Amodal3R: Amodal 3D Reconstruction from Occluded 2D Images

Tianhao Wu[1*], Chuanxia Zheng[2†], Frank Guan[3], Andrea Vedaldi[2], Tat-Jen Cham[1]

[*]S-Lab, [1]Nanyang Technological University, [2]VGG, University of Oxford

[3]Singapore Institute of Technology

https://sm0kywu.github.io/Amodal3R/

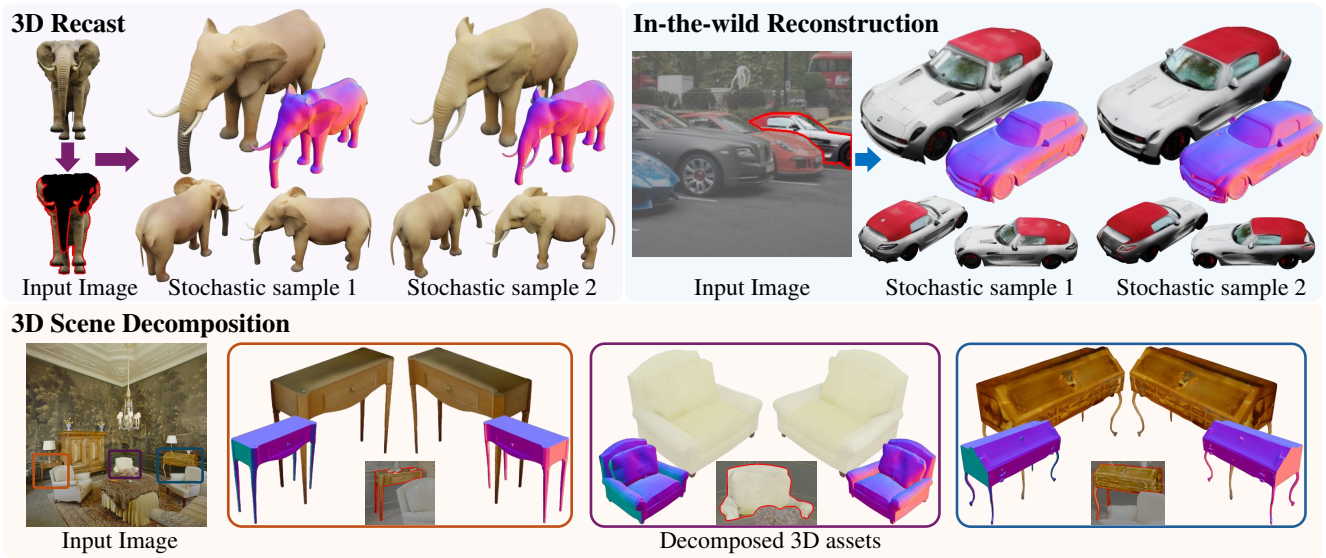


Figure 1. **Example results of Amodal3R.** Given partially visible objects within images (occluded regions are shown in black, visible areas in red outlines), our Amodal3R generates *diverse* semantically meaningful 3D assets with reasonable geometry and plausible appearance. *We sample multiple times to get diverse results from the same occluded input.* Trained on synthetic datasets, it generalizes well to out-of-domain test examples from real scenes, where most objects are partially visible, and reconstructs reasonable 3D assets.

Abstract

Most existing image-to-3D models assume that objects are fully visible, ignoring occlusions that commonly occur in real-world scenarios. In this paper, we introduce Amodal3R, a conditional image-to-3D model designed to reconstruct plausible 3D geometry and appearance from partial observations. We extend a "foundation" 3D generator by introducing a visible mask-weighted attention mechanism and an occlusion-aware attention layer that explicitly leverage visible and occlusion priors to guide the reconstruction process. We demonstrate that, by training solely on synthetic data, Amodal3R learns to recover full 3D objects even in the presence of occlusions in real scenes. It substantially outperforms state-of-the-art methods that independently perform 2D amodal completion followed by 3D reconstruction, thereby establishing a new benchmark for occlusion-aware 3D reconstruction.

[†]Project Lead.

1. Introduction

Humans possess a remarkable ability to infer the complete 3D shape and appearance of objects from single views, even when those objects are partly hidden behind occluders. This ability, known as *amodal completion*, is key to operating in complex real-world environments, where objects are often partially occluded by surrounding clutter. However, existing image-based 3D reconstruction models [5, 6, 23, 41, 47, 48, 65, 70, 71, 75, 77, 80, 88, 89] fail to recover full 3D assets when the object is partially occluded. We thus consider the problem of reconstructing 3D objects from one or more partially occluded views, a novel task that we call *amodal 3D reconstruction*.

Amodal 3D reconstruction is a challenging task that requires reconstructing an object's photorealistic 3D geometry and appearance, including the occluded regions, which are inherently ambiguous. Previous approaches [7, 54] typ-

ically decompose the task into a *two-stage* pipeline: first performing 2D amodal completion [85, 86, 91], followed by conventional 3D reconstruction [41, 75, 89]. While these two-stage pipelines are straightforward to implement, they have some drawbacks. First, 2D amodal completion methods rely predominantly on appearance-only priors, without incorporating explicit information on the 3D structure. This *lack of 3D geometric reasoning* limits their ability to generate physically plausible completions. Second, some 3D reconstruction methods depend on, or benefit from, multi-view inputs. In this case, 2D amodal completion may lack *multi-view consistency*, particularly when performed independently for different views, which confuses the 3D reconstruction process.

In this paper, we present **Amodal3R**, a novel *single-stage* paradigm for amodal 3D reconstruction that surpasses previous state-of-the-art approaches. Amodal3R fine-tunes a pre-trained 3D generator by introducing a dedicated branch that explicitly models occlusions. Its key advantage is performing reconstruction, completion, and occlusion reasoning directly in a 3D latent space rather than using a *two-stage* scheme. With this, the model can reconstruct both visible and occluded regions of the object coherently and plausibly. Specifically, we introduce a *mask-weighted* cross-attention to highlight visible information, along with a new *occlusion-aware* attention layer to guide completion in occluded areas. Together, these components enable the model to focus more on the visible parts of the object without significantly perturbing the pre-trained model.

We designed several benchmarks to evaluate the performance of Amodal3R, including object-level Google Scanned Objects [16] and Toys4K [67] augmented with occlusions, 3D scenes from Replica [68], and real-world monocular images. Without relying on additional heuristics, Amodal3R achieves significantly superior performance compared to state-of-the-art models and generalizes well to different datasets, including real ones.

In summary, our main contributions are as follows:

- We propose a feed-forward 3D reconstruction model that directly reconstructs complete and high-quality 3D objects from one or more partially occluded views.
- We introduce a mask-weighted cross-attention mechanism and an occlusion-aware layer to inject occlusion awareness into an existing 3D reconstruction model, improving both the geometry and appearance of reconstructed objects when they are partially occluded.
- We demonstrate, via experiments on the 3D scene dataset Replica and real-world images, that our one-stage pipeline significantly outperforms existing two-stage approaches, establishing a new benchmark for amodal 3D reconstruction.

2. Background

2D Amodal Completion. Recent advances in 2D amodal completion [2, 54, 83, 85, 86, 91] have achieved significant success in reconstructing occluded regions of objects in 2D images. While these methods show promise for 3D generation pipelines [2, 54], they still have limitations. First, 2D amodal completion models have limited 3D geometric priors, as they interpret images as 2D patterns. While they excel at completing textures, they may generate physically implausible geometries when handling complex occlusions. This stems from their lack of explicit 3D reasoning and reliance on 2D appearance priors, without true volumetric understanding. Second, for models that use multi-view images as input, the results of 2D amodal completion are often inconsistent across views. Inconsistent views need to be handled by the 3D reconstructors, which causes confusion (see Sec. 4). Although there has been significant work on multi-view consistent generation [7, 62, 70, 72, 81, 89, 94], multi-view consistent completion is less explored. RenderDiffusion [1] contains an example, but the results are often blurry or lack details. This motivates our 3D-centric reconstruction framework.

3D Shape Completion. Several methods [12, 14, 69, 95] start from a partial 3D reconstruction, then complete it in 3D. They use encoder-decoder architectures [24] or volumetric representations [10, 29] to robustly recover 3D shape, but not texture, which is necessary in many applications. They also still require recovering the partial 3D geometry from an occluded image, a challenge in itself. Furthermore, they ignore the input image when completing the object in 3D, disregarding the occlusion pattern that caused the 3D geometry to be recovered only partially. Recent work [11, 82] utilizes the partially visible object as input specifically for 3D shape completion. However, it focuses solely on geometry reconstruction, without recovering the object's texture and appearance. In contrast, our approach is end-to-end, recovering the complete 3D shape and appearance of the object from the occluded image.

3D Generative Models. Early advancements in 3D generation are based on GANs [19], exploring various 3D representations such as point clouds [25, 39], voxel grids [84, 96], view sets [48, 55], NeRF [4, 5, 50, 60], SDF [18], and 3D Gaussian mixtures [78]. While GANs can capture complex 3D structures, they struggle with stability, scalability, and data diversity. The focus then shifted to diffusion models [22, 59, 66], which are more capable and versatile. They were first applied to novel view synthesis [76] in image space, before expanding to model a variety of 3D representations, including point clouds [44, 45, 49, 79, 93], voxel grids [37, 46], triplanes [64, 98], meshes [20, 43], and 3D Gaussian mixtures [9, 35, 38, 87]. Autoregressive models [8, 47, 65] have also been explored for mesh generation, focusing on improving the topology of the final 3D mesh.

