DeOcc-1-to-3: 3D De-Occlusion from a Single Image via Self-Supervised Multi-View Diffusion

Yansong Qu[1], Shaohui Dai[1], Xinyang Li[1], Yuze Wang[2], You Shen[1], Shengchuan Zhang[1*], Liujuan Cao[1]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China
[2]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
quyans@stu.xmu.edu.cn, zsc_2016@xmu.edu.cn

Abstract

Reconstructing 3D objects from a single image is a long-standing challenge, particularly under real-world occlusions. While recent diffusion-based view synthesis models can generate consistent novel views from a single RGB image, they generally assume fully visible inputs and struggle when parts of the object are occluded, leading to inconsistent views and degraded 3D reconstruction quality. To address this limitation, we propose DeOcc-1-to-3, an end-to-end framework for occlusion-aware multi-view generation. Our method directly synthesizes six structurally consistent novel views from a single partially occluded image, enabling downstream 3D reconstruction without requiring prior inpainting or manual annotations. We design a self-supervised training pipeline that leverages occluded–unoccluded image pairs and pseudo-ground-truth views to guide structure-aware completion and view consistency. Without modifying the original architecture, we fully fine-tune the diffusion model to jointly learn completion and multi-view generation. Additionally, we introduce the first benchmark for occlusion-aware reconstruction, covering diverse occlusion levels, object categories, and mask patterns, providing a standardized evaluation protocol.

Project page — https://quyans.github.io/DeOcc123/

1 Introduction

Reconstructing a complete 3D object from a single image remains a long-standing challenge in computer vision. Recent advances in diffusion-based multi-view generation models (Liu et al. 2023; Shi et al. 2023a; Xu et al. 2024; Shi et al. 2023b) enable the synthesis of structurally consistent novel views from a single RGB image, providing high-quality inputs for downstream 3D reconstruction. However, these models typically assume the object is fully visible. In real-world scenarios, images often contain partial occlusions caused by clutter, object interactions, or limited viewpoints, rendering this assumption impractical. Occlusions present a significant challenge for both view synthesis and 3D modeling. When parts of an object are occluded, existing methods often fail to accurately infer its geometry and appearance, leading to inconsistent novel views and incomplete or broken 3D reconstructions.

A common workaround is a two-stage pipeline: apply 2D inpainting (Ozguroglu et al. 2024; Xu, Zhang, and Shi 2024; Dogaru, Özer, and Egger 2024; Zhan et al. 2020, 2024) to complete the occluded regions, then apply view synthesis or 3D reconstruction to the completed image. However, this approach suffers from three key limitations: (1) 2D inpainting lacks 3D priors, often producing geometrically inconsistent completions that propagate into the view synthesis stage, compromising structural consistency across views; (2) view synthesis models remain unaware of the hallucinated regions' uncertainty, making them prone to artifacts and implausible geometry in occluded areas; (3) the decoupled design prevents joint optimization of completion and reconstruction, leading to error accumulation across stages.

To address these challenges, we propose an end-to-end occlusion-aware multi-view generation framework that directly synthesizes six structurally consistent views from a single occluded image. Without modifying the original architecture, we fully fine-tune the model to jointly learn completion and view synthesis in a unified process. We introduce a self-supervised training strategy by constructing paired occluded and unoccluded images through random 2D occlusions. A pretrained multi-view diffusion model provides six-view pseudo-ground-truths to supervise training, enabling structure-aware completion and view-consistent generation. We also establish a comprehensive benchmark for occlusion-aware reconstruction, covering diverse occlusion levels, object categories, and masking patterns, with standardized protocols for quantitative evaluation.

In summary, our main contributions are as follows:

- We introduce the task of occlusion-aware multi-view generation and propose an end-to-end framework that directly synthesizes structurally consistent novel views from a single occluded image, enabling 3D de-occlusion.
- We develop a self-supervised training paradigm using paired occluded-unoccluded images with pseudo-ground-truth supervision, enabling structure-aware learning without manual annotations. Our method fully fine-tunes the multi-view diffusion model without architectural changes, ensuring seamless compatibility with downstream 3D reconstruction frameworks.

*Corresponding author.

