This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Dehallu3D: Hallucination-Mitigated 3D Generation from Single Image via Cyclic View Consistency Refinement

Xiwen Wang Shichao Zhang Ruowei Wang Mao Li Chenyu Zhou
Ji-Zhe Zhou* Qijun Zhao* Hailun Zhang
Sichuan University, China

Abstract

*Large 3D reconstruction models have revolutionized the 3D content generation field, enabling broad applications in virtual reality and gaming. Just like other large models, large 3D reconstruction models suffer from hallucinations as well, introducing structural outliers (e.g., odd holes or protrusions) that deviate from the input data. However, unlike other large models, hallucinations in large 3D reconstruction models remain severely underexplored, leading to malformed 3D-printed objects or insufficient immersion in virtual scenes. Such hallucinations majorly originate from that existing methods reconstruct 3D content from sparsely generated multi-view images which suffer from large viewpoint gaps and discontinuities. To mitigate hallucinations by eliminating the outliers, we propose **Dehallu3D** for 3D mesh generation. Our key idea is to design a balanced multi-view continuity constraint to enforce smooth transitions across dense intermediate viewpoints, while avoiding over-smoothing that could erase sharp geometric features. Therefore, Dehallu3D employs a plug-and-play optimization module with two key constraints: (i) adjacent consistency to ensure geometric continuity across views, and (ii) adaptive smoothness to retain fine details. We further propose the Outlier Risk Measure (ORM) metric to quantify geometric fidelity in 3D generation from the perspective of outliers. Extensive experiments show that Dehallu3D achieves high-fidelity 3D generation by effectively preserving structural details while removing hallucinated outliers.*

1. Introduction

Generating 3D contents from 2D images is crucial for applications ranging from augmented reality and virtual reality, to 3D printing. Despite significant advancements [6, 8, 20, 29, 33, 44], the process of generating accurate 3D content from a single image continues to pose substantial challenges.

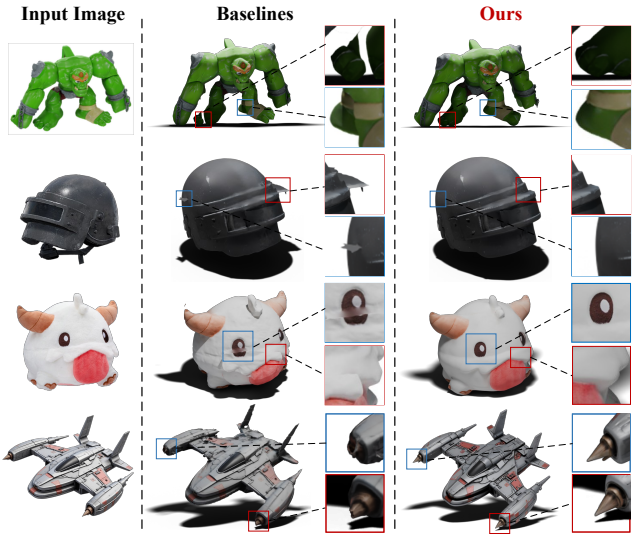


Figure 1. Our **Dehallu3D** generates high-quality and high-fidelity 3D meshes, effectively mitigating mesh outliers while adaptively preserving sharp features. The red and blue boxes highlight noticeable outlier regions in the Baseline method, contrasted with the corresponding regions in Dehallu3D (Ours) to showcase improvements.

Diffusion models [24] have achieved profound milestones in 2D generation, paving the way for 3D synthesis. DreamFusion and subsequent work [11, 22, 25] introduced Score Distillation Sampling (SDS) to adapt pretrained 2D diffusion models for 3D object generation. However, SDS-based methods suffer from issues like the Janus problem [22] due to the absence of robust 3D priors. To address this, transformer-based Large Reconstruction Models (LRMs) [4, 17, 31, 40] integrate 3D data for feedforward generation, but still face challenges in achieving accurate reconstruction. As a solution to the above issues, a common paradigm trains 2D diffusion models to generate multi-view images for 3D reconstruction via neural rendering or multi-view stereo. [15, 30, 37, 42, 43].

*Corresponding authors.

These large 3D reconstruction models, similar to other large models, frequently encounter hallucinations. Such hallucinations typically manifest as outputs that diverge from ground truth, as shown in Figure 1, where they give rise to outliers in reconstructed contents. However, in contrast to other large models, the hallucinations observed in 3D reconstruction models remain underexplored by current research. Consequently, these outliers pose significant challenges for downstream applications. In 3D printing, even minor geometric deviations can lead to manufacturing failures. Similarly, in gaming experiences, visual inconsistencies introduced by such outliers severely disrupt user immersion. We argue that such hallucinations originates from the inconsistent sparse multi-view images, which suffer from large viewpoint gaps and poor continuity. This misleads the model into introducing intricate details that are not supported by the input data. Inspired by the principle of interpolation, we address this by inserting dense intermediate viewpoints to bridge gaps and enforce continuity across views.

To this end, we propose **Dehallu3D**, a novel framework for single-image 3D mesh generation that effectively eliminates hallucination-induced outliers. The core design is a plug-and-play module termed Cyclic View Consistency Refinement (CVCR). Specifically, CVCR renders dense views beyond the orthographic views and enforces consistency across adjacent views to bridge the discontinuities between sparse views. This consistency supervision helps mitigate hallucinations by eliminating outliers. As increasing view density inevitably leads to higher computational cost, we count on a balanced angular interval to reach both reconstruction quality and efficiency. Moreover, to prevent oversmoothing during view consistency enforcement, we apply adaptive smoothness constraints to preserve sharp geometric features such as spikes.

