Direct and Explicit 3D Generation from a Single Image

Haoyu Wu[1*] Meher Gitika Karumuri[2] Chuhang Zou[2] Seungbae Bang[2]
Yuelong Li [2] Dimitris Samaras [1] Sunil Hadap [2]
[1]Stony Brook University [2]Amazon Inc.

Abstract

Current image-to-3D approaches suffer from high computational costs and lack scalability for high-resolution outputs. In contrast, we introduce a novel framework to directly generate explicit surface geometry and texture using multi-view 2D depth and RGB images along with 3D Gaussian features using a repurposed Stable Diffusion model. We introduce a depth branch into U-Net for efficient and high quality multi-view, cross-domain generation and incorporate epipolar attention into the latent-to-pixel decoder for pixel-level multi-view consistency. By back-projecting the generated depth pixels into 3D space, we create a structured 3D representation that can be either rendered via Gaussian splatting or extracted to high-quality meshes, thereby leveraging additional novel view synthesis loss to further improve our performance. Extensive experiments demonstrate that our method surpasses existing baselines in geometry and texture quality while achieving significantly faster generation time.

1. Introduction

The task of generating 3D assets from single images [29, 43, 46, 51, 54, 57] is pivotal in several application domains, such as 3D content creation, virtual reality, augmented reality, as well as 3D aware image generation and editing. However, building a 3D model from a sparse set of images, let alone just one, is a highly ill-posed problem. There are inherent ambiguities in this inverse rendering problem, and the greatest challenge is effectively "hallucinating" unseen portions of the object in terms of geometry and texture. Recent cutting-edge generative AI approaches (e.g. diffusion models, transformers) attempt to overcome these obstacles by learning powerful 3D priors and show promising results.

A major technical challenge for 3D generation is how to represent 3D objects / scenes that can be easily modeled. Recent techniques seek for volumetric representation, such as Neural Radiance Fields (NeRF) [29, 52] or triplane [3, 22]. Such volumetric representation typically has high computational and memory complexity which inhibits its scalability towards high-quality and high-resolution generation. In pursuit of efficient explicit representation, recent works directly generate point clouds [50, 54, 76, 102] or meshes [48]. Although explicit generations significantly reduce computational complexity compared to volumetric representations, they represent 3D quantities in a do-

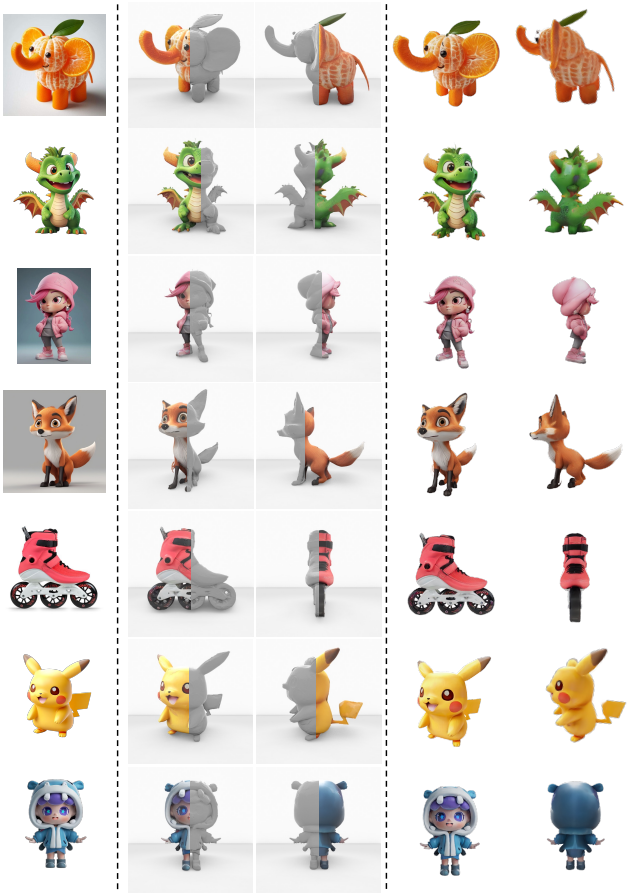

Figure 1. We present our approach that generates high-resolution (x512), textured 3D asset from a single image. From left to right: input in-the-wild images downloaded from the internet, generated 3D textured meshes, novel view synthesis via Gaussian splatting.

* This work was done as part of Haoyu's internship at Amazon.

