InstantMesh: Efficient 3D Mesh Generation from a Single Image with Sparse-view Large Reconstruction Models

Jiale Xu[1,2] Weihao Cheng[1] Yiming Gao[1] Xintao Wang[1*†] Shenghua Gao[2*] Ying Shan[1]
[1]ARC Lab, Tencent PCG [2]ShanghaiTech University

https://github.com/TencentARC/InstantMesh

Figure 1. Given a single image as input, our InstantMesh framework can generate high-quality 3D meshes within 10 seconds.

Abstract

We present InstantMesh, a feed-forward framework for instant 3D mesh generation from a single image, featuring state-of-the-art generation quality and significant training scalability. By synergizing the strengths of an off-the-shelf multiview diffusion model and a sparse-view reconstruction model based on the LRM [14] architecture, InstantMesh is able to create diverse 3D assets within 10 seconds. To enhance the training efficiency and exploit more geometric supervisions, e.g.*, depths and normals, we integrate a differentiable iso-surface extraction module into our framework and directly optimize on the mesh representation. Experimental results on public datasets demonstrate that InstantMesh significantly outperforms other latest image-to-3D baselines, both qualitatively and quantitatively. We release all the code, weights, and demo of InstantMesh, with the intention that it can make substantial contributions to the community of 3D generative AI and empower both researchers and content creators.*

*Corresponding Authors.
†Project Lead.

1. Introduction

Crafting 3D assets from single-view images can facilitate a broad range of applications, eg, virtual reality, industrial design, gaming and animation. We have witnessed a revolution on image and video generation with the emergence of large-scale diffusion models [37, 38] trained on billion-scale data, which is able to generate vivid and imaginative contents from open-domain prompts. However, duplicating this success on 3D generation presents challenges due to the limited scale and poor annotations of 3D datasets.

To circumvent the problem of lack of 3D data, previous works have explored distilling 2D diffusion priors into 3D representations with a per-scene optimization strategy. DreamFusion [34] proposes score distillation sampling (SDS) which makes a breakthrough in open-world 3D synthesis. However, SDS with text-to-2D models frequently encounter the multi-face issue, *i.e.*, the "Janus" problem. To improve 3D consistency, later work [35] proposes to distill from Zero123 [23] which is a novel view generator fine-tuned from Stable Diffusion [37]. A series of works [24, 26, 42, 47, 50] further propose multi-view generation models, thereby the optimization processes can be guided by multiple novel views simultaneously.

2D distillation based methods exhibit strong zero-shot generation capability, but they are time-consuming and not practical for real-world applications. With the advent of large-scale open-world 3D datasets [8, 9], pioneer works [13, 14, 45] demonstrate that image tokens can be directly mapped to 3D representations (*e.g.*, triplanes) via a novel large reconstruction model (LRM). Based on a highly scalable transformer architecture, LRMs point out a promising direction for the fast creation of high-quality 3D assets. Concurrently, Instant3D [19] proposes a diagram that predicts 3D shapes via an enhanced LRM with multi-view input generated by diffusion models. The method marries LRMs with image generation models, which significantly improves the generalization ability.

LRM-based methods use triplanes as the 3D representation, where novel views are synthesized using an MLP. Despite the strong geometry and texture representation capability, decoding triplanes requires a memory-intensive volume rendering process, which significantly impedes training scales. Moreover, the expensive computational overhead makes it challenging to utilize high-resolution RGB and geometric information (*e.g.*, depths and normals) for supervision. To boost the training efficiency, recent works seek to utilize Gaussians [18] as the 3D representation, which is effective for rendering but not suitable for geometric modeling. Several concurrent works [54, 63] opt to apply supervisions on the mesh representation directly using differentiable surface optimization techniques [39, 40]. However, they adopt CNN-based architectures, which limit their flexibility to deal with varying input viewpoints and training scalability on larger datasets that may be available in the future.

In this work, we present InstantMesh, a feed-forward framework for high-quality 3D mesh generation from a single image. Given an input image, InstantMesh first generates 3D consistent multi-view images with a multi-view diffusion model, and then utilizes a sparse-view large reconstruction model to predict a 3D mesh directly, where the whole process can be accomplished in seconds. By integrating a differentiable iso-surface extraction module, our reconstruction model applies geometric supervisions on the mesh surface directly, enabling satisfying training efficiency and mesh generation quality. Building upon an LRM-based architecture, our model offers superior training scalability to large-scale datasets. Experimental results demonstrate that InstantMesh outperforms other latest image-to-3D approaches significantly. We hope that InstantMesh can serve as a powerful image-to-3D foundation model and make substantial contributions to the field of 3D generative AI.