To the best of our knowledge, we are the first to explicitly cope with the outliers in the context of 3D reconstruction. Since there are no available evaluation metrics to assess the geometry fidelity from this perspective, we design an Outlier Risk Measure (ORM) metric to assess the quality of reconstructed 3D contents, particularly from the viewpoint of outliers. Overall, the contributions of this work are:

- We propose **Dehallu3D**, a novel framework for high-fidelity 3D mesh generation from a single image that effectively mitigates hallucinations by eliminating outliers.
- We introduce a plug-and-play module **CVCR** that ensures the consistency of adjacent views and smoothness variations in the mesh, thereby preserving fine details while suppressing hallucinated outliers.
- We devise the **Outlier Risk Measure (ORM)**, a new metric tailored to evaluate geometric fidelity.
- Our experiments show that Dehallu3D achieves superior performance, enabling high-quality single-image-to-3D generation.

2. Related work

2.1. Singe image-to-3D generation

Due to the inherent lack of geometric information in 2D images, 3D generation from a single image is an ill-posed problem. DreamFusion [22] adopts SDS techniques to generate 3D assets. It distills 3D geometry and appearance from large and pre-trained image diffusion models [24]. Thus, it exploits the powerful 2D priors while avoiding the reliance on 3D assets which are less abundant than 2D image datasets. However, SDS-based methods [11, 13, 25, 26, 39] may face the multi-face problems, also known as the "Janus" problem due to the lack of 3D priors. Zero123 [14] finetunes Stable Diffusion [23] to synthesis novel views on given camera poses and address this problem. These Optimization-based methods are constrained by slow generation speeds, and the following works [2, 9] improve the results in speed. 3D native generators [4, 17, 31, 32, 38, 40, 41] directly generate 3D content from a single 2D image, bypassing multi-view supervision. While they avoid multi-view consistency issues, they rely heavily on the model's ability to learn complex 3D structures from limited 2D input. This can lead to challenges in complex or uncommon scenarios, where the model may struggle to produce high-quality results and the generated 3D content may lack detail or accuracy. In contrast, One-2-3-45 [12] proposes to integrate 2D generative models with multi-view 3D reconstruction, thus achieving a better performance in both quality and efficiency. Many other works follow this paradigm and improve the results a lot, in either speed or quality [8, 21, 30, 35, 37, 43]. This paradigm has gained popularity due to its high reconstruction quality. However, it still faces a critical issue: large gaps and discontinuities between the multi-views. This leads to a degradation in the quality of subsequent reconstruction. Therefore, we propose to improve the reconstruction quality by bridging the gap between multi-view images.

2.2. Mesh representation for reconstruction

3D representations are significant for 3D reconstruction tasks, as they directly influence the efficiency and quality of reconstructed geometry and texture. Considerable progress has been made in various 3D representation techniques, such as Neural Radiance Fields (NeRF) [18, 19], point clouds [10], and 3D Gaussian [5, 17]. However, meshes continue to dominate as the preferred representation, owing to their well-established rendering pipeline. Some studies [1, 15, 16, 37, 42, 43] utilize techniques such as Structure from Motion (SfM), Multi-View Stereo (MVS), and mesh surface reconstruction for 3D modeling. Subsequently, differentiable meshes [27, 28, 31, 36] are proposed for 3D op-

timization tasks. A differentiable mesh is a hybrid 3D representation that combines both implicit and explicit surface representations, such as Signed Distance Functions (SDFs) and meshes. Although alternative 3D representations offer distinct benefits, they incur significant memory overhead. In contrast, our approach retains meshes as the 3D representation, leveraging their efficient rendering and compact storage properties.

3. Method

3.1. Overview

In this section, we propose **Dehallu3D**, a novel single-image 3D mesh generation framework that produces high-fidelity meshes while effectively mitigating outliers during reconstruction. As illustrated in Figure 2, the pipeline of Dehallu3D begins with high-resolution multi-view generation. Given an input image, we first generate four orthographic multi-view color images along with their corresponding normal maps. We then apply a fast initialization method to generate a rough mesh. To progressively refine the mesh, Dehallu3D adopts a two-stage optimization pipeline: (1) **Coarse Mesh Reconstruction.** This stage focuses on globally correcting the mesh topology using differentiable rendering. We introduce a surface exposure-weighted normal loss to prioritize geometry constraints from high-visibility regions, ensuring reliable global structure. (2) **Cyclic View Consistency Refinement (CVCR).** To mitigate mesh outliers and recover fine details, we propose CVCR, a plug-and-play refinement module that enforces consistency across cyclically adjacent viewpoints and adaptive smoothness noise to prevent over-smoothing. Through this coarse-to-fine strategy, Dehallu3D achieves robust and detailed mesh reconstruction from a single image.

3.2. Mesh initialization

The quality of mesh initialization critically influences subsequent optimization stages and final results. Therefore, an effective mesh initialization method should establish approximately correct topological structures. To this end, following [37], we adopt a fast initialization process to directly capture the complete topological connection features of the visible area through the front and back views. Specifically, given an input image, we first use the existing high-resolution multi-view generators to obtain multi-view color images and their corresponding normal maps. Next, we utilize the normal maps from the front and back views to improve the reliability of depth estimation through integration and random rotations, and combine this with Poisson reconstruction technology to generate a high-quality initialized mesh.

3.3. Coarse mesh reconstruction

To effectively improve the quality of the mesh and mitigate outliers caused by error accumulation due to multi-view inconsistencies during reconstruction, we adopt a two-stage optimization pipeline. In the coarse reconstruction stage, rapidly correcting the overall shape of the mesh is crucial. We perform mesh reconstruction and optimization based on differentiable rendering. The loss function at the current stage is defined as

$$\mathcal{L}_{coarse} = \mathcal{L}_{mask} + \mathcal{L}_{normal} + \mathcal{L}_{SE}. \tag{1}$$

In $\mathcal{L}_{coarse}$, we propose a *surface exposure-weighted normal loss* $\mathcal{L}_{SE}$ to effectively fuse normal information from multiple views, which is defined as

$$\begin{aligned} \mathcal{L}_{SE} &= \sum_{v \in \mathcal{V}} \sum_{i}^{4} \epsilon_i^v \cdot |N_v^R - N_i^v|_2^2, \\ \epsilon_i^v &= m_i^v \cdot \frac{A_i^v}{\sum_j m_j^v A_j^v}. \end{aligned} \tag{2}$$

Here, $\mathcal{V}$ denotes the set of mesh vertices, N_v^R is the extracted surface normal of vertex v, and N_i^v represents the reference normal of vertex v in view i. The visibility mask $m_i^v \in 0, 1$ indicates whether vertex v is visible in view i. The projected surface area A_i^v is the sum of the projected areas of the triangular faces associated with vertex v in view i, reflecting the degree of observability of the vertex in that view. Views with larger projected areas typically correspond to core regions of the mesh, providing stronger geometric constraints that help quickly correct the coarse mesh's global structure. The weighting term ϵ_i^v is derived from the projected area of vertex v in view i, dynamically prioritizing views with higher projected areas while suppressing the influence of views with low visibility. This enables robust global shape optimization during the coarse reconstruction stage.