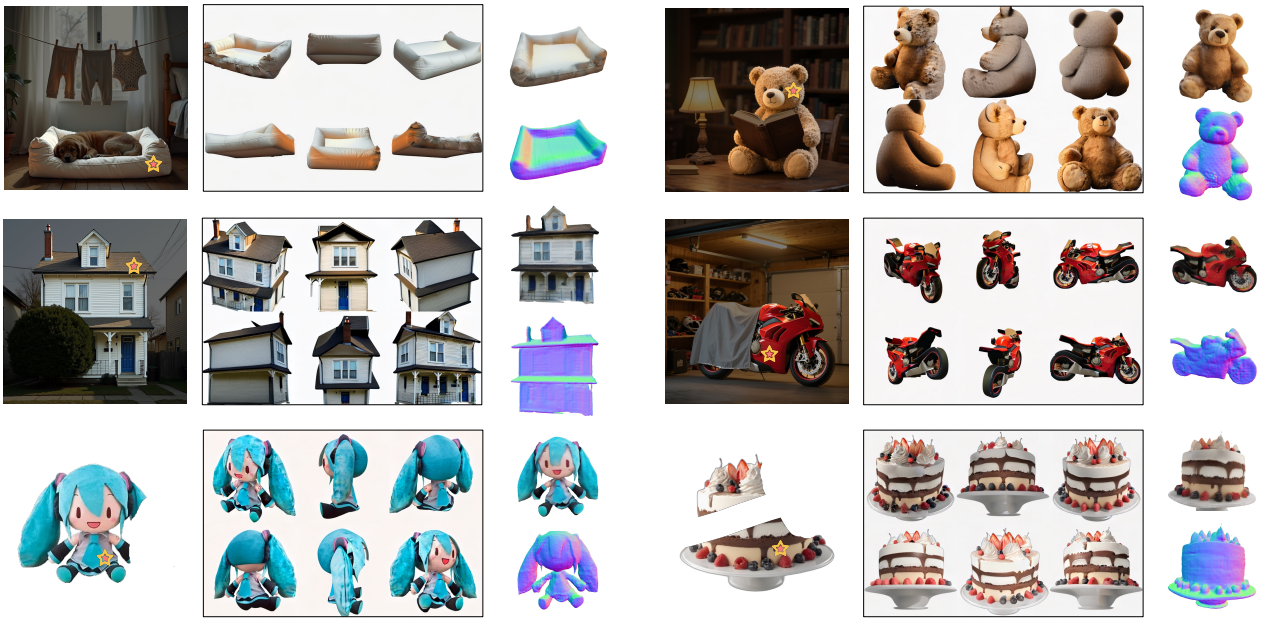

Figure 1: **DeOcc-1-to-3** takes a single occluded image (left) as input and synthesizes structurally consistent multi-view de-occluded images (middle). These outputs can be seamlessly integrated into various 3D reconstruction or generation frameworks to produce accurate meshes and surface normals (right). The proposed pipeline demonstrates generalization across diverse object categories and occlusion scenarios.

- We establish the first benchmark for occlusion-aware 3D reconstruction, covering diverse occlusion levels, object categories, and masking patterns, with standardized evaluation metrics.

2 Related Works

2.1 Single-image to 3D Representations

Early methods for single-image 3D reconstruction aimed to predict explicit or implicit representations, such as meshes, point clouds, or signed distance fields (SDFs)(Newcombe et al. 2011; Park et al. 2019). However, these approaches often struggle with complex geometries and occlusions. With the advent of neural rendering, NeRF(Mildenhall et al. 2021; Wang et al. 2023a; Qu, Wang, and Qi 2023; Huang et al. 2024) enabled high-fidelity 3D reconstruction via differentiable volumetric rendering, and PixelNeRF (Yu et al. 2021) extended this capability to single-image inputs. To improve efficiency, recent methods like 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023; Shen et al. 2025; Wang et al. 2025; Qu et al. 2024; Dai et al. 2025; Wang, Wang, and Qi 2024) leverage explicit Gaussian primitives for real-time rendering and editing. Nevertheless, directly regressing NeRF or 3DGS representations from a single view remains challenging due to inherent geometric ambiguities and limited multi-view consistency.

With the rise of powerful image-conditioned diffusion models, DreamFusion (Poole et al. 2022) and its successors (Li et al. 2024; Qu et al. 2025; Wang et al. 2023b) introduced Score Distillation Sampling (SDS), which leverages a pretrained 2D text-to-image diffusion model as an implicit energy function to optimize 3D representations. This paradigm enables text-driven 3D generation without explicit 3D supervision. However, SDS-based methods often suffer from high computational costs, multi-face inconsistencies, and saturated appearances. An alternative direction directly fine-tunes diffusion models to synthesize structurally consistent multi-view images from a single input (Liu et al. 2023), exemplified by Zero-1-to-3 (Liu et al. 2023), Zero123++ (Shi et al. 2023a), and MVDream (Shi et al. 2023b). These approaches enable efficient and high-quality 3D reconstruction via standard mesh-based pipelines. Nevertheless, they typically assume fully visible inputs and fail when the target object is partially occluded. This reveals a key limitation: existing models lack the capacity to infer and complete occluded structures—an essential ability for real-world scenarios where occlusions are prevalent.

2.2 2D Amodal Completion

2D amodal completion aims to recover the full shape and appearance of objects partially occluded in images (Li et al. 2025). Early approaches (Kimia, Frankel, and Popescu 2003; Silberman et al. 2014) rely on geometric heuristics, such as Euler spirals and Bézier curves, to extrapolate occluded boundaries based on predefined occlusion orders. However, these methods are restricted to simple shapes and lack robustness in complex real-world scenarios. Later

works (Yan et al. 2019; Zhou et al. 2021) employ supervised learning on synthetic datasets but are typically constrained to specific object categories and occlusion patterns. More recently, advances in generative models have enabled several methods (Zhan et al. 2024; Ozguroglu et al. 2024; Lee, Benes, and Yeh 2025; Li et al. 2025) to tackle amodal completion via powerful image generation frameworks, such as Pix2Gestalt (Ozguroglu et al. 2024) and SynergyAmodal (Li et al. 2025), achieving promising zero-shot performance.

2.3 3D De-occlusion

Early attempts at 3D de-occlusion relied on 2.5D depth completion (Zhang and Funkhouser 2018; Ma, Cavalheiro, and Karaman 2018) or template-based human mesh fitting (Alldieck et al. 2018; Pavlakos et al. 2019), which typically require RGB-D inputs, body priors, or multi-frame observations. While effective in constrained settings, these methods struggle to generalize across diverse object categories and complex occlusion patterns. Recent approaches explore generative strategies. CHROME (Dutta et al. 2025) employs pose-conditioned diffusion to synthesize multi-view images of occluded humans, followed by Gaussian splatting for 3D reconstruction. OccFusion (Sun et al. 2024) renders coarse human meshes and refines them using diffusion-based inpainting. Slice3D (Wang et al. 2024) predicts cross-sectional slices from occluded inputs and assembles them into 3D volumes.

Despite these advances, prior methods often depend on dense supervision, task-specific priors, or specialized architectures, limiting their scalability and generalization. In contrast, our method directly generates structure-consistent novel views from a single occluded image and integrates seamlessly with off-the-shelf reconstruction pipelines.

