scale quantitative results[3] on different datasets across different papers. This makes it difficult to track progress and identify directions for future research. To resolve this and standardize evaluation, we develop a protocol based on data generated from existing datasets of 3D object assets. Our benchmark includes thousands of common objects from hundreds of different categories and multiple data sources. We consider real images paired with 3D meshes [51, 52], and also generate photorealistic renders of 3D object scans [66]. Our large scale quantitative evaluation provides a rigorous perspective on the current state-of-the-art.

In summary, our contributions are:

- ZeroShape: A regression-based zero-shot 3D shape reconstruction method with state-of-the-art performance at a fraction of the compute and data budget of prior work.
- A unified large-scale evaluation benchmark for zero-shot 3D shape reconstruction, generated by standardized processing and rendering of existing 3D datasets.

2. Related Work

Estimating the 3D shape of an object from a single is a complex inverse problem: while the shape of the visible object can be estimated from shading, estimating the shape of the occluded portion requires prior knowledge about object ge-

[3] On the order of hundreds of objects from tens of categories at best, to just a few dozen objects at worst.

ometry. This is one of the marvels of human perception and achieving it computationally is a major goal for our field. We review regression and generative methods for this task.

Regression-based Methods. These works investigate different ways to represent 3D object shapes and the architectures to produce them from a single image, e.g., meshes [23, 58, 62] or implicit representations like discrete [8, 54] or continuous [35, 39] occupancy, signed distance fields [21, 56, 68], point clouds [2, 14], or sets of parametric surfaces [15, 69]. A major limitation of these works is the limited generalization beyond the categories of the training set. The improvements of decomposing the problem into predicting the depth and then estimating the complete shape [50, 56, 64, 67, 70], and representing 3D in a viewer centered rather than object centered reference frame [50, 56, 70] allowed for improved zero-shot generalization. Most architectures follow an encoder/decoder design, where the encoder produces a feature map from which the decoder predicts the 3D shape. While early works produced a single feature vector for the entire image, it was later identified that using local features from a 2D feature map improved the detail of the predicted shapes [58, 68] and improved generalization to unseen categories [5, 67]. This culminated with the current state-of-the art regression method, MCC [63], which takes an RGB-D image as input, and uses a transformer-based encoder-decoder setup to produce a "shell occupancy" prediction [4]. Our approach incorporates all these prior findings for improved generalization, and builds upon them with a new module for estimating the visible shape of the object that estimates depth and camera intrinsics, which is processed with a cross attention-based decoder to produce an occupancy prediction.

3D Generative Methods This category of methods does zero-shot 3D shape reconstruction using a learned 3D generative prior, where the 3D generation is conditioned on one or few input images. Given image or text conditioning, early work [65] used GANs to generate voxels, whereas more recent works use diffusion models to generate point clouds [37], or function parameters for implicit 3D representations [22]. Another related type of generative framing is conditional view synthesis. Works in this direction fine-tune 2D generative models [31], or train them from scratch [60, 71], to synthesize novel views conditioned on single images and viewing angles. This results in an implicit 3D prior, from which a 3D shape can then be extracted by fitting a 3D neural representation to the synthesized images, or predicting its parameters [30].

3D from 2D Generative Models There have been efforts to use the real-world 2D image prior from text-to-2D generative models [1, 44, 48, 49] to reconstruct 3D shape from a single image. This category of works [11, 34, 53] often uses techniques such as the SDS loss [40, 57] and generates 3D assets from images by optimizing for each object separately. The prolonged optimization time prevents these works from being evaluated at scale or applied in many real-world applications. Orthogonal to the optimization-based approaches, we focus on learning a 3D shape prior that generalizes across instance. We do not perform any per-instance optimization at test time.

3. Method

Our goal is to achieve state of the art zero-shot performance for estimating the complete 3D shape of an object from a single image. Formally, given an object-centric single-view RGB image $I \in \mathbb{R}^{h\times w\times 3}$, we regress a function that takes I as input and predicts the shape. We represent shape using an implicit occupancy representation, where we model the shape surface as the isosurface of occupancy function $f(\boldsymbol{x}|I;\theta)$. Here, $\boldsymbol{x} \in \mathbb{R}^3$ denotes the query point's coordinates—when the query point lies within the surface $f(\boldsymbol{x}|I;\theta) = 1$, otherwise $f(\boldsymbol{x}|I;\theta) = 0$.

3.1. Architecture

We now present our architecture (see Fig. 3) for shape reconstruction. Our architecture is based on two established practices from prior works in this field: 1) usage of intermediate geometric representation [33, 56, 64, 67, 70] and 2) explicit reasoning with spatial feature maps [5, 63, 68]. Specifically, our model consists of three submodules: a depth and camera estimator, a geometric unprojection unit and a projection-guided shape reconstructor.

Depth and camera estimator. We propose to estimate the 3D visible object surface as an intermediate representation. To infer the full shape of an object, one must understand the visible surface—not only because the visible surface is often a large part of the full surface, but also because an accurate visible surface facilitates geometric reasoning of the full object reconstruction. This is because cues for reconstruction that allow for generalization, such as symmetry, curvature, and repetition, can be more effectively detected and leveraged in the 3D space. For example, if an object is symmetric, then accurately inferring the 3D symmetry planes from a partial 3D surface is much easier than from 2D RGB or relative depth.

Inspired by this, we estimate 3D visible surface as our intermediate representation instead of the commonly used depth maps [56, 64, 70]. MCC [63] also uses visible surface to estimate the full shape, but they assume the visible surface to be given as input. When inferring on in-the-wild images, they use fixed intrinsics to unproject depth maps into the 3D surface. Erroneous intrinsics lead to skewed 3D visible surfaces (see Fig. 4), resulting in inaccurate 3D cues for the complete object shape. Therefore, we propose

[4] Traditionally occupancy is formulated as predicting whether a point in 3D is inside/outside a watertight mesh, whereas MCC predicts whether it is within an ε wide shell representing the surface of the object.

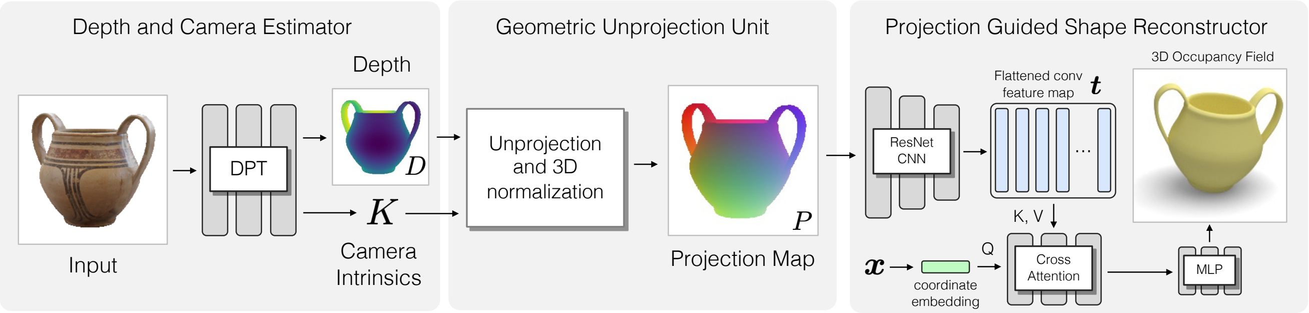

Depth and Camera Estimator
Depth
Input
DPT
D
K
Camera Intrinsics
Geometric Unprojection Unit
Unprojection and 3D normalization
P
Projection Map
Projection Guided Shape Reconstructor
3D Occupancy Field
Flattened conv feature map
t
ResNet CNN
K, V
Q
x
coordinate embedding
Cross Attention
MLP