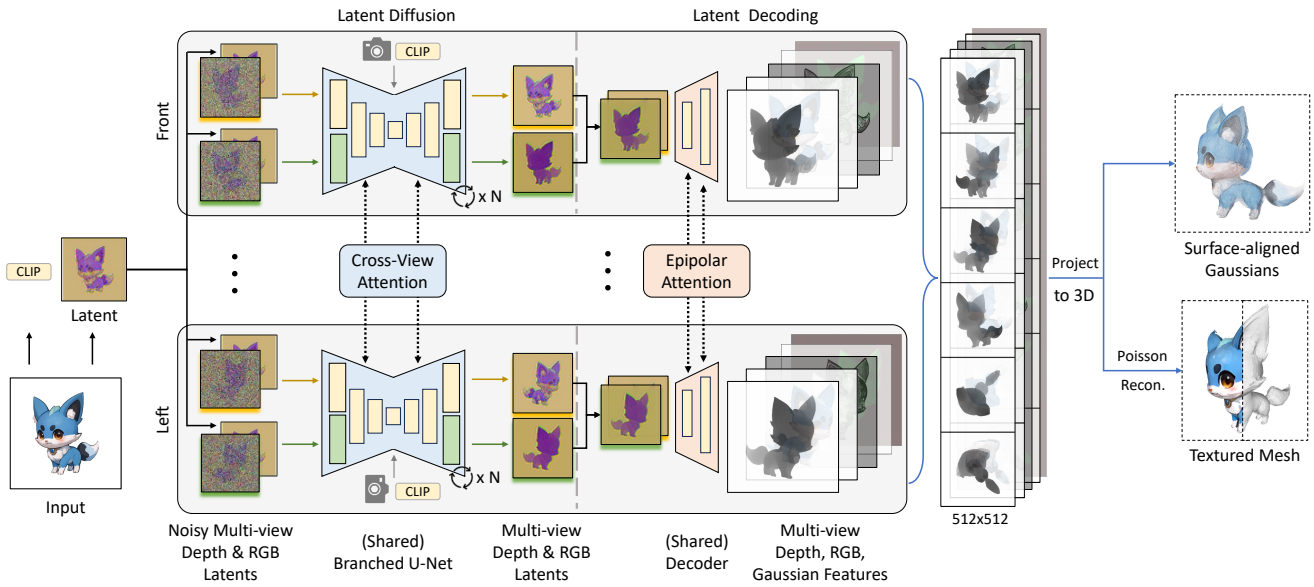

Figure 2. **Overview**. Our method is a feed-forward image-to-3D model. Given an input image, we generate depth and RGB latent images from six orthographic views via simultaneous multi-view diffusion. The process is conditioned on input latent, input CLIP embedding, and cameras. We incorporate a branched U-Net for efficient and high-quality cross-domain diffusion. For each view, we channel-concatenate depth and RGB latents and decode it to depth, RGB, and Gaussian features in pixel space (512x512 resolution). We add epipolar attention in the decoder, which is crucial for generating pixel-level multi-view consistent depths. We lift our output (RGB and opacity from Gaussians) into 3D space via depth unprojection, creating high-quality textured mesh via Poisson surface reconstruction. Additionally, our lifted surface-aligned 3D Gaussians enable novel view synthesis via Gaussian splatting, allowing additional gradient decent loss from NVS.

main which significantly deviates from that of natural images, posing great challenges in borrowing strong 2D priors from contemporary foundation models such as Stable Diffusion (SD) [62]. Recently, an alternative two-stage strategy [18, 46, 49] was proposed that first generates multi-view images and subsequently fits a 3D representation out of them. However, effectively enforcing consistency across multiple views remains a challenge. Methods that only generate RGB images [18, 46] typically require dense views to form extensive coverage of the scene/object, which can be computationally expensive.

To address the aforementioned challenges, we propose to directly generate explicit surface geometry and texture using multi-view 2D RGB, depth and Gaussian feature images. We believe that this representation offers a more scalable approach towards high-resolution and detail-preserving generation, and we hypothesize that depth maps capture geometry information more effectively than other alternatives like normals. We repurpose the Stable Diffusion [36, 62] model and additionally introduce a branched U-Net with expert blocks for efficient and high quality multi-view image generation. To ensure pixel-level multi-view consistent depth maps, we incorporate epipolar attention [24, 80, 92] in the latent decoding process. Our representation is versatile and can be easily converted to other formats. In particular, we immediately obtain a structured 3D representation when we back-project multi-view depth, RGB images and additional Gaussian features to 3D, *i.e.* dense surface-aligned point cloud or Gaussians. This 3D representation can be either extracted to high-quality textured mesh via Poisson surface reconstruction [32] or rendered efficiently using Gaussian Splatting [34] for novel view synthesis (NVS), thereby leveraging additional NVS loss to further improve our performance

In summary, our main **contributions** are:

- We propose a novel streamlined framework to predict multi-view depth maps along with RGB and Gaussian features by fine-tuning a pre-trained 2D diffusion model. Our representation offers a compact encoding of the 3D models and scales well towards high-resolution generation, enabling high-fidelity content creation.
- We develop a branched U-Net that learns to efficiently generate multi-view RGB and depth images while leveraging cross-domain similarities.
- To enforce pixel-level multi-view consistent depths, we incorporate epipolar attention in latent-to-pixel decoder following geometric insights.
- We demonstrate significant improvements in quality and speed over existing benchmarks, marking a major advancement in single-view 3D generation.

2. Related Work

Representation of 3D models. A fundamental issue to address when formulating 3D generation is how to repre-

sent the 3D model, which can be either explicit surface-like representation or implicit volumetric representation. Typical examples of explicit representation include Point-E [54] and LION [76] that train diffusion models to generate point clouds, or MeshDiffusion [48] that generates parametrized 3D meshes. One drawback of directly generating explicit representation is that it becomes difficult to reuse 2D image priors due to the domain gap between images and explicit representation formats such as point clouds or meshes. In contrast, our method predicts depth images which enable a shared 2D latent representation and facilitate incorporation of 2D priors.

Since the popularity of NeRF [52], many recent techniques adopt volumetric representation such as neural fields [1, 6, 10, 17, 20, 22, 29, 31, 35, 53, 55, 79, 97], triplane [3, 25, 26, 84, 90, 102], etc. However, these approaches typically require expensive volumetric rendering, posing great challenges on scaling up to high-resolution predictions. For example, LRM [26] has triplane resolution of just 64x64, relying on low resolution features to render high resolution images. In contrast, our method predicts multiple sparse views which naturally scale towards high-resolution.

Some recent feed-forward methods [70, 72, 89, 91, 98, 102] also propose to generate Gaussians as 3D representation. However, our approach primarily focuses on generating pixel-perfect multi-view depth maps, leading to significantly improved generation quality.