The mask-based loss is defined as

$$\mathcal{L}_{mask} = \sum_{i=1}^{4} \|M_i - M_i^R\|_2^2. \tag{3}$$

Here, M_i and M_i^R denote the alpha channel values of the generated color image under view i and the corresponding rendered color image, respectively.

The normal-based loss is defined as

$$\mathcal{L}_{normal} = \sum_{i=1}^{4} \|N_i - N_i^R\|_2^2. \tag{4}$$

Here, N_i denotes the generated normal map under view i and N_i^R denotes the rendered normal map under view i. $\mathcal{L}_{mask}$ and $\mathcal{L}_{normal}$ are both common MSE losses in

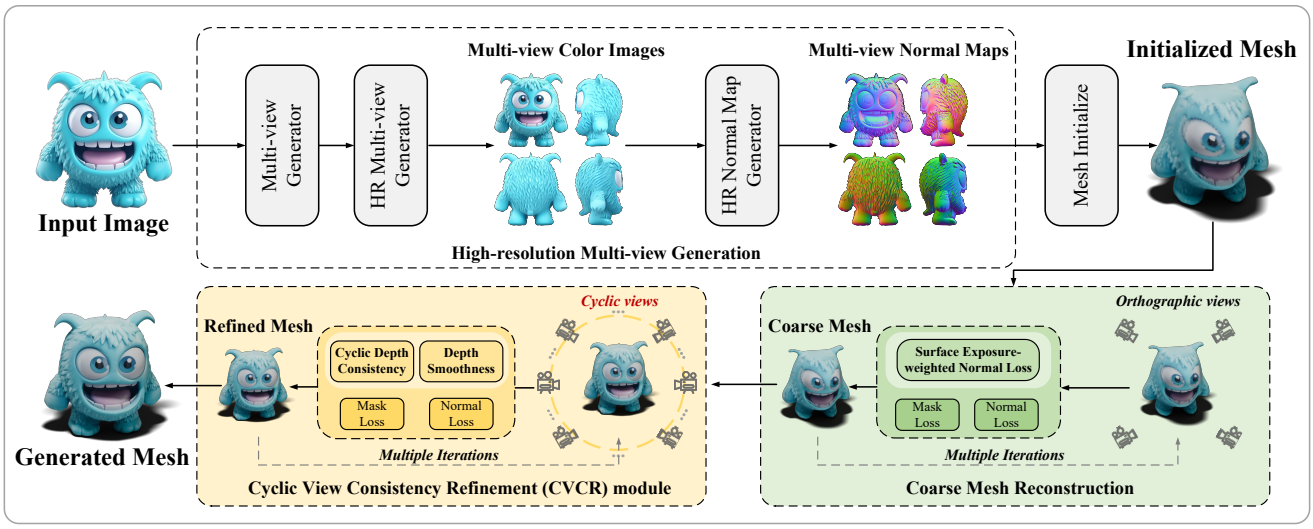


Figure 2. **Overview of Dehallu3D.** Dehallu3D first generate orthographic multi-view color images and corresponding normal maps, which are used to initialize a coarse mesh. Next, a globally plausible mesh is quickly constructed through the Coarse Mesh Reconstruction stage. Finally, the proposed **Cyclic View Consistency Refinement (CVCR)** module is employed to mitigate outliers and further refine the mesh.

3D mesh reconstruction. After the above optimization, the overall geometric structure of the mesh has been significantly improved. Next, we refine the mesh to further mitigate outliers during reconstruction.

3.4. Cyclic view consistency refinement

To further enhance geometric fidelity and mitigate outliers, we propose a **Cyclic View Consistency Refinement (CVCR)** module. Unlike general orthographic view-consistent methods, CVCR explicitly models the cyclic relationship among adjacent views over a full 360° rotation, enforcing pairwise alignment between neighboring depth maps. This module is designed to be *plug-and-play*, making it easily integrable into a wide range of mesh reconstruction pipelines, regardless of the mesh initialization strategy.

The CVCR module optimizes the mesh using the following objective:

$$\mathcal{L}_{CVCR} = \mathcal{L}_{mask} + \mathcal{L}_{normal} + \lambda_1 \mathcal{L}_{DC} + \lambda_2 \mathcal{L}_{DS}. \quad (5)$$

Here, $\mathcal{L}_{mask}$ and $\mathcal{L}_{normal}$ are inherited from the coarse reconstruction stage. The two newly introduced terms, $\mathcal{L}_{DC}$ and $\mathcal{L}_{DS}$, are described below.

Empirical evidence shows that depth maps of real-world objects maintain consistency across adjacent small-angle views when observed through a full 360° rotation. In contrast, abrupt depth discontinuities in such views frequently lead to obvious outliers in 3D meshes. Building on this, we introduce a *cyclic depth consistency loss* $\mathcal{L}_{DC}$ to encourage the reconstructed mesh to better adhere to real-world physical properties by enforcing consistency between depth maps rendered from adjacent viewpoints with small angular differences. This helps mitigate outliers in the mesh. The loss is defined as

$$\begin{aligned} \mathcal{L}_{DC} &= \sum_{i=1}^{V} \left(1 - \Delta(D_i^R, D_{i \bmod V+1}^R)\right), \\ \Delta(D_i^R, D_j^R) &= \mathrm{SSIM}(D_i^R, D_j^R) \cdot \mathrm{CS}(D_i^R, D_j^R). \end{aligned} \quad (6)$$

Here, V denotes the total number of views (we set $V = 72$, yielding $360°/72 = 5°$ angular differences). D_i^R is the rendered depth map from view i, and $D_{i \bmod V+1}^R$ is its adjacent cyclic view. The similarity function $\Delta(\cdot)$ integrates the Structural Similarity Index Measure (SSIM) [34] and Cosine Similarity (CS) to evaluate both structural and directional alignment. This design ensures robustness against pixel misalignment in adjacent depth maps caused by angular deviations between views.