3 Method

This section introduces our occlusion-aware multi-view generation framework, covering training data construction (Sec. 3.3), self-supervised fine-tuning (Sec. 3.4), integration with 3D reconstruction pipelines (Sec. 3.5), and the construction of a benchmark dataset for evaluating 3D de-occlusion reconstruction (Sec. 3.6).

3.1 Preliminaries: Zero123++

We build upon Zero123++ (Shi et al. 2023a), a diffusion-based multi-view generation model that synthesizes six novel views from a single RGB image. The output is formatted as a 3×2 tiled image corresponding to six predefined camera poses. Given an input image I_{input}, the model is trained with a velocity-based denoising objective under the diffusion framework:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t \mid I_{\text{input}})\|^2\right], \quad (1)$$

where x_t is the noisy latent at timestep t, and ϵ_θ denotes the noise predicted by the model.

Zero123++ integrates local conditioning via scaled reference attention and global conditioning via CLIP image embeddings, ensuring strong spatial coherence and semantic alignment across views. Moreover, as it is trained with explicit multi-view supervision, Zero123++ inherently learns to maintain view consistency, making it a robust and architecture-agnostic backbone for our occlusion-aware fine-tuning framework.

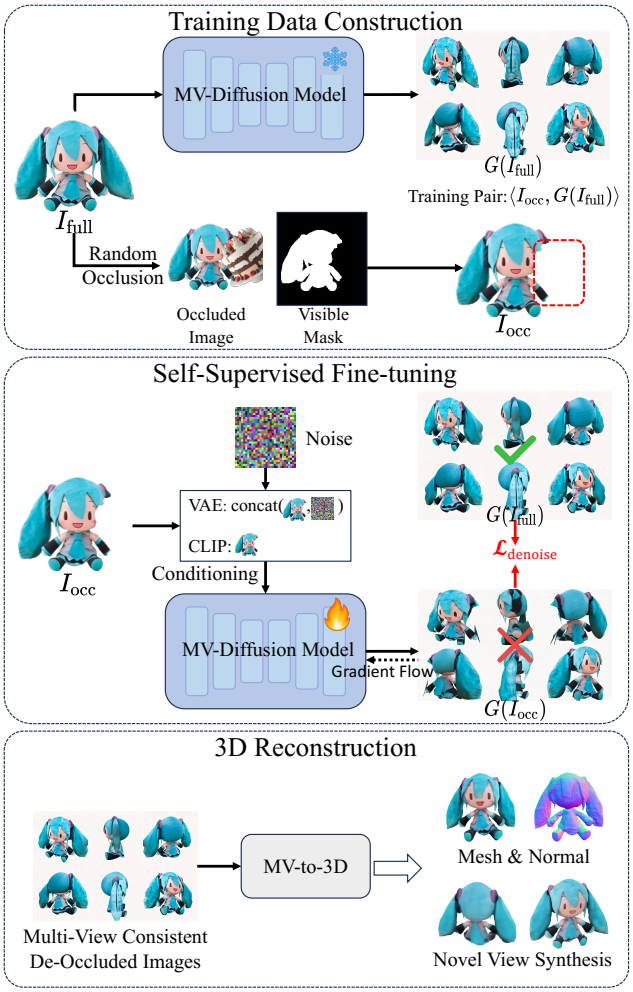

Figure 2: Overview of DeOcc-1-to-3. **Top**: Occluded images I_{occ} are generated by applying random occlusions to full images I_{full}. A frozen multi-view diffusion model produces six-view pseudo-ground-truths $G(I_{\text{full}})$, forming training pairs $\langle I_{\text{occ}}, G(I_{\text{full}})\rangle$. **Middle**: The student model is fully fine-tuned to predict consistent novel views $G(I_{\text{occ}})$. **Bottom**: The predicted six-view images are fed into downstream reconstruction models for 3D reconstruction.

3.2 Overall Framework

Previous works (Ozguroglu et al. 2024; Liu et al. 2024) typically follow a two-stage pipeline that first completes occluded images via 2D inpainting, then reconstructs 3D models. However, these approaches often yield incomplete 3D reconstructions due to the lack of explicit 3D structural reasoning in the 2D completion stage. Inspired by MVDream (Shi et al. 2023b) and Zero123++(Shi et al. 2023a), we propose a native 3D de-occlusion framework

based on a multi-view diffusion model. Our method first constructs occlusion-augmented training pairs with pseudo multi-view supervision, then fine-tunes a view synthesis model to produce occlusion-aware novel views. The synthesized images are subsequently passed to a 3D reconstruction module (e.g., InstantMesh(Xu et al. 2024)) to recover a complete 3D mesh. Our pipeline is fully annotation-free, category-agnostic, and robust across various occlusion patterns. Specifically, the model takes a single occluded image as input and generates six predefined novel views with structural and appearance consistency, which facilitates downstream 3D reconstruction.

As illustrated in Figure 2, the overall pipeline comprises three stages: (1) Construct occlusion-augmented training pairs with pseudo multi-view supervision; (2) Fine-tune a multi-view diffusion model to enable occlusion-aware generation; (3) Feed the generated views into a 3D reconstruction module to obtain a complete 3D model.

3.3 Training Data Construction

The scarcity of real-world 3D data makes it difficult to obtain large-scale, consistent occlusion datasets for training. Synthetic data (Hu et al. 2019) can be produced, but domain gaps hinder generalization to real scenes. In contrast, real 2D image data is abundant and diverse. To leverage this, we adopt a two-stage strategy: first constructing a large-scale 2D occlusion dataset, and then using a pretrained multi-view diffusion model to generate 3D-consistent pseudo-ground-truth views for supervision.