Figure 3. **Overview of our model.** Our consists of three modules: a depth and camera estimator, a geometric unprojection unit and a projection-guided shape reconstructor. The depth and camera estimator predicts the depth and camera intrinsics from the input image with a DPT backbone. The geometric unprojection unit converts the depth and intrinsics estimation into a normalized 3D visible surface, which is parameterized by a three-channel projection map. The shape reconstructor finally reconstructs the full occupancy field by fetching localized information from projection map through cross attention.

to jointly estimate depth and intrinsics before predicting the full shape. Note that learning to estimate depth and intrinsics can be fully supervised with synthetic data. Specifically, our depth and camera estimator estimates the depth map of the object $D \in \mathbb{R}^{h \times w}$ and the camera intrinsics $K \in \mathbb{R}^{3 \times 3}$ from the image I. We used a shared DPT [46] backbone for the depth and camera estimator, and use two different shallow heads to predict D from the local tokens and K from the global token.

Geometric unprojection unit. Given the intrinsics K and the depth map D, the geometric unprojection unit unprojects them into a projection map $P \in \mathbb{R}^{h \times w \times 3}$. The projection map encodes the visible surface of the object, where each pixel value P_{ij} represents the coordinate of the unprojected 3D point at the pixel location (i, j). Formally, the geometric unprojection can be written as

$$P_{ij} = D_{ij} K^{-1} [i, j, 1]^T. \tag{1}$$

We use a view-centric coordinate system, because prior works show that view-centric learning is beneficial to generalization [55, 56]. Therefore the camera coordinate frame is the "world" coordinate frame for shape reconstruction, which means that only the camera intrinsics matrix is required to unproject pixels to 3D. Note that unprojection is fully differentiable w.r.t. D and K, so we can easily use it as a module in an end-to-end learning-based model. Additionally, the projection maps are foreground-segmented, and the represented visible surface is normalized in the 3D space to be zero-mean and unit-scale before being fed into the next module.

Projection-guided shape reconstructor. Using the estimated projection map P, our projection-guided shape reconstructor recovers the full object shape. Specifically, the projection-guided shape reconstructor first uses a ResNet [16] encoder to encode and reshape the projection map P into a set of d-dimensional vectors, $\boldsymbol{t} \in \mathbb{R}^{k \times d}$. Each of the k vectors encodes a localized patch in the projection map. To facilitate an explicit spatial reasoning, we use the cross-attention-based approach proposed in MCC [63]. We linearly map every query point $\boldsymbol{x} \in \mathbb{R}^3$ to the same dimension of feature vectors, d. Then we use two cross attention layers to fetch relevant patch encodings from $\boldsymbol{t}$ and fuse them with each query separately. Finally, the reconstructor predicts the occupancy value of each $\boldsymbol{x}$ using the fused feature vector via an MLP.

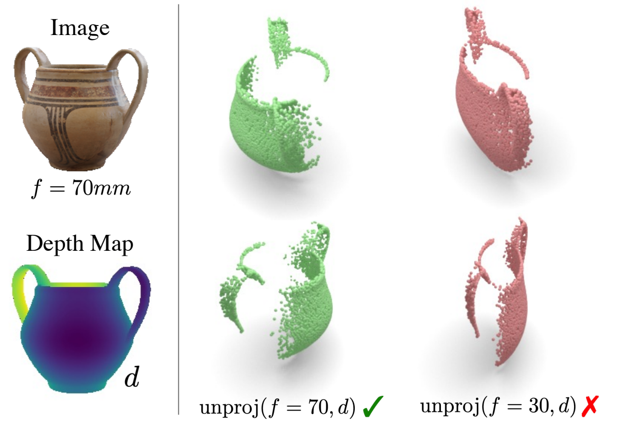


Figure 4. **Effect of Intrinsics.** Unprojecting an accurate depth map into a 3D surface surface with erroneous intrinsics leads to skewed shape with wrong 3D aspect ratio.

3.2. Loss

We use a two-stage training paradigm for our model, where we first pretrain the depth and camera estimator and then fine-tune the whole model with 3D supervision. For depth and camera pretraining, we use a depth loss $\mathcal{L}_{depth}$ and projection-based intrinsics loss $\mathcal{L}_{proj}$. For the depth loss, we use the SSIMAE loss from [45]. Note that the SSI-

MAE loss is scale-invariant, meaning that the depth estimator trained using this loss will be metrically correct up to an unknown scale factor. Therefore, directly regressing the absolute intrinsics is suboptimal due to the uncertainty in the absolute scale. Instead, we observe that the only factor that impacts shape reconstruction is whether the visible surface recovered using Eq. (1) is accurate. Therefore, we directly minimize the MSE loss between the predicted projection map P and the ground truth projection map P^*, and backpropagate to the camera and depth estimators.

Once the depth and camera estimator are learned, we jointly train the whole model using the 3D occupancy loss $\mathcal{L}_{occ}$, which is a standard binary cross entropy between the predicted occupancy values $f(\boldsymbol{x}|I;\theta)$ and ground truth in the viewer-centric coordinate frame.

3.3. Implementation Details

We train our model with the Adam [25] optimizer. During depth and camera pretraining, we use a learning rate of 3×10^{-5}, a batch size of 44, a weight decay of 0.05 and momentum parameters of (0.9, 0.95). We train our model for 15 epochs and initialize the depth estimator with the Omnidata [13] weights. During the joint training stage, we use a learning rate of 3×10^{-5} for the projection-guided shape reconstructor, and a learning rate of 10^{-5} for the pretrained depth and camera estimator (geometric unprojection unit does not have learnable parameters). We use a batch size of 28, a weight decay of 0.05 and momentum parameters of (0.9, 0.95). At every iteration, we randomly sample 4096 points to compute the occupancy loss. We train our model on $4\times$ NVIDIA GeForce RTX 2080 Ti, which takes $\sim$2 days for pretraining and $\sim$3 days for joint training.

4. Data Curation

4.1. Training Dataset

We use all the 55 categories of ShapeNetCore.v2 [6] for a total of about 52K meshes, as well as over 1000 categories from the Objaverse-LVIS [9] subset. This subset of Objaverse has been manually filtered by crowd workers to primarily include meshes of objects instead of other assets like scans of large scenes and buildings. After filtering Objaverse-LVIS to remove objects with minimal geometry (e.g. objects consisting of a single plane) this dataset has 42K meshes. Pooling these two data sources gives us a total of over 90K 3D object meshes from over 1000 categories.

We use Blender [41] to generate synthetic images from the 3D meshes, and to extract a variety of useful annotations: depth maps, camera intrinsics, and object and camera pose. Because the object distribution of ShapeNet is highly skewed (67% of data is 7 categories), we generate 1 to 20 images per object, scaled inversely from the number of meshes in the category of the object, resulting in a total of 159K images. For Objaverse we generate 25 images per object resulting in 939K images. Our traning set consists of slightly less than 1.1M images.