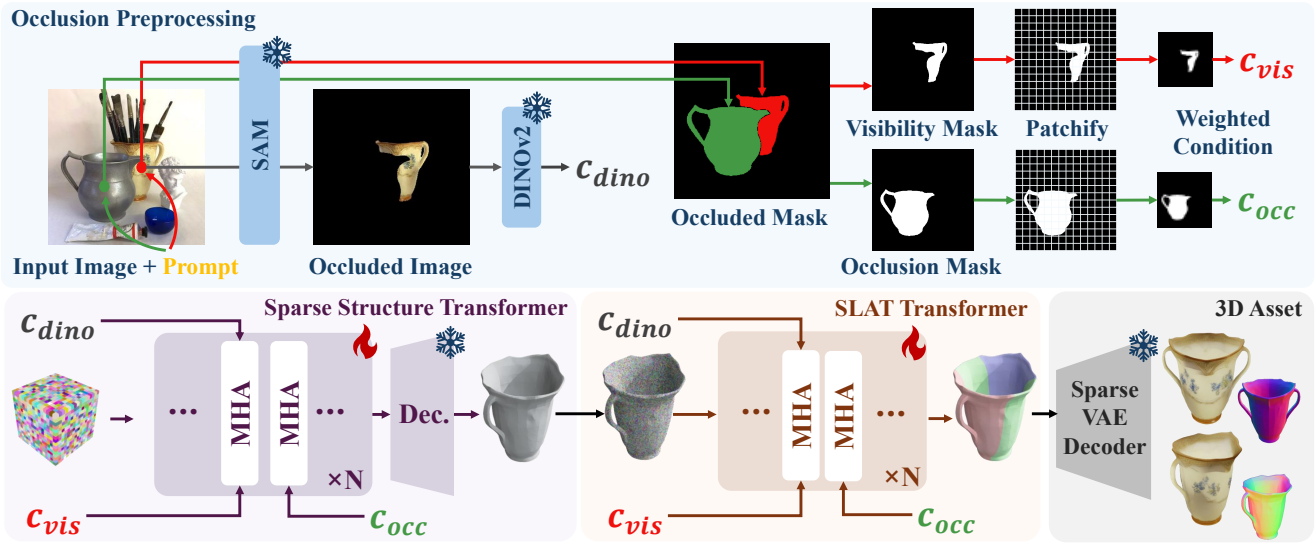


Figure 2. **Overview of Amodal3R.** Given an image as input and the regions of interest, Amodal3R first extracts the partially visible target object, along with the visibility and occlusion masks using an off-the-shelf 2D segmenter. It then applies DINOv2 [53] to extract features $\boldsymbol{c}_{\text{dino}}$ as additional conditioning for the 3D reconstructor. To enhance occlusion reasoning, each transformer block incorporates a *mask-weighted cross-attention* (via $\boldsymbol{c}_{\text{vis}}$) and *occlusion-aware attention layer* (via $\boldsymbol{c}_{\text{occ}}$), ensuring the 3D reconstructor accurately perceives visible information while effectively inferring occluded parts. For conditioning details, see Sec. 3.1.

An alternative to performing diffusion in 2D image space [62, 70, 89] is to use 3D latent spaces [28, 51, 58, 63, 74, 80]. Such methods typically consist of two stages: the first for generating geometry and the second for generating appearance, and are capable of high-quality image-to-3D generation. However, they assume that input objects are fully visible, which limits their application to real-world scenes, where occlusions are frequent. Here, we extend such approaches to generate high-quality 3D assets from occluded input images — a challenging task that requires inferring complete 3D objects from partial information.

3. Method

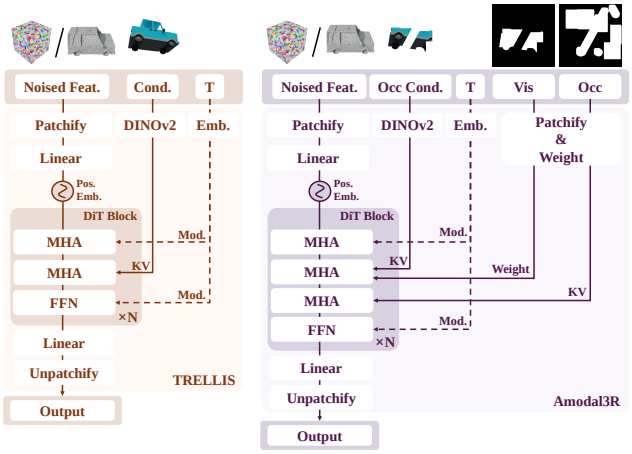
Consider an image x containing a partially occluded view of an object of interest $\boldsymbol{o}_i$. We wish to reconstruct the *complete* 3D shape and appearance y of the object $\boldsymbol{o}_i$. This task is conceptually similar to existing image-to-3D reconstruction, except that here the object is partially occluded instead of being fully visible in x. We call this problem *amodal 3D reconstruction*.

We introduce Amodal3R (Fig. 2), a new method for amodal 3D reconstruction. Unlike recent two-stage methods [7, 54] that first perform 2D amodal completion followed by 3D reconstruction, Amodal3R is an end-to-end occlusion-aware 3D reconstruction model that generates *complete* geometry and appearance directly within the 3D space. Formally, Amodal3R is a conditional generator $v_\theta(y|o_i, M_{\text{vis}}, M_{\text{occ}})$ that takes as input the image x, centered on a partially visible object $\boldsymbol{o}_i$, as well as the visibility mask M_{vis} and occlusion mask M_{occ}. The visibility mask M_{vis} highlights the pixels in x that correspond to object $\boldsymbol{o}_i$, while M_{occ} marks the pixels occupied by occluders, *i.e.*, all other objects that potentially obscure $\boldsymbol{o}_i$. For real images, these masks can be efficiently obtained using pre-trained 2D segmentation models like Segment Anything [34]. By providing point coordinate prompts for the object of interest $\boldsymbol{o}_i$ and its occluders, the segmentation model can generate the corresponding masks. In cases where multiple occluders are present, or when occluders fragment the visible area of the target object, a sequential process is employed. Specifically, point prompts for each visible/occluding region are provided to the 2D segmenter one at a time, with the model generating an individual mask for each region. These masks are then aggregated to form comprehensive visibility and occlusion masks.

The challenges for Amodal3R are how to: **(1)** produce a *plausible* and *complete* 3D shape and appearance from partial observations, even in the absence of multi-view information; and **(2)** ensure 3D *consistency* in terms of geometry and photometry, maintaining seamless visual coherence between visible and generated regions.

3.1. Mask-Conditional Generative Models

To fine-tune an image-to-3D model for partially occluded inputs, the key change is to condition the transformer v_θ not only on the object of interest $\boldsymbol{o}_i$, but also on the visibility and occlusion masks, *i.e.*, M_{vis} and M_{occ}. A naïve approach is to patchify or embed the masks into tokens $(\boldsymbol{c}_{\text{vis}}, \boldsymbol{c}_{\text{occ}})$,



(a) TRELLIS [80] (b) Amodal3R (Ours)

Figure 3. **The transformer structure of Amodal3R.** Compared with the original TRELLIS [80] design, we further introduce the mask-weighted cross-attention and occlusion-aware layer. It applies to both sparse structure and SLAT diffusion models.

which can be concatenated to the image tokens $\boldsymbol{c}_{\text{dino}}$ and processed by cross-attention as before. However, the RGB $\boldsymbol{o}_i$ contains significantly more information than the binary masks M_{vis} and M_{occ}. The transformer, which is initially trained with RGB only, may simply ignore the mask information. This is compounded by the fact that learning to use this new information, which involves a new type of token that is incompatible with image tokens, may require aggressive fine-tuning of the transformer.

To sidestep this problem, inspired by masked attention approaches in 2D completion [92], we introduce *mask-weighted cross-attention* and an *occlusion-aware attention* layer to better utilize the visibility mask M_{vis} and the occlusion mask M_{occ} without disrupting the pre-trained 3D generator. These are described next.

Mask-weighted Cross-Attention. A key novel component of Amodal3R is *mask-weighted cross-attention*, which allows the model to focus its attention on the visible parts of the object. We implement this mechanism in each attention block in the transformers υ_θ of a pre-trained 3D generator. Given the latent tokens $\boldsymbol{\ell} \in \mathbb{R}^{L\times C}$ input to a cross-attention layer as well as the image features $\boldsymbol{c}_{\text{dino}} \in \mathbb{R}^{K\times C'}$ of the partially visible object, cross-attention computes the similarity score matrix via

$$\boldsymbol{q} = W_q\boldsymbol{\ell}, \quad [\boldsymbol{k}, \boldsymbol{v}] = W_{kv}\boldsymbol{c}_{\text{dino}}, \quad S = \boldsymbol{q}\boldsymbol{k}^\top/\sqrt{D}, \quad (1)$$

where W_q and W_{kv} are learnable projections that map the latents $\boldsymbol{\ell}$ to the query $\boldsymbol{q}$, and the conditioning image feature $\boldsymbol{c}_{\text{dino}}$ to the key $\boldsymbol{k}$ and the value $\boldsymbol{v}$, respectively, while D is the dimension of the query and key vectors.