2D Occlusion Data Construction. We construct occlusion-aware image pairs using the SA-1B dataset (Kirillov et al. 2023) and the Segment Anything Model (SAM) (Kirillov et al. 2023) for foreground segmentation. A randomly selected segmented object is overlaid onto a natural background to synthesize occluded images. This process produces: Raw image containing the complete foreground object: I_{raw}; Foreground object mask: M_{full}; Occluded composite image: I_{mix}; Mask of the visible (unoccluded) part of the target object: M_{occ}. We then compute the paired foreground images as:

$$I_{\text{full}} = I_{\text{raw}} \odot M_{\text{full}},$$
$$I_{\text{occ}} = I_{\text{mix}} \odot M_{\text{occ}},$$

where $\odot$ denotes element-wise multiplication.

To ensure data quality, we filter out samples where the foreground object is inherently incomplete (Ozguroglu et al. 2024), or where the complete foreground lies at the image boundary. This curation ensures compatibility with the input distribution of the pretrained multi-view diffusion model. The final 2D dataset comprises 100K samples.

3D Occlusion Data Construction. We feed the clean image I_{full} into a pretrained multi-view diffusion model G (used as a teacher) to generate six-view pseudo-ground-truth images $G(I_{\text{full}})$. This enables the construction of paired training samples $\langle I_{\text{occ}}, G(I_{\text{full}})\rangle$ for training. Then, we use I_{occ} as the conditional input to fine-tune the same model (student) to synthesize the corresponding novel views.

To enhance robustness and prevent overfitting to any specific occlusion patterns, we further augment the training set in two ways: (1) We apply both dilation and erosion to the occlusion masks, simulating a wider range of occlusion severities and boundary shapes. This allows the model to learn from both heavier and lighter occlusion scenarios; these augmented samples are included in the training set to improve robustness; (2) We include a subset of unoccluded foreground pairs $\langle I_{\text{full}}, G(I_{\text{full}})\rangle$ as identity pairs in the training data. This prevents the model from over-predicting completions when the input object is already complete, thereby preserving the fidelity of fully visible inputs.

After filtering out defective or ambiguous samples, the final 3D de-occlusion dataset comprises approximately 40K high-quality training pairs, covering a broad range of object categories, occlusion levels, and masking patterns. This pipeline is entirely self-supervised and requires no manual annotations such as segmentation masks or depth maps.

3.4 Model Structure and Training Strategy

As illustrated in Figure 2, we adopt a diffusion-based multi-view generation backbone G (e.g., Zero123++) without altering its architecture. The model takes a single occluded RGB image I_{occ} as input and outputs a 3×2 concatenated image representing six predefined novel views, denoted as $G(I_{\text{occ}})$. Notably, no text prompts are used during occlusion-aware generation, adhering to our design principle of maximizing usability and generalization across diverse categories and occlusion types.

Training is performed with a standard denoising objective:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t \mid I_{\text{occ}})\|^2\right],$$

where x_t is a noisy version of the pseudo-ground-truth $G(I_{\text{full}})$ at timestep t, and ϵ_θ is the predicted noise.

We fine-tune the entire U-Net, including residual blocks and attention modules, with the AdamW optimizer. The noise schedule and both local (reference attention) and global (CLIP-based image encoder) conditioning remain identical to the original model, ensuring stability and consistency during training.

To further improve the robustness and visual fidelity of the generated views, we maintain an Exponential Moving Average (EMA) of the model weights θ during training:

$$\theta_{\text{EMA}} \leftarrow \beta \cdot \theta_{\text{EMA}} + (1-\beta)\cdot\theta, \tag{2}$$

with a decay rate of $\beta = 0.9999$. The EMA weights are used at inference time to enhance generation stability and quality.

3.5 3D Reconstruction Integration

The proposed occlusion-aware multi-view generation framework synthesizes six-view RGB images that are both geometrically consistent and structurally complete. These images can serve as universal inputs for a variety of 3D reconstruction or generation methods (Xu et al. 2024; Xiang et al. 2025; Zhao et al. 2025; Zhang et al. 2025b), enabling the recovery of 3D models in diverse representations, such as NeRF (Mildenhall et al. 2021), 3D Gaussian Splatting (Kerbl et al. 2023), and mesh-based surfaces (Zhang et al. 2024, 2025a).