We generate images with varying focal lengths, from 30mm to 70mm for a 35mm image sensor size equivalent. We generate diverse object-camera geometry: rather than the common approach of always pointing the camera at the middle of the object at a fixed distance, we vary the object camera distance and vary the LookAt point of the camera. This allows us to capture a wide range of variability in how 3D shape projects to 2D. We follow the convention to use center cropped and foreground segmented images for training and testing. We provide more details in the supplement.

4.2. Evaluation Benchmark

We use three different real-world dataset evaluation: OmniObject3D [66], Ocrtoc3D [51], and Pix3D [52]. Because our testing images images come from the real world, or are renders of real 3D object scans distinct from our training set, they are a good test set for zero-shot generalization.

OmniObject3D. OmniObject3D is a large and diverse dataset of 3D scans and videos of objects from 216 categories, including household objects and products, food and toys. Because the foreground segmentations are noisy, we follow convention and render the 3D scans to generate test images [30, 31]. We improve the default material shader which generates glass-like surface appearance to appear more natural. We use Blender and HDR environment maps to generate realistic images with diverse lighting. We randomly sample camera viewpoint, distance and focal length.

Ocrtoc3D. Ocrtoc3D is a real-world object dataset that contains object-centric videos and full 3D annotations from 15 coarse categories. Some coarse categories contain many subcategories (e.g. toy animals contain various species). For each video the mesh (3D scan) and the viewpoint information are provided. We clean up this dataset by manually removing outliers (e.g. empty meshes/wrong object scales) and use the full filtered dataset consisting of 749 unique image-object pairs.

Pix3D. Pix3D is a real-world object dataset that contains 3D annotations from 9 categories. For each image in this dataset, an object mask, a CAD model, and the input viewpoint information are provided. These 3D annotations come from manual alignment between shapes and images. We follow the split of [19] and use 1181 images.

Benchmark curation. To create an easy-to-use benchmark, we convert the three heterogeneous datasets into a unified format. This includes aligning and converting the camera intrinsics and extrinsics, and object poses, to a standardized convention across the test datasets and our synthetic dataset. This is often a tedious obstacle in 3D vision research. We also organize images, masks and other metadata in a standardized manner. The release of our training data, data gen-

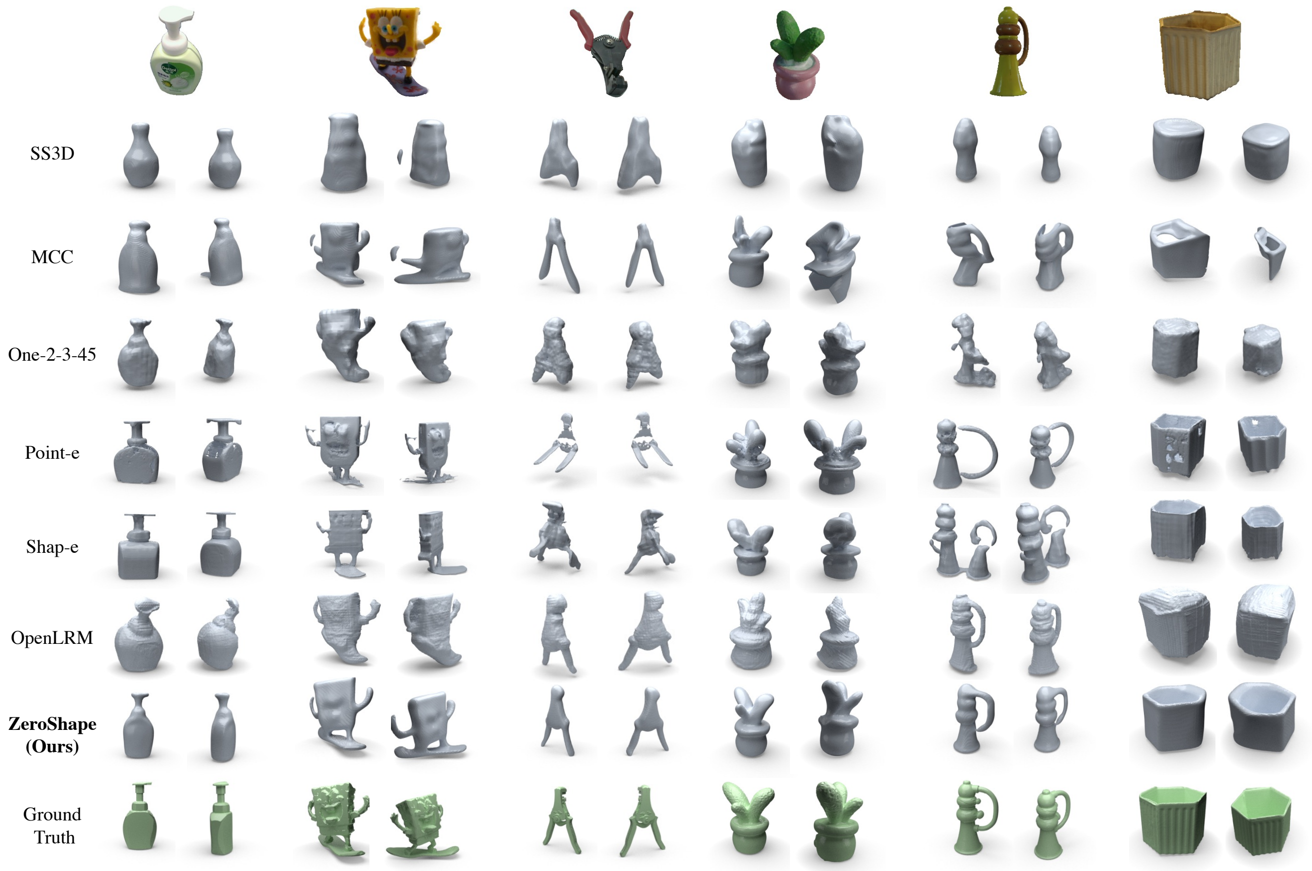


Figure 5. **Qualitative results.** We compare ZeroShape to other SOTA methods on our curated benchmark (first three columns are from Ocrtoc3D [51], last three are from OmniObject3D [66]). Our reconstruction not only better aligns with the visible surfaces from images, but also recovers a faithful global structure of the reconstructed objects.

erating pipeline, and benchmark will benefit the community by providing a unified setup for large scale training on synthetic data and large scale testing on real data.

5. Experiments

In this section we present our experiments, which include state-of-the-art comparisons and ablations. We first describe the baselines we implemented on our benchmark, and then show detailed quantitative and qualitative results.

5.1. Metrics

We evaluate the shape reconstruction models using Chamfer Distance (CD) and F-score as our quantitative metrics following [15, 19, 20, 28, 55, 56].

Chamfer Distance. Chamfer Distance (CD) measures the alignment between two pointclouds. Following [20], CD is defined as an average of accuracy and completeness. Denoting pointclouds as X and Y, CD is defined as:

$$\mathrm{CD}(X,Y) = \frac{1}{2|X|}\sum_{x \in X} \min_{y \in Y} \|x-y\|_2 + \frac{1}{2|Y|}\sum_{y \in Y} \min_{x \in X} \|x-y\|_2 \tag{2}$$

F-score. F-score measures the pointcloud alignment under a classification framing. By selecting a rejection threshold d, F-score@d (FS@d) is the harmonic mean of precision@d and recall@d. Specifically, precision@d is the percentage of predicted points that lies close to GT point cloud within distance d. Similarly, recall@d is the percentage of ground truth points that have neighboring predicted points within distance d. FS@d can be intuitively interpreted as the percentage of surface that is correctly reconstructed under a certain threshold d that defines correctness.
Evaluation protocol. To compute CD and F-score, we grid-sample the implicit function and extract the isosurface via Marching Cubes [32] for methods using implicit representation. Then we sample 10K points from the surfaces for the evaluation of CD and F-scores. Because most methods cannot predict view-centric shapes, we use brute-force search to align the coordinate frame of prediction with

Table 1. **Quantitative comparison on OmniObject3D.** Our method performs favorably to other state-of-the-art methods.