While $\mathcal{L}_{DC}$ enforces global consistency, it may overconstrain regions with valid sharp features, leading to oversmoothing. To mitigate this, we introduce a complementary *depth smoothness loss* $\mathcal{L}_{DS}$ that adaptively preserves sharp features in the mesh while mitigating hallucinated outliers based on image gradients, which is defined as

$$\begin{aligned} \mathcal{L}_{DS} &= \sum_{i=1}^{V} \sum_{j,k}^{pixel} |\nabla D_i^{R(j,k)}| \cdot w_i^{j,k}, \\ w_i^{j,k} &= \exp(-\|\nabla I_i^{R(j,k)}\|_2). \end{aligned} \quad (7)$$

Here, I_i^R represents the rendered color image for view i, and D_i^R denotes its corresponding depth map. The gradient magnitude of the depth map at pixel (j,k) is defined as $|\nabla D_i^{R(j,k)}| = \sqrt{(\partial_x D_i^{R(j,k)})^2 + (\partial_y D_i^{R(j,k)})^2}$, capturing the intensity of depth changes. Similarly, the gradient

magnitude of the color image at pixel (j,k) is defined as $\|\nabla I_i^{R(j,k)}\|_2 = \sqrt{\|\partial_x I_i^{R(j,k)}\|_2^2 + \|\partial_y I_i^{R(j,k)}\|_2^2}$, reflecting color variations associated with geometric features. The weight $w_i^{j,k}$ dynamically adjusts the strength of the smoothness constraint. In regions with high color image gradients where indicates potential sharp features, the weight reduces the smoothness penalty, preserving depth discontinuities.

4. Experiment

4.1. Datasets and evaluation metrics

Our experiments focus on both mesh appearance quality and mesh geometric quality. Following previous work [19, 35, 37, 43], we perform our experiments on the Google Scanned Objects (GSO) dataset [3]. For a fair experimental comparison, all objects in the dataset are rendered at a resolution of 512×512 with Blender Cycles as input for all methods. All generated mesh results are normalized to the bounding box of $[-0.5, 0.5]$ to ensure alignment. For visual quality evaluation, we select elevation angles from $\{0, 15, 30\}$ and 8 evenly distributed azimuth angles to render each object generated from different methods, resulting in 24 views per object. We employ PSNR, SSIM, LPIPS, and Clip-Similarity (Clip-Sim) as metrics to evaluate the visual quality. For geometric quality evaluation, we utilize Chamfer Distance (CD) and F-Score as metrics. More implementation details are demonstrated in the supplementary material. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU.

4.2. Outlier risk measure

In 3D mesh reconstruction, mesh outliers impair geometric consistency and cause significant local errors. However, existing metrics often fail to capture the impact of these outliers.

To this end, we propose a novel evaluation metric based on Conditional Value-at-Risk (CVaR) [7] to measure the outlier degrees of the meshes. When computing CVaR, it is first necessary to determine the Value at Risk (VaR), that is, the maximum risk that may occur at a given probability level. Given a set of risk values $\varphi = \{r_i\}_{i=1}^{\mathcal{R}}$ under a discrete risk distribution $\mathcal{D}$, the discrete CVaR is defined as

$$\begin{aligned} \mathrm{VaR}_\xi(\varphi) &= \min\{r_i \in \varphi : \sum_{r_j \in \varphi} \mathbb{I}(r_i \geq r_j)\mathcal{D}(r_j) \geq \xi\}, \\ \mathrm{CVaR}_\xi(\varphi) &= \frac{1}{1-\xi} \sum_{r_i \in \varphi} \mathbb{I}(r_i \geq \mathrm{VaR}_\xi(\varphi))\mathcal{D}(r_i) r_i. \end{aligned} \tag{8}$$

Here, ξ denotes the confidence level, $\mathbb{I}$ denotes the indicator function, and $\mathcal{D}$ is set to the empirical distribution $\mathcal{D}(r_i) = 1/\mathcal{R}$. Referring to the hypothesis in [7], *the presence of outliers in point clouds will increase the level of tail risk in its point risk distribution*, we convert the mesh into point clouds and measure the outlier degrees of different meshes by computing the tail risk in their point clouds.

In CVaR, φ is used to quantify the risk of points in the point clouds. Given a 3D mesh $\mathcal{M}$ and its point cloud is denoted as $\mathcal{P} = \{p_i\}_{i=1}^{P}$. To evaluate this risk, we introduce an outlier scoring function $S(\mathcal{P})$. Specifically, $S(\mathcal{P})$ integrates both global and local outlier measures in the point cloud. The global measure, denoted as $S_g(\mathcal{P})$, is derived from the reconstruction loss of the VAE. Meanwhile, the local measure, $S_l(\mathcal{P})$, is computed based on the neighborhood density of the internal points, which is defined as

$$S_l(\mathcal{P}) = \frac{1}{P} \sum_{i=1}^{P} \frac{d_i^k}{\frac{1}{|\mathcal{N}_i|} \sum_{p_j \in \mathcal{N}_i} d_j^k}. \tag{9}$$

Here, $\mathcal{N}_i \subseteq \mathcal{P}$ denotes the nearest neighbors of $p_j \in \mathcal{P}$ and d_j^k denotes the distance from p_j to its k-th nearest neighbor. We compute the ratio of the average local density of a point and its neighbors to reflect the local outlier degree of the point in $\mathcal{P}$. The average outlier degrees of all points in $\mathcal{P}$ is taken as the local outlier degree $S_l(\mathcal{P})$ of the point cloud. We use the k-th distance to reflect the density of the point. The larger the k-th distance, the smaller the density of the point. The lower the local density of the point p_i, the more likely it is to be an outlier.