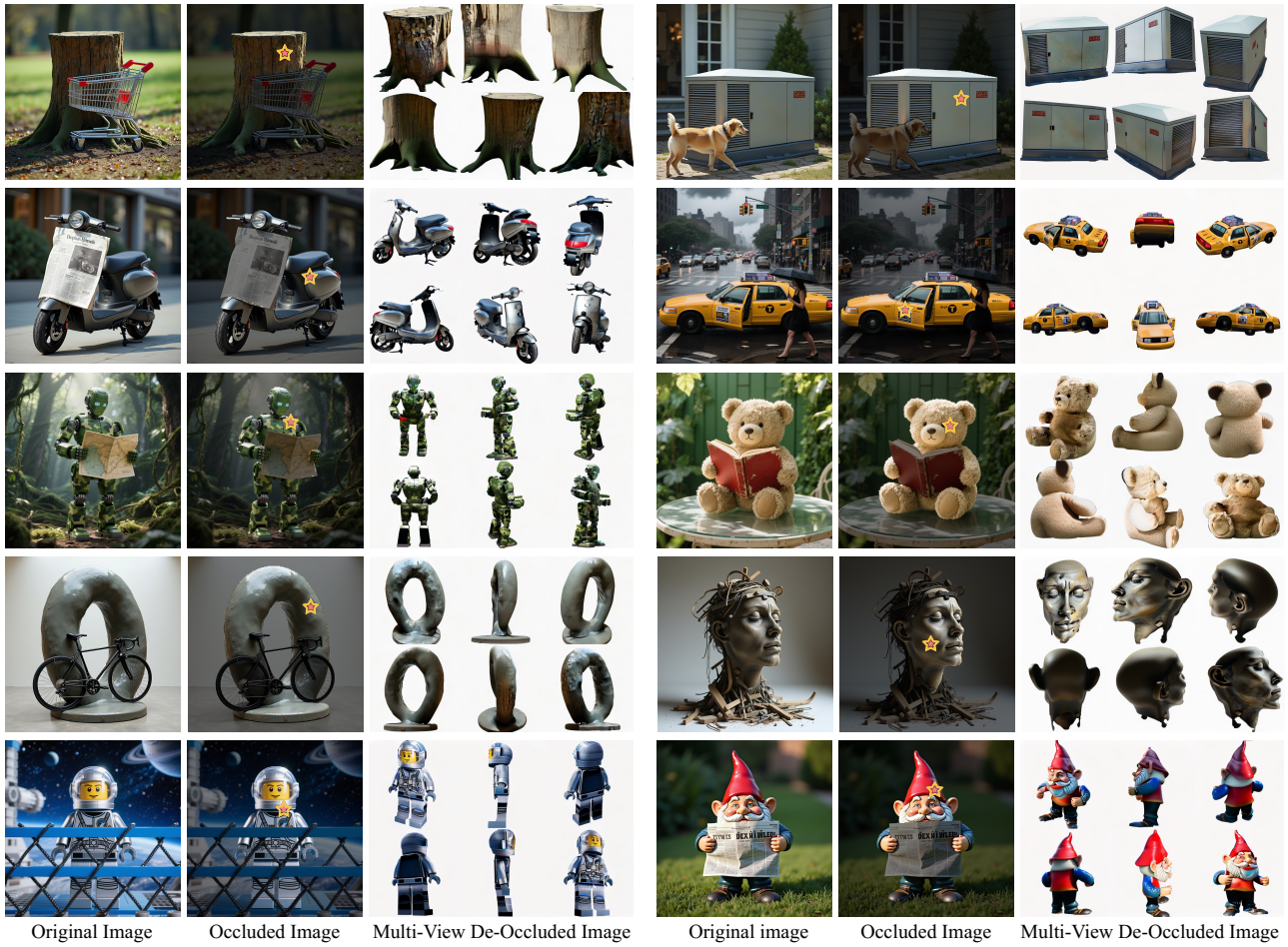

Figure 3: Qualitative de-occlusion results on diverse objects. Each triplet shows (left) the original image, (middle) the occluded input, and (right) our multi-view de-occluded output (six views). Our method recovers coherent geometry and texture across various shapes, materials, and occlusion types.

In our implementation, we adopt InstantMesh (Xu et al. 2024) as a representative reconstruction backend for its high efficiency. InstantMesh extracts tri-plane features from the input view, fuses these into a volumetric representation, and decodes the volume into a 3D mesh. Notably, since our model maintains the original multi-view diffusion output format, it integrates seamlessly into existing pipelines without architectural modifications. Experiments show that our occlusion-aware generator significantly improves reconstruction quality under occlusions, producing more complete, consistent, and smooth surfaces.

3.6 Occ-LVIS Benchmark

To systematically evaluate 3D object reconstruction under occlusions, we construct a dedicated benchmark derived from the Objaverse-LVIS dataset (Deitke et al. 2023). We select high-quality assets from the LVIS subset and render them into multi-view image sequences under controlled occlusion settings.

Each object is first rendered from a canonical frontal view and then occluded by randomly selected foreground objects. We subsequently obtain: (1) Six canonical views: with azimuth angles of $\{30°, 90°, 150°, 210°, 270°, 330°\}$ and alternating elevation angles of $\{30°, -20°\}$; (2) Four random views: to evaluate generalization; (3) Geometry-complete mesh ground truths: for evaluating 3D reconstruction accuracy and completeness.

Level	Occlusion Ratio Range	Proportion (%)
L0	0%	8.0%
L1	0% – 10%	11.3%
L2	10% – 20%	25.5%
L3	20% – 30%	28.6%
L4	30% – 40%	19.4%
L5	$\geq 40\%$	7.2%

Table 1: Occlusion level statistics in the Occ-LVIS.

To facilitate fine-grained analysis, we stratify the benchmark into six occlusion levels based on the occlusion ra-

tio—the proportion of the target object's visible area. As detailed in Table 1, these levels range from fully visible objects to cases with more than 40% occlusion.

Each object in the benchmark is accompanied by its occlusion-level label, enabling fine-grained analysis of model robustness across different occlusion scenarios. The benchmark and evaluation protocols are designed to provide a standardized and comprehensive platform for occlusion-aware 3D generation methods.

4 Experiments

4.1 Implementation Details

We fine-tune the entire U-Net of the multi-view diffusion backbone with all residual blocks and attention modules updated. Training follows the original noise schedule and conditioning mechanisms (reference attention and CLIP-based global conditioning). We use the AdamW optimizer with an initial learning rate of 2×10^{-5}, a batch size of 32, and train for 15k steps. An Exponential Moving Average (EMA) of model weights is maintained with a decay rate of $\beta = 0.9999$, and the EMA weights are used for inference.

All experiments are conducted on NVIDIA A100-40G GPUs, with the full training process requiring approximately 96 A100 GPU-hours.

4.2 Evaluation Setup

Datasets. We conduct quantitative experiments on the proposed Occ-LVIS benchmark (Sec. 3.6). Additionally, we perform qualitative assessments on Internet-sourced images and synthetically generated samples to demonstrate generalization to diverse scenarios.

Baselines. We compare our method against the following baselines: (1) Image-to-3D pipeline (*3D-R*): Vanilla Zero123++ (Shi et al. 2023a) followed by InstantMesh (Xu et al. 2024), without any occlusion handling. (2) 2D de-occlusion method + Image-to-3D pipeline (*P2G-R*): Pix2Gestalt (Ozguroglu et al. 2024) for 2D amodal completion, followed by Zero123++ and InstantMesh. (3) 3D de-occlusion pipeline (*Ours*).

