SAM 3D: 3Dfy Anything in Images

Xingyu Chen* Fu-Jen Chu* Pierre Gleize* Kevin J Liang* Alexander Sax* Hao Tang* Weiyao Wang*
Michelle Guo Thibaut Hardin Xiang Li° Aohan Lin Jiawei Liu Ziqi Ma°
Anushka Sagar Bowen Song° Xiaodong Wang Jianing Yang° Bowen Zhang°
Piotr Dollár† Georgia Gkioxari† Matt Feiszli†§ Jitendra Malik†§

*Core Contributor (Alphabetical, Equal Contribution) °Intern †Project Lead §Equal Contribution
Meta Superintelligence Labs

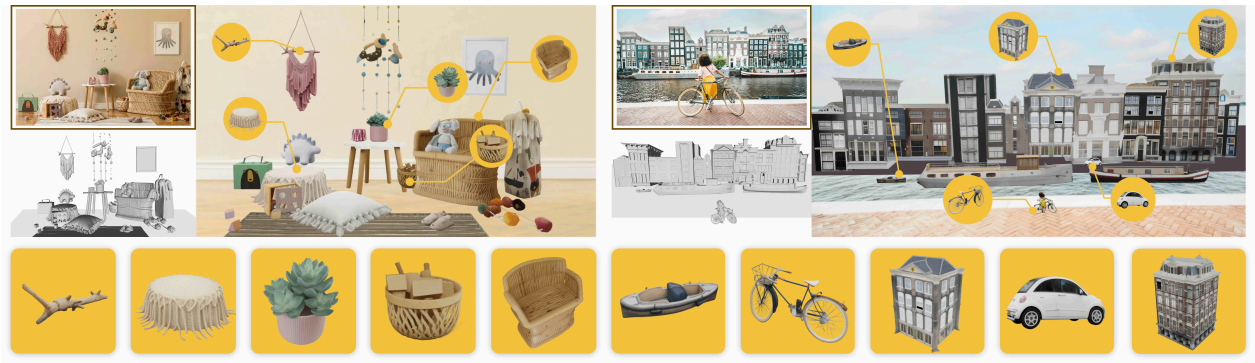


Figure 1. **SAM 3D converts a single image into a composable 3D scene made of individual objects**: Our method predicts per-object geometry, texture, and layout, enabling full scene reconstruction. Bottom: high-quality 3D assets recovered for each object.

Abstract

We present SAM 3D, a generative model for visually grounded 3D object reconstruction, predicting geometry, texture, and layout from a single image. SAM 3D excels in natural images, where occlusion and scene clutter are common and visual recognition cues from context play a larger role. We achieve this with a human- and model-in-the-loop pipeline for annotating object shape, texture, and pose, providing visually grounded 3D reconstruction data at unprecedented scale. We learn from this data in a modern, multi-stage training framework that combines synthetic pretraining with real-world alignment, breaking the 3D "data barrier". We obtain significant gains over recent work, with at least a $5:1$ *win rate in human preference tests on real-world objects and scenes. We release our code, model weights, an online demo, and a new challenging benchmark for in-the-wild 3D object reconstruction at* ***https://ai.meta.com/sam3d*** .

1. Introduction

In this paper (see Fig. 1) we present SAM 3D, a generative neural network for 3D reconstruction from a single image. The model can reconstruct 3D shape and texture for any object, as well as its layout with respect to the camera, even in complex scenes with significant clutter and occlusion. As the reconstruction is of full 3D shape, not just of the visible 2.5D surface, one can then re-render the object from any desired viewpoint.

Computer vision has traditionally focused on multi-view geometry as providing the primary signal for 3D shape. However psychologists (and artists before them) have long known that *humans* can perceive depth and shape from a single image, *e.g.* Koenderink et al. [45] demonstrated this elegantly by showing that humans can estimate surface normals at probe points on an object's image, which can then be integrated to a full surface. In psychology textbooks these single image cues to 3D shape are called "pictorial cues", and include information such as in shading and texture patterns, but also recognition - the "familiar object" cue. In

computer vision, this line of research dates back to Roberts [80], who showed that once an image pattern was recognized as a known object, its 3D shape and pose could be recovered. The central insight is that recognition enables 3D reconstruction, an idea that has since resurfaced in different technical instantiations [9, 15, 30, 41, 108, 112]. Note that this permits generalization to novel objects, because even if a specific object has not been seen before, it is made up of parts seen before.

A fundamental challenge for learning such models is the lack of data: specifically, natural images paired with 3D ground truth are difficult to obtain at scale. Recent work [112, 120] has shown strong reconstruction from single images. However, these models are trained on isolated objects and struggle with objects in natural scenes, where they may be distant or heavily occluded. To add such images to the training set, we need to find a way to associate specific objects in such images with 3D shape models, acknowledging that generalist human annotators find it hard to do so (unlike, say, attaching a label like "cat" or marking its boundary). Two insights made this possible:

- We can create synthetic scenes where 3D object models are rendered and pasted into images (inspired by Dosovitskiy et al. [21]).
- While humans can't easily *generate* 3D shape models for objects, they can *select* the likely best 3D model from a set of proffered choices and align its pose to the image (or declare that none of the choices is good).

We design a training pipeline and data engine by adapting modern, multistage training recipes pioneered by LLMs [65, 66]. As in recent works, we first train on a large collection of rendered synthetic objects. This is supervised pretraining: our model learns a rich vocabulary for object shape and texture, preparing it for real-world reconstruction. Next is mid-training with semi-synthetic data produced by pasting rendered models into natural images. Finally, post-training adapts the model to real images, using both a novel model-in-the-loop (MITL) pipeline and human 3D artists, and aligns it to human preference. We find that synthetic pretraining generalizes, given adequate post-training on natural images.

Our post-training data, obtained from our MITL data pipeline, is key to obtaining good performance in natural images. Generalist human annotators aren't capable of producing 3D shape ground truth; hence our annotators select and align 3D models to objects in images from the output of modules – computational and retrieval-based – that produce multiple initial 3D shape proposals. Human annotators select from these proposals, or route them to human artists for a subset of hard instances. The vetted annotations feed back into model training, and the improved model is reintegrated into the data engine to further boost annotation quality. This virtuous cycle steadily improves the quality of 3D annotations, labeling rates, and model performance.

Due to the lack of prior benchmarks for real-world 3D reconstruction of object shape and layout, we propose a new evaluation set of $1,000$ image and 3D pairs: SAM 3D Artist Objects (SA-3DAO). The objects in our benchmark range from churches, ski lifts, and large structures to animals, everyday household items, and rare objects, and are paired with the real-world images in which they naturally appear. Professional 3D artists create 3D shapes from the input image, representing an expert human upper bound for visually grounded 3D reconstruction. We hope that contributing such an evaluation benchmark helps accelerate subsequent research iteration of real-world 3D reconstruction models.
We summarize our contributions as follows:

- We introduce **SAM 3D**, a new foundation model for 3D that predicts object shape, texture, and pose from a single image. By releasing code, model weights, and a demo, we hope to stimulate further advancements in 3D reconstruction and downstream applications of 3D.
- We build a MITL pipeline for annotating shape, texture, and pose data, providing visually grounded 3D reconstruction data at unprecedented scale.
- We exploit this data via LLM-style pretraining and post-training in a novel framework for 3D reconstruction, combining synthetic pretraining with real-world alignment to overcome the orders of magnitude data gap between 3D and domains such as text, images, or video.
- We release a challenging benchmark for real-world 3D object reconstruction, SA-3DAO. Experiments show SAM 3D's significant gains via metrics and large-scale human preference.