Methods	FS@1↑	FS@2↑	FS@5↑	CD↓
SS3D [4]	0.1515	0.3482	0.6618	0.482
MCC [63]	0.1362	0.3215	0.6015	0.551
One-2-3-45 [30]	0.1532	0.3585	0.6882	0.446
OpenLRM [18]	0.1683	0.3848	0.7204	0.407
Point-E [37]	0.1505	0.3598	0.6932	0.448
Shap-E [22]	0.1483	0.3650	0.7029	0.434
ZeroShape (ours)	**0.2297**	**0.4927**	**0.8169**	**0.310**

Table 2. **Quantitative comparison on Pix3D.** Our method performs favorably to other state-of-the-art methods.

Methods	FS@1↑	FS@2↑	FS@5↑	CD↓
SS3D [4]	0.1326	0.2998	0.6316	0.485
MCC [63]	0.1754	0.3386	0.6165	0.514
One-2-3-45 [30]	0.1364	0.3137	0.6666	0.443
OpenLRM [18]	0.1458	0.3190	0.6440	0.492
Point-E [37]	0.1779	0.3830	0.7255	0.403
Shap-E [22]	**0.2016**	0.4287	**0.7833**	**0.340**
ZeroShape (ours)	0.1928	**0.4290**	0.7759	0.345

Table 3. **Quantitative comparison on Ocrtoc3D.** Our method performs favorably to other state-of-the-art methods.

Methods	FS@1↑	FS@2↑	FS@5↑	CD↓
SS3D [4]	0.1271	0.2910	0.5963	0.543
MCC [63]	0.1994	0.4098	0.7135	0.411
One-2-3-45 [30]	0.1323	0.3076	0.6325	0.492
OpenLRM [18]	0.1552	0.3481	0.6885	0.432
Point-E [37]	0.1589	0.3591	0.6968	0.423
Shap-E [22]	0.1725	0.3939	0.7315	0.395
ZeroShape (ours)	**0.2410**	**0.5091**	**0.8459**	**0.286**

ground truth before calculating the metrics. This evaluation protocol ensures a fair comparison between methods with different shape representation and coordinate conventions.

5.2. Baselines

We consider five state-of-the-art baselines for shape reconstruction, SS3D [4], MCC [63], Point-E [37], Shap-E [22], One-2-3-45 [30] and OpenLRM [17, 18].

SS3D learns implicit shape reconstruction by first pretraining on ShapeNet GT, and then finetuning on real-world single-view images. The finetuning is performed in a category specific way, and then a single unified model is distilled from all category-specific experts. We compare our model with their final distilled model.

MCC learns shell occupancy reconstruction using multi-view estimated point clouds from CO3D [47]. Their model assumes known depth and intrinsics during inference. To evaluate their model on RGB images, we use the DPT-estimated depth and fixed intrinsics as MCC's input following their pipeline.

Point-E is a point cloud diffusion model that generates point clouds from text prompts or RGB images. They additionally train a separate model that converts point clouds into meshes. We compare our model with Point-E by combining their image-to-point and point-to-mesh models.

Shap-E is another diffusion model that learns conditioned shape generation from text or images. Different from Point-E, Shap-E uses a latent diffusion setup and can directly generate implicit shapes. The final mesh reconstruction are extracted with marching cubes.

One-2-3-45 learns implicit shape reconstruction by breaking it down into a generative view synthesis step and a multiview-to-3D reconstruction step. The view synthesis is achieved with Zero-123 [31], a diffusion model that generates novel-view images conditioned on the original images and poses. Based on the synthesized multi-view images, a cost-volume-based module reconstructs the full 3D mesh of the object.

LRM is a concurrent work that learns to predict NeRF [36] from single images using transformer-based architecture. Since the authors have not released the code, we use the code and weights from **OpenLRM** [5]. The mesh is extracted via Marching Cubes [32] from the triplane NeRF.

5.3. Comparison to SOTA Methods

We compare our approach to other state-of-the-art methods on the benchmark we curated. We now present and analyze the quantitative results for each dataset.

Results on OmniObject3D. We present our main quantitative comparison results on OmniObject3D, which covers a great variety of object types. The results are shown in Tab. 1. Comparing with other SOTA zero-shot 3D reconstruction methods, we see our approach achieves significantly better performance.

Results on Ocrtoc3D. We present additional quantitative comparison results on Ocrtoc3D. Ocrtoc is smaller than OmniObject, but still covers many object types, and the input images are real photos. The results are shown in Tab. 3. Similar to the results on OmniObject3D, our approach outperforms previous SOTA methods by a large margin.

Results on Pix3D. We also present quantitative comparison results on Pix3D. Unlike OmniObject3D and Ocrtoc3D, the object variety of this evaluation dataset is much lower — all objects are furniture and more than two third of the images are chairs and sofas. Therefore, the evaluation results are highly bias towards this specific class of objects. The results are shown in Tab. 2, and our method still achieves state-of-the-art performance. It is worth noting that Point-E

[5] https://github.com/3DTopia/OpenLRM

Table 4. **Ablation study on OmniObject3D.** The design choices of our architecture are quantitatively justified: enforcing explicit geometric reasoning, and implementing it through unprojection with estimated depth and intrinsics is essential.

Methods	FS@1↑	FS@2↑	FS@5↑	CD↓
Ours *w/o geo*	0.2110	0.4572	0.7797	0.347
Ours *w/o unproj*	0.2135	0.4738	0.8053	0.323
Ours *w/o intr*	0.2158	0.4742	0.8039	0.324
Ours	**0.2297**	**0.4927**	**0.8169**	**0.310**

and Shap-E also perform well on this dataset. We hypothesize this is might relate to the abundance of similar furniture categories in their training set.

5.4. Qualitative Results

We show qualitative results of different methods in Fig. 5. Generative approaches such as Point-E and Shap-E tend to have sharper surfaces and contain more details in their generation. However, many details are erroneous hallucination that do not accurately follow the input image, and the visible surfaces are often reconstructed incorrectly. Previous regression-based approaches such as MCC better follow the input cues in the input images, but the hallucination of the occluded surfaces is often inaccurate. We observe that One-2-3-45, OpenLRM and SS3D cannot always accurately capture details and concavities. Comparing with prior arts, the reconstruction of ZeroShape not only faithfully capture the global shape structure, but also accurately follows the local geometry cues from the input image. More qualitative results are included in the supplement.

5.5. Ablation Study

We analyze our method by ablating the design choices we made. We consider baselines by modifying different modules correspondingly. The results are shown in Tab. 4.

Explicit geometric reasoning. We first consider the baseline without any geometric reasoning (Ours *w/o geo*). We remove the projection unit together with the depth and camera pretraining losses. The number of parameters is controlled to be the same, and we train the model for the same number of total iterations. Comparing the first row to the last row, we see that enforcing explicit geometric reasoning in our model positively affects performance.