We wish to bias the token similarity matrix $S \in \mathbb{R}^{L\times K}$ towards the visible parts of the object. Recall that K is the number of tokens $\boldsymbol{c}_{\text{dino}}$ extracted by DINOv2 [52] from the occluded image $x \cdot M_{\text{vis}}$. Each token thus corresponds to a $P \times P$ patch in the input image (where $P = 14$). We extract analogous flattened 1-D tokens $\boldsymbol{c}_{\text{vis}} = [\rho(M_{\text{vis},1}), \ldots, \rho(M_{\text{vis},K})]$ from the visibility mask by calculating the fraction of $P \times P$ pixels that are visible in the j-th image patch $M_{\text{vis},j}$. $\boldsymbol{c}_{\text{vis},j} = \rho(M_{\text{vis},j}) \in [0, 1]$. We then use these quantities to bias the computation of the attention matrix $A = \text{softmax}(S + \log c_{\text{vis}}) \in [0, 1]^{L\times M}$ via broadcasting. Hence:

$$A_{ij} = \frac{\boldsymbol{c}_{\text{vis},j}\exp(S_{ij})}{\sum_{k=1}^{K}\boldsymbol{c}_{\text{vis},k}\exp(S_{ik})}. \quad (2)$$

In this manner, the visibility flag modulates the attention matrix A smoothly. If there are no visible pixels in a patch j, then $A_{ij} = 0$, so the corresponding image tokens are skipped in cross-attention. While we have illustrated how this works for a single head, in practice Amodal3R uses a multi-head transformer, to which Eq. (2) extends trivially. Please see the supplementary materials for more details.

Occlusion-aware Attention Layer. To prevent the network from arbitrarily extending the geometry beyond the occluded regions, we further introduce an occlusion-aware attention layer. The key insight is that specifying only visible information is insufficient; it is equally important to *distinguish foreground occluders from the background*, as this explicitly identifies the potential regions requiring completion. Namely, when a pixel is marked as invisible in the mask M_{vis}, this might be (i) because there is an occluder in front of that pixel (so the pixel *could plausibly* belong to the object of interest if the occlusion were removed), or (ii) because the pixel is entirely *off* the object. This information is encoded by the mask M_{occ}.

To help the model distinguish between visible, occluded, and background areas, we add one more cross-attention layer placed immediately after the mask-weighted cross-attention layer, and pool solely the occlusion mask M_{occ}. We encode the occlusion mask M_{occ} as a sequence of flattened 1-D tokens $\boldsymbol{c}_{\text{occ}} = [\rho(M_{\text{occ},1}), \ldots, \rho(M_{\text{occ},M})]$, and then simply pool $\boldsymbol{c}_{\text{occ}}$ using a cross-attention layer.

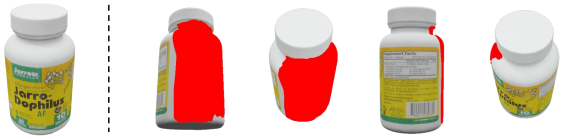


Figure 4. **3D-consistent mask example.** Given a 3D mesh, we render *selected triangles* in a distinct color from the others to generate multi-view consistent masks. It allows the evaluation of multi-view methods in handling contact occlusion. (The occluded regions are shown in red.)

3.2. Simulating Occluded 3D Data

To train and evaluate our model, we require a dataset of 3D assets imaged in scenarios with clutter and occlusions. We curated a large synthetic 3D dataset to train our model, as collecting such 3D assets in the real world is challenging.

Random 2D Occlusions. Each training sample $(\boldsymbol{o}, x, M_{\text{vis}}, M_{\text{occ}})$ consists of a 3D object $\boldsymbol{o}$ (from which ground truth latents can be obtained by using the encoders of Appendix A.1), an image x with partial occlusions, and corresponding visibility and occlusion masks M_{vis} and M_{occ}. In a real scenario, occlusions arise from other objects in the scene. In a multi-view setting, like the one discussed below, these occlusions must be consistent across views, reflecting the underlying scene geometry. However, because our model is trained for single-view reconstruction, we can simulate occlusions by randomly masking parts of the object after rendering it in 2D. This is simpler and allows us to generate fresh occlusion patterns each time a view is sampled for training.

Thus, given $\boldsymbol{o}$ and an image x rendered from a random viewpoint, we generate random occlusion masks as follows. Inspired by work on 2D completion [26, 40, 56, 90, 92], we randomly place lines, ellipses, and rectangles, simulating diverse masking patterns. The union of these random shapes gives us the occlusion mask M_{occ}. Then, if M_{obj} is the mask of the unoccluded object, the visible mask is given by $M_{\text{vis}} = M_{\text{obj}} \odot (1 - M_{\text{occ}})$. Examples of such patterns are given in the supplementary material.

3D-consistent Occlusions. In a real scene imaged from multiple views, occlusions are not random but geometrically consistent across different viewpoints, as they result from other objects. This is especially true for *contact occlusions*, where parts of an object remain occluded by other objects from *all* viewpoints. To evaluate the model's performance under such challenging conditions, 3D-consistent masks are required.

To efficiently generate such contact occlusion masks, we leverage 3D meshes during rendering. Starting from a randomly chosen triangle, we apply a random-walk strategy to iteratively select neighboring triangles, forming continuous occluded regions until the predefined mask ratio is met. By rendering these masked meshes using the same camera parameters as the RGB images, we ensure multi-view consistency in the generated occlusion masks (see Fig. 4).

3.3. Reconstruction with Multi-view Input

Since our flow-based model performs multiple denoising steps and does *not* require known camera poses for input views, Amodal3R can benefit from *multi-view* reference images $\mathcal{X} = \{x_i\}_{i=1}^{N}$ as conditioning inputs at different steps of the denoising process. While multi-view conditioning naturally improves reconstruction performance, a potential issue is that an image used earlier in the denoising process has a stronger influence on the final 3D geometry. This is because early denoising steps establish the coarse geometry, whereas later steps refine higher-frequency details [32, 33]. Therefore, we prioritize input images based on their visibility. Specifically, when experimenting with multi-view inputs, we sort the images in proportion to the object visibility $|M_{\text{vis}}|$, ensuring that images with higher visibility are used as primary inputs.

4. Experiments

4.1. Experiment Settings

Datasets. Amodal3R is trained on a combination of 3D synthetic datasets: 3D-FUTURE (9,472 objects [17]), ABO (4,485 objects [13]), and HSSD (6,670 objects [31]). Once trained, we first assessed its effectiveness on Toys4K (randomly sampling 1,500 objects [67]) and Google Scanned Objects (GSO) (1,030 objects [16]), which were excluded from the training data for both our model and the baselines. During inference, a 3D-consistent mask occludes each input object, and each view is augmented with additional randomly generated occlusion areas. This ensures that the model cannot simply extract the region to be completed from the occlusion masks. To further assess *out-of-domain* generalization in practical applications, we also evaluated all models on the 3D scene dataset Replica [68] as well as on in-the-wild images.

Metrics. To measure the quality of the rendered images, we use the Fréchet Inception Distance (FID) [21] and the Kernel Inception Distance (KID) [3]. To assess the quality of the reconstructed 3D geometry, we use the Coverage Score (COV), the Point cloud FID (P-FID) [49], and the Minimum Matching Distance (MMD) using the Chamfer Distance (CD). COV measures the diversity of the results, while P-FID and MMD measure the quality of the 3D reconstruction. We also use the CLIP score [57] to evaluate the consistency between each generated and ground-truth object pair.

Baselines. Most image-to-3D generators are trained on fully visible inputs. To ensure fair comparison, we completed the partially visible 2D objects before passing them to 3D generators. In particular, we use pix2gestalt [54], a state-of-the-art 2D amodal completion network. Using this protocol, we compared Amodal3R to state-of-the-art methods such as TRELLIS [80], Real3D [27], GaussianAnything [36]$_{\text{ICLR'25}}$, and LaRa [6]$_{\text{ECCV'24}}$.

Implementation Details. Amodal3R is trained on 4 A100 GPUs (40G) for 20K steps with a batch size of 16, taking approximately one day. We select TRELLIS [80] as underlying model because it is the latest 3D generator and is publicly available. However, our method is not specific to this model.

Method	V-num	2D-Comp	Appearance			Geometry		
			FID ↓	KID(%) ↓	CLIP ↑	P-FID ↓	COV(%) ↑	MMD(‰) ↓
GaussianAnything [36]	1	pix2gestalt [54]	92.26	1.30	0.74	34.69	35.92	5.03
Real3D [27]	1	pix2gestalt [54]	91.21	2.02	0.75	23.92	19.61	9.21
TRELLIS [80]	1	pix2gestalt [54]	58.82	5.87	0.76	26.43	31.65	4.17
Amodal3R (Ours)	1	—	**30.64**	**0.35**	**0.81**	**7.69**	**39.61**	**3.62**
LaRa [6]	4	pix2gestalt [54]	172.84	4.54	0.70	66.34	24.56	8.11
LaRa [6]	4	pix2gestalt [54]+MV	97.53	2.63	0.75	21.80	26.21	8.61
TRELLIS [80]	4	pix2gestalt [54]	65.69	6.92	0.78	24.64	32.33	4.26
TRELLIS [80]	4	pix2gestalt [54]+MV	60.37	1.85	0.83	19.68	31.75	4.21
Amodal3R (Ours)	4	—	**26.27**	**0.22**	**0.84**	**5.03**	**38.74**	**3.61**

Table 1. **Amodal 3D reconstruction results on GSO [16].** Here, V-num denotes the number of input views, and 2D-Comp refers to the 2D completion method. For single-view image-to-3D, we first complete occluded objects using pix2gestalt [54] before passing them to various 3D models. For multi-view image-to-3D, we adopt two variants: (1) pix2gestalt [54] is applied independently on each view; (2) pix2gestalt [54] + MV: a single-view completion is generated first, followed by multi-view diffusion [61] to synthesize 4 views as inputs. Without bells and whistles, Amodal3R consistently outperforms state-of-the-art models across all evaluation metrics.