The final outlier scoring function is $S(\mathcal{P}) = S_l(\mathcal{P}) + \lambda S_g(\mathcal{P})$. We measure the risk of points ϕ by $S(\mathcal{P})$ and then compute the tail risk in ϕ as the final outlier score of $\mathcal{M}$. We refer to the outlier score obtained through the above process as the **Outlier Risk Measure (ORM)**. The greater the ORM, the more outliers the meshes have. We hope to generate meshes with low ORM values, that is, meshes with relatively few outliers.

4.3. Comparisons

4.3.1. Quantitative comparison

To validate the superiority of our method, we conduct comprehensive comparisons with SOTA methods. Considering the 3D representation method and open-source availability, we select SF3D [1], Unique3D [37], CRM [35], InstantMesh [40], TripoSR [31], Wonder3D [15] as comparison methods in the experiments. The quantitative comparison results are shown in Table 1. Specifically, Dehallu3D achieves the best performance across multiple metrics for both visual quality and geometric quality. The best performance of our method on all metrics highlights Dehallu3D, which exploits depth consistency and depth smoothness to boost the quality of the generated 3D meshes.

We also compare the ORM results across all methods, as shown in Figure 5. Our method achieves the lowest ORM value, while Unique3D exhibits the highest, consistent with

	Appearance metrics				Geometry metrics	
Methods	PSNR↑	SSIM↑	LPIPS↓	Clip-Sim↑	CD↓	F-Score↑
Wonder3D [15]	20.4963	0.8908	0.1851	0.6970	0.02183	0.3580
TripoSR [31]	20.5309	0.8897	0.1841	0.7146	0.02241	0.3847
InstantMesh [40]	20.8954	0.8903	0.1749	<u>0.7538</u>	0.02198	0.4046
CRM [35]	21.1265	0.8889	0.1720	0.7191	0.02163	0.3967
Unique3D [37]	20.9795	0.8882	0.1742	0.7493	0.02175	<u>0.4073</u>
SF3D [1]	<u>21.3257</u>	<u>0.8912</u>	<u>0.1537</u>	0.7463	<u>0.02144</u>	0.3765
Dehallu3D (Ours)	**21.8407**	**0.8966**	**0.1453**	**0.7753**	**0.02023**	**0.4212**

Table 1. Quantitative comparison results. We mark the best scores in **bold** and the second-best scores with an <u>underline</u>.

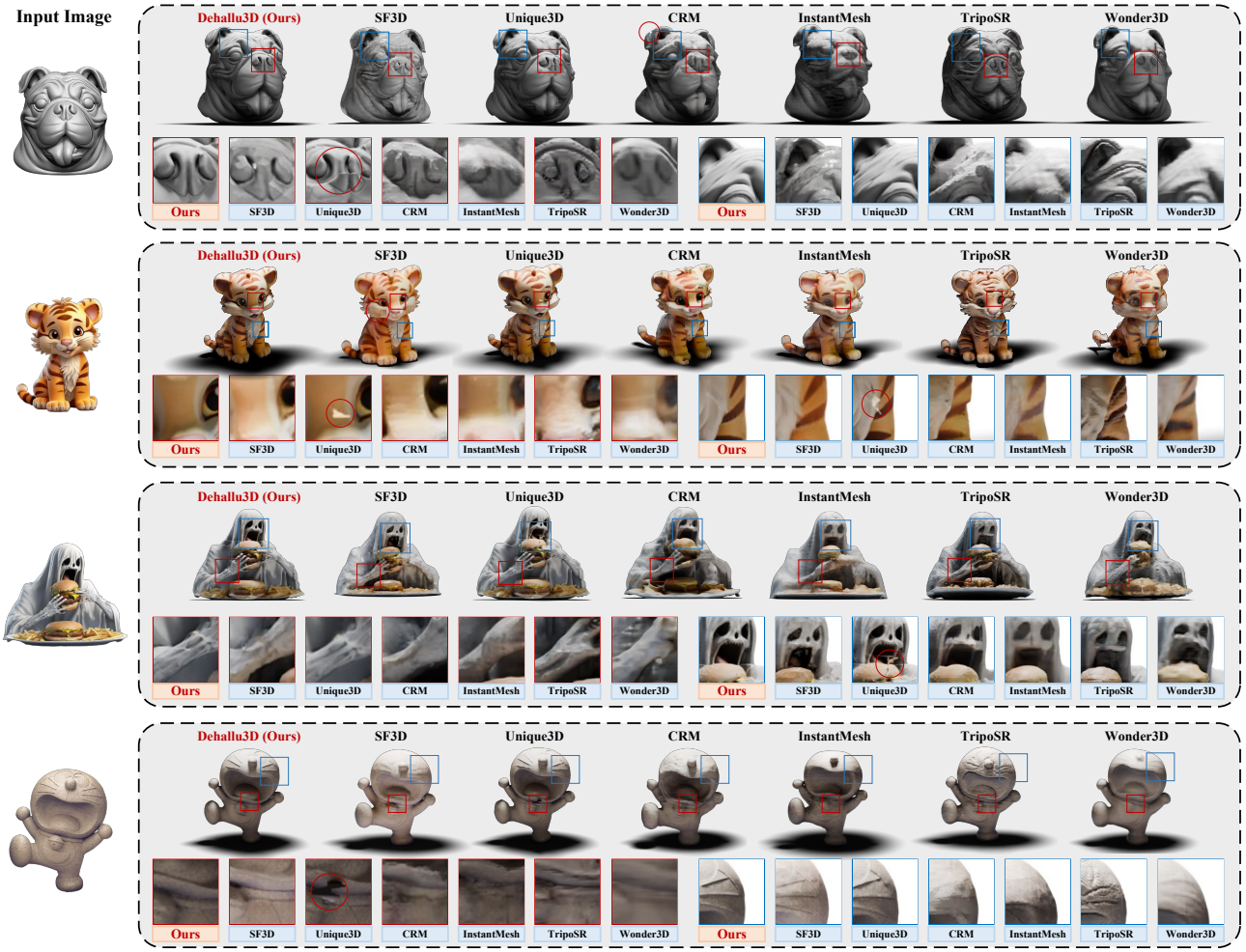


Figure 3. Qualitative comparison of different methods in mesh reconstruction. The red and blue boxes mark specific regions in the meshes of other methods for comparison with the corresponding optimized regions in the mesh generated by Dehallu3D, while red circles highlight defects in meshes.

its performance in both quantitative and qualitative experiments. We conducted experiments on the ORM-human perception correlation in the supplementary materials.