2. Related Work

Image-to-3D. Early attempts on image-to-3D mainly focus on the single-view reconstruction task [4, 28, 32, 33, 49, 64]. With the rise of diffusion models, pioneer works have investigated image-conditioned 3D generative modeling on various representations, *e.g.*, point clouds [27, 31, 46, 56, 64], meshes [1, 25], SDF grids [6, 7, 43, 62] and neural fields [11, 16, 30, 52, 61]. Despite the promising progress these methods have made, they are hard to generalize to open-world objects due to the limited scale of training data.

The advent of powerful text-to-image diffusion models [37, 38] inspires the idea of distilling 2D diffusion priors into 3D neural radiance fields with a per-scene optimization strategy. The score distillation sampling (SDS) proposed by DreamFusion [34] exhibits superior performance on zero-shot text-to-3D synthesis and outperforms CLIP-guided alternatives [15, 36, 58] significantly. However, SDS-based methods [3, 20, 48, 53] frequently encounter the multi-face issue, also known as the "Janus" problem. Zero123 [23] demonstrates that Stable Diffusion can be fine-tuned to synthesize novel views by conditioning on relative camera poses. Leveraging the novel view guidance provided by Zero123, recent image-to-3D methods [22, 35, 57] show improved 3D consistency and can generate plausible shapes from open-domain images.

Multi-view Diffusion Models. To address the inconsistency among multiple generated views of Zero123, some works [24, 26, 41, 50] try to fine-tune 2D diffusion models to synthesize multiple views for the same object simultaneously. With 3D consistent multi-view images, various techniques can be applied to obtain the 3D object, *e.g.*, SDS optimization [50], neural surface reconstruction meth-

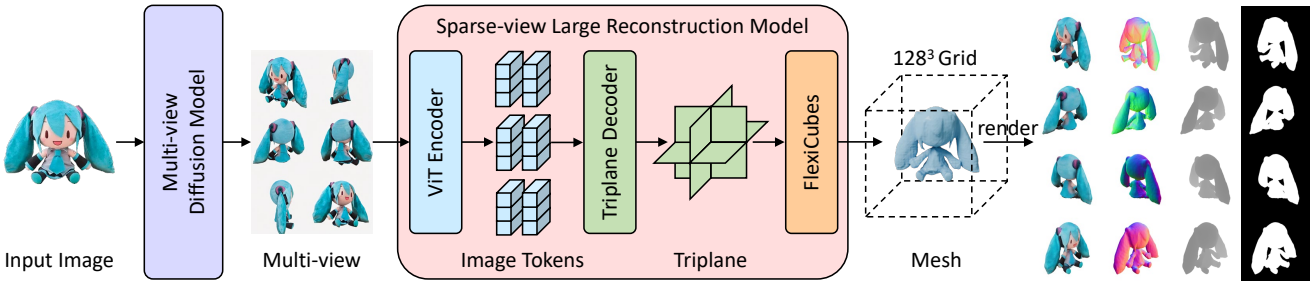

Figure 2. The overview of our InstantMesh framework. Given an input image, we first utilize a multi-view diffusion model to synthesize 6 novel views at fixed camera poses. Then we feed the generated multi-view images into a transformer-based sparse-view large reconstruction model to reconstruct a high-quality 3D mesh. The whole image-to-3D generation process takes only around 10 seconds. By integrating an iso-surface extraction module, *i.e.*, FlexiCubes, we can render the 3D geometry efficiently and apply geometric supervisions like depths and normals directly on the mesh representation to enhance the results.

ods [24, 26], multi-view-conditioned 3D diffusion models [21]. To further enhance the generalization capability and multi-view consistency, some recent works[5, 12, 47, 66] exploit the temporal priors in video diffusion models for multi-view generation.

Large Reconstruction Models. The availability of large-scale 3D datasets [8, 9] enables training highly generalizable reconstruction models for feed-forward image-to-3D creation. Large Reconstruction Model [14, 19, 51, 60] (LRM) demonstrates that the transformer backbone can effectively map image tokens to implicit 3D triplanes with multi-view supervision. Instant3D [19] further extends LRM to sparse-view input, significantly boosting the reconstruction quality. By combining with multi-view diffusion models, Instant3D can achieve highly generalizable and high-quality single-image to 3D generation. Inspired by Instant3D, LGM [44] and GRM [59] replace the triplane NeRF [29] representation with 3D Gaussians [18] to enjoy its superior rendering efficiency and circumvent the need for memory-intensive volume rendering process. However, Gaussians fall short on explicit geometry modeling and high-quality surface extraction. Given the success of neural mesh optimization methods [39, 40], concurrent works MVD^2 [63] and CRM [54] opt to optimize on the mesh representation directly for efficient training and high-quality geometry and texture modeling. Different from their convolutional network architecture, our model is built upon LRM and opts for a purely transformer-based architecture, offering superior flexibility and training scalability.