Method	V-num	2D-Comp	Appearance			Geometry		
			FID ↓	KID(%) ↓	CLIP ↑	P-FID ↓	COV(%) ↑	MMD(‰) ↓
GaussianAnything [36]	1	pix2gestalt [54]	57.17	1.22	0.80	21.97	33.56	7.23
Real3D [27]	1	pix2gestalt [54]	59.92	1.63	0.79	23.31	24.35	9.53
TRELLIS [80]	1	pix2gestalt [54]	43.05	6.83	0.80	26.04	26.28	6.87
Amodal3R (Ours)	1	—	**23.45**	**0.42**	**0.83**	**5.00**	**37.09**	**5.89**
LaRa [6]	4	pix2gestalt [54]	123.52	3.61	0.75	45.91	27.89	9.67
LaRa [6]	4	pix2gestalt [54]+MV	75.33	4.14	0.80	13.00	24.82	10.93
TRELLIS [80]	4	pix2gestalt [54]	46.34	8.77	0.81	28.76	25.35	7.13
TRELLIS [80]	4	pix2gestalt [54]+MV	43.00	7.53	0.81	24.41	26.55	7.05
Amodal3R (Ours)	4	—	**20.93**	**0.50**	**0.85**	**3.78**	**39.03**	**5.75**

Table 2. **Amodal 3D reconstruction results on Toys4K [67].** The experimental setting is the same to Tab. 1, except for the dataset.

Figure 5. **Single-view amodal 3D reconstruction.** The occlusion regions are shown in black and the visible regions are highlighted with red outlines. More examples are provided in supplementary material Fig. D.4

4.2. Main Results

Quantitative Results. We compare Amodal3R with prior methods on *amodal 3D reconstruction* in Tabs. 1 and 2. Amodal3R significantly outperforms the baselines across all metrics in both single- and multi-view settings, demonstrating its effectiveness. Notably, two-stage methods using multiple views ("4 V-num + pix2gestalt") perform worse than their single-view counterparts ("1 V-num + pix2gestalt"). This shows that *inconsistent 2D completion* degrades 3D reconstruction (Sec. 2). In contrast, Amodal3R avoids this issue by operating directly in 3D space and benefits from additional occluded views as expected.

Qualitative Results. Qualitative results are shown in Figs. 5 and 6 and in the supplementary material Figs. D.4 and D.5. For all baselines, pix2gestalt is first applied for 2D amodal completion (second column), and the corresponding completed images are passed to each baseline image-to-3D generator. Amodal3R consistently produces high-quality 3D assets under both single- and multi-view settings, even

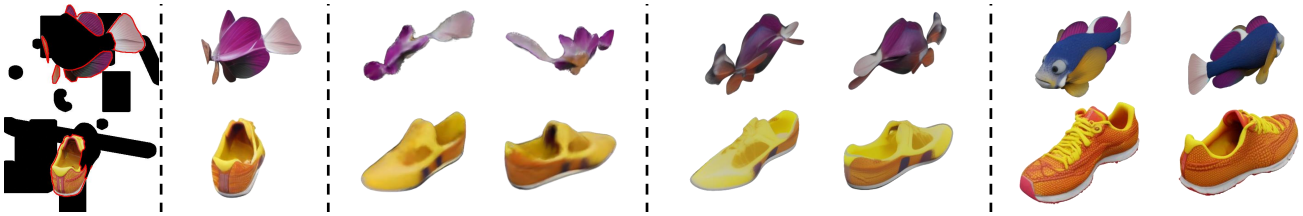


Figure 6. **Multi-view amodal 3D reconstruction.** The above results are reconstructed using 4 occluded input views. Due to limited space, we present only the best results for LaRa and TRELLIS under the "pix2gestalt+MV" setting. We apply 3d-consistent mask and random extended occlusion areas. More examples are provided in supplementary material Fig. D.5.

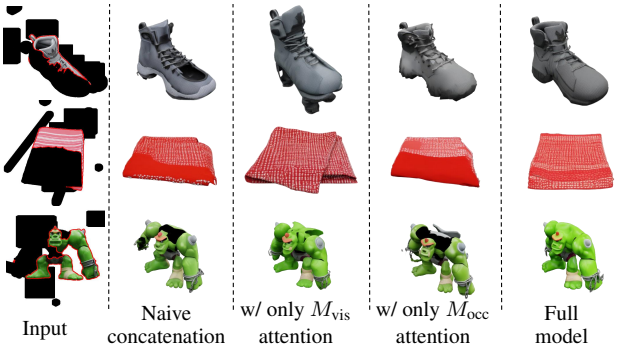


Figure 7. **Ablation examples.** Naive concatenation fails to reconstruct complete shape and appearance. Mask-weighted attention alone extends geometry into background regions, while occlusion-aware attention alone cannot guarantee photorealistic appearance.

in challenging scenarios. In contrast, 2D completion inconsistencies accumulate as more views are added, especially when pix2gestalt is more uncertain, confusing the downstream reconstructor models. For instance, in the giraffe example in Fig. 5, the pix2gestalt completion fails to capture the overall structure of the 3D object, which in turn leads the 3D generator models to misinterpret it as a woodstick-like shape. Instead, Amodal3R reconstructs the 3D geometry and appearance well, maintaining good alignment with the partially visible inputs. These results underscore the effectiveness of Amodal3R and highlight the benefit of integrating occlusion reasoning directly in 3D space, eliminating the reliance on separate, potentially inconsistent, 2D completion stages.

4.3. Ablation Study

We conducted several ablations to study the impact of the different components of Amodal3R and report the results in Tab. 3 and Fig. 7. For these experiments, we test single-view image-to-3D reconstruction on the GSO dataset.

Naive Conditioning. We first evaluated a version of the model that still conditions the reconstruction on the visibility and occlusion masks, but without using the modules of Sec. 3.1. Instead, we directly concatenate the tokens $\boldsymbol{c}_{\text{vis}}$ to $\boldsymbol{c}_{\text{dino}}$ to form the condition for the cross-attention layer. The results (first row in Tab. 3 and second column in Fig. 7) show that, while the resulting model can still perform basic completion, the textures in the occluded regions are inconsistent with those in the visible ones, and the reconstructed geometry is inaccurate, *e.g.* the hole in the shoe.

Method	Appearance		Geometry	
	FID ↓	KID(%) ↓	COV(%) ↑	MMD(‰) ↓
naive conditioning	31.96	0.49	37.96	3.61
w/ only mask-weighted attention	**30.53**	0.38	36.90	3.69
w/ only occlusion-aware layer	31.77	0.57	**40.19**	**3.51**
full model (**Ours**)	30.64	**0.35**	39.61	3.62

Table 3. **Ablations** on different mask conditioning designs.

Mask-weighted Attention. To evaluate the effectiveness of our proposed mask-weighted attention mechanism, we integrate it into the training while omitting the occlusion-aware layer. The results demonstrate improved visual quality, especially in capturing texture details in the visible areas. However, the geometries exhibit deficiencies, as seen in the problematic shoe in Fig. 7 (first row), and the toy monster with a broken back (third row).

Occlusion-aware Layer. The integration of the occlusion-aware layer improves the geometry both quantitatively and qualitatively. This aligns with our motivation for introducing an additional cross-attention layer, aimed at reconstructing the visible areas via the image-conditioned layer and occluded areas via the subsequent layer. However, the occlusion-aware layer alone still results in unsatisfactory appearances, which again indirectly attests to the effectiveness of the mask-weighted attention mechanism.

Full Model. Consequently, the full model integrating both modules achieves optimal 3D generation results characterized by complete geometry and consistent textures.