Evaluation Metrics. We assess the quality of the generated results from both 2D and 3D perspectives. We first align the generated meshes with the ground truth within a unit sphere centered at the origin under consistent coordinate systems. The aligned meshes are then rendered into multi-view images. For 2D visual evaluation, we compare these rendered images with the ground truth using CLIP Score (Radford et al. 2021), FID (Heusel et al. 2017), and KID (Bińkowski et al. 2018). For 3D geometric evaluation, we report Chamfer Distance (CD), F1-Score, and Volume IoU (V-IoU).

4.3 Qualitative Results

As shown in Figure 3, we present qualitative results demonstrating the effectiveness of our occlusion-aware multi-view generation framework across a wide range of object categories and occlusion scenarios. The experiments are conducted on unseen test samples, including both synthetic occlusions and real-world occluded images. Notably, several cases involve severe occlusion exceeding 40% of the object area, as well as challenging visual structures. For instance, the third column shows a "teddy bear holding a book", and the fourth displays an "art installation stacked with firewood". Despite these complexities, our method successfully recovers coherent object geometry and appearance across all six views. The synthesized outputs exhibit strong multi-view consistency, minimal artifacts, and plausible completions of the occluded regions. These results highlight the robustness and generalization ability of our fine-tuned model, even under high occlusion and out-of-distribution conditions.

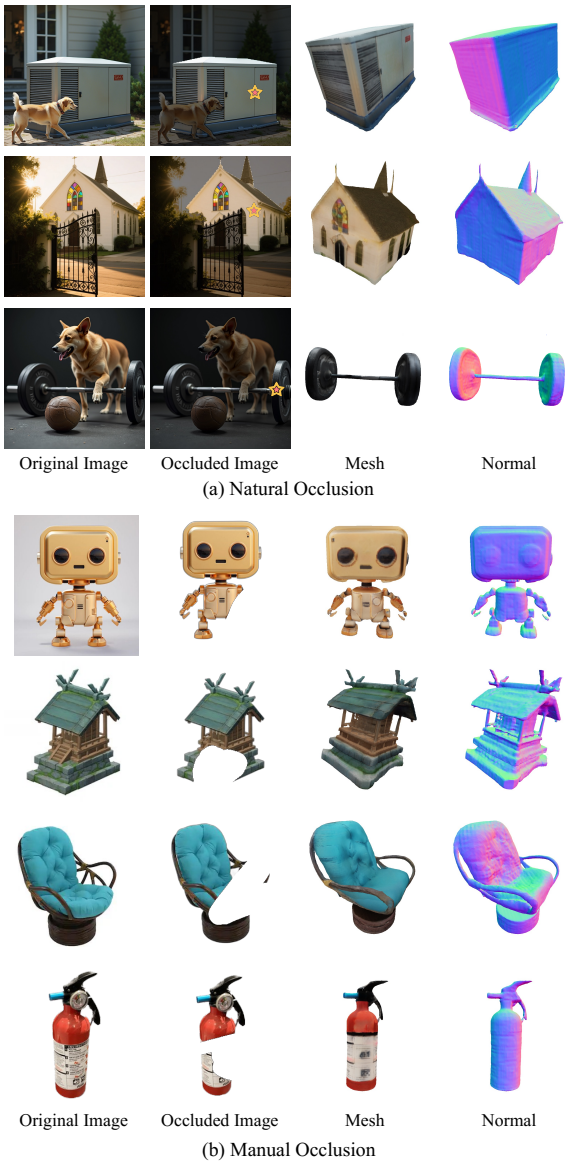

Figure 4: 3D reconstruction results using predicted de-occlusion images under (a) natural and (b) manual occlusions. For each object, we show the reconstructed mesh, and the surface normals. Our method recovers complete geometry and accurate normals under both real and synthetic occlusions, demonstrating multi-view consistent de-occlusion.

Method	2D Metrics			3D Metrics			Efficiency			User Study
	CLIP ↑	FID ↓	KID ↓	CD ↓	F1-Score ↑	V-IoU ↑	Time (s) ↓	FLOPs (T) ↓	Params (G) ↓	Preference (%)
3D-R	0.741	21.749	0.012	0.011	0.382	0.195	**11.09**	**325**	**1.5**	-
P2G-R	0.777	12.838	0.003	0.009	0.462	0.341	20.38	332	2.7	37.2
Ours	**0.785**	**11.757**	**0.002**	**0.007**	**0.491**	**0.361**	**11.09**	**325**	**1.5**	**62.8**
Abl-A	0.734	23.029	0.013	0.011	0.369	0.195	-	-	-	-
Abl-B	0.777	14.539	0.004	0.009	0.458	0.334	-	-	-	-

Table 2: Quantitative, efficiency, and user study comparison of our method, baselines (3D-R, P2G-R), and ablation variants (Abl-A, Abl-B) on the Occ-LVIS benchmark. Abl-A: without mask dilation-erosion and whole-object augmentation. Abl-B: without whole-object augmentation. The best results are highlighted in bold.

We present qualitative comparison results on the proposed Occ-LVIS benchmark.[1] The *3D-R* baseline fails to recover missing content due to the lack of an occlusion completion mechanism, resulting in broken or incomplete 3D reconstructions. While the *P2G-R* pipeline can complete most occluded regions, it suffers from error accumulation across stages. This leads to noticeable texture degradation (e.g., distorted car paint in the first row), as well as occasional failure cases (e.g., missing water bottle content in the sixth row). In contrast, our method consistently generates high-quality, structure-consistent de-occluded views across diverse occlusion types and object categories, enabling more faithful and complete 3D reconstruction.