Alternative intermediate representations. Prior works [56, 64, 65] typically consider depth as the 2.5D intermediate representation. To compare this to our projection-based representation, we consider a baseline where the latent vectors directly come from the depth map instead of a 3D projection map. As shown in Tab. 4 (Ours *w/o unproj*), depth leads to inferior performance to our intrinsic-guided projection map representation.

Intrinsic-guided projection. We propose joint learning of intrinsics with depth to more accurately estimate the 3D shape of the visible object surface. To study the impact of this, we compare our full model with a baseline without intrinsics learning, where the unprojection to 3D is done via a fixed intrinsics during both training and testing. This baseline (Ours *w/o intr*) leads to indifferent performance to using depth intermediate representation and is worse than our full model. We also show qualitative examples of the estimated surface using our pretrained intrinsics estimator in Fig. 6. Compared with fixed intrinsics, unprojection with our estimated intrinisics leads to more accurate reconstruction of the visible surface.

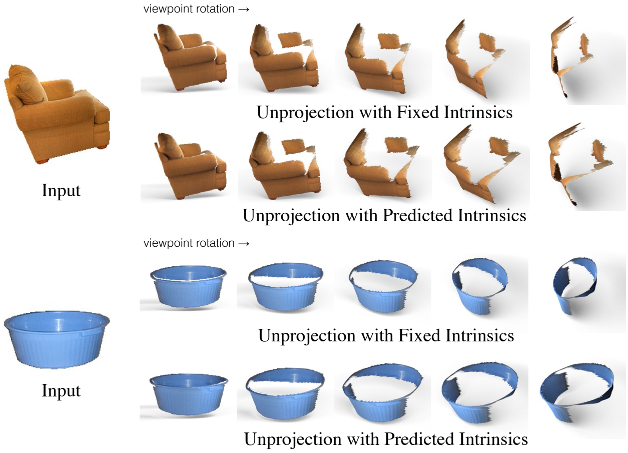

viewpoint rotation →
Input
Unprojection with Fixed Intrinsics
Unprojection with Predicted Intrinsics
viewpoint rotation →
Input
Unprojection with Fixed Intrinsics
Unprojection with Predicted Intrinsics

Figure 6. **Benefits of intrinsics learning.** We show the reconstructed visible surfaces for two real image inputs. The visible surface is unprojected from estimated depths, with either fixed intrinsics or predicted intrinsics. Using fixed intrinsics cause unrealistic deformations in the 3D aspect ratio of the visible object surface (e.g. objects appear to be compressed).

6. Limitations and Discussion

Due to computational resource limitations, we are not able to process and train our model on the full Objaverse dataset. Currently, the meshes from Objaverse we use only consist of 5% of Objaverse and 0.4% of Objaverse-XL objects. Based on the promising scaling properties of recent foundation models [12, 24, 61], we believe it will be valuable to explore the scaling properties of method.

Another limitation of our work is that we have not considered the modeling of object texture. Predicting textures of unseen surfaces is highly ill-posed and can greatly benefit from a strong 2D prior. Given the recent success of 2D diffusion models [48] and their application in optimization-based 3D generation methods [7, 11, 29, 34, 40, 59], we think it will be promising to initialize or regularize these methods with our shape prior, potentially boosting both the optimization efficiency and generation quality.

7. Conclusion

We present a strong regression-based model for zero-shot shape reconstruction. The core of our model is an intermediate representation of the 3D visible surface which facilitates effective explicit 3D geometric reasoning. We also curate a large real-world evaluation benchmark to test zero-shot shape reconstruction methods. Our benchmark pools data from three different real-world 3D datasets and has an order of magnitude larger scale than the test sets used by prior work. Tested on our benchmark, our model significantly outperforms other SOTA methods and achieves higher computational efficiency, despite being trained with much less 3D data. We hope our effort is a meaningful step towards building zero-shot generalizable 3D reconstruction models.

Acknowledgement. This work was supported by NIH R01HD104624-01A1.

References

[1] Midjourney: https://www.midjourney.com/showcase/recent/. Accessed: May 5, 2023. 1, 3

[2] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 3

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1

[4] Kalyan Vasudev Alwala, Abhinav Gupta, and Shubham Tulsiani. Pre-train, self-train, distill: A simple recipe for supersizing 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3773–3782, 2022. 7

[5] Jan Bechtold, Maxim Tatarchenko, Volker Fischer, and Thomas Brox. Fostering generalization in single-view 3d reconstruction by learning a hierarchy of local and global shape priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15880–15889, 2021. 2, 3

[6] 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. 5

[7] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 8

[8] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016. 3

[9] 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. *arXiv preprint arXiv:2212.08051*, 2022. 2, 5

[10] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 2

[11] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. *arXiv preprint arXiv:2212.03267*, 2022. 3, 8

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 8

[13] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 5

[14] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 3

[15] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3, 6

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[17] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023. 7

[18] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 7

[19] Zixuan Huang, Stefan Stojanov, Anh Thai, Varun Jampani, and James M Rehg. Planes vs. chairs: Category-guided 3d shape learning without any 3d cues. In *European Conference on Computer Vision*, pages 727–744. Springer, 2022. 5, 6

[20] Zixuan Huang, Varun Jampani, Anh Thai, Yuanzhen Li, Stefan Stojanov, and James M Rehg. Shapeclipper: Scalable 3d shape learning from single-view images via geometric and clip-based consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 6

[21] Muhammad Zubair Irshad, Sergey Zakharov, Rares Ambrus, Thomas Kollar, Zsolt Kira, and Adrien Gaidon. Shapo: Implicit representations for multi-object shape appearance and pose optimization. 2022. 3

[22] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 1, 3, 7

[23] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 3

[24] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 8

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 1

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1

[28] Chen-Hsuan Lin, Chaoyang Wang, and Simon Lucey. Sdf-srn: Learning signed distance 3d object reconstruction from static images. *arXiv preprint arXiv:2010.10505*, 2020. 6

[29] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 8

[30] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 1, 2, 3, 5, 7

[31] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *arXiv preprint arXiv:2303.11328*, 2023. 1, 2, 3, 5, 7

[32] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 6, 7

[33] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. 2, 3

[34] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360 reconstruction of any object from a single image. In *CVPR*, 2023. 3, 8

[35] 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. 3

[36] 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. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 7

[37] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 1, 3, 7

[38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[39] 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. 3

[40] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3, 8

[41] Blender Project. Blender, https://blender.org/. 5

[42] Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Weidong Guo, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High quality entity segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4047–4056, 2023. 1

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[45] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 4

[46] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 4

[47] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 7

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 3, 8

[49] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 3

[50] Daeyun Shin, Charless C Fowlkes, and Derek Hoiem. Pixels, voxels, and views: A study of shape representations for single view 3d object shape prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3069, 2018. 2, 3

[51] Rakesh Shrestha, Siqi Hu, Minghao Gou, Ziyuan Liu, and Ping Tan. A real world dataset for multi-view 3d reconstruction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 56–73. Springer, 2022. 1, 2, 5, 6

[52] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2974–2983, 2018. 1, 2, 5

[53] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023. 3

[54] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *Proceedings of the IEEE international conference on computer vision*, pages 2088–2096, 2017. 3

[55] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2019. 4, 6

[56] Anh Thai, Stefan Stojanov, Vijay Upadhya, and James M Rehg. 3d reconstruction of novel object shapes from single images. In *2021 International Conference on 3D Vision (3DV)*, pages 85–95. IEEE, 2021. 2, 3, 4, 6, 8