4.4. Real-World Generation / Completion

Amodal3R demonstrates strong generalization to *out-of-domain* test examples, primarily because it builds upon a large-scale pre-trained 3D generator and is fine-tuned with diverse categories. To demonstrate this advantage,

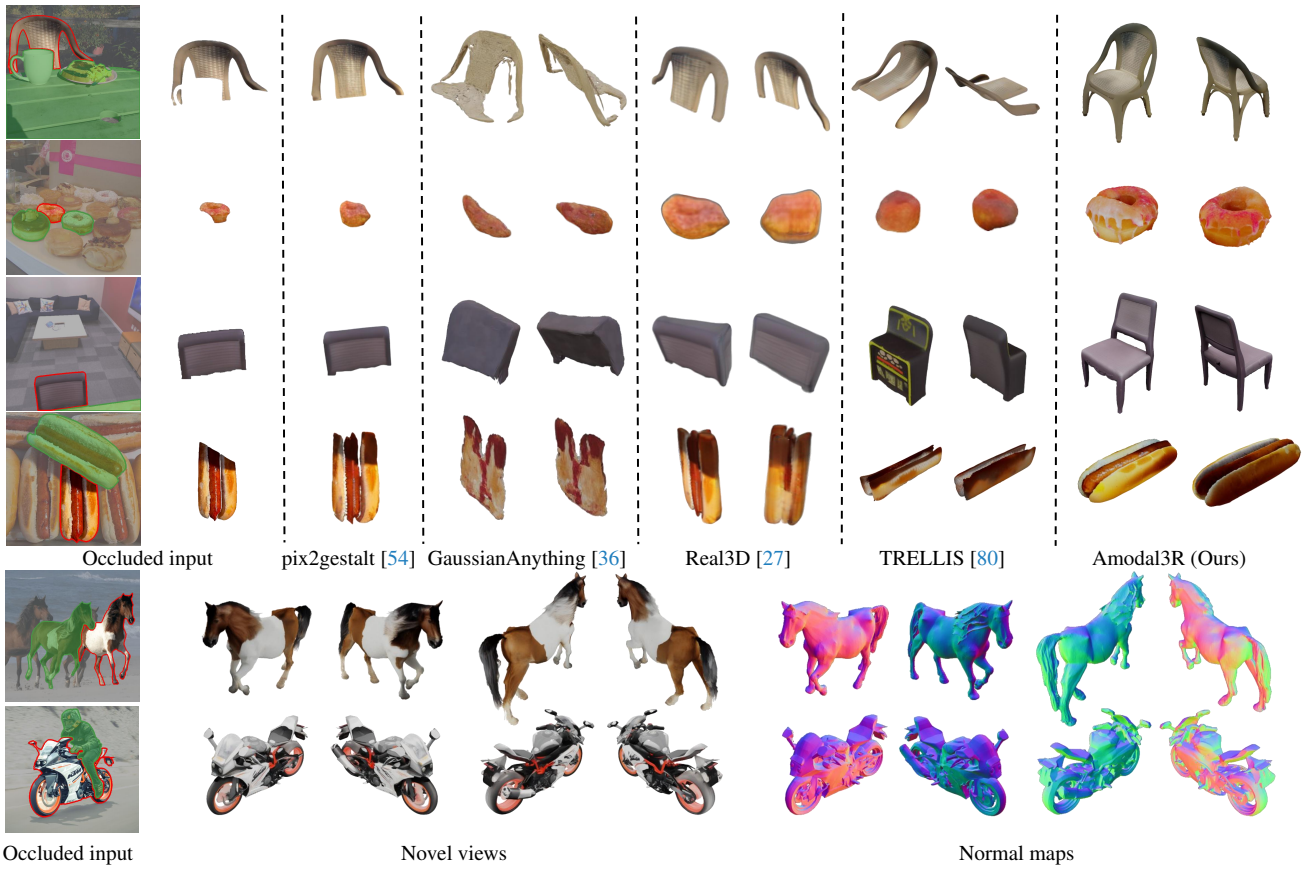


Figure 8. **Examples on Replica [68] and in-the-wild images.** The target objects and the occluders are shown in red and green outlines. More examples are provided in supplementary material Figs. D.7 to D.9.

Method	**P-FID**↓	**COV**(%) ↑	**MMD**(‰) ↓
GaussianAnything	91.37	**61.22**	25.00
Real3D	<u>79.55</u>	55.10	<u>18.28</u>
TRELLIS	128.52	59.17	23.95
Amodal3R	**61.46**	<u>59.18</u>	**10.98**

Table 4. **Quantitative results on the Replica dataset [68].**

we conducted scene-level evaluations. Qualitative results are shown in Fig. 8, which include COCOA [97] (first two rows), Replica [68] (third and fourth rows), and in-the-wild images (fifth and sixth rows). Here, we used Segment Anything [34] to obtain the visibility and occlusion masks. The results show that Amodal3R generates photorealistic 3D assets, whereas pix2gestalt fails to infer complete shapes from the same inputs, leading to unsatisfactory results from subsequent image-to-3D models. We also visualize the colored normal maps, which show that the results of Amodal3R are simultaneously rich in geometric and textural detail.

Quantitative results are presented in Tab. 4. As real-world datasets with instance 3D amodal annotations for both geometry and appearance are limited, here we focus on geometry evaluation by curating a set of 49 completed 3D instances from the Replica scenes. Furthermore, we used the Umeyama algorithm [73] to match the generated meshes with the ground truth meshes in terms of scale and pose. It can be observed that our Amodal3R is still on par with or better than the baselines on the scene-level dataset.

5. Conclusion

We have introduced Amodal3R, a feed-forward approach to reconstruct the complete photorealistic 3D assets from partially visible images. The key to our success is that we construct a mask-weighted cross-attention mechanism and an occlusion-aware layer to effectively exploit visible and occluded information. Compared to the state-of-the-art methods that rely on sequential 2D completion followed by 3D generation, our Amodal3R achieves remarkably better performance by operating directly in 3D space. Furthermore, results on in-the-wild images indicate its potential for subsequent applications in 3D decomposition, marking a step towards robust 3D asset reconstruction in real-world environments with complex occlusion.

Acknowledgements

This study is supported under the RIE2020 Industry Alignment Fund-Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contributions from the industry partner(s). This research is also supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 RG107/24. Dr. Chuanxia Zheng is supported by EPSRC SYN3D EP/Z001811/1.

References

[1] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12608–12618, 2023. 2

[2] Jiayang Ao, Yanbei Jiang, Qiuhong Ke, and Krista A Ehinger. Open-world amodal appearance completion. *arXiv preprint arXiv:2411.13019*, 2024. 2

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 5, 3

[4] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 5799–5809, 2021. 2

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. 1, 2

[6] Anpei Chen, Haofei Xu, Stefano Esposito, Siyu Tang, and Andreas Geiger. Lara: Efficient large-baseline radiance fields. In *European Conference on Computer Vision*, pages 338–355. Springer, 2024. 1, 5, 6, 7

[7] Minghao Chen, Roman Shapovalov, Iro Laina, Tom Monnier, Jianyuan Wang, David Novotny, and Andrea Vedaldi. Partgen: Part-level 3d generation and reconstruction with multi-view diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2, 3

[8] Yiwen Chen, Tong He, Di Huang, Weicai Ye, Sijin Chen, Jiaxiang Tang, Xin Chen, Zhongang Cai, Lei Yang, Gang Yu, et al. Meshanything: Artist-created mesh generation with autoregressive transformers. *arXiv preprint arXiv:2406.10163*, 2024. 2

[9] Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. In *Neural Information Processing Systems (NeurIPS)*, 2024. 2

[10] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023. 2

[11] Junhyeong Cho, Kim Youwang, Hunmin Yang, and Tae-Hyun Oh. Robust 3d shape reconstruction in zero-shot from a single image in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[12] Ruihang Chu, Enze Xie, Shentong Mo, Zhenguo Li, Matthias Nießner, Chi-Wing Fu, and Jiaya Jia. Diffcomplete: Diffusion-based generative 3d shape completion. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[13] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21126–21136, 2022. 5

[14] Ruikai Cui, Weizhe Liu, Weixuan Sun, Senbo Wang, Taizhang Shang, Yang Li, Xibin Song, Han Yan, Zhennan Wu, Shenzhou Chen, et al. Neusdfusion: A spatial-aware generative model for 3d shape completion, reconstruction, and generation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 2

[15] 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, 2024. 3

[16] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 2, 5, 6, 3

[17] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 5

[18] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35:31841–31854, 2022. 2

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[20] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023. 2

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5, 3

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6840–6851, 2020. 2

[23] 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. In *International Conference on Learning Representations (ICLR)*, 2024. 1

[24] Tao Hu, Zhizhong Han, and Matthias Zwicker. 3d shape completion with multi-view consistent inference. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10997–11004, 2020. 2

[25] Zitian Huang, Yikuan Yu, Jiawen Xu, Feng Ni, and Xinyi Le. Pf-net: Point fractal network for 3d point cloud completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 7662–7670, 2020. 2

[26] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):107, 2017. 5

[27] Hanwen Jiang, Qixing Huang, and Georgios Pavlakos. Real3d: Scaling up large reconstruction models with real-world images. *arXiv preprint arXiv:2406.08479*, 2024. 5, 6, 8, 2

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

[29] Yoni Kasten, Ohad Rahamim, and Gal Chechik. Point cloud completion with pretrained text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[30] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[31] 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. 5

[32] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 5

[33] Yulhwa Kim, Dongwon Jo, Hyesung Jeon, Taesu Kim, Daehyun Ahn, Hyungjun Kim, et al. Leveraging early-stage robustness in diffusion models for efficient and high-quality image synthesis. *Advances in Neural Information Processing Systems*, 36:1229–1244, 2023. 5

[34] 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. 3, 8

[35] Yushi Lan, Feitong Tan, Di Qiu, Qiangeng Xu, Kyle Genova, Zeng Huang, Sean Fanello, Rohit Pandey, Thomas Funkhouser, Chen Change Loy, et al. Gaussian3diff: 3d gaussian diffusion for 3d full head synthesis and editing. *arXiv preprint arXiv:2312.03763*, 2023. 2

[36] Yushi Lan, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. Gaussiananything: Interactive point cloud latent diffusion for 3d generation. *arXiv preprint arXiv:2411.08033*, 2024. 5, 6, 8, 2, 3

[37] Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Diffusion-sdf: Text-to-shape via voxelized diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12642–12651, 2023. 2