To further validate the quality and multi-view consistency of our multi-view de-occluded images, we leverage InstantMesh (Xu et al. 2024) to reconstruct 3D geometry from the predicted six-view outputs. We also visualize the surface normals of the resulting mesh for qualitative inspection. As shown in Figure 4, our method enables high-quality 3D reconstruction across various scenarios, producing geometrically consistent multi-view completions and realistic object surfaces. These results confirm that our generated views not only complete occluded content plausibly, but also serve as reliable inputs for downstream mesh-based 3D modeling.

4.4 Quantitative Results

2D and 3D Results. As shown in Table 2, our method achieves the best performance across all metrics. For 2D evaluation, it obtains the lowest FID and KID, indicating superior perceptual quality and distribution alignment with real images. Additionally, our method achieves the highest CLIP similarity score, demonstrating stronger semantic alignment between the generated views and the input. For 3D evaluation, our method achieves the lowest CD, indicating accurate surface reconstruction, and the highest F1-Score and V-IoU, reflecting better geometric completeness and volumetric consistency. Compared to the baseline (*3D-R*) and the two-stage completion pipeline (*P2G-R*), our occlusion-aware generation framework consistently improves both multi-view quality and 3D reconstruction performance. These results confirm that our multi-view generation preserves structural integrity across views and provides high-quality supervision for downstream 3D reconstruction.

Efficiency Results. Our method achieves multi-view consistent de-occlusion and 3D reconstruction from a single occluded image in just 11 seconds, as shown in Table 2. It maintains efficiency comparable to the *3D-R* pipeline while reducing time cost, FLOPs, and parameter count compared to the *P2G-R* pipeline, providing an efficient and practical solution for real-world 3D de-occlusion applications.

User Study. We collected survey responses from 41 users, with each questionnaire containing comparisons of 30 cases, totaling 1230 responses. Users were asked to select the more visually pleasing and complete de-occluded objects. Our method achieves 62.8% preference score, indicating that it is preferred by users.

4.5 Ablation Study

As shown in Table 2, we conduct an ablation study with two variants to evaluate the effects of mask dilation-erosion and whole-object augmentation on performance. The mask dilation-erosion strategy helps the model better handle occlusion extents and boundary details, while whole-object augmentation enhances the model's robustness when the input is already a fully visible object. This prevents the model from overfitting during training and indiscriminately producing completed results for any input, which would otherwise undermine the capabilities of the original base model.

5 Limitations and Conclusion

Limitations. While our method performs well across diverse occlusion scenarios, it still struggles with extremely heavy occlusions where structural cues are severely lacking. Additionally, when the input contains occluded text, the generated results may exhibit meaningless or distorted characters, inherited from the pretrained model's limited capacity for textual representation.

Conclusion. We present DeOcc-1-to-3, a self-supervised framework for multi-view 3D de-occlusion from a single occluded image. By fine-tuning a multi-view diffusion model, it jointly completes missing structures and generates consistent novel views. Our method integrates seamlessly with existing 3D reconstruction pipelines and generalizes across occlusion scenarios. We also introduce the first benchmark for standardized evaluation of 3D de-occlusion.

[1] Additional qualitative examples are available in the extended arXiv version: https://arxiv.org/abs/2506.21544.

Acknowledgments

This work was supported by the National Science Fund for Distinguished Young Scholars (No.62025603 and No.62525605), National Natural Science Foundation of China (No. U21B2037, U22B2051, No. U23A20383, No. 62176222, No. 62176226, No. 62272401, No. 62576300).

References

Alldieck, T.; Magnor, M. A.; Xu, W.; Theobalt, C.; and Pons-Moll, G. 2018. Video based reconstruction of 3D people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8387–8397.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Dai, S.; Qu, Y.; Li, Z.; Li, X.; Zhang, S.; and Cao, L. 2025. Training-Free Hierarchical Scene Understanding for Gaussian Splatting with Superpoint Graphs. *arXiv preprint arXiv:2504.13153*.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13142–13153.

Dogaru, A.; Özer, M.; and Egger, B. 2024. Generalizable 3D Scene Reconstruction via Divide and Conquer from a Single View. *arXiv preprint arXiv:2404.03421*.

Dutta, A.; Zheng, M.; Gao, Z.; Planche, B.; Choudhuri, A.; Chen, T.; Roy-Chowdhury, A. K.; and Wu, Z. 2025. CHROME: Clothed Human Reconstruction with Occlusion-Resilience and Multiview-Consistency from a Single Image. *arXiv preprint arXiv:2503.15671*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hu, Y.-T.; Chen, H.-S.; Hui, K.; Huang, J.-B.; and Schwing, A. G. 2019. Sail-vos: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3105–3115.

Huang, C.; Li, X.; Zhang, S.; Cao, L.; and Ji, R. 2024. NeRF-DetS: Enhancing Multi-View 3D Object Detection with Sampling-adaptive Network of Continuous NeRF-based Representation. *arXiv e-prints*, arXiv–2404.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kimia, B. B.; Frankel, I.; and Popescu, A.-M. 2003. Euler spiral for shape completion. *International journal of computer vision*, 54(1): 159–182.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Lee, J. J.; Benes, B.; and Yeh, R. A. 2025. Tuning-Free Amodal Segmentation via the Occlusion-Free Bias of Inpainting Models. *arXiv preprint arXiv:2503.18947*.

Li, X.; Lai, Z.; Xu, L.; Qu, Y.; Cao, L.; Zhang, S.; Dai, B.; and Ji, R. 2024. Director3d: Real-world camera trajectory and 3d scene generation from text. *Advances in Neural Information Processing Systems*, 37: 75125–75151.

Li, X.; Yi, C.; Lai, J.; Lin, M.; Qu, Y.; Zhang, S.; and Cao, L. 2025. SynergyAmodal: Deocclude Anything with Text Control. *arXiv preprint arXiv:2504.19506*.

Liu, R.; Wu, R.; Van Hoorick, B.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9298–9309.