3. InstantMesh

The architecture of InstantMesh is similar to Instant3D [19], consisting of a multi-view diffusion model G_M and a sparse-view large reconstruction model G_R. Given an input image I, G_M generates 3D consistent multi-view images from I, which are fed into G_R to reconstruct a high-quality 3D mesh. We now introduce our technical improvements on data preparation, model architecture and training strategies.

3.1. Multi-view Diffusion Model

Technically, our sparse-view reconstruction model accepts free-viewpoint images as input, so we can integrate arbitrary multi-view generation model into our framework, *e.g.*, MVDream [42], ImageDream [50], SyncDreamer [24], SPAD [17] and SV3D [47], to achieve both text-to-3D and image-to-3D assets creation. We opt for Zero123++ [41] due to its reliable multi-view consistency and tailored viewpoint distribution that covers both the upper and lower parts of a 3D object.

White-background Fine-tuning. Given an input image, Zero123++ generates a 960×640 gray-background image presenting 6 multi-view images in a 3×2 grid. In practice, we notice that the generated background is not consistent across different image areas and varies in RGB values, leading to floaters and cloud-like artifacts in the reconstruction results. And LRMs are often trained on white-background images too. To remove the gray background, we need to utilize third-party libraries or models that cannot guarantee the segmentation consistency among multiple views. Therefore, we opt to fine-tune Zero123++ to synthesize consistent white-background images, ensuring the stability of the latter sparse-view reconstruction procedure.

Data Preparation and Fine-tuning Details. We prepare the fine-tuning data following the camera distribution of Zero123++. Specifically, for each 3D model in the LVIS subset of Objaverse [8], we render a query image and 6 target images, all in white backgrounds. The azimuth, elevation and camera distance of the query image is randomly sampled from a pre-defined range. The poses of the 6 target images consist of interleaving absolute elevations of 20° and -10°, combined with azimuths relative to the query image that start at 30° and increase by 60° for each pose.

During fine-tuning, we use the query image as the con-

dition and stitch the 6 target images into a 3×2 grid for denoising. Following Zero123++, we adopt the linear noise schedule and v-prediction loss. We also randomly resize the conditional image to make the model adapt to various input resolutions and generate clear images. Since the goal of fine-tuning is a simple replacement of background color, it converges extremely fast. Specifically, we fine-tune the UNet for 1000 steps with a learning rate of 1.0×10^{-5} and a batch size of 48. The fine-tuned model can fully preserve the generation capability of Zero123++ and produce white-background images consistently.

3.2. Sparse-view Large Reconstruction Model

We present the details of the sparse-view reconstruction model G_R that predicts meshes given generated multi-view images. The architecture of G_R is modified and enhanced from Instant3D [19].

Data Preparation. Our training dataset is composed of multi-view images rendered from the Objaverse [8] dataset. Specifically, we render 512×512 images, depths and normals from 32 random viewpoints for each object in the dataset. Besides, we use a filtered high-quality subset to train our model. The filtering goal is to remove objects that satisfy any of the following criteria: (i) objects without texture maps, (ii) objects with rendered images occupying less than 10% of the view from any angle, (iii) including multiple separate objects, (iv) objects with no caption information provided by the Cap3D dataset, and (v) low-quality objects. The classification of "low-quality" objects is determined based on the presence of tags such as "lowpoly" and its variants (e.g., "low_poly") in the metadata. Specifically, by applying our filtering criteria, we curated approximately 270k high-quality instances from the initial pool of 800k objects in the Objaverse dataset.

Input Views and Resolution. During training, we randomly select a subset of 6 images as input and another 4 images as supervision for each object. To be consistent with the output resolution of Zero123++, all the input images are resized to 320×320. During inference, we feed the 6 images generated by Zero123++ as the input of the reconstruction model, whose camera poses are fixed. To be noted, our transformer-based architecture makes it natural to utilize varying number of input views, thus it is practical to use less input views for reconstruction, which can alleviate the multi-view inconsistency issue in some cases.