4.3.2. Qualitative comparison

In this section, we visually evaluate the performance of the proposed method and its effectiveness in 3D mesh generation tasks. The primary goal of the qualitative experiments

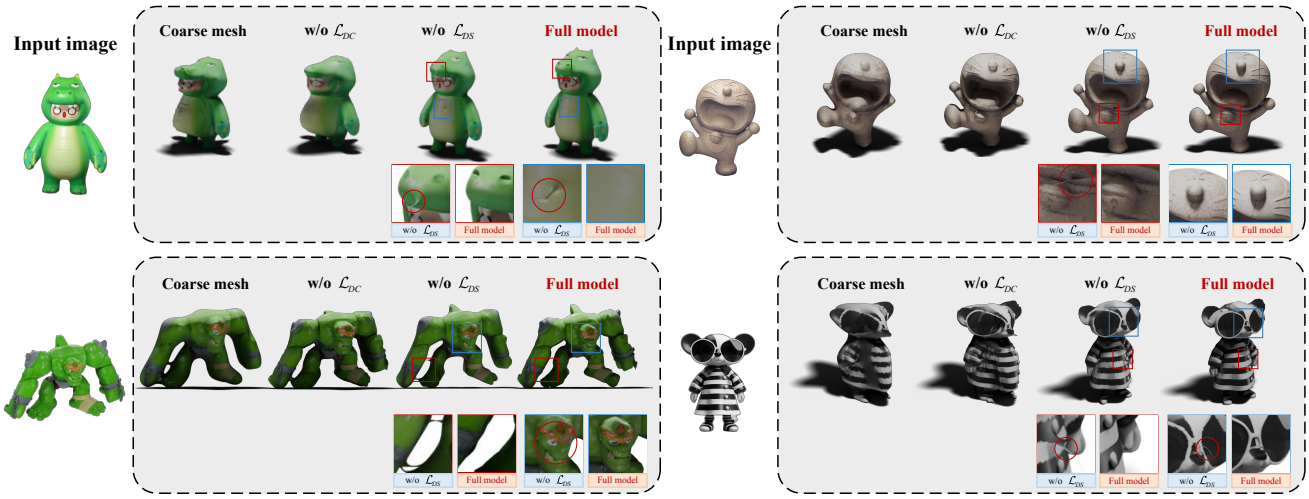


Figure 4. Qualitative comparison results for ablation study on $\mathcal{L}_{DC}$ and $\mathcal{L}_{DS}$ in CVCR module.

			Appearance metrics				Geometry metrics	
$\mathcal{L}_{SE}$	$\mathcal{L}_{DC}$	$\mathcal{L}_{DS}$	PSNR↑	SSIM↑	LPIPS↓	Clip-Sim↑	CD↓	F-Score↑
✗	✗	✗	16.8972	0.8474	0.2283	0.6315	0.02572	0.3201
✔	✗	✗	17.6534	0.8615	0.2102	0.6678	0.02449	0.3468
✗	✔	✗	20.9276	0.8823	0.1697	0.7392	0.02217	0.3879
✗	✗	✔	18.3124	0.8749	0.2019	0.6834	0.02332	0.3545
✗	✔	✔	21.1973	0.8894	0.1531	0.7403	0.02114	0.4097
✔	✔	✔	**21.8407**	**0.8966**	**0.1453**	**0.7753**	**0.02023**	**0.4212**

Table 2. Quantitative comparison results for ablation study on the proposed losses.

	Appearance metrics				Geometry metrics		Time
Angles	PSNR↑	SSIM↑	LPIPS↓	Clip-Sim↑	CD↓	F-Score↑	seconds↓
CVCR (3°)	**21.8628**	**0.8971**	0.1455	**0.7764**	0.02030	0.4203	208.1
CVCR (5°)	21.8407	0.8966	**0.1453**	0.7753	**0.02023**	**0.4212**	163.3
CVCR (10°)	21.3687	0.8879	0.1527	0.7591	0.02134	0.4118	129.5
CVCR (15°)	20.8415	0.8763	0.1671	0.7459	0.02221	0.3936	**110.8**

Table 3. Ablation study of angular intervals between adjacent views in the CVCR module.

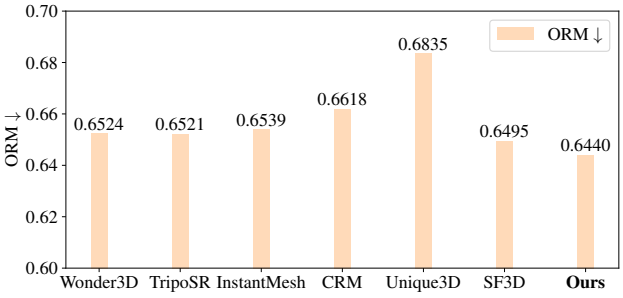


Figure 5. Comparison of ORM results across all methods.

is to demonstrate the model's capability in handling mesh outliers and to highlight its strengths in maintaining geometric consistency and preserving fine details through visual results. We select a set of example samples from existing studies. A comparative visualization of the meshes reconstructed by our method against other methods is presented in Figure 3. SF3D generally achieves desirable 3D mesh generation from input images, but the texture of the entire mesh remains relatively blurry. Compared to SF3D, Dehallu3D excels in capturing texture details in the mesh, enabling better reconstruction of detailed features in the input images.