Generation with Score Distillation Sampling. Generation of 3D objects / scenes from single or sparse input views is an inherently ambiguous problem. Recent advancements in 3D generation endeavored to leverage large 2D generative models [60, 62, 63] as strong image priors by reformulating the problem into 2D domains. These 2D generative models are trained on extensive internet-scale image datasets, which generalizes over diverse scenarios. A notable innovation in this area has been introduced by DreamFusion [56] which introduced Score Distillation Sampling (SDS). The core methodology involves optimizing a parameterized 3D representation, *e.g.* NeRF [52], SDF, or mesh using a pre-trained 2D prior model to supervise rendered views. This technique has been later successfully applied to both text-to-3D and image-to-3D synthesis tasks [2, 7, 9, 28, 41, 58, 65, 66, 74, 78, 83, 87, 95, 101], demonstrating its versatility and effectiveness. Despite the early promise of this approach, the generated 3D objects often lack 3D consistency. In addition, it requires a time-consuming per-scene optimization process which limits its application in real-time scenarios.

Generation and Fusion of Multiple Views. Another research direction focuses on generating multi-view images from a single image [4, 15, 18, 19, 38, 43, 44, 71, 73, 75, 77, 85, 88, 93, 96, 100]. Zero123 [43] pioneers in adapting the pre-trained 2D diffusion model for multi-view image synthesis by incorporating camera conditions into the model. While this approach delivers promising results, it lacks constraints across different views and hence struggles with consistency across the generated multi-view images. To mitigate this, SyncDreamer [46] introduces a volume-based multi-view information aggregation module using 3D CNN and spatial attention. One-2-3-45 [42] aims to combine 2D generative models and multi-view 3D reconstruction. Some other works [49, 67] leverage 2D dense cross-view attention to enhance 3D consistency. With the generated multi-view images available, these methods then fit a 3D representation via the reconstruction loss. While these methods offer qualitative improvement and an alternative path to SDS, they still struggle with multi-view consistency and may fail to produce high-quality meshes.

In addition, many such techniques only predict RGB images and thus require abundant views to form a dense coverage of the 3D models. Therefore, the subsequent optimization process can be time-consuming, typically lasting at least several minutes per scene. Among recent attempts [33, 69] to embrace depth prediction, MVD-Fusion [27] was proposed that leverages intermediate low-resolution depth maps as fusion guidance. However, it does not produce high-resolution depth maps as end predictions, which we believe is important for high-quality 3D reconstruction.

Other methods explore alternative 3D related images [40, 47] including normals [49]. For normal maps, we find it is difficult to derive depth discontinuities out of normal values, and 3D reconstruction from multiple normal maps typically require a dedicated optimization process. In contrast, depth maps faithfully capture geometric details, enabling high-fidelity reproduction of object shapes.

3. Preliminaries

3D Gaussian splatting performs novel view synthesis using Gaussians as a 3D representation. Each Gaussian in 3D is defined by a center $\mathbf{x} \in \mathbb{R}^3$, a color feature $\mathbf{c} \in \mathbb{R}^C$, an opacity value $\alpha \in \mathbb{R}$, a scaling factor $\mathbf{s} \in \mathbb{R}^3$, and a rotation quaternion $\mathbf{q} \in \mathbb{R}^4$. Renderings are performed via alpha composition using differentiable rasterization in real time.

Pixel-aligned 3D Gaussians. The Gaussian center $\mathbf{x}$ can be replaced by depth pixels $\mathbf{t} \in \mathbb{R}$. Suppose that ray_o and ray_d represent the ray origin and ray direction, respectively; the Gaussian center is then inferred as $\mathbf{x} = \text{ray}_o + \mathbf{t} \cdot \text{ray}_d$. The final output can be obtained by merging the 3D Gaussians from N views, resulting in $N \cdot HW$ Gaussians.

4. Method

Given an input image, our goal is to generate the 3D shape with high-resolution textures and high-fidelity geometry. We propose to directly generate explicit surface geometry by decomposing the 3D representation into multi-view

consistent outputs, each of which contains a depth map, an RGB image, and a Gaussian feature map. Predictions from different views are used to construct a 3D point cloud, which can be extracted into a textured mesh or rendered using 3D Gaussian splatting [34], enabling natural support for novel view synthesis (NVS) and taking advantage of additional NVS loss to further improve our performance.

An overview of our method is illustrated in Fig. 2.

4.1. Multi-view Generation in Latent Space

Reconstructing 3D objects from one single input view is an ill-posed problem in general. Injecting prior knowledge is critical to resolving the inherent ambiguity. To this end, we re-purpose the Stable Diffusion model [62], which possesses rich domain knowledge through internet-scale pre-training, to predict multi-view latents. Since Stable Diffusion generates each view independently, it could potentially generate inconsistent views, causing blurriness or distortions in the generated 3D objects. We follow prior works [46, 49, 67] to extend the self-attention layer in the Stable Diffusion U-Net into dense cross-attention among different views, so that features are not only attended spatially but also across views. Thus, we obtain multi-view consistent generation in latent space.

4.2. Enforcing Cross-view Depth Consistency with Epipolar Attention

Using the pre-trained VAE decoder in the Stable Diffusion framework, we decode multi-view latents back to multi-view RGB images. Although the dense cross-view attention is done in the latent space, we and prior works [46, 49, 67] also observe good consistency for the decoded multi-view RGB images.

Thus, a natural solution to predict multi-view depth maps is to follow the same framework: one could simply finetune the decoder to decode the multi-view latents back to multi-view depth maps. However, as shown in Fig. 7 and Tab. 3 ("w/o Epipolar Attention"), we find this naive approach fails to produce 3D consistent depth maps, resulting in bad mesh extractions. The reason is that we use decoded images for explicit geometry representation, which requires higher pixel-level value accuracy and 3D consistency. This suggests dense cross-view attention in the latent space achieves consistency in RGB images but still falls short to estimate pixel-level accurate and multi-view consistent depth maps.