2. The SAM 3D Model

2.1. Problem Formulation

The act of taking a photograph maps a 3D object to a set of 2D pixels, specified by a mask M in an image I. We seek to invert this map. Let the object have shape S, texture T, and rotation, translation and scale (R, t, s) in camera coordinates. Since the 3D to 2D map is lossy, we model the reconstruction problem as a conditional distribution $p(S, T, R, t, s|I, M)$. Our goal is to train a generative model $q(S, T, R, t, s|I, M)$ that approximates p as closely as possible.

2.2. Architecture

We build upon recent SOTA two-stage latent flow matching architectures [112]. SAM 3D first jointly predicts object pose and coarse shape, then refines the shapes by integrating pictorial cues (see Fig. 2). Unlike Xiang et al. [112] that reconstructs isolated objects, SAM 3D predicts object layout, creating coherent multi-object scenes.

Input encoding. We use DINOv2 [70] as an encoder to extract features from two pairs of images, resulting in 4 sets of conditioning tokens:

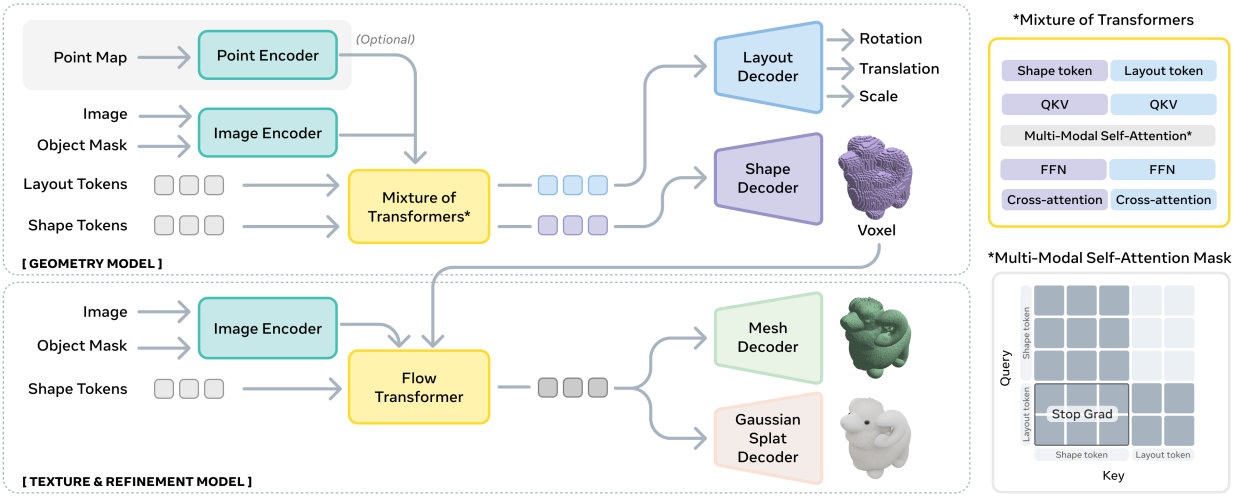


Figure 2. **SAM 3D architecture.** (**top**) SAM 3D first predicts coarse shape and layout with the Geometry model; (**right**) the mixture of transformers architecture apply a two-stream approach with information sharing in the multi-modal self-attention layer. (**bottom**) The voxels predicted by the Geometry model are passed to the Texture & Refinement model, which adds higher resolution detail and textures.

- **Cropped object**: We encode the cropped image I by mask M and its corresponding *cropped binary mask*, providing a focused, high-resolution view of the object.
- **Full image**: We encode the full image I and its *full image binary mask*, providing global scene context and recognition cues absent from the cropped view.

Optionally, the model supports conditioning on a coarse scene point map, P obtained via hardware sensors (*e.g.*, LiDAR on an iPhone), or monocular depth estimation [100, 119], enabling SAM 3D to integrate with other pipelines.

The Geometry Model models the conditional distribution $p(O, R, t, s|I, M)$, where $O \in \mathbb{R}^{64^3}$ is coarse shape, $R \in \mathbb{R}^6$ the 6D rotation [127], $t \in \mathbb{R}^3$ the translation, and $s \in \mathbb{R}^3$ the scale. Conditioned on the input image and mask encodings, we employ a 1.2B parameter flow transformer with the Mixture-of-Transformers (MoT) architecture [18, 53], modeling geometry O and layout (R, t, s) using the attention mask in Fig. 2. See Sec. D.1 for details.

The Texture & Refinement Model learns the conditional distribution $p(S, T|I, M, O)$. We first extract active voxels from the coarse shape O predicted by Geometry model. A 600M parameter sparse latent flow transformer [73, 112] refines geometric details and synthesizes object texture.

3D Decoders. The latent representations from the Texture & Refinement Model can be decoded to either mesh or 3D Gaussian splats via a pair of VAE decoders $\mathcal{D}_m$, $\mathcal{D}_g$. These separately-trained decoders share the same VAE encoder and hence the same structured latent space [112]. We also detail several improvements in Sec. D.6.

3. Training SAM 3D

SAM 3D breaks the 3D data barrier using a recipe that progresses from synthetic pretraining to natural post-training, adapting the playbook from LLMs, robotics, and other large generative models. We build capabilities by stacking different training strategies in pre- and mid-training, and then align the model to real data and human-preferred behaviors through a post-training data flywheel. SAM 3D uses the following approach:

Step 1: Pretraining. This phase builds foundational capabilities, such as shape generation, into a base model.

Step 1.5: Mid-Training. Sometimes called continued pretraining, mid-training imparts general skills such as occlusion robustness, mask-following, and using visual cues.

Step 2: Post-Training. Post-training elicits target behavior, such as adapting the model from synthetic to real-world data or following human aesthetic preferences. We collect training samples $(I, M) \rightarrow (S, T, R, t, s)$ and preference data from humans and use them in both supervised finetuning (SFT) and direct preference optimization (DPO) [75].

This alignment (step 2) can be repeated, first collecting data with the current model and then improving the model with the new data. This creates a virtuous cycle with humans providing the supervision. Fig. 9b shows that as we run the data engine longer, model performance steadily improves; dataset generation emerges as a byproduct of this alignment.

The following sections detail the training objectives and data sources used in SAM 3D. We focus on the Geometry model; Texture & Refinement is trained similarly (details in Sec. D.5). Training hyper-parameters are in Sec. D.7.