[38] Xinhai Li, Huaibin Wang, and Kuo-Kun Tseng. Gaussiandiffusion: 3d gaussian splatting for denoising diffusion probabilistic models with structured noise. *arXiv preprint arXiv:2311.11221*, 2023. 2

[39] Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning efficient point cloud generation for dense 3d object reconstruction. In *proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 2

[40] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5

[41] 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. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 9298–9309, 2023. 1, 2

[42] 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. 1

[43] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 9970–9980, 2024. 2

[44] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2837–2845, 2021. 2

[45] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12923–12932, 2023. 2

[46] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4328–4338, 2023. 2

[47] Charlie Nash, Yaroslav Ganin, SM Ali Eslami, and Peter Battaglia. Polygen: An autoregressive generative model of 3d meshes. In *International Conference on Machine Learning (ICML)*, pages 7220–7229. PMLR, 2020. 1, 2

[48] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 7588–7597, 2019. 1, 2

[49] 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. 2, 5, 3

[50] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 11453–11464, 2021. 2

[51] Evangelos Ntavelis, Aliaksandr Siarohin, Kyle Olszewski, Chaoyang Wang, Luc V Gool, and Sergey Tulyakov. Autodecoding latent 3d diffusion models. *Advances in Neural Information Processing Systems*, 36:67021–67047, 2023. 3

[52] 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. 4, 1

[53] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023. 3, 1

[54] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3931–3940. IEEE Computer Society, 2024. 1, 2, 3, 5, 6, 7, 8, 4

[55] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *Proceedings of the ieee conference on computer vision and pattern recognition (CVPR)*, pages 3500–3509, 2017. 2

[56] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544. IEEE, 2016. 5

[57] 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 (ICML)*, pages 8748–8763. PmLR, 2021. 5, 3

[58] Barbara Roessle, Norman Müller, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, Angela Dai, and Matthias Nießner. L3dg: Latent 3d gaussian diffusion. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 10684–10695, 2022. 2

[60] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:20154–20166, 2020. 2

[61] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 6, 4

[62] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2, 3

[63] Wu Shuang, Youtian Lin, Yifei Zeng, Feihu Zhang, 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, 2025. 3

[64] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20875–20886, 2023. 2

[65] Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19615–19625, 2024. 1, 2

[66] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning (ICML)*, pages 2256–2265. pmlr, 2015. 2

[67] Stefan Stojanov, Anh Thai, and James M Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1798–1808, 2021. 2, 5, 6, 3

[68] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 5, 8

[69] David Stutz and Andreas Geiger. Learning 3d shape completion from laser scan data with weak supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1955–1964, 2018. 2

[70] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion: (0-)image-conditioned 3d generative models from 2d data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8863–8873, 2023. 1, 2, 3

[71] Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), pages 10208–10217, 2024. 1

[72] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. In *European Conference on Computer Vision*, pages 175–191. Springer, 2024. 2

[73] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4): 376–380, 2002. 8

[74] Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039, 2022. 3

[75] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision (ECCV)*, pages 439–457. Springer, 2024. 1, 2

[76] 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. 2

[77] Daniel Watson, William Chan, Ricardo Martin Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 1

[78] Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 456–473. Springer, 2024. 2

[79] Zijie Wu, Yaonan Wang, Mingtao Feng, He Xie, and Ajmal Mian. Sketch and text guided diffusion model for colored point cloud generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8929–8939, 2023. 2

[80] 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. *arXiv preprint arXiv:2412.01506*, 2024. 1, 3, 4, 5, 6, 7, 8, 2

[81] Desai Xie, Jiahao Li, Hao Tan, Xin Sun, Zhixin Shu, Yi Zhou, Sai Bi, Sören Pirk, and Arie E Kaufman. Carve3d: Improving multi-view reconstruction consistency for diffusion models with rl finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6369–6379, 2024. 2

[82] Yiheng Xiong and Angela Dai. Pt43d: A probabilistic transformer for generating 3d shapes from single highly-ambiguous rgb images. *arXiv preprint arXiv:2405.11914*, 2024. 2

[83] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Amodal completion via progressive mixed context diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9099–9109, 2024. 2

[84] Bo Yang, Hongkai Wen, Sen Wang, Ronald Clark, Andrew Markham, and Niki Trigoni. 3d object reconstruction from a single depth view with adversarial learning. In *Proceedings of the IEEE international conference on computer vision workshops (ICCVW)*, pages 679–688, 2017. 2

[85] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. Amodal ground truth and completion in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28003–28013, 2024. 2

[86] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene deocclusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3784–3792, 2020. 2

[87] Bowen Zhang, Yiji Cheng, Jiaolong Yang, Chunyu Wang, Feng Zhao, Yansong Tang, Dong Chen, and Baining Guo. Gaussiancube: Structuring gaussian splatting using optimal transport for 3d generative modeling. *arXiv preprint arXiv:2403.19655*, 2024. 2

[88] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 1

[89] Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9720–9731, 2024. 1, 2, 3

[90] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1438–1447, 2019. 5

[91] Chuanxia Zheng, Duy-Son Dao, Guoxian Song, Tat-Jen Cham, and Jianfei Cai. Visiting the invisible: Layer-by-layer completed scene decomposition. *International Journal of Computer Vision (IJCV)*, 129:3195–3215, 2021. 2

[92] Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai, and Dinh Phung. Bridging global context interactions for high-fidelity image completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 11512–11522, 2022. 4, 5

[93] Xiao Zheng, Xiaoshui Huang, Guofeng Mei, Yuenan Hou, Zhaoyang Lyu, Bo Dai, Wanli Ouyang, and Yongshun Gong. Point cloud pre-training with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22935–22945, 2024. 2

[94] Xin-Yang Zheng, Hao Pan, Yu-Xiao Guo, Xin Tong, and Yang Liu. Mvdˆ 2: Efficient multiview 3d reconstruction for multiview diffusion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[95] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5826–5835, 2021. 2

[96] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. *Advances in neural information processing systems (NeurIPS)*, 31, 2018. 2

[97] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1464–1472, 2017. 8

[98] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10324–10335, 2024. 2

A. Implementation Details

A.1. Preliminaries: TRELLIS

We begin by briefly reviewing the TRELLIS [80] model on which our method is based. We chose TRELLIS because it was the state-of-the-art image-to-3D generator at the time of our study and is publicly available. However, our design is not specific to this model. TRELLIS is a conditional 3D diffusion model that performs denoising in a sparse 3D latent space. It introduces a transformer-based variational autoencoder (VAE) $(\mathcal{E}, \mathcal{D})$, where the encoder $\mathcal{E}$ maps sparse voxel features to structured latents $\boldsymbol{z}$, and the decoder $\mathcal{D}$ converts them into desired output representations, including 3D Gaussians [30], radiance fields, and meshes. A 3D object $\boldsymbol{o}_i$ is encoded using its *structured latent variables* (SLAT) defined as $\boldsymbol{z} = \{(\boldsymbol{z}_i, \boldsymbol{p}_i)\}_{i=1}^{L}$, where $\boldsymbol{z}_i \in \mathbb{R}^C$ is a local latent feature attached to the voxel at position $\boldsymbol{p}_i \in \{0, 1, \ldots, N-1\}^3$, N is the spatial resolution of the grid, and $L \ll N^3$ is the number of active voxels intersecting the object's surface. This representation encodes both coarse geometric structures and fine appearance details by associating local latents with active voxels.

TRELLIS comprises two diffusion models: one to predict the active voxel centers $\{\boldsymbol{p}_i\}_{i=1}^{L}$ (stage 1) and the other to recover the corresponding latents $\{\boldsymbol{z}_i\}_{i=1}^{L}$ (stage 2). Each model can be viewed as a denoising neural network υ_θ operating in a latent space $\boldsymbol{\ell}$, trained to remove Gaussian noise $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$ added to the latent code, *i.e.*, $\boldsymbol{\ell}^{(t)} = (1-t)\boldsymbol{\ell}^{(0)} + t\boldsymbol{\epsilon}$, where $t \in [0, 1]$ is the noise level [42]. The denoising network υ_θ is trained to minimize the flow loss:

$$\min_\theta \mathbb{E}_{(\boldsymbol{\ell}^{(0)}, x), t, \boldsymbol{\epsilon}} \|\upsilon_\theta(\boldsymbol{\ell}^{(t)}, x, t) - (\boldsymbol{\epsilon} - \boldsymbol{\ell}^{(0)})\|_2^2, \tag{A.1}$$

where x is the conditional input. In stage 1, the latent code is a compressed version of the $N \times N \times N$ occupancy volume, where the spatial resolution is reduced from $N = 64$ to $r = 16$. Hence, the latent vector is a matrix $\boldsymbol{\ell} \in \mathbb{R}^{L' \times C'}$ of $L' = r^3 = 4096$ C'-dimensional tokens. In stage 2, the latent code $\boldsymbol{\ell} = \{\boldsymbol{z}_i\}_{i=1}^{L} \in \mathbb{R}^{L \times C}$ is a matrix of L C-dimensional tokens, where L is now the number of active voxels. Similar transformers are implemented for the corresponding denoising networks υ_θ (Fig. 3). The conditioning image x is read via cross-attention layers that pool information from the tokens $\boldsymbol{c}_{\text{dino}}$ extracted by DINO [53] from the image x.