To address this problem, we propose adding *epipolar attention* [24, 30, 80, 82, 92], which is much more computationally efficient than the cross-view attention used in U-Net. It borrows insights from multi-view geometry, which dictates that corresponding pixels from different views are constrained to lie on certain lines called epipolar lines [23]. In implementation, we modify the attention transformer component to reduce the number of computations, as depicted in Fig. 3. Each token in the query image only attends to tokens from other views along the epipolar line. Since our approach predicts six othogonal views of front, back, left, right, top and bottom, the epipolar attention can be simplified as efficient row or column attention [39] because any pair of our output views is either parallel or perpendicular to each other.

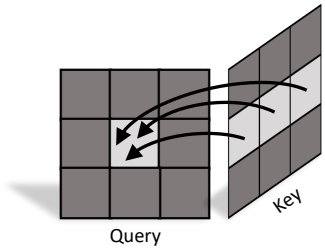

Figure 3. An illustration of the epipolar attention in the decoder. We utilize epipolar geometry to facilitate efficient multi-view information exchange among views.

4.3. Efficient Cross-domain Denoising via Expert Branch

Although epipolar attention greatly improves 3D consistency for the decoded depth maps, we find that the resulting depth maps lack details, as shown in Fig. 7 and Tab. 3 ("w/o Depth Latent"). This suggests that relying solely on the RGB latent for decoding both RGB and depth, is insufficient for producing detailed depth maps. To mitigate this problem, we modify the latent U-Net to also output latent depth images along with latent RGB images. In practice, we find that channel-concatenating RGB and depth latents and feeding them to the decoder produces high-quality RGB and depth maps.

We observe that following Wonder3D's [49] architecture of using domain switch to infer RGB and depth images, doubles the required diffusion inference resources by employing separate labels for RGB and depth. As this process is time and memory intensive, instead, we implement a modified version of expert branch strategy proposed in HyperHuman [45] where we train for both RGB and depth latents simultaneously instead of successively. This improves the training and inference time significantly. As demonstrated in Fig. 4, we modify the U-Net so that each domain (RGB and depth) has its dedicated expert branch for the first DownBlock and the last UpBlock. We input noisy RGB and depth latents into their respective first DownBlocks. Then, the extracted features from RGB domain traverse the U-Net middle layers. These RGB features are then fed into the last UpBlock of different domain expert branches along with residual features from respective first DownBlocks. This produces individual branch outputs, *i.e.* denoised RGB and depth latent. Additionally, we find this individual emphasis on RGB features effectively prevents catastrophic model

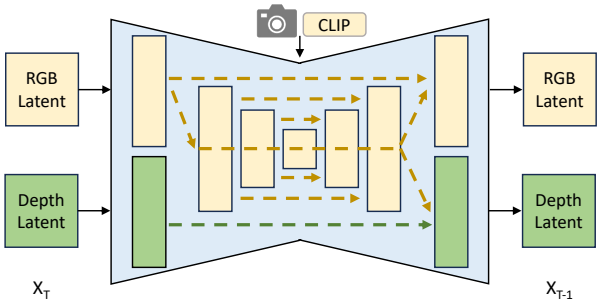

Figure 4. An illustration of the branched U-Net for cross-domain latent diffusion. We add two blocks (green) as the depth branch for latent depth prediction. By sharing most weights with the original model that predicts RGB latents, we achieve efficient and high-quality depth prediction in a single inference.

forgetting in training. After obtaining RGB and depth latents, we channel-concatenate them together and forward it to the decoder.

4.4. Efficient Rendering with 3D Gaussian Splatting

We find that when we extend the decoder to predict additional Gaussian features and assign them to our colored point cloud, we essentially create pixel-aligned 3D Gaussians [5, 8, 70, 72, 86, 91, 98] that are also surface-aligned due to our multi-view consistent depth design. We find this simple decoder extension enables high-quality novel view synthesis through Gaussian splatting [34]. Thus, our decoder training can benefits from additional supervision via novel view synthesis loss, enhancing the performance.

Predicting pixel-aligned Gaussians. Using the generated RGB and depth images, we can create a dense colored point cloud by back-projecting all pixels into 3D space. We choose to further expand the output of the decoder to include additional 8 channels (1-channel opacity, 3-channel scale, and 4-channel rotation quaternion). Because we produce surface-aligned Gaussians [21], we restrict their scales to be near 1 pixel using $0.01 \cdot \text{Sigmoid}(s) + 2.5 \cdot (1 - \text{Sigmoid}(s))$ where s is the output scale without activation, similar to [91].

4.5. Textured Mesh Extraction

From the generated multi-view outputs, we utilize RGB, depth, and opacity from Gaussian features for textured mesh extraction. First, we find the gradients of our generated depth map provides good approximation of surface normals. In addition, we mask out pixels with corresponding opacity values smaller than 0.1 and back-project the masked RGB/depth/normal pixels into 3D to create an oriented colored point cloud. We apply screened Poisson surface reconstruction [32] to extract the mesh. Laplacian smoothing is applied to smooth the stair-case appearance. We generate texture coordinate using xatlas [94], then assign color values on texture map by projecting color values from point cloud. Because our method produces pixel-perfect depth maps, our mesh extraction from point cloud (3D Gaussians) is accomplished without the need for any complex neural optimization processes, unlike methods such as LGM [72].

4.6. Implementation Details

We fine-tune the U-Net from Stable Diffusion Image Variation [36, 62]. We initialize the depth branch of the U-Net with the weights of the first DownBlock and last UpBlock of the pre-trained U-Net and fine-tune all parameters of the U-Net. To enable high-resolution training with faster convergence, we first fine-tune U-Net on 256x256 resolution with a batch size of 512 for 30K iterations, and then fine-tune it on 512x512 resolution with a batch size of 96 for 100K iterations.