[57] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 3

[58] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 3

[59] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 8

[60] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 3

[61] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 8

[62] Chao Wen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2mesh++: Multi-view 3d mesh generation via deformation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1042–1051, 2019. 3

[63] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3d reconstruction. *arXiv preprint arXiv:2301.08247*, 2023. 2, 3, 4, 7

[64] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. *Advances in neural information processing systems*, 30, 2017. 2, 3, 8

[65] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning shape priors for single-view 3d completion and reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 646–662, 2018. 3, 8

[66] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. *arXiv preprint arXiv:2301.07525*, 2023. 1, 2, 5, 6

[67] Yongqin Xiang, Julian Chibane, Bharat Lal Bhatnagar, Bernt Schiele, Zeynep Akata, and Gerard Pons-Moll. Any-shot gin: Generalizing implicit networks for reconstructing novel classes. In *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022. 2, 3

[68] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in neural information processing systems*, 32, 2019. 2, 3

[69] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018. 3

[70] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Josh Tenenbaum, Bill Freeman, and Jiajun Wu. Learning to reconstruct shapes from unseen classes. *Advances in neural information processing systems*, 31, 2018. 2, 3

[71] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. *arXiv preprint arXiv:2212.00792*, 2022. 3

A. Additional Qualitative Comparison

We show additional qualitative results on OmniObject3D, Ocrtoc3D and Pix3D in Fig. 7, Fig. 8 and Fig. 9, respectively. Comparing with prior arts, the reconstruction of ZeroShape better captures the global shape structure and visible geometric details.

B. Inference on AI-generated Images

We present additional results of ZeroShape using images generated with DALL·E 3. To test the out-of-domain generalization ability, we generate images of imaginary objects as the input to our model (see Fig. 10). Despite the domain gap to realistic or rendered images, ZeroShape can faithfully recover the global shape structure and accurately follow the local geometry cues from the input image. These results also demonstrate the potential of using ZeroShape in a text-based 3D generation workflow.

C. Data Curation Details

In this section we describe our data generation procedure for training and for rendering the object scans from OmniObject3D to generate one of our benchmark test sets.

C.1. Synthetic Training Dataset Generation

Image Rendering. For an arbitrary 3D mesh asset, our Blender-based rendering pipeline first loads it into a scene and normalizes it to fit inside a unit cube. Our scene consists of a large rectangular bowl with a flat bottom, a common scene setup that 3D artists use for rendering to allow for realistic shading, and 4 point light sources and one area light source. We randomly place cameras around the object with 30mm to 70mm focal length for a 35mm sensor size equivalent. We randomly vary the distance, elevation (from 5 to 65 degrees), the LookAt point of the camera and generate images of 600 × 600 resolution (see Fig. 11). This variation in object/camera geometry allows capturing the variability of projective geometry in real world scenarios, coming from different capture devices and camera poses. This is in contrast with prior work that uses fixed intrinsics, fixed distance, and LookAt pointed at the center of the object.

In addition to RGB images, we extract segmentation masks, depth maps, intrinsics, extrinsics and object pose. We center crop the objects, mask out the background, resize images to 224 × 224 and process the additional annotations to account for the crop, segmentation and resize.

Ground Truth Occupancy Extraction. To obtain ground truth occupancy, we first extract watertight meshes using the code from occupancy networks [6], and then extract SDF for 32^3 query points per mesh following DISN [7]. The SDF is converted to occupancy during training.

C.2. Generating the OmniObject3D Testing Set

The original videos released by the OmniObject3D dataset have noisy foreground masks and are mostly taken indoor on a tabletop. To improve the lighting variability and ensure accurate segmentations, we follow the rendering procedure described in the previous section to generate testing data. Different from our training set generation, we use HDRI environment maps to generate scene lighting, which results in high lighting quality and diversity (see Fig. 12).

[6] https://github.com/autonomousvision/occupancy_networks

[7] https://github.com/laughtervv/DISN

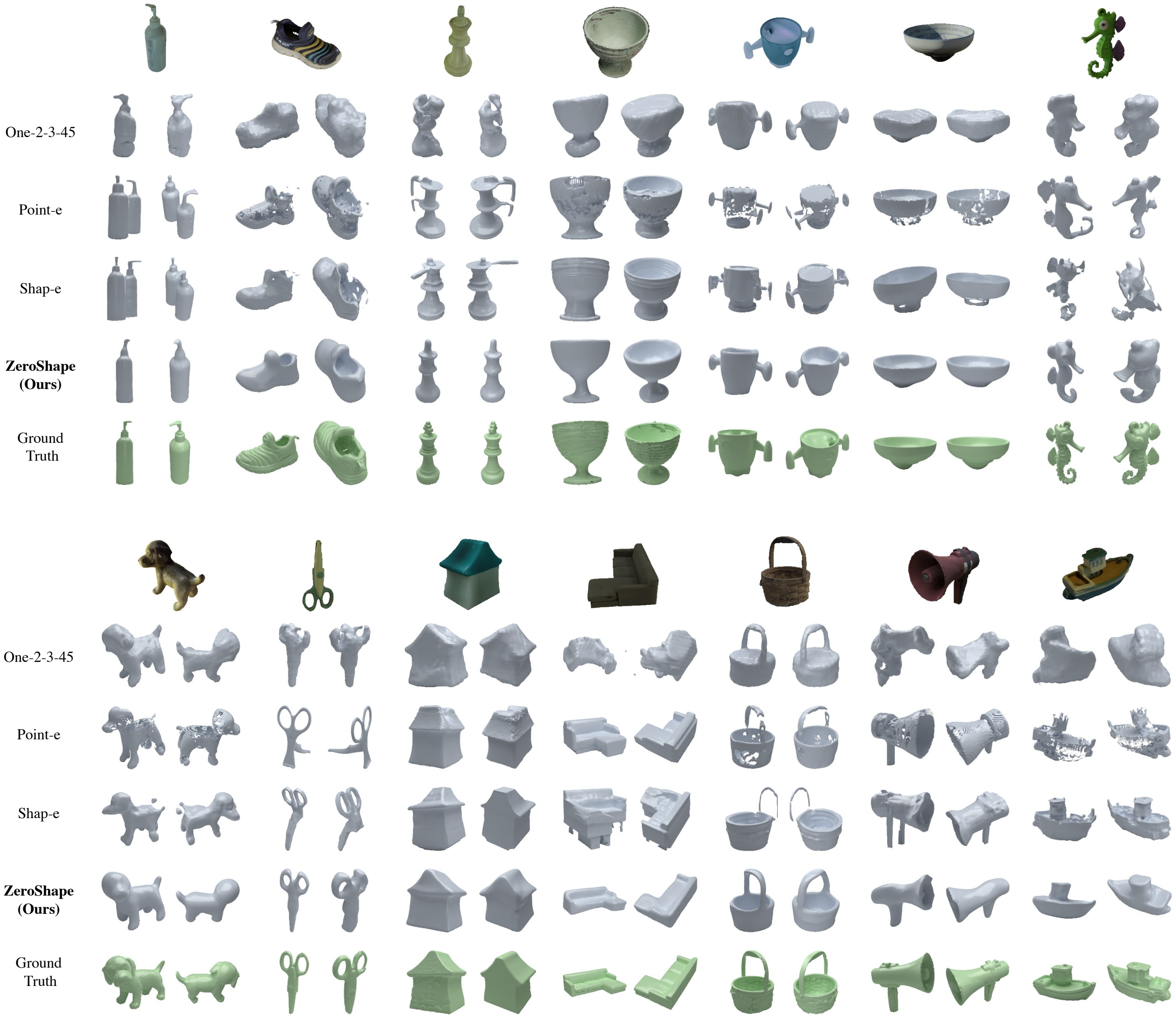


Figure 7. Additional qualitative results and comparison on OmniObject3D.