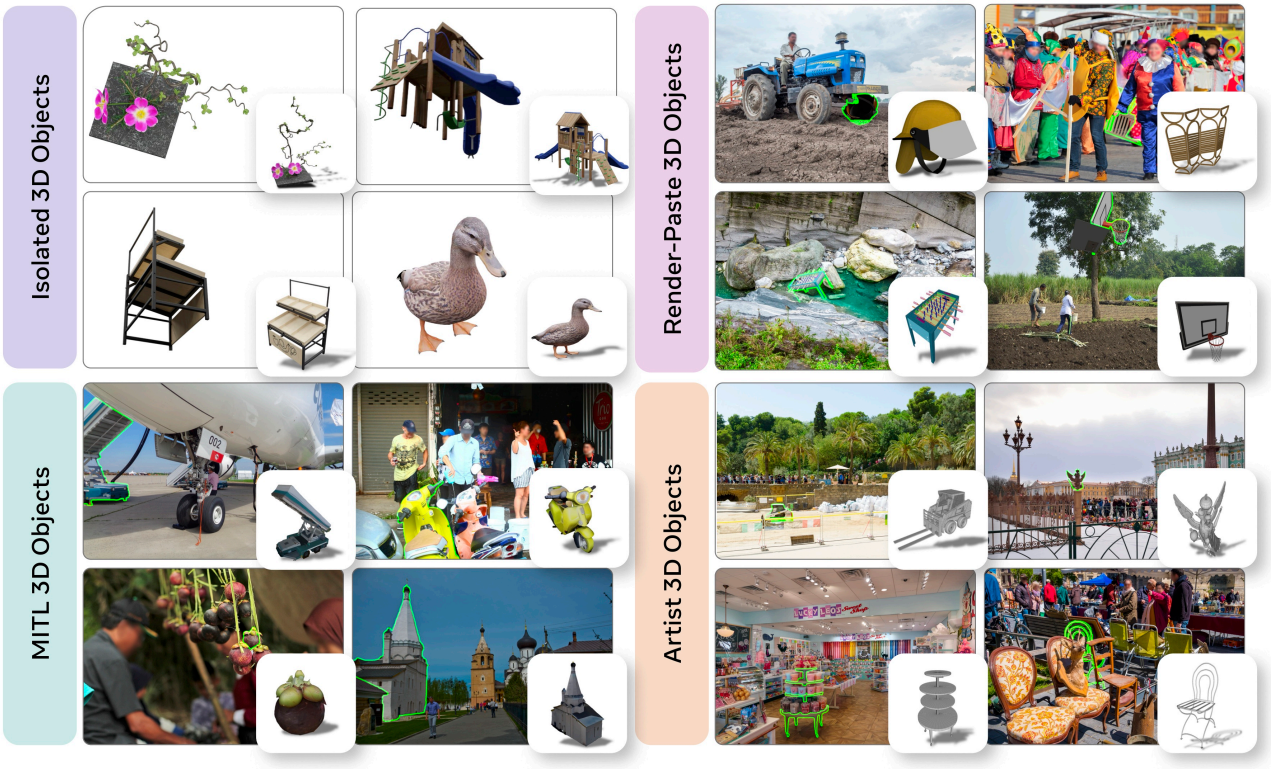


Figure 3. **SAM 3D data,** with a green outline around the target object, and the ground truth mesh shown in the bottom right. Samples are divided into four rows, based on type. Art-3DO meshes are untextured, while the rest may be textured or not, depending on the underlying asset (Iso-3DO, RP-3DO) or if the mesh was annotated for texture (MITL-3DO).

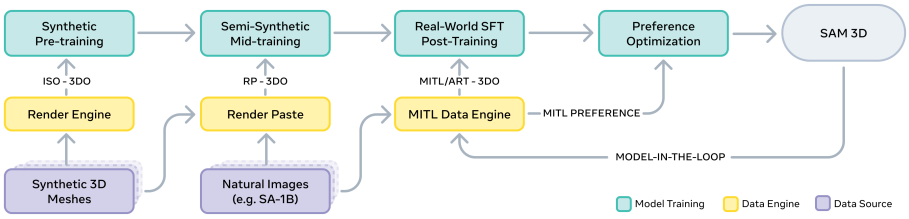


Figure 4. **SAM 3D training paradigm.** We employ a multi-stage pipeline incrementally exposing the model to increasingly complex data and modalities.

Training stage	Modalities	Datasets	Condition input
Stage 1 Geometry model			
Pre-training	S, R	Iso-3DO	object-centric crop
Mid-training	S, R	RP-3DO†	full image
	S, R, t, s	ProcThor, RP-3DO‡	full image, pointmap*
SFT	S, R, t, s	MITL, SA-3DAO	full image, pointmap*
Alignment	S	MITL preference	full image, pointmap*
Stage 2 Texture & refinement model			
Pre-training	T	Trellis-500K [112]	object-centric crop
Mid-training	T	RP-3DO†,§	full image
SFT	T	MITL	full image
Alignment	T	MITL preference	full image

* optional.
† Flying occlusion (FO), ‡ Object Swap-Random (OS-R), § Object Swap-Annotated (OS-A) from RP-3DO. See Sec. C.2 for details.

Table 1. **SAM 3D training stages.**

3.1. Pre & Mid-Training: Building a Base Model

Training begins with synthetic pretraining and mid-training, leveraging available large-scale datasets to learn strong priors for shape and texture, and skills such as mask-following, occlusion handling, and pose estimation. The rich features learned here drastically reduce the number of labeled real-world samples required in post-training [37], which generally incur acquisition costs. In pre- and mid-training, models are trained with rectified conditional flow matching [57] to generate multiple 3D modalities (see Sec. D.2).

3.1.1. Pretraining: Single Isolated 3D Assets

Pretraining trains the model to reconstruct accurate 3D shapes and textures from renders of isolated synthetic objects, following the successful recipes from [109, 112, 120]. Specifically, we gather a set of image I, shape S, and texture T triplets, using 2.7 million object meshes from Objaverse-XL [17] and licensed datasets, and render them from 24 viewpoints, each producing a high-resolution image of a single centered object; more detail in Sec. C.1. We call this dataset *Iso-3DO* and train for 2.5 trillion training tokens.