We fine-tune the VAE decoder together with our epipolar attention design from Stable Diffusion [62]. The fine-tuning is done on 512x512 resolution with a batch size of 8 for 90K iterations. We utilize a combination of regression loss and rendering loss for decoder training. After decoding multi-view outputs, we compute Mean Square Error (MSE) loss and LPIPS loss [99] for decoded RGB images and $\mathcal{L}_{L_1}$ loss and gradient matching loss $\mathcal{L}_{\text{gm}}$ [61] for decoded depth maps. We also back-project depths and Gaussian features to 3D and render the Gaussians via differentiable Gaussian splatting [34]. We randomly render 10 novel views and compute MSE loss and LPIPS loss for rendered RGB images and MSE loss for rendered alpha images. The overall loss function is:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{Reg}} + \mathcal{L}_{\text{NVS}}, \\ \mathcal{L}_{\text{Reg}} &= \mathcal{L}^{\text{rgb}}_{\text{MSE}} + \lambda_{\text{LPIPS}}\mathcal{L}^{\text{rgb}}_{\text{LPIPS}} + \mathcal{L}^{\text{dep}}_{L_1} + \lambda_{\text{gm}}\mathcal{L}^{\text{dep}}_{\text{gm}}, \\ \mathcal{L}_{\text{NVS}} &= \mathcal{L}^{\text{rgb - nvs}}_{\text{MSE}} + \lambda_{\text{LPIPS}}\mathcal{L}^{\text{rgb - nvs}}_{\text{LPIPS}} + \mathcal{L}^{\text{alpha - nvs}}_{\text{MSE}}, \end{aligned} \tag{1}$$

where $\mathcal{L}^{\text{rgb}}$ and $\mathcal{L}^{\text{dep}}$ stand for losses over RGB and depth images across the six views, whereas $\mathcal{L}^{\text{rgb - nvs}}$, $\mathcal{L}^{\text{dep - nvs}}$ and $\mathcal{L}^{\text{alpha - nvs}}$ denote losses over synthesized novel views of RGB, depth and alpha images respectively. We set λ_{LPIPS} as 0.5 and λ_{gm} as 2.

For both U-Net and decoder training, we use a learning rate of 1e-4. Our U-Net is conditioned by the CLIP embedding [59] of the input image via cross attention. The noisy RGB and depth latents are channel-concatenated with the input latent and then sent to U-Net for denoising [43]. We learn camera embeddings, transform them with an MLP, and add them to U-Net's timestep embeddings [43, 49, 67]. We utilize Xformers [37] and FlashAttention [12, 13] in U-Net and decoder training to enable fast and memory-efficient attention. For latent diffusion inference, we use a guidance scale of 3 and the number of diffusion steps is set to 50 using the DDIM [68] scheduler. Our image camera view follows Wonder3D's input view related system.

Figure 5. **Qualitative results on GSO dataset.** We visualize the input single image and the resulting 3D mesh (with and without textures) for our method and baselines. Our approach achieves higher mesh quality in terms of both geometry and texture.

5. Experiments

Datasets. For model training, we leverage the LVIS subset of the Objaverse dataset [14], similar to previous works [42, 49]. This dataset contains around 46K objects in 1,156 categories. To generate ground-truth data, we normalize the objects to fit within a unit sphere. We use Blender [11] to render depth and RGB images from six orthogonal views: front, back, left, right, top and bottom. Additionally, random rotations are applied to objects to enrich the dataset. For evaluation, we follow prior research to use the Google Scanned Objects (GSO) dataset [16]. We employ the same evaluation dataset as previous works [46, 49], consisting of 30 common objects used in daily life. For each object, we render an image to serve as the input for our evaluation process. We also evaluate our method using additional object images collected by other methods [49, 51]. The image resolution is set to 512×512 for both training and evaluation.

Metrics. We follow the standard evaluation protocol in [46, 49, 84] for single-image 3D reconstruction. We report the Chamfer distance (CD) and the volume IoU between the reconstructed mesh and the ground-truth mesh. We also report depth map error (absolute distance) from mesh rendering as a measure of surface error. In experiment, we process each method's predictions using scale adaptive ICP [64] to refine the alignment of the reconstructed mesh to ground

Method	Chamfer Dist.↓	Volume IoU↑	Depth Error↓	PSNR↑	SSIM↑	LPIPS↓	Time ↓
Realfusion [51]	0.1015	0.2882	0.394	12.44	0.764	0.373	∼1 hour
Zero123 [43]	0.0627	0.4451	0.327	14.90	0.808	0.296	∼30 mins
Magic123 [57]	0.0564	0.3988	0.282	10.50	0.770	0.386	∼1 hour
Wonder3D [49]	0.0236	0.6731	0.134	15.21	0.824	0.269	2-3 mins
SyncDreamer [46]	0.0234	0.6464	0.134	15.92	0.833	0.202	5-10 mins
Point-E [54]	0.0520	0.2445	0.308	13.73	0.807	0.314	1-2 mins
Shap-E [29]	0.0438	0.3430	0.223	12.67	0.793	0.318	8-20 s
LGM [72]	0.0396	0.4538	0.210	14.09	0.833	0.328	∼1 min
CRM [84]	0.0334	0.5594	0.173	14.02	0.835	0.309	∼30 s
One-2-3-45 [42]	0.0282	0.6131	0.143	16.51	0.838	0.217	∼45s
InstantMesh [90]	0.0264	0.6584	0.143	16.51	0.842	0.205	∼30s
OpenLRM [25]	0.0186	0.7054	0.108	14.62	0.844	0.254	∼20s
MVD-Fusion [27]	0.0362	—	—	—	—	—	∼35s
Ours	**0.0135**	**0.7339**	**0.073**	**17.85**	**0.851**	**0.159**	15-25s

Table 1. **Quantitative evaluation on the GSO dataset.** We report performance of baselines and our results on textured mesh generation. We classify the baselines into three categories: SDS optimization-based methods (first three rows), multi-view image (normal) generation-based optimization (lines 4-5), and direct feed-forward 3D generation methods (lines 6-14). We mark the best scoring methods with bold.