Mesh as 3D Representation. Previous LRM-based methods output triplanes that require volume rendering to synthesize images. During training, volume rendering is memory expensive that hinders the use of high-resolution images and normals for supervision. To enhance the training efficiency and reconstruction quality, we integrate a differentiable iso-surface extraction module, *i.e.*, FlexiCubes [40], into our reconstruction model. Thanks to the efficient mesh rasterization, we can use full-resolution images and additional geometric information for supervision, *e.g.*, depths and normals, without cropping them into patches. Applying these geometric supervisions leads to smoother mesh outputs compared to the meshes extracted from the triplane NeRF. Besides, using mesh representation can also bring convenience to applying additional post-processing steps to enhance the results, such as SDS optimization [3, 20] or texture baking [22]. We leave it as a future work.

Different from the single-view LRM, our reconstruction model takes 6 views as input, requiring more memory for the cross-attention between the triplane tokens and image tokens. We notice that training such a large-scale transformer from scratch requires a significant period of time. For faster convergence, we initialize our model using the pre-trained weights of OpenLRM [13], an open-source implementation of LRM. We adopt a two-stage training strategy as described below.

Stage 1: Training on NeRF. In the first stage, we train on the triplane NeRF representation and reuse the prior knowledge of the pre-trained OpenLRM. To enable multi-view input, we add AdaLN camera pose modulation layers in the ViT image encoder to make the output image tokens pose-aware following Instant3D, and remove the source camera modulation layers in the triplane decoder of LRM. We adopt both image loss and mask loss in this training stage:

$$\begin{aligned} \mathcal{L}_1 = &\sum_i \left\| \hat{I}_i - I_i^{gt} \right\|_2^2 \\ &+ \lambda_{\text{lpips}} \sum_i \mathcal{L}_{\text{lpips}} \left(\hat{I}_i, I_i^{gt} \right) \\ &+ \lambda_{\text{mask}} \sum_i \left\| \hat{M}_i - M_i^{gt} \right\|_2^2, \end{aligned} \tag{1}$$

where $\hat{I}_i$, I_i^{gt}, $\hat{M}_i$ and M_i^{gt} denote the rendered images, ground truth images, rendered mask, and ground truth masks of the i-th view, respectively. During training, we set $\lambda_{\text{lpips}} = 2.0, \lambda_{\text{mask}} = 1.0$, and use a learning rate of 4.0×10^{-4} cosine-annealed to 4.0×10^{-5}. To enable high-resolution training, our model renders 192×192 patches which are supervised by cropped ground truth patches ranging from 192×192 to 512×512.

Stage 2: Training on Mesh. In the second stage, we switch to the mesh representation for efficient training and applying additional geometric supervisions. We integrate FlexiCubes [40] into our reconstruction model to extract mesh surface from the triplane implicit fields. The original triplane NeRF renderer consists of a density MLP and a color MLP, we reuse the density MLP to predict SDF instead, and add two additional MLPs to predict the deformation and weights required by FlexiCubes.

For a density field $f(\mathbf{x}) = d, \mathbf{x} \in \mathbb{R}^3$, points inside the object have larger values and points outside the object have

smaller values, while an SDF field $g(\mathbf{x}) = s$ is just the opposite. Therefore, we initialize the weight $\mathbf{w} \in \mathbb{R}^C$ and bias $b \in \mathbb{R}$ of the last SDF MLP layer as follows:

$$\begin{aligned} \mathbf{w} &= -\mathbf{w}_d, \\ b &= \tau - b_d, \end{aligned} \tag{2}$$

where $\mathbf{w}_d \in \mathbb{R}^C$ and $b_d \in \mathbb{R}$ are the weight and bias of the original density MLP's last layer, and τ denotes the iso-surface threshold used for density fields. Denoting the input feature of the last MLP layer as $\mathbf{f} \in \mathbb{R}^C$, we have

$$\begin{aligned} s &= \mathbf{w} \cdot \mathbf{f} + b \\ &= (-\mathbf{w}_d) \cdot \mathbf{f} + (\tau - b_d) \\ &= -(\mathbf{w}_d \cdot \mathbf{f} + b_d - \tau) \\ &= -(d - \tau), \end{aligned} \tag{3}$$