3.1.2. Mid-Training: Semi-Synthetic Capabilities

Next, mid-training builds up foundational skills that will enable the model to handle objects in real-world images:

- *Mask-following*: We train the model to reconstruct a target object, defined by a binary mask on the input image.
- *Occlusion robustness*: The artificial occluders in our

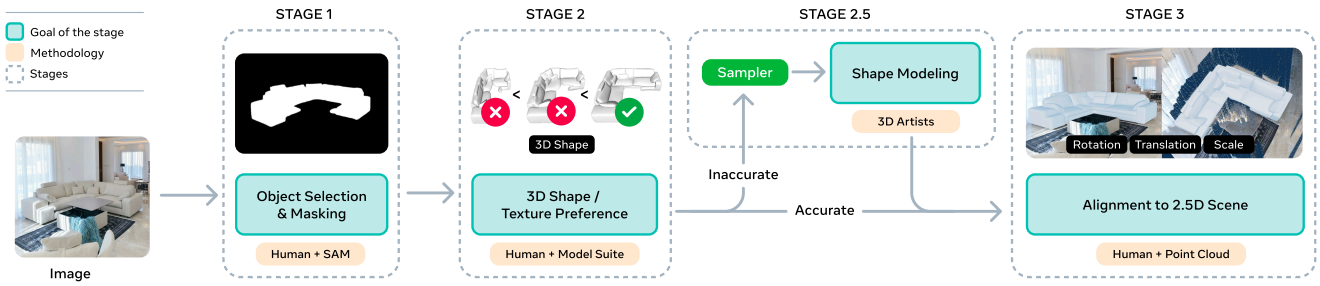


Figure 5. **Life of an example going through the data collection pipeline.** We streamline annotation by breaking it into subtasks: annotators first choose target objects (Stage 1); rank and select 3D model candidates (Stage 2); then pose these models within a 2.5D scene (Stage 3). Stages 2 and 3 use model-in-the-loop.

dataset incentivize learning shape completion.

- *Layout estimation*: We train the model to produce translation and scale in normalized camera coordinates.

We construct our data by rendering textured meshes into natural images using alpha compositing. This "render-paste" dataset contains one subset of occluder-occludee pairs, and another subset with real objects replaced by synthetic objects at similar location and scale, creating physically-plausible data with accurate 3D ground truth. We call these 61 million samples with 2.8 million unique meshes *RP-3DO*; Fig. 3 shows examples. See Sec. C.2 for more details.

After mid-training (2.7 trillion training tokens), the model has now been trained with all input and output modalities for visually grounded 3D reconstruction. However, all data used has been (semi-)synthetic; to both close the domain gap and fully leverage real-world cues, we need real images.

3.2. Post-Training: Real-World Alignment Loop

In post-training, we have two goals. The first is to close the domain gap between (semi-)synthetic data and natural images. The second is to align with human preference for shape quality. We adapt the model by using our data engine iteratively; we first **(i) collect training data** with the current model, and then **(ii) update our model** using multi-stage post-training on this collected data. We then repeat.

3.2.1. Post-Training: Collection Step

The core challenge with collecting data for 3D visual grounding is that most people cannot create meshes directly; this requires skilled 3D artists, who even then can take multiple hours. This is different from the segmentation masks collected in SAM [44]. However, given options, most people *can* choose which mesh resembles an object in the image most accurately. This fact forms the foundation of our data collection for SAM 3D. We convert preferences into training data as follows: sample from our post-trained model, ask annotators to choose the best candidate and then grade its overall quality according to a rubric which we define and update. If the quality meets the (evolving) bar, the candidate becomes a training sample.

Unfortunately at the first iteration, our initial model yields few high-quality candidates. This is because before the first collection step, very little real-world data for 3D visual grounding exists. We deal with this cold start problem by leveraging a suite of existing learned and retrieval-based models to produce candidates. In early stages, we draw mostly from the ensemble, but as training progresses our best model dominates, eventually producing about 80% of the annotated data seen by SAM 3D.

Our annotation pipeline collects 3D object shape S, texture T, orientation R, 3D location t, and scale s from real-world images. We streamline the process by dividing into subtasks and leveraging existing appropriate models and human annotators within each (see Fig. 5): identifying target objects, 3D model ranking and selection, and posing these within a 3D scene (relative to a point map). We outline each stage of the data engine below and present details in Sec. B. In total, we annotate almost 1 million images with $\sim$ 3.14 million untextured meshes and $\sim$ 100K textured meshes–unprecedented scale for 3D data paired with natural images.

Stage 1: Choosing target objects (I, M). The goal of this stage is to identify a large, diverse set of images I and object masks M to lift to 3D. To ensure generalization across objects and scenes, we sample images from several diverse real-world datasets, and utilize a 3D-oriented taxonomy to balance the object distribution. To obtain object segmentation masks, we use a combination of pre-existing annotations [44] and human labelers selecting objects of interest.

Stage 2: Object model ranking and selection (S, T). The goal of this stage is to collect image-grounded 3D shape S and texture T. As described above, human annotators choose shape and texture candidates which best match the input image and mask. Annotators rate the example r and reject chosen examples that do not meet a predefined quality threshold, *i.e.* $r < \alpha$. Bad candidates also become negative examples for preference alignment.

Our data engine maximizes the likelihood of a successful annotation, $r > \alpha$, by asking annotators to choose between $N = 8$ candidates from the ensemble; a form of best-of-

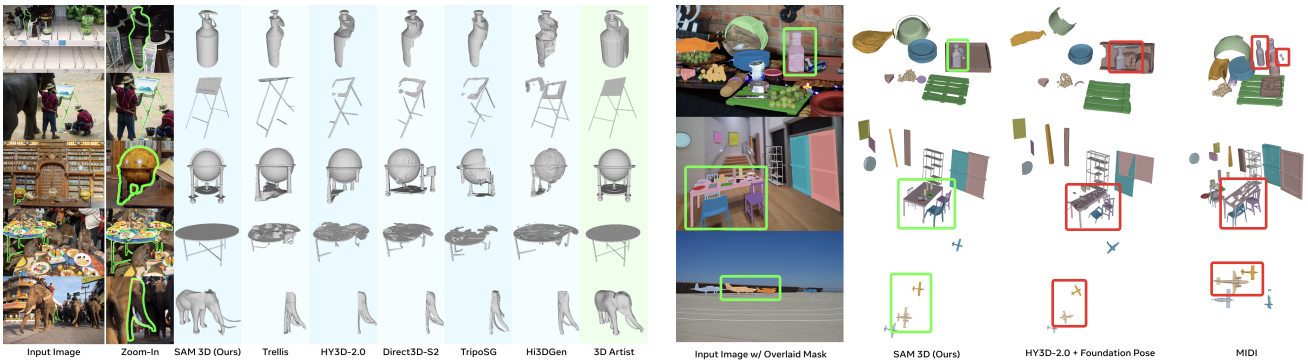


Figure 6. **Qualitative comparison to competing image-to-3D methods.** (**left**) We compare to the recent Trellis [112], Hunyuan3D-2.1 [39], Direct3D-S2 [110] and Hi3DGen [122] on the artist-generated SA-3DAO for single shape reconstruction; we provide the 3D artist-created ground truth mesh as reference. (**right**) We show SAM 3D's full 3D scene reconstructions versus alternatives [38, 103].