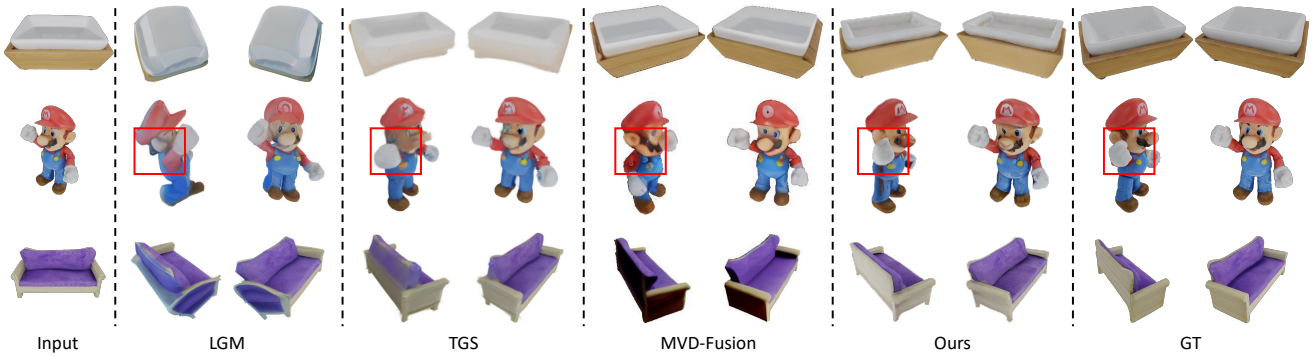

Figure 6. **Qualitative comparisons for novel view synthesis**. Our generated 3D Gaussians deliver higher visual quality and capture more intricate details.

truth. To measure texture quality, we render 512x512 images from the output mesh and the ground-truth mesh using 36 fixed camera views per object and report the average PSNR, SSIM [81] and LPIPS [99].

For novel view synthesis, we directly render 36 images via 3D Gaussian splatting from our output surface-aligned Gaussians and report PSNR, SSIM, and LPIPS metrics, following [34].

Baselines. We compare to state-of-the-art single image-to-3D methods including Zero123 [43], RealFusion [51], Magic123 [57], SyncDreamer [46], Wonder3D [49], Point-E [54], Shap-E [29], One-2-3-45 [42], CRM [84], InstantMesh [90], OpenLRM [25], LGM [72], TGS [102] and MVD-Fusion [27]. We use their official implementations, except for LRM [26] which only has open-sourced implementation OpenLRM [25] available. For MVD-Fusion, since it produces coarse point clouds from low-resolution depth maps (32x32) and the textured mesh generation is not available, we only report its Chamfer distance for 3D reconstruction, following the original paper.

Method	PSNR↑	SSIM↑	LPIPS↓
LGM [72]	14.3465	0.8191	0.2991
MVD-Fusion [27]	16.5586	0.8314	0.2071
TGS [102]	17.5151	0.8612	0.2234
Ours	**18.1698**	**0.8621**	**0.1586**

Table 2. **Quantitative comparison for novel view synthesis on the GSO dataset**. We report performance of baselines and our results and mark the best scoring methods with bold.

5.1. Comparisons with SOTA Methods

3D Reconstruction. Our approach demonstrates superior performance compared to state-of-the-art methods in single-image 3D reconstruction, both qualitatively as shown in Fig. 5 and quantitatively as shown in Tab. 1. Compared to other baselines, our method consistently generates meshes with better texture fidelity and geometric accuracy.

Method	CD↓	IoU↑	PSNR↑	LPIPS↓
Full	**0.0135**	**0.7339**	**17.85**	**0.159**
w/o Epipolar Attn.	0.0323	0.4518	15.12	0.262
w/o Depth Latent	0.0249	0.6010	15.40	0.238
w/o NVS loss	0.0136	0.7286	17.77	0.166

Table 3. **Quantitative ablations on our design choices.** We mark the best scoring methods with bold.

Notably, our approach performs exceptionally well in capturing intricate details because we generate high-resolution depth maps with multi-view consistency at the pixel level. Furthermore, our method also achieves competitive generation speed compared to other feed-forward methods, second only to Shap-E but showing much better geometric quality.

Novel View Synthesis. Our method also outperforms other baselines in novel view synthesis as shown in Tab. 2 and Fig. 6. Compared to LGM [72] and TGS [102], which also produce 3D Gaussians, our renderings exhibit superior visual quality. We observe that MVD-Fusion [27], despite utilizing depth-aware 3D attention, still results in multi-view inconsistencies. In contrast, our method generates dense, surface-aligned Gaussians, ensuring multi-view consistency and yielding detailed and accurate renderings.

5.2. Ablation Study

Method	CD↓	IoU↑	PSNR↑	GPU↓	Time↓
Ours	**0.0152**	**0.7303**	17.19	9 GB	20 s
w/o Branch	0.0166	0.6716	**17.20**	11 GB	24 s

Table 4. **Ablations on our branched U-Net design**. We replace our branched U-Net design with domain-switch [49] for cross-domain latent denoising. Note that the study was performed at 256x256 resolution due to compute resource constraints.

Method	CD↓	IoU↑	PSNR↑	LPIPS↓	Time↓
Wonder3D	0.0236	0.6731	15.21	0.269	3min
Ours$_{Normal}$	0.0194	0.6930	15.43	0.261	3min
Ours	**0.0152**	**0.7303**	**17.19**	**0.171**	20s

Table 5. **Depth maps vs. normal maps as the representation.** Note that the study was performed at 256x256 resolution due to compute resource constraints.