Given the active grid $\{\boldsymbol{p}_i\}_{i=1}^{K}$, the local latents $\{\boldsymbol{z}_i\}_{i=1}^{K}$ for each active voxel are predicted via iterative denoising using model (A.1), and these are subsequently decoded by the VAE to produce high-fidelity 3D outputs. The active grid $\{\boldsymbol{p}_i\}_{i=1}^{K}$ is obtained in a similar manner, by denoising encoded and compressed versions of the $N \times N \times N$ occupancy volume.

A.2. Network Design

We adopt the network design in TRELLIS [80] to load the pre-trained image-to-3D weights and integrate the mask-weighted cross-attention mechanism into each DiT block (24 blocks in total). Each image-conditioned cross-attention layer is immediately followed by an occlusion-aware cross-attention layer.

(a) Patchify and weighting of visibility/occlusion mask. The input condition image has a resolution of 512×512, which is resized to 518×518 to facilitate splitting into patches of size 14×14, as required by DINOv2 [52]. The resulting condition is then flattened into a tensor $\mathbf{c}_{dino} \in \mathbb{R}^{1374 \times 1024}$, where the sequence length corresponds to 37×37 patches plus 1 CLS token and 4 register tokens. To better align the visibility and occlusion masks with the DINOv2 features, we first split the masks into patches of the same size, then calculate the weight score for each patch using Eq. 4 and Eq. 5. The final $\mathbf{c}_{vis} \in \mathbb{R}^{1374 \times 1}$ and $\mathbf{c}_{occ} \in \mathbb{R}^{1374 \times 1}$ are obtained by flattening the weight scores, with a value of 1 assigned to the CLS and register token positions.

(b) Occlusion-aware cross-attention layer. We set the feature dimension of the occlusion-aware cross-attention layers to 1024, matching that of the image-conditioned cross-attention layers. To maintain consistent dimensions, we replicate the flattened occlusion masks to form a tensor $\mathbf{c}_{occ\_stack} \in \mathbb{R}^{1374 \times 1024}$.

(c) Multi-head Cross-Attention. Our mask-weighted multi-head cross-attention (MHA) layer, which encourages the model to focus its attention on the visible parts of the object, is an extension of the cross-attention described in the main paper. Specifically, H heads are run in parallel, resulting in H attention scores. For the mask-weighted attention mechanism, we impose $\boldsymbol{c}_{\text{vis}}$ simultaneously to each head:

$$\boldsymbol{A}_h = \text{softmax}(\boldsymbol{S}_h + \log \boldsymbol{c}_{\text{vis}}), \tag{A.2}$$

$$\text{MHA} = [\boldsymbol{A}_1\boldsymbol{v}; \boldsymbol{A}_2\boldsymbol{v}; \ldots; \boldsymbol{A}_H\boldsymbol{v}] \tag{A.3}$$

A.3. Training Details

(a) Pre-trained model loading. While TRELLIS is split into multiple modules, in our work we only train the sparse structure flow transformer and the structured latent flow transformer (see the overview figure, where the "fire" symbols indicate the parts that are fine-tuned, and "snowflake" symbols indicate that we directly use the pre-trained weights).
(b) Data Augmentation. As described in Sec. 3.3, we generate random masks during training for data augmentation. Specifically, we begin by randomly drawing 1 to 3 lines, circles, and ellipses in the mask image. Next, to ensure these regions connect—thereby better simulating real-world occlusions, where mask regions are typically not highly fragmented—we randomly add 3 to 7 rectangular regions that have undergone an expansion operation. This results in stable masking of the objects in the training data. Example inputs are presented in Fig. A.1.

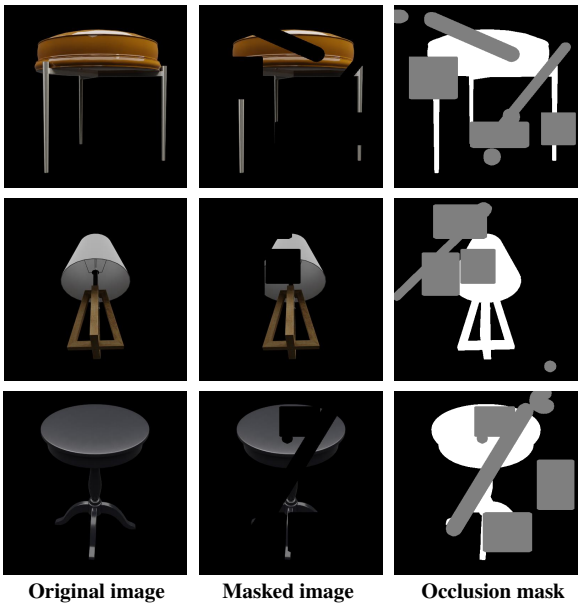


Figure A.1. **Examples of random mask generation.** The visible areas are shown in white, occluded areas in gray, and background in black.

A.4. Inference Details

(a) 3D-consistent mask ratio. For the multi-view 3D-consistent masks described in Sec. 3.3, we set the mask ratio to a random number between 0.4 and 0.6 for each object, resulting in a variety of reasonable mask areas.
(b) Time consumption. Despite the introduction of additional cross-attention layers, our inference time remains comparable to that of the baselines. Amodal3R can generate and render each object in under 10 seconds.

Figure A.2. **Examples using occluded images directly as the input of baseline models.**

B. Experimental Details

B.1. Evaluation Protocol

We evaluate the results using Google Scanned Objects (GSO) (1,030 objects) [16] and a randomly sampled subset of Toys4K [67] containing 1,500 objects. Here, we provide additional details regarding the computation of our evaluation metrics.

(a) Rendering quality and semantic consistency alignment. To assess overall rendering quality, we compute the Fréchet Inception Distance (FID) [21] and Kernel Inception Distance (KID) [3]. Moreover, we evaluate semantic consistency using the CLIP score [57] by measuring the cosine similarity between the CLIP features of each generated image and its corresponding ground truth. For each object, we render 4 views using cameras with yaw angles of $\{0°, 90°, 180°, 270°\}$ and a pitch angle of 30°. The camera is positioned with a radius of 2.0 and looks at the origin with a FoV of 40°, consistent with TRELLIS [80]. While FID and KID are calculated between the ground truth and generated sets (6,000 images for Toys4K and 4,120 images for GSO), the CLIP score is calculated in a pair-wise manner, and we report the mean value to evaluate semantic consistency.

(b) Geometry quality. For 3D geometry evaluation, we adopt Point cloud FID (P-FID) [49], Coverage Score (COV), and Minimum Matching Distance (MMD) using Chamfer Distance (CD). Following previous work [36, 49, 80], we sample 4096 points from each GT/generated point cloud, which are obtained from the unprojected multi-view depth maps using farthest point sampling.

C. More Results

In this section, we provide additional qualitative examples and comparison results to further demonstrate the performance of our Amodal3R.

C.1. Baselines using occluded input

As stated in the main paper, "occluded images will lead to incomplete or broken structures" in current 3D generative models. Here, we provide examples where pix2gestalt is omitted and the occluded images are directly used as input. As shown in Fig. A.2, when baseline models receive images of partially visible objects as input, they often fail to recover complete and intact 3D assets.

C.2. More single-view to 3D examples

Due to page restrictions, we only provide limited examples in the main paper. Here, we visualize more examples of single-view to 3D to further demonstrate the effectiveness of our method in Fig. D.4. The results show that, compared with the 2D amodal completion + 3D generation baselines, our Amodal3R yields higher quality 3D assets across multiple categories.

C.3. More multi-view to 3D examples

We first provide visualized examples to explicitly explain the difference between the "pix2gestalt" and "pix2gestalt + MV" settings in multi-view to 3D generation in Fig. C.3. For the "pix2gestalt" setting, we directly implement pix2gestalt for amodal completion independently under each view. For the "pix2gestalt + MV" setting, we first choose the view with the greatest visibility from the 4 occluded views, then use pix2gestalt to complete the object (shown in the pix2gestalt column in the qualitative result), and subsequently use Zero123++ to get the 4 consistent views as input to LaRa and TRELLIS. It can be observed that pix2gestalt alone results in obvious multi-view inconsistency, while with Zero123++ the consistency is significantly improved, leading to better 3D generation quality.

More multi-view to 3D examples are provided in Fig. D.5, where our Amodal3R again generates 3D assets with better quality than the baselines.

D. Discussion, Limitation, and Future Work

While Amodal3R achieves impressive 3D amodal completion, it comes with several limitations we hope to address in the future. **(1) Dataset expansion.** Due to computational resource limitations, Amodal3R is trained on a relatively small dataset, consisting of only 20,627 synthetic 3D assets, predominantly confined to the furniture category. Consequently, completions on some complex or out-of-distribution objects may fail or lead to unrealistic structures. We believe that training on larger datasets, *e.g.*, Objaverse-XL [15], could mitigate these issues. **(2) Real-world data adaptation.** Unlike pix2gestalt, Amodal3R is trained exclusively on synthetic data. As a result, it cannot leverage environmental cues and must rely solely

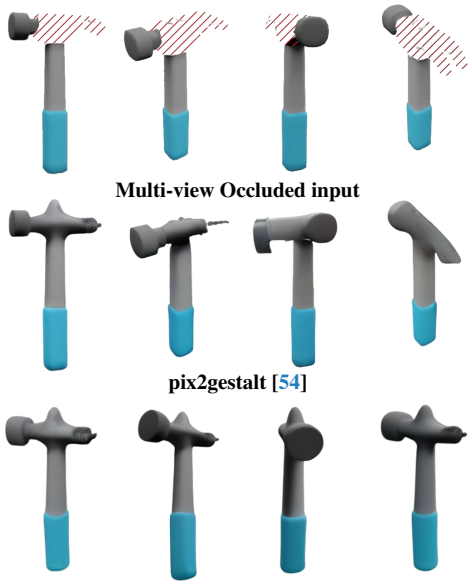


Figure C.3. **Example of "pix2gestalt" and "pix2gestalt + MV" input for multi-view to 3D evaluation.**

on the visible portions of occluded objects for reconstruction. Creating real-world 3D modal datasets will further enhance the ability to apply models to real scenes. **(3) Controllable completion.** Currently, how objects are completed is entirely up to the model and lacks user control. Enhancing the model to accept additional conditions, such as text, and allowing users to edit and control the style of the completion process will be an important research direction for us in the future.