With such an initialization, we reverse the "direction" of density field to match the SDF direction and ensure that the iso-surface boundary lies at the 0 level-set of the SDF field at the beginning. We empirically find that this initialization benefits the training stability and convergence speed of FlexiCubes. The loss function of the second stage is:

$$\begin{aligned} \mathcal{L}_2 = \mathcal{L}_1 &+ \lambda_{\text{depth}} \sum_i M^{gt} \otimes \left\| \hat{D}_i - D_i^{gt} \right\|_1 \\ &+ \lambda_{\text{normal}} \sum_i M^{gt} \otimes \left(1 - \hat{N}_i \cdot N_i^{gt} \right) \\ &+ \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}, \end{aligned} \tag{4}$$

where $\hat{D}_i$, D_i^{gt}, $\hat{N}_i$ and N_i^{gt} denote the rendered depth, ground truth depth, rendered normal and ground truth normal of the i-th view, respectively. $\otimes$ denotes element-wise production, and $\mathcal{L}_{\text{reg}}$ denotes the regularization terms of FlexiCubes. During training, we set $\lambda_{\text{depth}} = 0.5, \lambda_{\text{normal}} = 0.2, \lambda_{\text{reg}} = 0.01$, and use a learning rate of 4.0×10^{-5} cosine-annealed to 0. We train our model on 8 NVIDIA H800 GPUs in both stages.

Camera Augmentation and Perturbation. Different from view-space reconstruction models [13, 44, 45, 65], our model reconstruct 3D objects in a canonical world space where the z-axis aligns with the anti-gravity direction. To further improve the robustness on the scale and orientation of 3D objects, we perform random rotation and scaling on the input multi-view camera poses. Considering that the multi-view images generated by Zero123++ may be inconsistent with their pre-defined camera poses, we also add random noise to the camera parameters before feeding them into the ViT image encoder.

Model Variants. In this work, we provide 4 variants of the sparse-view reconstruction model, two from Stage 1 and two from Stage 2. We name each model according to its 3D representation ("NeRF" or "Mesh") and the scale of parameters ("base" or "large"). The details of each model are shown in Table 1. Considering that different 3D presentations and model scales can bring convenience to different application scenarios, we release the weights of all the 4 models. We believe our work can serve as a powerful image-to-3D foundation model and facilitate future research on 3D generative AI.

Table 1. Details of sparse-view reconstruction model variants.

Parameter	InstantNeRF		InstantMesh	
	base	large	base	large
Representation	NeRF	NeRF	Mesh	Mesh
Input views	6	6	6	6
Transformer dim	1024	1024	1024	1024
Transformer layers	12	16	12	16
Triplane size	64	64	64	64
Triplane dim	40	80	40	80
Samples per ray	96	128	-	-
Grid size	-	-	128	128
Input size	320	320	320	320
Render size	192	192	512	512

4. Experiments

In this section, we conduct experiments to compare our InstantMesh with existing state-of-the-art image-to-3D baseline methods quantitatively and qualitatively.

4.1. Experimental Settings

Datasets. We evaluate the quantitative performance using two public datasets, *i.e.*, Google Scanned Objects (GSO) [10] and OmniObject3D (Omni3D) [55]. GSO contains around 1K objects, from which we randomly pick out 300 objects as the evaluation set. For Omni3D, we select 28 common categories and then pick out the first 5 objects from each category for a total of 130 objects (some categories have less than 5 objects) as the evaluation set.

To evaluate the 2D visual quality of the generated 3D meshes, we create two image evaluation sets for both GSO and Omni3D. Specifically, we render 21 images of each object in an orbiting trajectory with uniform azimuths and varying elevations in $\{30°, 0°, -30°\}$. As Omni3D also includes benchmark views randomly sampled on the top semi-sphere of an object, we pick 16 views randomly and create an additional image evaluation set for Omni3D.

Baselines. We compare the proposed InstantMesh with 4 baselines: (i) TripoSR [45]: an open-source LRM implementation showing the best single-view reconstruction performance so far; (ii) LGM [44]: a unet-based Large Gaussian Model that reconstructs Gaussians from generated multi-view images; (iii) CRM [54]: a unet-based Convolutional Reconstruction Model that reconstructs 3D meshes from generated multi-view images and canonical coordinate

maps (CCMs). (iv) SV3D [47]: an image-conditioned diffusion model based on Stable Video Diffusion [2] that generates an orbital video of an object, we only evaluate it on the novel view synthesis task since generating 3D meshes from its output is not straight-forward.