We conduct ablation studies on the GSO dataset, as shown in Tab. 3 and Fig. 7. First, we show that removing epipolar attention in the depth decoder leads to a major drop of all quantitative metrics, due to the severe inconsistent multi-view depth maps and distorted mesh extraction, highlighting the indispensable role of epipolar attention. Second, without learning to generate the depth latent and provide it as additional input for latent decoding, *i.e.* only generating RGB latent and feed it to the decoder, we observe that the predicted depths are of lower quality

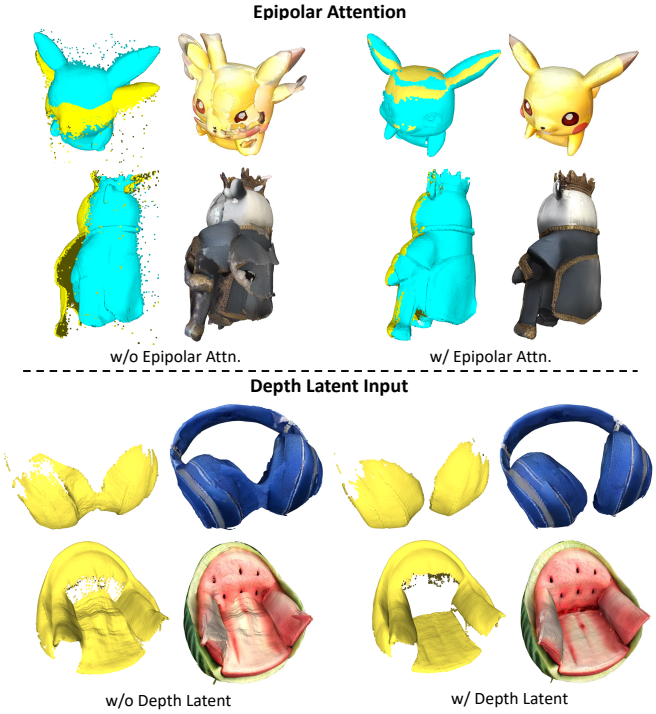

Figure 7. **Qualitative ablations on the GSO dataset.** We show back-projected depth points for 1 (yellow) or 2 views (yellow, cyan) and the final mesh.

and are inaccurate. In addition, to validate our branched U-Net design that performs simultaneous multi-domain latent denoising, we experiment replacing it with domain-switch technique from Wonder3D [49] which performs successive multi-domain denoising, as shown in Tab. 4. Our branched U-Net achieves 20% faster in inference and uses 18% less GPU memory, with no degradation in quality.

Is depth representation better than normal maps within the same framework? Our approach outperforms both in terms of generation quality and efficiency compared to predicting normal maps using our framework or Wonder3D (Tab. 5), which also predicts multi-view RGBD for 3D reconstruction. Normals do not represent depth discontinuities and require optimization to reconstruct the geometry, which often results in incorrect or blurry geometry.

6. Conclusion

We present an approach to directly predict explicit surface geometry and texture for single-image 3D reconstruction. Experiments show that our method significantly improves the speed and quality of 3D reconstructions compared to other benchmarks.

Acknowledgment. This research was supported by funding from Amazon, the NASA Biodiversity Program (Award 80NSSC21K1027), and the National Science Foundation (NSF) Grant IIS-2212046.

References

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

[2] Mohammadreza Armandpour, Huangjie Zheng, Ali Sadeghian, Amir Sadeghian, and Mingyuan Zhou. Reimagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*, 2023. 3

[3] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3

[4] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[5] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 5

[6] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2416–2425, 2023. 3

[7] 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. 3

[8] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *arXiv preprint arXiv:2403.14627*, 2024. 5

[9] 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. 3

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

[11] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 6

[12] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. 2023. 5

[13] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022. 5

[14] 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. 6

[15] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20637–20647, 2023. 3

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

[17] Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14300–14310, 2023. 3

[18] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 2, 3

[19] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, pages 11808–11826. PMLR, 2023. 3

[20] Jiatao Gu, Qingzhe Gao, Shuangfei Zhai, Baoquan Chen, Lingjie Liu, and Josh Susskind. Control3diff: Learning controllable 3d diffusion models from single-view images. In *2024 International Conference on 3D Vision (3DV)*, pages 685–696. IEEE, 2024. 3

[21] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5354–5363, 2024. 5

[22] 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. 1, 3

[23] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[24] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7779–7788, 2020. 2, 4

[25] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models, 2023. 3, 7

[26] 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. 3, 7

[27] Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, and Shubham Tulsiani. Mvd-fusion: Single-view 3d via depth-consistent multi-view generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9698–9707, 2024. 3, 7, 8

[28] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023. 3

[29] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 1, 3, 7

[30] Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. Spad: Spatially aware multi-view diffusers. In *CVPR*, 2024. 4

[31] Animesh Karnewar, Niloy J Mitra, Andrea Vedaldi, and David Novotny. Holofusion: Towards photo-realistic 3d generative modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[32] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 2, 5

[33] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 3

[34] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 4, 5, 7

[35] Seung Wook Kim, Bradley Brown, Kangxue Yin, Karsten Kreis, Katja Schwarz, Daiqing Li, Robin Rombach, Antonio Torralba, and Sanja Fidler. Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8496–8506, 2023. 3

[36] Lambda Labs. Stable diffusion image variations - a hugging face space. huggingface.co/lambdalabs/sd-image-variations-diffusers. 2, 5

[37] Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, et al. xformers: A modular and hackable transformer modelling library, 2022. 5

[38] Jiabao Lei, Jiapeng Tang, and Kui Jia. Generative scene synthesis via incremental view inpainting using rgbd diffusion models. *arXiv preprint arXiv:2212.05993*, 2022. 3