	SA-3DAO				ISO3D Eval Set	
Model	F1@0.01 (↑)	vIoU (↑)	Chamfer (↓)	EMD (↓)	ULIP (↑)	Uni3D (↑)
Trellis	0.1475	0.1392	0.0902	0.2131	0.1473	0.3698
HY3D-2.1	0.1399	0.1266	0.1126	0.2432	0.1293	0.3546
HY3D-2.0	0.1574	0.1504	0.0866	0.2049	0.1484	0.3662
Direct3D-S2	0.1513	0.1465	0.0962	0.2160	0.1405	0.3653
TripoSG	0.1533	0.1445	0.0844	0.2057	**0.1529**	0.3687
Hi3DGen	0.1629	0.1531	0.0937	0.2134	0.1419	0.3594
SAM 3D	**0.2344**	**0.2311**	**0.0400**	**0.1211**	0.1488	**0.3707**

Table 2. **3D shape quantitative comparison** to competing image-to-3D methods, including Trellis [112], HY3D-2.1 [39], HY3D-2.0 [92], Direct3D-S2 [110], TripoSG [52], Hi3DGen [122]. SA-3DAO shows metrics that measure accuracy against GT geometry; ISO3D [22] has no geometric GT and so we show perceptual similarities between 3D and input images (ULIP [118] and Uni3D [126]). TripoSG uses a significantly higher mesh resolution, which is rewarded in perceptual metrics.

N search [71] using humans. The expected quality of this best candidate improves with N, and we further increase N by first filtering using a model, and then filtering using the human [2]; we show results in Sec. B.7.

Stage 2.5: Hard example triage (Artists). When no model produces a reasonable object shape, our non-specialist annotators cannot correct the meshes, resulting in a lack of data precisely where the model needs it most. We route a small percentage of these hardest cases to professional 3D artists for direct annotation, and we denote this set *Art-3DO*.

Stage 3: Aligning objects to 2.5D scene (R, t, s). The previous stages produce a 3D shape for the object, but not its layout in the scene. For each Stage 2 shape, annotators label the object pose by manipulating the 3D object's translation, rotation, and scale relative to a point cloud. We find that point clouds provide enough structure to enable consistent shape placement and orientation.

In general, we can think of the data collection as an API that takes a current best model, $q(S, T, R, t, s \mid I, M)$, and returns (i) training samples $D^+ = (I, M, S, T, R, t, s)$, (ii) a quality rating $r \in [0, 1]$, and (iii) a set of less preferred candidates ($D^- = (I, M, S', T', R', t', s')$) that are all worse than the training sample.

3.2.2. Post-Training: Model Improvement Step

The model improvement step in SAM 3D uses these training samples and preference results to update the base model through multiple stages of finetuning and preference alignment. Within each post-training iteration we aggregate data from all previous collection steps; keeping only samples where D^+ is above some quality threshold α. As training progresses, α can increase over time, similar to the cross-entropy method for optimization [14]. Our final post-training iteration uses 0.5 trillion training tokens.

Supervised Fine-Tuning (SFT). When post-training begins, the base model has only seen synthetic data. Due to the large domain gap between synthetic and real-world data, we begin by finetuning on our aligned meshes from stage 3.

We begin SFT with the noisier non-expert labels (MITL-3DO), followed by the smaller, high-quality set from 3D artists (Art-3DO). The high quality Art-3DO data enhances model quality by aligning outputs with artists' aesthetic preferences. We find this helps suppress common failures, *e.g.* floaters, bottomless meshes, and missing symmetry.

Preference optimization (alignment). After fine-tuning, the model can robustly generate shape and layout for diverse objects and real-world images. However, humans are sensitive to properties like symmetry, closure, etc. which are difficult to capture with generic objectives like flow matching. Thus, we follow SFT with a stage of direct preference optimization (DPO) [75], using $D+/D^-$ pairs from Stage 2 of our data engine. We found this off-policy data was effective at eliminating undesirable model outputs, even after SFT on Art-3DO. DPO training details are in Sec. D.3.

Distillation. Finally, to enable sub-second shape and layout from the Geometry model, we finish a short distillation stage to reduce the number of function evaluations (NFE) required during inference from $25 \rightarrow 4$. We adapt Frans et al. [25] to

Generation	Model	SA-3DAO				Aria Digital Twin			
		3D IoU (↑)	ICP-Rot. (↓)	ADD-S (↓)	ADD-S @ 0.1 (↑)	3D IoU (↑)	ICP-Rot. (↓)	ADD-S (↓)	ADD-S @ 0.1 (↑)
Pipeline	Trellis + Megapose	0.2449	39.3866	0.5391	0.2831	0.2531	33.6114	0.4358	0.1971
Pipeline	HY3D-2.0 + Megapose	0.2518	33.8307	0.7146	0.3647	0.3794	29.0066	0.1457	0.4211
Pipeline	HY3D-2.0 + FoundationPose	0.2937	32.9444	0.3705	0.5396	0.3864	25.1435	0.1026	0.5992
Pipeline	HY3D-2.1 + FoundationPose	0.2395	39.8357	0.4186	0.4177	0.2795	33.1197	0.2135	0.4129
Pipeline	SAM 3D + FoundationPose	0.2837	32.9168	0.3848	0.5079	0.3661	18.9102	0.0930	0.6495
Joint	MIDI	-	-	-	-	0.0336	44.2353	2.5278	0.0175
Joint	**SAM 3D**	**0.4254**	**20.7667**	**0.2661**	**0.7232**	**0.4970**	**15.2515**	**0.0765**	**0.7673**

Table 3. **3D layout quantitative comparison** to competing layout prediction methods on SA-3DAO and Aria Digital Twin [72]. SAM 3D significantly outperforms both *pipeline* approaches used in robotics [47, 103] and *joint* generative models (MIDI [38]). Most SA-3DAO scenes only contain one object so we do not show MIDI results that require multi-object alignment. The metrics measure bounding box overlap, rotation error, and chamfer-like distances normalized by object diameter.

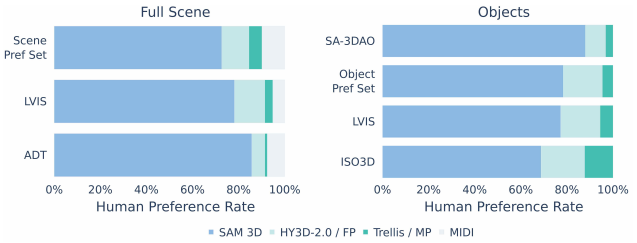


Figure 7. **Preference comparison on scene-level and object-level reconstruction.** Numbers indicate human preference rates. Objects comparisons are done on textured meshes. SAM 3D is significantly preferred over others on all fronts.

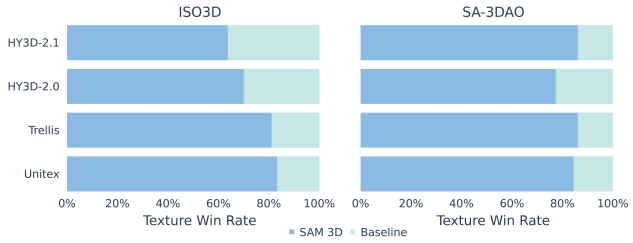


Figure 8. **Preference comparison on texture**. Since SAM 3D provides higher quality shape, we use the geometry results from SAM 3D and only perform texture generations for all methods. SAM 3D significantly outperforms others.

our setting, and describe the details in Sec. D.4.