Metrics. We evaluate both the 2D visual quality and 3D geometric quality of the generated assets. For 2D visual evaluation, we render novel views from the generated 3D mesh and compare them with the ground truth views, and adopt PSNR, SSIM, and LPIPS as the metrics. For 3D geometric evaluation, we first align the coordinate system of the generated meshes with the ground truth meshes, and then reposition and re-scale all meshes into a cube of size $[-1, 1]^3$. We report Chamfer Distance (CD) and F-Score (FS) with a threshold of 0.2, which are computed by sampling 16K points from the surface uniformly.

4.2. Main Results

Quantitative Results. We report the quantitative results on different evaluation sets in Table 2, 3, and 4, respectively. For each metric, we highlight the top three results among all methods, and a deeper color indicates a better result. For our method, we report the results of using different sparse-view reconstruction model variants (*i.e.*, "NeRF" and "Mesh").

From the 2D novel view synthesis metrics, we can observe that InstantMesh outperforms the baselines on SSIM and LPIPS significantly, indicating that its generation results have the best perceptually viewing quality. As Figure 3 shows, InstantMesh demonstrates plausible appearances, whereas the baselines frequently exhibit distortions in novel views. We can also observe that the PSNR of InstantMesh is slightly lower than the best baseline, suggesting that the novel views are less faithful to the ground truth at pixel level since they are "dreamed" by the multi-view diffusion model. However, we argue that the perceptual quality is more important than faithfulness, as the "true novel views" should be unknown and have multiple possibilities given a single image as reference.

As for the 3D geometric metrics, InstantMesh outperforms the baselines on both CD and FS significantly, which indicates a higher fidelity of the generated shapes. From Figure 3, we can observe that InstantMesh presents the most reliable geometries among all methods. Benefiting from the scalable architecture and tailored training strategies, InstantMesh achieves the state-of-the-art image-to-3D performance.

Qualitative Results. To compare our InstantMesh with other baselines qualitatively, we select two images from the GSO evaluation set and two images from Internet, and obtain the image-to-3D generation results. For each generated mesh, we visualize both the textured renderings (upper) and pure geometry (lower) from two different viewpoints. We use the "Mesh" variant of sparse-view reconstruction model to generate our results.

Table 2. Quantitative results on Google Scanned Objects (GSO) orbiting views.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CD ↓	FS ↑
TripoSR	23.373	0.868	0.213	0.217	0.843
LGM	21.538	0.871	0.216	0.345	0.671
CRM	22.195	0.891	0.150	0.252	0.787
SV3D	22.098	0.861	0.201	-	-
Ours (NeRF)	23.141	0.898	0.119	0.177	0.882
Ours (Mesh)	22.794	0.897	0.120	0.180	0.880

Table 3. Quantitative results on OmniObject3D (Omni3D) orbiting views.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CD ↓	FS ↑
TripoSR	21.996	0.877	0.198	0.245	0.811
LGM	20.434	0.864	0.226	0.382	0.635
CRM	21.630	0.892	0.147	0.246	0.802
SV3D	21.510	0.866	0.186	-	-
Ours (NeRF)	22.635	0.903	0.110	0.199	0.869
Ours (Mesh)	21.954	0.901	0.112	0.203	0.864

Table 4. Quantitative results on OmniObject3D (Omni3D) benchmark views.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CD ↓	FS ↑
TripoSR	19.977	0.859	0.206	0.221	0.847
LGM	18.665	0.832	0.250	0.356	0.653
CRM	19.422	0.865	0.172	0.274	0.778
SV3D	20.294	0.853	0.176	-	-
Ours (NeRF)	19.752	0.869	0.150	0.206	0.863
Ours (Mesh)	19.552	0.868	0.150	0.204	0.866

As depicted in Figure 3, the generated 3D meshes of InstantMesh present significantly more plausible geometry and appearance. TripoSR can generate satisfactory results from images that have a similar style to the Objaverse dataset, but it lacks the imagination ability and tends to generate degraded geometry and textures on the back when the input image is more free-style (Figure 3, 3rd row, 1st column). Thanks to the high-resolution supervision, InstantMesh can also generate sharper textures compared to TripoSR. LGM and CRM share a similar framework to ours by combining a multi-view diffusion model with a sparse-view reconstruction model, thus they also enjoy the imagination ability. However, LGM exhibits distortions and obvious multi-view inconsistency, while CRM has difficulty in generating smooth surfaces.