Next, we further compare the meshes generated by Dehallu3D and Unique3D. While Unique3D demonstrates the

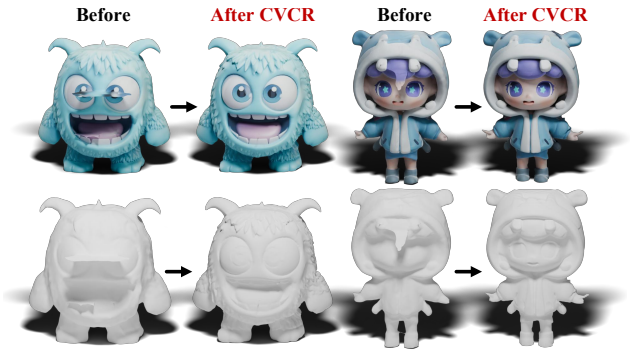


Figure 6. After applying the CVCR module to Unique3D.

ability to reconstruct texture details, it fails to consider the overall depth consistency and depth smoothness of the mesh, making it susceptible to mesh outliers during reconstruction. This issue in turn causes a decrease in the overall quality of the mesh. The remaining methods including CRM, InstantMesh, TripoSR, and Wonder3D, generate meshes that are inferior in quality compared to our method. The reconstruction results of these methods lack fine texture details and precise geometry. More qualitative experimental results are presented in the supplementary material.

4.4. Ablation study

The quantitative results in Table 2 demonstrate the synergistic impact of each proposed loss. Using only $\mathcal{L}_{SE}$ fails to mitigate mesh outliers. Introducing $\mathcal{L}_{DC}$ reduces outliers but can lead to over-smoothing of depth details due to the lack of $\mathcal{L}_{DS}$. Combining $\mathcal{L}_{DS}$ and $\mathcal{L}_{DC}$ significantly enhances mesh quality. However, omitting $\mathcal{L}_{SE}$ results in a poor-quality coarse mesh, limiting the effectiveness of subsequent optimization.

As shown in Figure 4, we further demonstrate the synergistic effect of $\mathcal{L}_{DC}$ and $\mathcal{L}_{DS}$ through qualitative comparisons, highlighting their contributions to reducing mesh outliers and enhancing mesh quality. Without $\mathcal{L}_{DS}$, the generated mesh exhibits noticeable protrusions and surface irregularities. Without $\mathcal{L}_{DC}$, the mesh quality is significantly degraded due to the absence of cross-view consistency constraints. Combining $\mathcal{L}_{DC}$ and $\mathcal{L}_{DS}$ enables synergistic optimization, substantially improving mesh quality.

As presented in Table 3, we analyze the impact of the angular interval for the dense views in our CVCR module. While larger intervals (e.g., 10° or 15°) degrade performance due to sparse supervision, a very dense 3° interval also leads to flawed results. Specifically, the 3° interval shows slightly worse geometric fidelity (CD, F-Score) and perceptual quality (LPIPS) compared to the 5° interval. We attribute this to potential over-smoothing; at such a small gap, the cyclic depth consistency loss ($\mathcal{L}_{DC}$) becomes overly strict, penalizing valid sharp geometric features by mistaking them for discontinuities. This effect appears to overpower the protection from our adaptive smoothness loss ($\mathcal{L}_{DS}$). Furthermore, a smaller angular interval directly increases the number of required rendering views, consequently leading to a higher time overhead. Critically, this increased time expenditure does not proportionally guarantee superior results, a trend consistent with our prior analysis. Therefore, prioritizing a trade-off between quality and efficiency, we adopted a balanced angular interval of $\mathbf{5}^\circ$, which offers an acceptable computational overhead and delivers satisfactory performance.

Finally, we integrate the proposed plug-and-play CVCR module into the Unique3D method, which suffers from noticeable mesh outliers caused by multi-view inconsistencies. As shown in Figure 6, the quality of the meshes in Unique3D significantly improved after incorporating the CVCR module.

5. Limitations

To achieve high fidelity, Dehallu3D introduces dense view rendering within its CVCR module, which inevitably leads to additional time overhead. This is a deliberate trade-off that prioritizes geometric accuracy over inference speed, making our method particularly suitable for accuracy-critical applications, such as 3D printing.

Nevertheless, we recognize the importance of efficiency in certain contexts. In the future, we plan to focus on improving reconstruction efficiency. Our efforts will target the optimization of this dense refinement stage, for instance, by enhancing parallel processing or developing more efficient algorithms for view rendering and comparison.

6. Conclusion

In this paper, we present **Dehallu3D** for single-image 3D reconstruction that mitigates model hallucinations by eliminating outliers by enforcing adjacent view consistency. The CVCR module is designed to be plug-and-play, rendering dense viewpoints to bridge gaps across sparse views, a common issue in multi-view generation methods. In addition, adaptive smoothness is applied to prevent over-smoothing caused by enforcing view consistency. To quantify geometric fidelity, we introduce a novel metric specifically designed to evaluate 3D quality by detecting outliers in meshes. Extensive experiments demonstrate that our method achieves the SOTA performance in both visual quality and geometric accuracy, significantly reducing the occurrence of outliers. And we validated that the proposed metric ORM aligns well with human perception in terms of hallucinated outlier assessment.

Acknowledgment

This work is supported by the National Natural Science Foundation of China, Young Scientists Fund (C Class)(No.62506251) and the National Natural Science Foundation of China (No.61773270, No.62176170).