4. Experiments

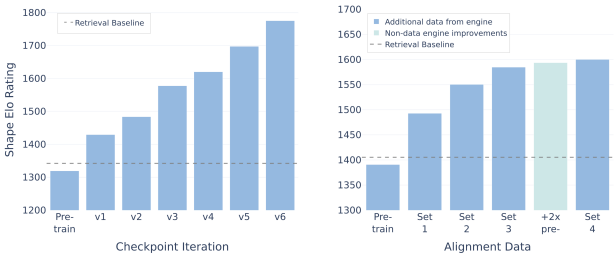
Dataset. To comprehensively evaluate the model capability under real-world scenarios, we carefully build a new benchmark **SA-3DAO**, consisting of 1K 3D artist-created meshes created from natural images. We also include **ISO3D** from 3D Arena [22] for quantitatively evaluating shape and texture, and Aria Digital Twin (**ADT**) [72] for layout. We further conduct human preference evaluation on two curated sets for both scene-level and object-level reconstruction. The **Pref Set** uses real-world images from MetaCLIP [115] and SA-1B [44], as well as a set based on LVIS [35]. Refer to Sec. E for details on evaluation sets.

Settings. We conduct experiments with a Geometry model trained to condition on pointmaps. When pointmaps are unavailable, we estimate them with Wang et al. [100]. We found that shape and texture quality do not depend on whether the model is trained with pointmap conditioning (see Sec. F.5), but layout (translation/scale) evaluation in Tab. 3 requires ground-truth depth/pointmap as reference.

4.1. Comparison with SOTA

3D shape and texture. We evaluate single-object generation by comparing SAM 3D with prior state-of-the-art (SOTA) methods. In human preference studies, SAM 3D achieves an $5:1$ head-to-head win rate on real images (see Fig. 7). Tab. 2 presents quantitative evaluation on shape quality, where SAM 3D matches or exceeds previous SOTA performance on isolated object images (**ISO3D**), and significantly outperforms all baselines on challenging real-world inputs (**SA-3DAO**). Qualitative examples in Fig. 6 further illustrate the model's strong generalization under heavy occlusion. In Fig. 8, we compare SAM 3D texture vs. other texture models, given SAM 3D shapes (SAM 3D's improved shape actually benefits other methods in this eval). Annotators significantly prefer SAM 3D texture (details in Sec. F.2).

3D scene reconstruction. In preference tests on three evaluation sets, users prefer scene reconstructions from SAM 3D by $6:1$ over prior SOTA (Fig. 7). Fig. 6 and Fig. 21 in the appendix shows qualitative comparisons, while Tab. 3 shows quantitative metrics for object layout. On real-world data like **SA-3DAO** and **ADT**, the improvement is significant and persists even when *pipeline* approaches use SAM 3D meshes. SAM 3D introduces a new real-world capability to generate shape and layout *jointly* (ADD-S @ 0.1 $2\% \rightarrow 77\%$), and a sample-then-optimize approach, as in the render-and-compare approaches [47, 103] can further improve performance (Sec. F.3). The strong results for layout and scene reconstruction demonstrate that SAM 3D can robustly handle both RGB-only inputs (*e.g.*, **SA-3DAO**, **LVIS**, **Pref Set**) as



(a) Historical Elo in data engine (b) Expanding training data alone

Figure 9. **Data engine with additional iterations.** The plots show Elo scores of different models; a 400 point Elo difference corresponds to 10:1 odds in a preference test. Models were **(a)** checkpoints around 3 weeks apart, indicating cumulative improvements as we scale and add different stages and **(b)** post-trained (SFT) using expanded training data.

well as provided pointmaps (*e.g.*, **ADT**). Moreover, SAM 3D is modular w.r.t. the depth estimator: annotators preferred outputs using an improved estimator not seen during training

4.2. Analysis Studies

Post-training iterations steadily improve performance. We observed steady improvements as we ran the data engine for longer, with near-linear Elo scaling shown in the historical comparisons from Stage 2 of our data engine (Fig. 9a). We found it important to scale all stages simultaneously. The cumulatively linear effect results from more data engine iterations, along with scaling up pretraining, mid-training, and adding new post-training stages. Fig. 9b shows that iterating MITL-3DO data alone yields consistent improvements but with decreasing marginal impact.

Multi-stage training improves performance. SAM 3D's real-world performance emerges through multi-stage training. Tab. 4 reveals near-monotonic 3D shape improvements as each training stage is added, validating the approach that leads to the final model (last row). In the appendix, Fig. 16 shows similar results for texture and Tab. 7 shows the contribution of each individual real-world data stage, by knocking out the MITL-3DO, Art-3DO data, or DPO stages.

Additional ablations. Please see the appendix for additional ablations on rotation representation (Sec. F.4), DPO (Sec. D.3), distillation (Sec. D.4), pointmaps (Sec. F.5), and scaling best-of-N in the data engine (Sec. B.7).

5. Related Work

3D reconstruction from images is a longstanding challenge of computer vision, spanning multi-view methods [36, 64], single-view approaches using direct 3D supervision or generative models [92, 112, 125], and layout estimation that extends to full scenes [6, 38]. Recent generative methods show strong results on isolated synthetic objects [11, 17],

	SA-3DAO				Preference set
Training Stage	F1 @ 0.01 (↑)	vIoU (↑)	Chamfer (↓)	EMD (↓)	Texture WR (↑)
Pre-training (Iso-3DO)	0.1349	0.1202	0.1036	0.2396	-
+ Mid-training (RP-3DO)	0.1705	0.1683	0.0760	0.1821	60.7
+ SFT (MITL-3DO)	0.2027	0.2025	0.0578	0.1510	66.9
+ DPO (MITL-3DO)	0.2156	0.2156	0.0498	0.1367	66.4
+ SFT (Art-3DO)	0.2331	**0.2337**	0.0445	0.1257	-
+ DPO (Art-3DO)	**0.2344**	0.2311	**0.0400**	**0.1211**	-

Table 4. **Cumulative improvements from multi-stage training on 3D shape and texture**. For texture, we report win rates (WR) between each row and the row *above* it.

but training such models requires large-scale 3D data, which remains scarce for real-world images; SAM 3D addresses this through multi-stage training and a post-training data engine that provides diverse real-world 3D supervision. We provide a full discussion of related work in Sec. A.

6. Conclusion

We share SAM 3D: a new foundation model for full reconstruction of 3D shape, texture, and layout of objects from natural images. SAM 3D's robustness on in-the-wild images is made possible by an innovative post-training data engine combining model-in-the-loop proposals with human annotation, and a multi-stage training recipe that bridges synthetic pretraining with real-world alignment. In human preference studies, SAM 3D achieves a 5:1 win rate over prior state of the art on real-world objects and 6:1 on scenes. We also contribute SA-3DAO, a new benchmark of 1K artist-created 3D meshes from natural images, to support future research on real-world 3D reconstruction. With the release of our model, code, and benchmark, we expect SAM 3D to unlock new capabilities across diverse applications, such as robotics, AR/VR, gaming, film, and interactive media.