Comparison between "NeRF" and "Mesh" variants. We

Figure 3. The 3D meshes generated by InstantMesh demonstrate significantly better geometry and texture compared to the other baselines. The results of InstantMesh are rendered at a fixed elevation of 20°, while the results of other methods are rendered at a fixed elevation of 0° since they reconstruct objects in the view space.

also compare the "Mesh" and "NeRF" variants of our sparse-view reconstruction model quantitatively and qualitatively. From Table 2, 3, and 4, we can see that the "NeRF" variant achieves slightly better metrics than the "Mesh" variant. We attribute this to the limited grid resolution of FlexiCubes, resulting in the lost of details when extracting

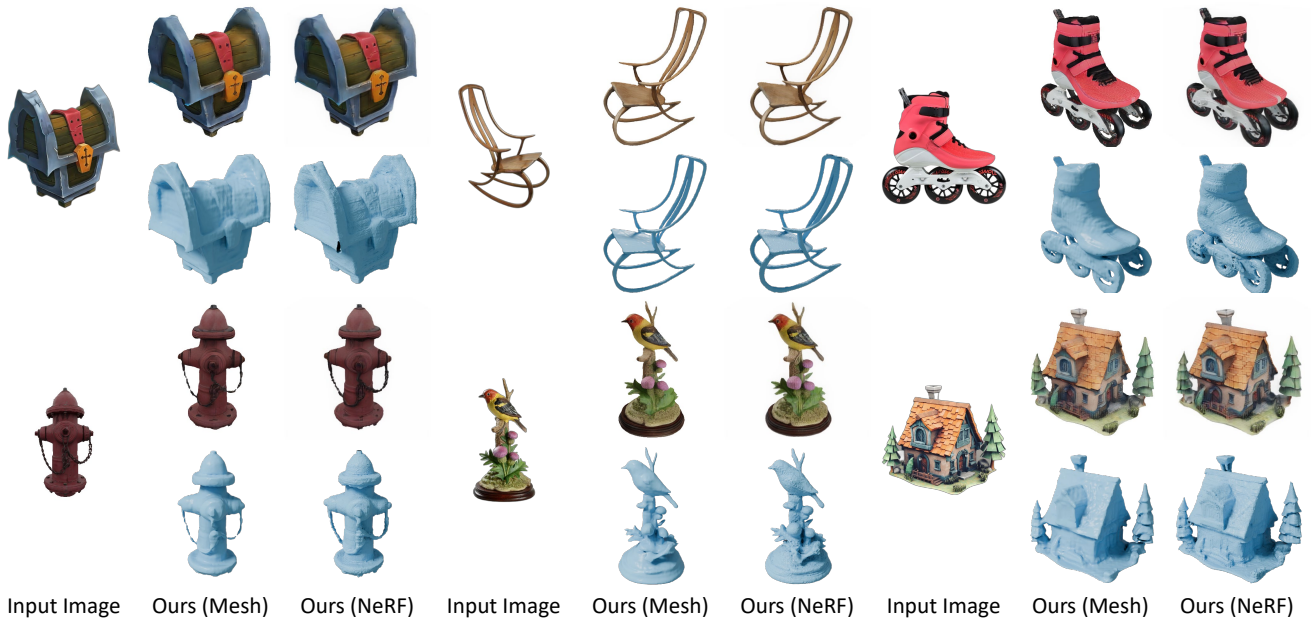

Figure 4. Image-to-3D generation results using different sparse-view reconstruction model variants. For each generated mesh, we visualize both the textured rendering (upper) and untextured geometry (lower). All images are rendered at a fixed elevation of 20°.

mesh surfaces. Nevertheless, the drop in metrics is marginal and negligible considering the convenience brought by the efficient mesh rendering compared to the memory-intensive volume rendering of NeRF. Besides, we also visualize some image-to-3D generation results of the two model variants in Figure 4. By applying explicit geometric supervisions, *i.e.*, depths and normals, the "Mesh" model variant can produce smoother surfaces compared to the meshes extracted from the density field of NeRF, which are generally more desirable in practical applications.

5. Conclusion

In this work, we present InstantMesh, an open-source instant image-to-3D framework that utilizes a transformer-based sparse-view large reconstruction model to create high-quality 3D assets from the images generated by a multi-view diffusion model. Building upon the Instant3D framework, we introduce mesh-based representation and additional geometric supervisions, significantly boosting the training efficiency and reconstruction quality. We also make improvements on other aspects, such as data preparation and training strategy. Evaluations on public datasets demonstrate that InstantMesh outperforms other latest image-to-3D baselines both qualitatively and quantitatively. InstantMesh is intended to make substantial contributions to the 3D Generative AI community and empower both researchers and creators.