[39] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *arXiv preprint arXiv:2405.11616*, 2024. 4

[40] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arXiv preprint arXiv:2310.02596*, 2023. 3

[41] 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. 3

[42] 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. 3, 6, 7

[43] 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. 1, 3, 5, 7

[44] Xinhang Liu, Shiu-hong Kao, Jiaben Chen, Yu-Wing Tai, and Chi-Keung Tang. Deceptive-nerf: Enhancing nerf reconstruction using pseudo-observations from diffusion models. *arXiv preprint arXiv:2305.15171*, 2023. 3

[45] Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Yanyu Li, Dahua Lin, Xihui Liu, Ziwei Liu, and Sergey Tulyakov. Hyperhuman: Hyper-realistic human generation with latent structural diffusion. *arXiv preprint arXiv:2310.08579*, 2023. 4

[46] 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. 1, 2, 3, 4, 6, 7

[47] Ying-Tian Liu, Yuan-Chen Guo, Guan Luo, Heyi Sun, Wei Yin, and Song-Hai Zhang. Pi3d: Efficient text-to-3d generation with pseudo-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19915–19924, 2024. 3

[48] Zhen Liu, Yao Feng, Michael J Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffusion: Score-based generative 3d mesh modeling. *arXiv preprint arXiv:2303.08133*, 2023. 1, 3

[49] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3, 4, 5, 6, 7, 8

[50] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 1

[51] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2023. 1, 6, 7

[52] 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. 1, 3

[53] 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. 3

[54] 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. 1, 3, 7

[55] Evangelos Ntavelis, Aliaksandr Siarohin, Kyle Olszewski, Chaoyang Wang, Luc V Gool, and Sergey Tulyakov. Autodecoding latent 3d diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

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

[57] 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. 1, 7

[58] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *CVPR*, 2024. 3

[59] 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. 5

[60] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[61] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 5

[62] 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, 3, 4, 5

[63] 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. 3

[64] Yusuf Sahillioğlu and Ladislav Kavan. Scale-adaptive icp. *Graphical Models*, 116:101113, 2021. 6

[65] Hoigi Seo, Hayeon Kim, Gwanghyun Kim, and Se Young Chun. Ditto-nerf: Diffusion-based iterative text to omni-directional 3d model. *arXiv preprint arXiv:2304.02827*, 2023. 3

[66] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023. 3

[67] 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. 3, 4, 5

[68] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[69] Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, et al. Ldm3d: Latent diffusion model for 3d. *arXiv preprint arXiv:2305.10853*, 2023. 3

[70] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. *arXiv preprint arXiv:2312.13150*, 2023. 3, 5

[71] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. In *ICCV*, 2023. 3

[72] 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, 7, 8

[73] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Josh Tenenbaum, Frédo Durand, Bill Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[74] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. In *2024 International Conference on 3D Vision (3DV)*, pages 1554–1563. IEEE, 2024. 3

[75] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *CVPR*, 2023. 3

[76] Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039, 2022. 1, 3

[77] 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. 3

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

[79] 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. 3

[80] Xiaofeng Wang, Zheng Zhu, Guan Huang, Fangbo Qin, Yun Ye, Yijia He, Xu Chi, and Xingang Wang. Mvster: Epipolar transformer for efficient multi-view stereo. In *European Conference on Computer Vision*, pages 573–591. Springer, 2022. 2, 4

[81] 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. 7

[82] Zhen Wang, Qiangeng Xu, Feitong Tan, Menglei Chai, Shichen Liu, Rohit Pandey, Sean Fanello, Achuta Kadambi, and Yinda Zhang. Mvdd: Multi-view depth diffusion models. *arXiv preprint arXiv:2312.04875*, 2023. 4

[83] 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. 3

[84] 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. 3, 6, 7

[85] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 3

[86] Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. *arXiv preprint arXiv:2403.16292*, 2024. 5

[87] Jinbo Wu, Xiaobo Gao, Xing Liu, Zhengyang Shen, Chen Zhao, Haocheng Feng, Jingtuo Liu, and Errui Ding. Hd-fusion: Detailed text-to-3d generation leveraging multiple noise estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. 3

[88] Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 3d-aware image generation using 2d diffusion models. In *ICCV*, 2023. 3

[89] Dejia Xu, Ye Yuan, Morteza Mardani, Sifei Liu, Jiaming Song, Zhangyang Wang, and Arash Vahdat. Agg: Amortized generative 3d gaussians for single image to 3d. *arXiv preprint arXiv:2401.04099*, 2024. 3

[90] 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. 3, 7

[91] 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, 5

[92] Zhenpei Yang, Zhile Ren, Qi Shan, and Qixing Huang. Mvs2d: Efficient multi-view stereo via attention-driven 2d convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 4

[93] Paul Yoo, Jiaxian Guo, Yutaka Matsuo, and Shixiang Shane Gu. Dreamsparse: Escaping from plato's cave with 2d diffusion model given sparse views. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[94] Jonathan Young. https://github.com/jpcy/xatlas, 2021. 5

[95] Chaohui Yu, Qiang Zhou, Jingliang Li, Zhe Zhang, Zhibin Wang, and Fan Wang. Points-to-3d: Bridging the gap between sparse points and shape-controllable text-to-3d generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6841–6850, 2023. 3

[96] Jason J Yu, Fereshteh Forghani, Konstantinos G Derpanis, and Marcus A Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *ICCV*, 2023. 3

[97] 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. 3

[98] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. *arXiv preprint arXiv:2404.19702*, 2024. 3, 5

[99] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 7

[100] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 3

[101] Joseph Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance. *arXiv preprint arXiv:2305.18766*, 2023. 3

[102] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10324–10335, 2024. 1, 3, 7, 8