References

[1] Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16240–16250, 2025. 2, 5, 6

[2] Yiwen Chen, Chi Zhang, Xiaofeng Yang, Zhongang Cai, Gang Yu, Lei Yang, and Guosheng Lin. It3d: Improved text-to-3d generation with explicit view synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1237–1244, 2024. 2

[3] 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. 5

[4] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 1, 2

[5] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[6] Jiyang Li, Lechao Cheng, Zhangye Wang, Tingting Mu, and Jingxuan He. Loopgaussian: creating 3d cinemagraph with multi-view images via eulerian motion field. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 476–485, 2024. 1

[7] Xinke Li, Junchi Lu, Henghui Ding, Changsheng Sun, Joey Tianyi Zhou, and Yeow Meng Chee. Pointcvar: Risk-optimized outlier removal for robust 3d point cloud classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21340–21348, 2024. 5

[8] Xingyi Li, Yizheng Wu, Jun Cen, Juewen Peng, Kewei Wang, Ke Xian, Zhe Wang, Zhiguo Cao, and Guosheng Lin. icontrol3d: An interactive system for controllable 3d scene generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10814–10823, 2024. 1, 2

[9] Yuhan Li, Yishun Dou, Yue Shi, Yu Lei, Xuanhong Chen, Yi Zhang, Peng Zhou, and Bingbing Ni. Focaldreamer: Text-driven 3d editing via focal-fusion assembly. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3279–3287, 2024. 2

[10] Stefan Lionar, Xiangyu Xu, Min Lin, and Gim Hee Lee. Nu-mcc: Multiview compressive coding with neighborhood decoder and repulsive udf. *Advances in Neural Information Processing Systems*, 36:63011–63022, 2023. 2

[11] Fangfu Liu, Diankun Wu, Yi Wei, Yongming Rao, and Yueqi Duan. Sherpa3d: Boosting high-fidelity text-to-3d generation via coarse 3d prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20763–20774, 2024. 1, 2

[12] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[13] Pengkun Liu, Yikai Wang, Hang Xiao, Hongxiang Xue, Xinzhou Wang, and Fuchun Sun. Isotropic3d: Image-to-3d generation based on a single clip embedding. *Knowledge-Based Systems*, page 115367, 2026. 2

[14] 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*, pages 9298–9309, 2023. 2

[15] 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9970–9980. IEEE, 2024. 1, 2, 5, 6

[16] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 2

[17] Longfei Lu, Huachen Gao, Tao Dai, Yaohua Zha, Zhi Hou, Junta Wu, and Shu-Tao Xia. Large point-to-gaussian model for image-to-3d generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10843–10852, 2024. 1, 2

[18] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[19] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2, 5

[20] Xiaochao Pan, Jiawei Yao, Hongrui Kou, Tong Wu, and Canran Xiao. Harmonicnerf: Geometry-informed synthetic view augmentation for 3d scene reconstruction in driving scenarios. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5987–5996, 2024. 1

[21] Hao-Yang Peng, Jia-Peng Zhang, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. Charactergen: Efficient 3d character generation from single images with multi-view pose canonicalization. *ACM Transactions on Graphics (TOG)*, 43(4): 1–13, 2024. 2

[22] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 2

[23] 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*, pages 10684–10695, 2022. 2

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1, 2

[25] Liao Shen, Xingyi Li, Huiqiang Sun, Juewen Peng, Ke Xian, Zhiguo Cao, and Guosheng Lin. Make-it-4d: Synthesizing a consistent long-term dynamic scene video from a single image. In *Proceedings of the 31st ACM international conference on multimedia*, pages 8167–8175, 2023. 1, 2

[26] Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261*, 2023. 2

[27] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021. 2

[28] Tianchang Shen, Jacob Munkberg, Jon Hasselgren, Kangxue Yin, Zian Wang, Wenzheng Chen, Zan Gojcic, Sanja Fidler, Nicholas Sharp, and Jun Gao. Flexible isosurface extraction for gradient-based mesh optimization. *ACM Transactions on Graphics (ToG)*, 42(4):1–16, 2023. 2

[29] Yu-Pei Song, Yuan-Tong Liu, Xiao Wu, Qi He, Zhaoquan Yuan, and Ao Luo. Magiccartoon: 3d pose and shape estimation for bipedal cartoon characters. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8219–8227, 2024. 1

[30] Zhenyu Tang, Junwu Zhang, Xinhua Cheng, Wangbo Yu, Chaoran Feng, Yatian Pang, Bin Lin, and Li Yuan. Cycle3d: High-quality and consistent image-to-3d generation via generation-reconstruction cycle. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7320–7328, 2025. 1, 2

[31] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 1, 2, 5, 6

[32] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. In *ICLR*, 2024. 2

[33] Ruowei Wang, Jiaqi Li, Dan Zeng, Xueqi Ma, Zixiang Xu, Jianwei Zhang, and Qijun Zhao. Genudc: High quality 3d mesh generation with unsigned dual contouring representation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 10862–10871, New York, NY, USA, 2024. Association for Computing Machinery. 1

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[35] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *European conference on computer vision*, pages 57–74. Springer, 2024. 2, 5, 6

[36] Xinyue Wei, Fanbo Xiang, Sai Bi, Anpei Chen, Kalyan Sunkavalli, Zexiang Xu, and Hao Su. Neumanifold: Neural watertight manifold reconstruction with efficient and high-quality rendering support. *arXiv preprint arXiv:2305.17134*, 2023. 2

[37] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *Advances in Neural Information Processing Systems*, 37:125116–125141, 2024. 1, 2, 3, 5, 6

[38] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21469–21480, 2025. 2

[39] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4479–4489, 2023. 2

[40] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 1, 2, 5, 6

[41] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217*, 2023. 2

[42] Haibo Yang, Yang Chen, Yingwei Pan, Ting Yao, Zhineng Chen, Chong-Wah Ngo, and Tao Mei. Hi3d: Pursuing high-resolution image-to-3d generation with video diffusion models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 6870–6879, 2024. 1, 2

[43] Qiao Yu, Xianzhi Li, Yuan Tang, Xu Han, Long Hu, Yixue Hao, and Min Chen. Fancy123: One image to high-quality 3d mesh generation via plug-and-play deformation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 595–604, 2025. 1, 2, 5

[44] Wangguandong Zheng, Haifeng Xia, Rui Chen, Libo Sun, Ming Shao, Siyu Xia, and Zhengming Ding. Sketch3d: Style-consistent guidance for sketch-to-3d generation. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 3617–3626, 2024. 1