Limitations. We notice that some limitations still exist in our framework and leave them for future work. (i) Following LRM [14] and Instant3D [19], our transformer-based triplane decoder produces 64×64 triplanes, whose resolution may be a bottleneck for high-definition 3D modeling. (ii) Our 3D generation quality is inevitably influenced by the multi-view inconsistency of the diffusion model, while we believe this issue can be alleviated by utilizing more advanced multi-view diffusion architectures in the future. (iii) Although FlexiCubes can improve the smoothness and reduce the artifacts of the mesh surface due to the additional geometric supervisions, we notice that it is less effective on modeling tiny and thin structures compared to NeRF (Figure 4, 2nd row, 1st column).

References

[1] Antonio Alliegro, Yawar Siddiqui, Tatiana Tommasi, and Matthias Nießner. Polydiff: Generating 3d polygonal meshes with diffusion models. *arXiv preprint arXiv:2312.11417*, 2023. 2

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 6

[3] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22246–22256, 2023. 2, 4

[4] Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. Bsp-net: Generating compact meshes via binary space partitioning. In

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 45–54, 2020. 2

[5] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024. 3

[6] 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

[7] Gene Chou, Yuval Bahat, and Felix Heide. Diffusion-sdf: Conditional generative modeling of signed distance functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2262–2272, 2023. 2

[8] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 3, 4

[9] 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. 2, 3

[10] 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

[11] 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

[12] Junlin Han, Filippos Kokkinos, and Philip Torr. Vfusion3d: Learning scalable 3d generative models from video diffusion models. *arXiv preprint arXiv:2403.12034*, 2024. 3

[13] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023. 2, 4, 5

[14] 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 *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3, 8

[15] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 867–876, 2022. 2

[16] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 2

[17] Yash Kant, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, Igor Gilitschenski, and Aliaksandr Siarohin. Spad: Spatially aware multiview diffusers. *arXiv preprint arXiv:2402.05235*, 2024. 3

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):1–14, 2023. 2, 3

[19] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 4, 8

[20] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 2, 4

[21] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. *arXiv preprint arXiv:2311.07885*, 2023. 3

[22] 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, 4

[23] 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

[24] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 3

[25] Zhen Liu, Yao Feng, Michael J. Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffusion: Score-based generative 3d mesh modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[26] 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. *arXiv preprint arXiv:2310.15008*, 2023. 2, 3

[27] 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

[28] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2

[29] 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. 3

[30] 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*, pages 4328–4338, 2023. 2

[31] 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

[32] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3504–3515, 2020. 2

[33] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. Deep mesh reconstruction from single rgb images via topology modification networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9964–9973, 2019. 2

[34] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[35] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 2

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[37] 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

[38] 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. 2

[39] 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, 3

[40] 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, 3, 4

[41] 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. 2, 3

[42] 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

[43] Jaehyeok Shim, Changwoo Kang, and Kyungdon Joo. Diffusion-based signed distance fields for 3d shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20887–20897, 2023. 2

[44] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 3, 5

[45] 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. 2, 5

[46] Michał J Tyszkiewicz, Pascal Fua, and Eduard Trulls. Gecco: Geometrically-conditioned point diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2128–2138, 2023. 2

[47] 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. *arXiv preprint arXiv:2403.12008*, 2024. 2, 3, 6

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

[49] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 2

[50] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 2, 3

[51] 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 *The Twelfth International Conference on Learning Representations*, 2024. 3

[52] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 4563–4573, 2023. 2

[53] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[54] 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. *arXiv preprint arXiv:2403.05034*, 2024. 2, 3, 5

[55] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 5

[56] 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

[57] 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

[58] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20908–20918, 2023. 2

[59] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024. 3

[60] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, and Kai Zhang. DMV3d: Denoising multi-view diffusion using 3d large reconstruction model. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[61] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023. 2

[62] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional sdf diffusion for controllable 3d shape generation. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023. 2

[63] Xin-Yang Zheng, Hao Pan, Yu-Xiao Guo, Xin Tong, and Yang Liu. Mvd2: Efficient multiview 3d reconstruction for multiview diffusion. *arXiv preprint arXiv:2402.14253*, 2024. 2, 3

[64] 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

[65] 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. *arXiv preprint arXiv:2312.09147*, 2023. 5

[66] Qi Zuo, Xiaodong Gu, Lingteng Qiu, Yuan Dong, Zhengyi Zhao, Weihao Yuan, Rui Peng, Siyu Zhu, Zilong Dong, Liefeng Bo, et al. Videomv: Consistent multi-view generation based on large video generative model. *arXiv preprint arXiv:2403.12010*, 2024. 3