Acknowledgements

We thank the following individuals for their contributions to this work. For their contributions to SAM Playground Engineering we thank: *Robbie Adkins, Rene de la Fuente, Facundo Figueroa, Alex He, Dex Honsa, Alex Lende, Jonny Li, Peter Park, Don Pinkus, Roman Radle, Phillip Thomas*, and *Meng Wang*. We thank our excellent XFN team for leadership and support: *Kris Kitani, Vivian Lee, Sasha Mitts, George Orlin, Nikhila Ravi*, and *Andrew Westbury*. Thanks to *Helen Klein, Mallika Malhotra*, and *Azita Shokrpour* for support with Legal, Privacy, and Integrity. We thank *Michelle Chan, Kei Koyama, William Ngan, Yael Yungster* for all the design support throughout the project. Thanks to *Arpit Kalla* for work on model efficiency. We thank *Faye Ma* and *Kehan Lyu* for data engineering support and tooling, and *Emmanuel Hernandez, Robert Kuo* for pipeline development. We thank *Nan Yang* for support with egocentric video data efforts. Thanks to our two interns *Cem Gokmen, Jasmine Shone* for their work on 3D and *Lea Wilken* for feedback on the manuscript. Thanks to our fantastic data operations team: *Paris Baptiste, Karen Bergan, Kai Brown, Ida Cheng, Khadijat Durojaiye, Patrick Edwards, Daniella Factor, Eva Galper, Leonna Jones, Zayida Suber, Tatum Turner, Joseph Walker*, and *Claudette Ward*.

References

[1] Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. 2

[2] Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 5366–5376, Red Hook, NY, USA, 2017. Curran Associates Inc. 6, 1, 4

[3] Andreea Ardelean, Mert Özer, and Bernhard Egger. Generalizable 3d scene reconstruction via divide and conquer from a single view. In *International Conference on 3D Vision (3DV)*, 2025. 1

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery. 2

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, 2014. 2

[6] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13154–13164, 2023. 8, 1

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[8] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026. 15

[9] Thomas J Cashman and Andrew W Fitzgibbon. What shape are dolphins? building 3d morphable models from 2d images. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):232–244, 2012. 2

[10] Jose A Castellanos, José MM Montiel, José Neira, and Juan D Tardós. The SPmap: A probabilistic framework for simultaneous localization and map building. *IEEE Transactions on robotics and Automation*, 1999. 1

[11] 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. 8, 1, 11

[12] Tianyuan Dai, Josiah Wong, Yunfan Jiang, Chen Wang, Cem Gokmen, Ruohan Zhang, Jiajun Wu, and Li Fei-Fei. Automated creation of digital cousins for robust policy learning. *arXiv preprint arXiv:2410.07408*, 2024. 1

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020. 15

[14] Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005. 6

[15] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 465–474. 2023. 2

[16] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. In *NeurIPS*, 2022. Outstanding Paper Award. 1

[17] 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. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023. 4, 8, 1, 11

[18] Yuxuan Deng, Yujia Zhu, Jiahui Chen, Yuan Wang, Yifei Li, Haotian Li, Junnan Li, Jinsheng Zhang, Wenhui Liu, Yuzheng Zhang, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 3

[19] Kristin Diehl and Cait Poynor. Great expectations?! assortment size, expectations, and satisfaction. *Journal of Marketing Research*, 47(2):312–322, 2010. 5

[20] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun SHUM, and Tong Zhang. RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*, 2023. 1, 6

[21] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 2, 7

[22] Dylan Ebert. 3d arena: An open platform for generative 3d evaluation. *arXiv preprint arXiv:2506.18787*, 2025. 6, 7, 12, 13

[23] Carlos Hernández Esteban and Francis Schmitt. Silhouette and stereo fusion for 3D object modeling. *Computer Vision and Image Understanding*, 2004. 1

[24] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3D object reconstruction from a single image. In *CVPR*, 2017. 1

[25] Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. *arXiv preprint arXiv:2410.12557*, 2024. 6, 9

[26] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia,

Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 1

[27] Zheng Geng, Nan Wang, Shaocong Xu, Chongjie Ye, Bohan Li, Zhaoxi Chen, Sida Peng, and Hao Zhao. One view, many worlds: Single-image to 3d object meets generative domain randomization for one-shot 6d pose estimation. *arXiv preprint arXiv:2509.07978*, 2025. 1

[28] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016. 1

[29] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013. 1

[30] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9785–9795, 2019. 2, 1

[31] Aaron Grattafiori et al. The llama 3 herd of models, 2024. 2

[32] Kristen Grauman, Andrew Westbury, et al. Ego4d: Around the world in 3,000 hours of egocentric video. *International Journal of Computer Vision (IJCV)*, 2022. 2

[33] Kristen Grauman et al. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. *arXiv preprint arXiv:2401.10889*, 2024. 2

[34] Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (rest) for language modeling, 2023. 1

[35] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 7, 2, 12, 14

[36] Richard Hartley and Andrew Zisserman. Multiple view geometry in computer vision, 2003. 8, 1

[37] Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021. 4, 1

[38] Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. Midi: Multi-instance diffusion for single image to 3d scene generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23646–23657, 2025. 6, 7, 8, 1

[39] Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, Xin Huang, Sheng Zhang, Zebin He, Di Luo, Haolin Liu, Yunfei Zhao, et al. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. *arXiv preprint arXiv:2506.15442*, 2025. 6, 13

[40] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *SIGGRAPH*, 1984. 1

[41] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1966–1974, 2015. 2

[42] Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X Chang, and Manolis Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16384–16393, 2024. 1, 11

[43] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[44] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 5, 7, 2, 4, 12, 15

[45] Jan J Koenderink, Andrea J Van Doorn, and Astrid ML Kappers. Surface perception in pictures. *Perception & psychophysics*, 52(5):487–496, 1992. 1

[46] Nilesh Kulkarni, Justin Johnson, and David F. Fouhey. What's behind the couch? directed ray distance functions for 3D scene reconstruction. In *ECCV*, 2022. 1

[47] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022. 7, 1, 14

[48] Nathan Lambert. *Reinforcement Learning from Human Feedback*. Online, 2025. 2

[49] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling, 2024. 15

[50] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report, 2023. 2

[51] Yanghao Li, Haoqi Fan, Rohit Girdhar, and Alexander Kirillov. Segment anything in videos. *arXiv preprint arXiv:2305.06500*, 2023. 2

[52] Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025. 6

[53] Weixin Liang, LILI YU, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025. 3

[54] Yixun Liang, Kunming Luo, Xiao Chen, Rui Chen, Hongyu Yan, Weiyu Li, Jiarui Liu, and Ping Tan. Unitex: Universal high fidelity generative texturing for 3d shapes. *arXiv preprint arXiv:2505.23253*, 2025. 13

[55] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 2020. 1

[56] 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. 1

[57] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 4, 8

[58] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024. 2

[59] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *the 18th European Conference on Computer Vision (ECCV)*, 2024. 2

[60] Kevis-Kokitsi Maninis, Stefan Popov, Matthias Nießner, and Vittorio Ferrari. Vid2cad: Cad model alignment using multi-view constraints from videos. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 1