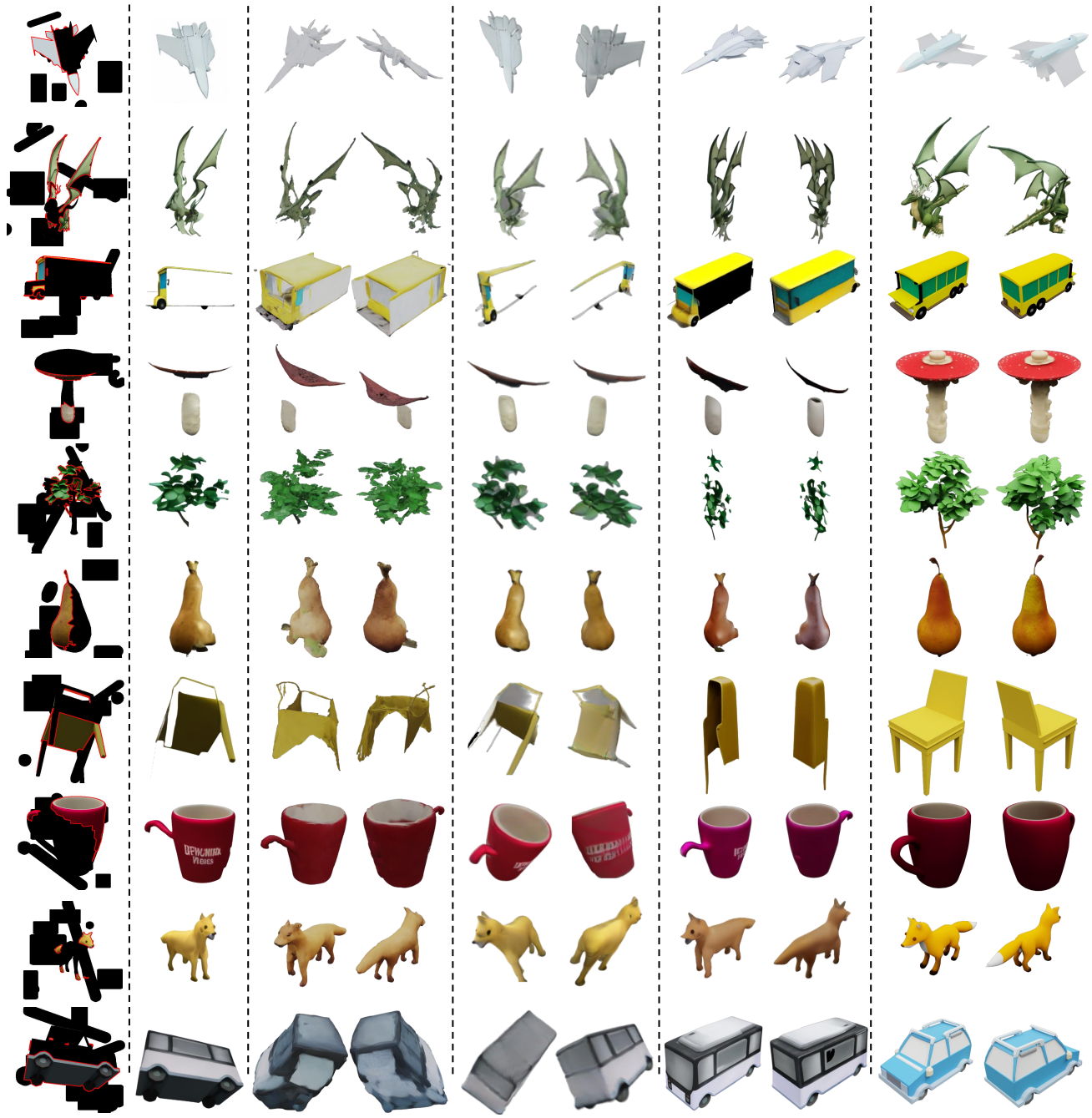


Figure D.4. **Additional single-view to 3D comparison examples.** The occluders are shown in black and the visible regions are highlighted with red outlines.

Figure D.5. **Additional multi-view to 3D comparison examples.** The occluders are shown in black and the visible regions are highlighted with red outlines.

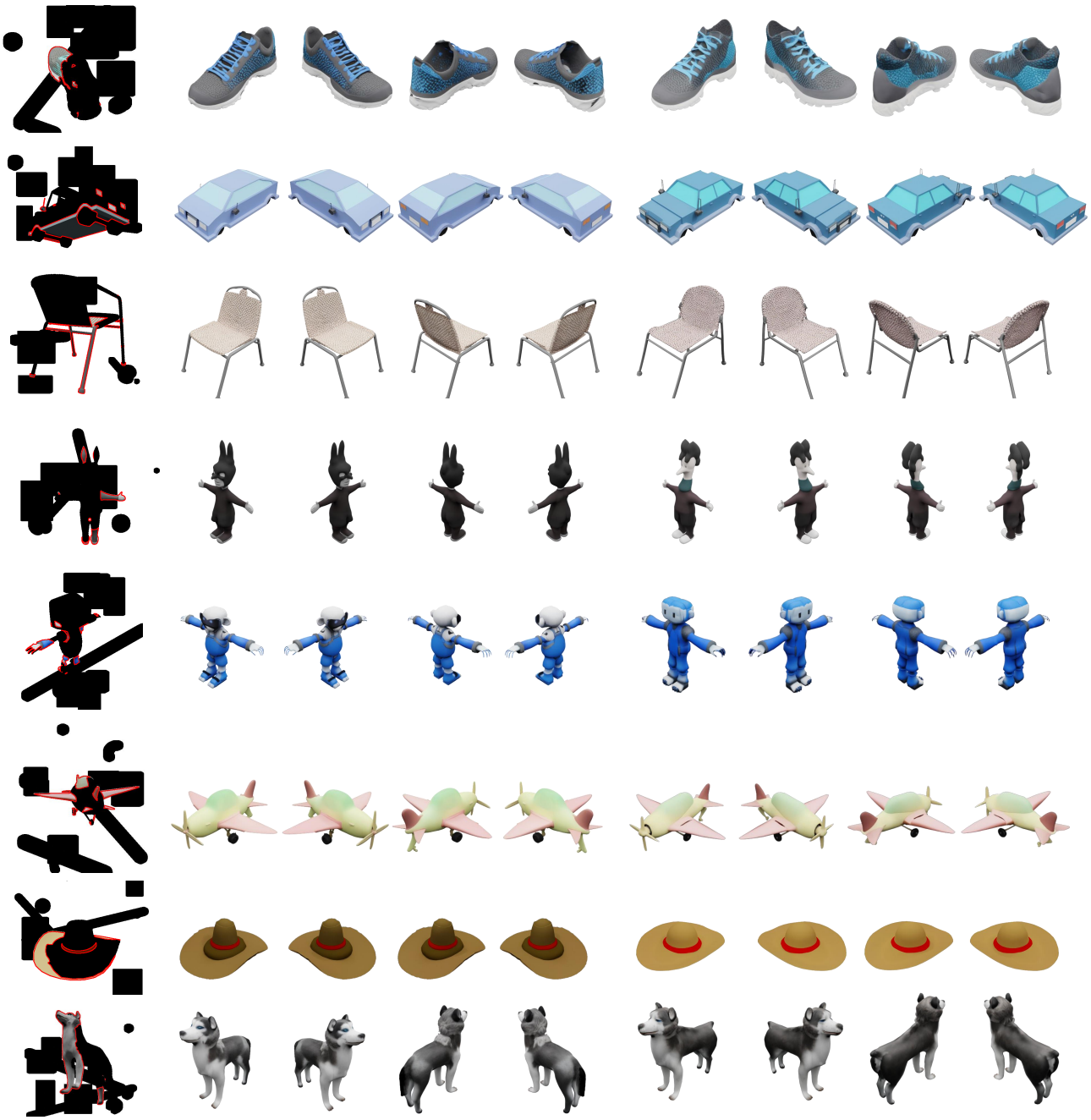


Figure D.6. **Additional diverse examples.** The occluders are shown in black and the visible regions are highlighted with red outlines.

Figure D.7. **Additional in-the-wild examples compared with pix2gestalt + TRELLIS.** The target objects and occluders are marked with the red and green outlines.

Input image **Reconstruction result**

Figure D.8. **Additional in-the-wild examples.** The target objects and occluders are marked with the red and green outlines.

Figure D.9. **Additional in-the-wild examples.** The target objects and occluders are marked with the red and green outlines.