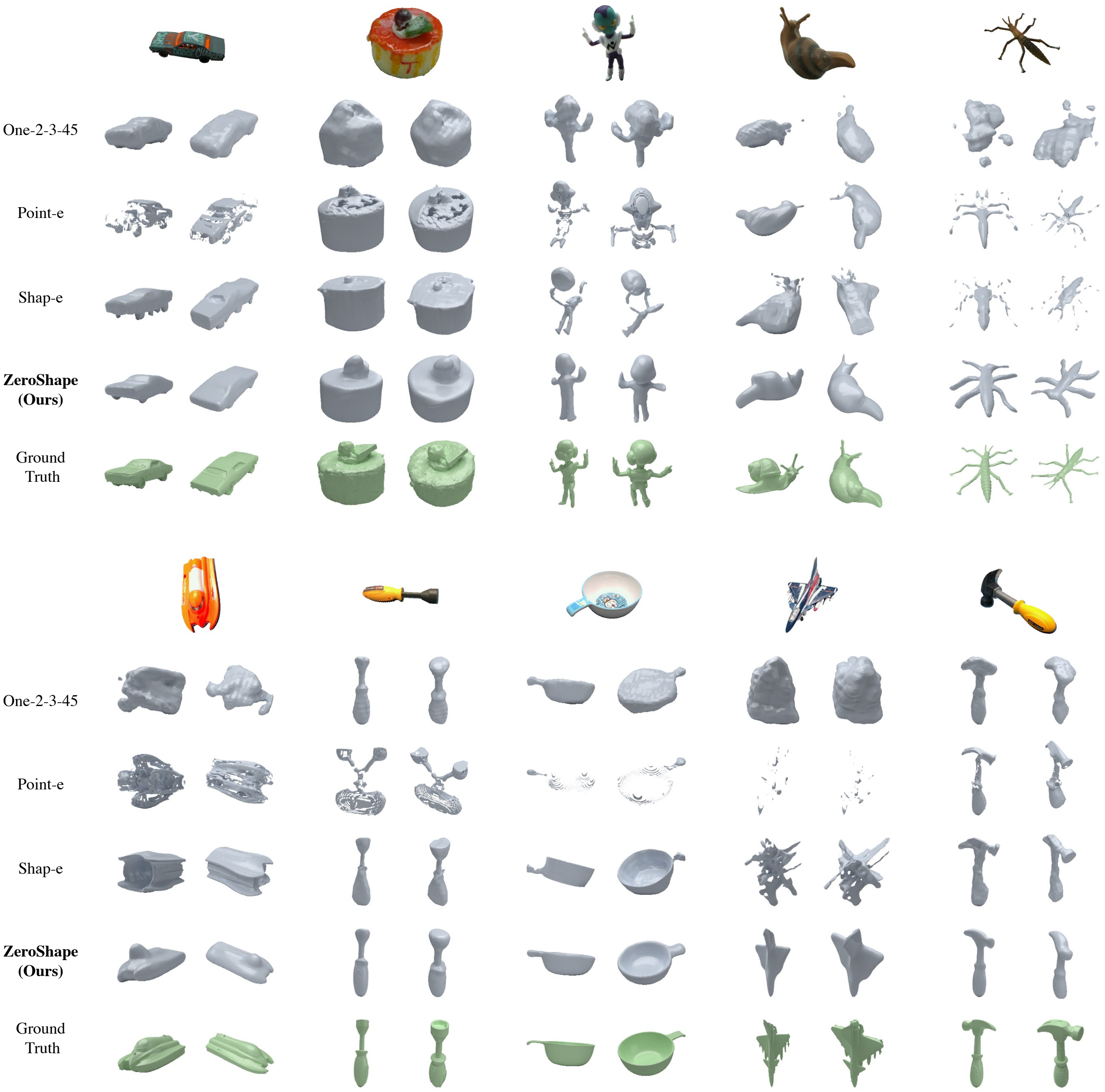


Figure 8. Additional qualitative results and comparison on Ocrtoc3D.

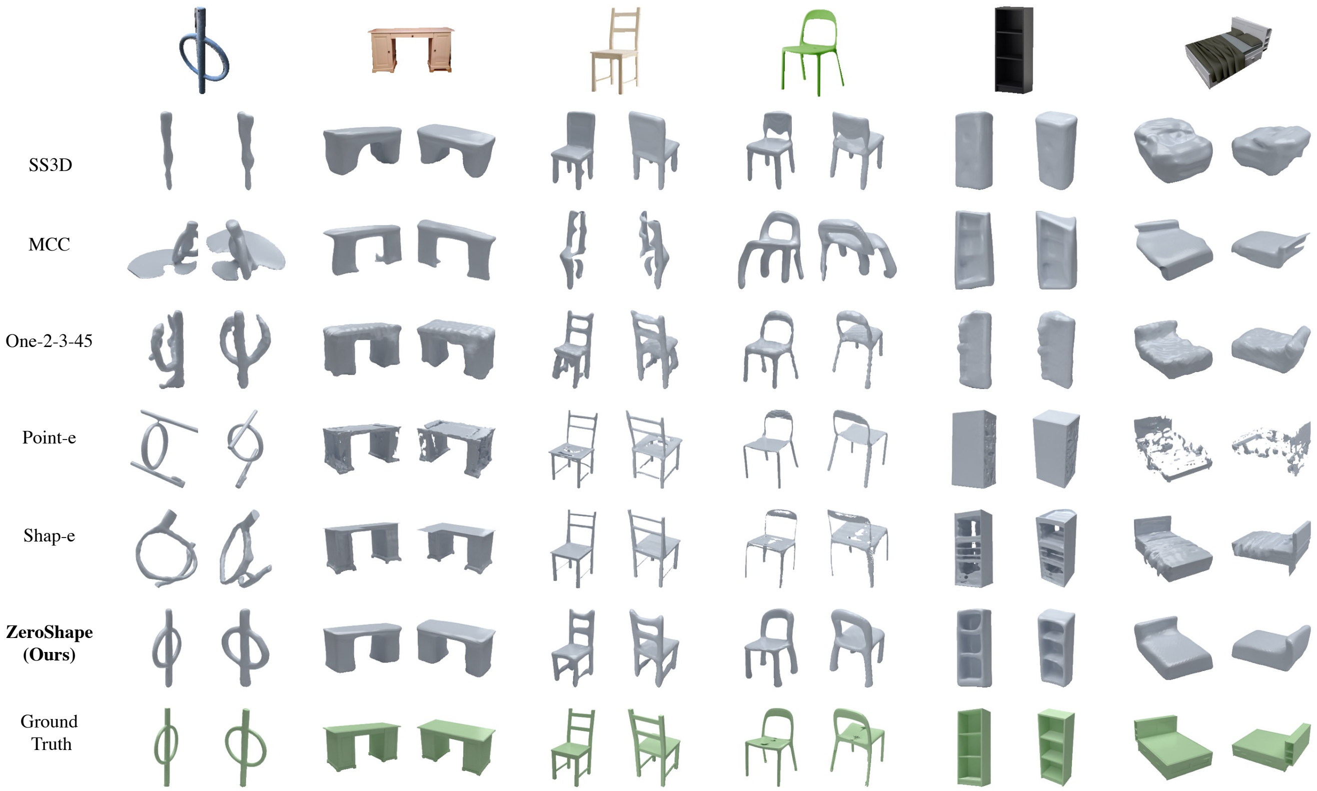


Figure 9. Qualitative results and comparison on Pix3D.