Liu, Z.; Liu, Q.; Chang, C.; Zhang, J.; Pakhomov, D.; Zheng, H.; Lin, Z.; Cohen-Or, D.; and Fu, C.-W. 2024. Object-level Scene Deocclusion. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Ma, F.; Cavalheiro, G.; and Karaman, S. 2018. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 1–8. IEEE.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Newcombe, R. A.; Izadi, S.; Hilliges, O.; Molyneaux, D.; Kim, D.; Davison, A. J.; Kohi, P.; Shotton, J.; Hodges, S.; and Fitzgibbon, A. 2011. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, 127–136. Ieee.

Ozguroglu, E.; Liu, R.; Surís, D.; Chen, D.; Dave, A.; Tokmakov, P.; and Vondrick, C. 2024. pix2gestalt: Amodal segmentation by synthesizing wholes. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3931–3940. IEEE Computer Society.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 165–174.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10975–10985.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Qu, Y.; Chen, D.; Li, X.; Li, X.; Zhang, S.; Cao, L.; and Ji, R. 2025. Drag your gaussian: Effective drag-based editing with score distillation for 3d gaussian splatting. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, 1–12.

Qu, Y.; Dai, S.; Li, X.; Lin, J.; Cao, L.; Zhang, S.; and Ji, R. 2024. Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5328–5337.

Qu, Y.; Wang, Y.; and Qi, Y. 2023. SG-NeRF: Semantic-guided Point-based Neural Radiance Fields. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 570–575. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Shen, Y.; Zhang, Z.; Li, X.; Qu, Y.; Lin, Y.; Zhang, S.; and Cao, L. 2025. Evolving High-Quality Rendering and Reconstruction in a Unified Framework with Contribution-Adaptive Regularization. *arXiv preprint arXiv:2503.00881*.

Shi, R.; Chen, H.; Zhang, Z.; Liu, M.; Xu, C.; Wei, X.; Chen, L.; Zeng, C.; and Su, H. 2023a. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*.

Shi, Y.; Wang, P.; Ye, J.; Long, M.; Li, K.; and Yang, X. 2023b. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*.

Silberman, N.; Shapira, L.; Gal, R.; and Kohli, P. 2014. A contour completion model for augmenting surface reconstructions. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, 488–503. Springer.

Sun, A.; Xiang, T.; Delp, S.; Fei-Fei, L.; and Adeli, E. 2024. OccFusion: Rendering Occluded Humans with Generative Diffusion Priors. *arXiv preprint arXiv:2407.00316*.

Wang, Y.; Lira, W.; Wang, W.; Mahdavi-Amiri, A.; and Zhang, H. 2024. Slice3d: Multi-slice occlusion-revealing single view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9881–9891.

Wang, Y.; Wang, J.; Gao, R.; Qu, Y.; Duan, W.; Yang, S.; and Qi, Y. 2025. Look at the Sky: Sky-Aware Efficient 3D Gaussian Splatting in the Wild. *IEEE Transactions on Visualization and Computer Graphics*, 31(5): 3481–3491.

Wang, Y.; Wang, J.; and Qi, Y. 2024. WE-GS: An In-the-wild Efficient 3D Gaussian Representation for Unconstrained Photo Collections. arXiv:2406.02407.

Wang, Y.; Wang, J.; Qu, Y.; and Qi, Y. 2023a. Rip-nerf: Learning rotation-invariant point-based neural radiance field for fine-grained editing and compositing. In *Proceedings of the 2023 ACM international conference on multimedia retrieval*, 125–134.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023b. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36: 8406–8441.

Xiang, J.; Lv, Z.; Xu, S.; Deng, Y.; Wang, R.; Zhang, B.; Chen, D.; Tong, X.; and Yang, J. 2025. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21469–21480.

Xu, J.; Cheng, W.; Gao, Y.; Wang, X.; Gao, S.; and Shan, Y. 2024. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*.

Xu, K.; Zhang, L.; and Shi, J. 2024. Amodal completion via progressive mixed context diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9099–9109.

Yan, X.; Wang, F.; Liu, W.; Yu, Y.; He, S.; and Pan, J. 2019. Visualizing the invisible: Occluded vehicle segmentation and recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7618–7627.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4578–4587.

Zhan, G.; Zheng, C.; Xie, W.; and Zisserman, A. 2024. Amodal ground truth and completion in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28003–28013.

Zhan, X.; Pan, X.; Dai, B.; Liu, Z.; Lin, D.; and Loy, C. C. 2020. Self-supervised scene de-occlusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3784–3792.

Zhang, X.; Liu, Y.; Li, Y.; Zhang, R.; Liu, Y.; Wang, K.; Ouyang, W.; Xiong, Z.; Gao, P.; Hou, Q.; et al. 2025a. Tar3d: Creating high-quality 3d assets via next-part prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5134–5145.

Zhang, X.; Yin, B.-W.; Chen, Y.; Lin, Z.; Li, Y.; Hou, Q.; and Cheng, M.-M. 2024. Temo: Towards text-driven 3d stylization for multi-object meshes. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 19531–19540.

Zhang, X.; Zhou, Y.; Wang, K.; Wang, Y.; Li, Z.; Jiao, S.; Zhou, D.; Hou, Q.; and Cheng, M.-M. 2025b. AR-1-to-3: Single Image to Consistent 3D Object via Next-View Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 26273–26283.

Zhang, Y.; and Funkhouser, T. 2018. Deep depth completion of a single RGB-D image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 175–185.

Zhao, Z.; Lai, Z.; Lin, Q.; Zhao, Y.; Liu, H.; Yang, S.; Feng, Y.; Yang, M.; Zhang, S.; Yang, X.; et al. 2025. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*.

Zhou, Q.; Wang, S.; Wang, Y.; Huang, Z.; and Wang, X. 2021. Human de-occlusion: Invisible perception and recovery for humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3691–3701.