[61] Kevis-Kokitsi Maninis, Stefan Popov, Matthias Nießner, and Vittorio Ferrari. Cad-estate: Large-scale cad model annotation in rgb videos. In *International Conference on Computer Vision*, 2023. 1

[62] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 7

[63] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *CVPR*, 2019. 1

[64] 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 *ECCV*, 2020. 8, 1

[65] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2025. 2

[66] Kaixiang Mo, Yuxin Shi, Weiwei Weng, Zhiqiang Zhou, Shuman Liu, Haibo Zhang, and Anxiang Zeng. Mid-training of large language models: A survey, 2025. 2

[67] NVIDIA, :, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv: 2503.14734*, 2025. 4

[68] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. 2

[69] Abby O'Neill et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024. 4

[70] 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. 2

[71] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv: 2203.02155*, 2022. 6, 1

[72] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. 7, 1, 11, 12, 14

[73] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3

[74] Stefan Popov, Pablo Bauszat, and Vittorio Ferrari. Corenet: Coherent 3d scene reconstruction from a single rgb image. In *European Conference on Computer Vision*, 2020. 1

[75] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a

reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 3, 6, 1

[76] Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21783–21794, 2024. 1

[77] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. 15

[78] 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. 1

[79] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4209–4219, 2024. 1

[80] Lawrence G Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963. 2

[81] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code, 2024. 2

[82] Denys Rozumnyi, Stefan Popov, Kevis-Kokitsi Maninis, Matthias Nießner, and Vittorio Ferrari. Estimating generic 3d room structures from 2d annotations. *arXiv preprint arXiv:2306.09077*, 2023. 1

[83] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002. 1

[84] Wubin Shi, Shaoyan Gai, Feipeng Da, and Zeyu Cai. Sampose: Generalizable model-free 6d object pose estimation via single-view prompt. *IEEE Robotics and Automation Letters*, 2025. 1

[85] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. DeepVoxels: Learning persistent 3D feature embeddings. In *CVPR*, 2019. 1

[86] Randall Smith, Matthew Self, and Peter Cheeseman. Estimating uncertain spatial relationships in robotics. In *Autonomous robot vehicles*, 1990. 1

[87] Julian Straub, Daniel DeTone, Tianwei Shen, Nan Yang, Chris Sweeney, and Richard Newcombe. Efm3d: A benchmark for measuring progress towards 3d egocentric foundation models. In *arXiv preprint arXiv:2406.10224*, 2024. 2

[88] 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, 11

[89] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022. 1

[90] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in neural information processing systems*, 34:251–266, 2021. 1

[91] Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, and Will Dabney. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*, 2024. 1

[92] Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025. 6, 8

[93] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 1992. 1

[94] Lorenzo Torresani, Aaron Hertzmann, and Chris Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *PAMI*, 2008. 1

[95] Michał J Tyszkiewicz, Kevis-Kokitsi Maninis, Stefan Popov, and Vittorio Ferrari. Raytran: 3d pose estimation and shape reconstruction of multiple objects from videos with ray-traced transformers. In *European Conference on Computer Vision*, 2022. 1

[96] Basile Van Hoorick, Purva Tendulkar, Dídac Surís, Dennis Park, Simon Stent, and Carl Vondrick. Revealing occlusions with 4d neural fields. In *CVPR*, 2022. 1

[97] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, 2024. 8

[98] Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z Jane Wang, and Rabab Ward. Multi-view 3d reconstruction with transformers. In *ICCV*, 2021. 1

[99] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2Mesh: Generating 3D mesh models from single RGB images. In *ECCV*, 2018. 1

[100] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5261–5271, 2025. 3, 7

[101] Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training incentivizes reinforcement learning scaling, 2025. 2

[102] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652, 2021. 1

[103] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024. 6, 7, 1, 14

[104] Chao Wen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2Mesh++: Multi-view 3D mesh generation via deformation. In *ICCV*, 2019. 1

[105] Charles Wheatstone. Contributions to the physiology of vision.—part the first. on some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical transactions of the Royal Society of London*, 1838. 1

[106] Markus Worchel, Rodrigo Diaz, Weiwen Hu, Oliver Schreer, Ingo Feldmann, and Peter Eisert. Multi-view mesh reconstruction with neural deferred shading. In *CVPR*, 2022. 1

[107] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. MarrNet: 3D shape reconstruction via 2.5D sketches. *NeurIPS*, 2017. 1

[108] Jane Wu, Georgios Pavlakos, Georgia Gkioxari, and Jitendra Malik. Reconstructing hand-held objects in 3d. *arXiv preprint arXiv:2404.06507*, 2024. 2

[109] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *Advances in Neural Information Processing Systems*, 37: 121859–121881, 2024. 4

[110] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Philip Torr, Xun Cao, and Yao Yao. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. *arXiv preprint arXiv:2505.17412*, 2025. 6

[111] Tianhao Wu, Chuanxia Zheng, Frank Guan, Andrea Vedaldi, and Tat-Jen Cham. Amodal3r: Amodal 3d reconstruction from occluded 2d images. *arXiv preprint arXiv:2503.13439*, 2025. 1

[112] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21469–21480, 2025. 2, 3, 4, 6, 8, 1, 9, 10, 13

[113] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems (RSS)*, 2018. 13

[114] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3D reconstruction from single and multi-view images. In *ICCV*, 2019. 1

[115] Hu Xu, Nikhila Goyal, Mitchell Wortsman, Gabriel Ilharco, Ozan Sener, Aniruddha Kembhavi, Ali Farhadi, and Rohit Girdhar. Metaclip: How to make clip efficiently. *arXiv preprint arXiv:2404.07143*, 2024. 7, 2, 12

[116] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Amodal completion via progressive mixed context diffusion. In *Conference on Computer Vision and Pattern Recognition*, 2024. 1

[117] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. DISN: Deep implicit surface network for high-quality single-view 3D reconstruction. *NeurIPS*, 2019. 1

[118] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1179–1189, 2023. 6, 13

[119] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024. 3

[120] Xianghui Yang, Huiwen Shi, Bowen Zhang, Fan Yang, Jiacheng Wang, Hongxu Zhao, Xinhai Liu, Xinzhou Wang, Qingxiang Lin, Jiaao Yu, et al. Hunyuan3d 1.0: A unified framework for text-to-3d and image-to-3d generation. *arXiv preprint arXiv:2411.02293*, 2024. 2, 4

[121] Kaixin Yao, Longwen Zhang, Xinhao Yan, Yan Zeng, Qixuan Zhang, Lan Xu, Wei Yang, Jiayuan Gu, and Jingyi Yu. Cast: Component-aligned 3d scene reconstruction from an rgb image. *ACM Transactions on Graphics (TOG)*, 2025. 1

[122] Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. *arXiv preprint arXiv:2503.22236*, 3:2, 2025. 6

[123] Fisher Yu, Haofeng Chen, Xin Wang, Wei Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[124] Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models, 2023. 1, 6

[125] 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. 8, 1

[126] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023. 6, 13

[127] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 3, 14