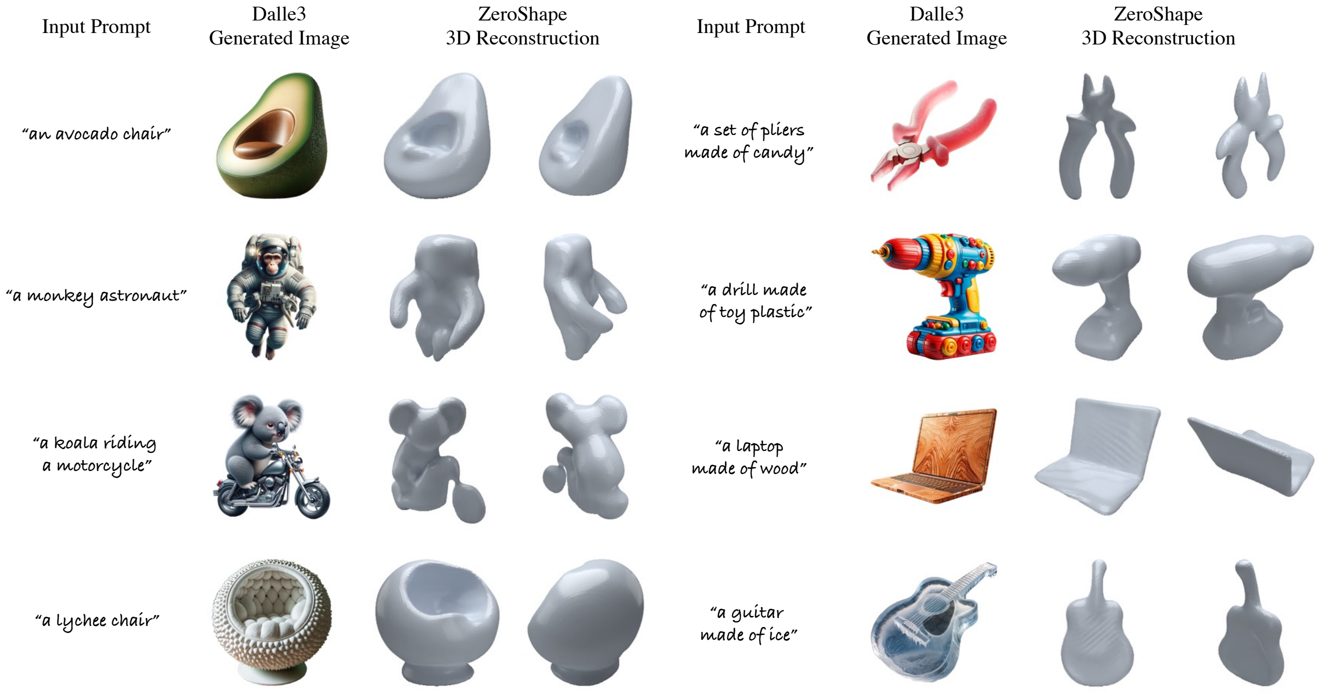
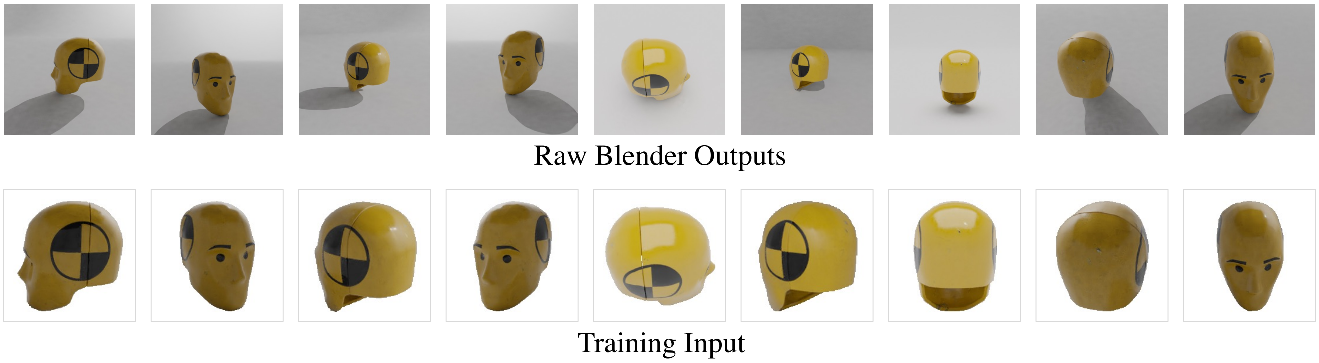


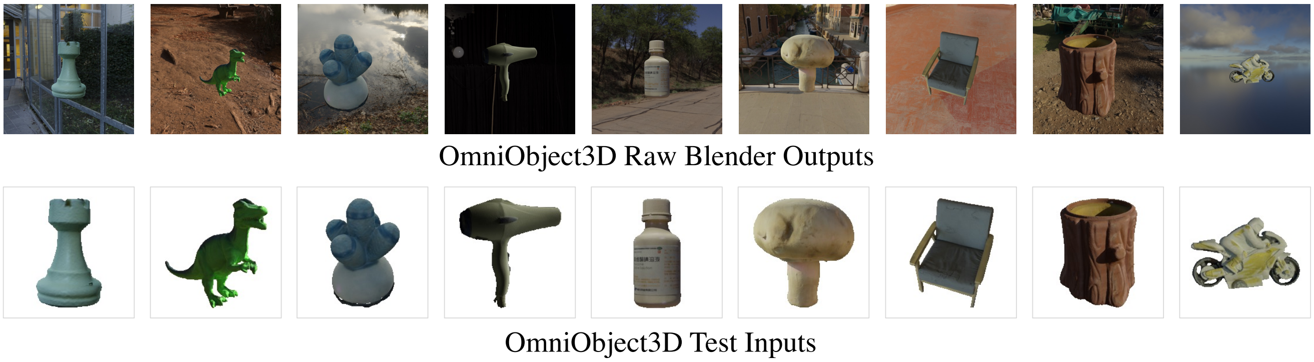
Figure 10. Qualitative results on images generated with DALL·E 3. These results demonstrate the zero-shot generalization ability of ZeroShape to complex novel images.



Raw Blender Outputs

Training Input

Figure 11. **Synthetic Training Data Generation.** We render training images with varying lighting, camera intrinsics and extrinsics. The images are center-cropped, foreground-segmented and resized before being used as training input.



OmniObject3D Raw Blender Outputs

OmniObject3D Test Inputs

Figure 12. **OmniObject3D Testing Data Generation.** For OmniObject3D, we generate realistic testing images with varying lighting, camera intrinsics and extrinsics. To increase rendering realism and diversity, we use diverse HDRI environment maps for scene lighting.