CraftsMan3D: High-fidelity Mesh Generation with 3D Native Diffusion and Interactive Geometry Refiner

Weiyu Li[1,2*], Jiarui Liu[1,2*], Hongyu Yan[1,2*], Rui Chen[1], Yixun Liang[1]
Xuelin Chen[3], Ping Tan[1,2], Xiaoxiao Long[1,2†]

[1]HKUST [2] LightIllusions [3]Adobe Research
[*]Core contributions [†]Corresponding author

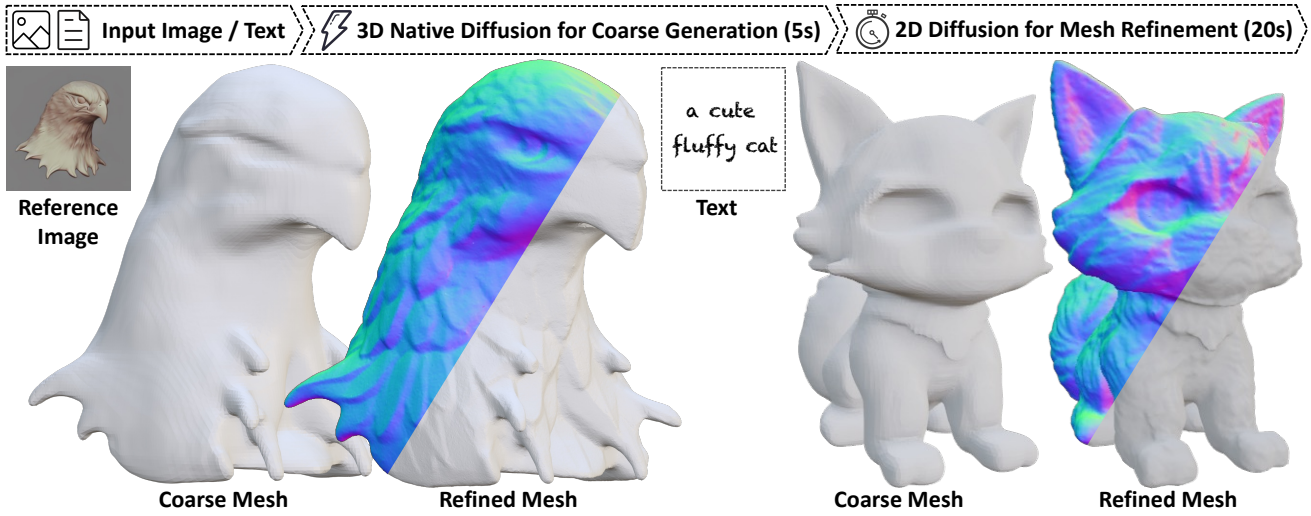


Figure 1. Our method, given a single reference image or text prompt, can generate intricate 3D shapes with high fidelity in just 25 seconds. Drawing inspiration from the typical workflow of the craftsman, we start by creating a coarse shape using a 3D native DiT model. We then enhance the surface details using either an automatic global geometry refiner or, more intriguingly, an interactive geometry refiner that allows for users to edit. For more visually compelling results, please refer to the supplementary video.

Abstract

We present a novel generative 3D modeling system, coined CraftsMan3D, which can generate high-fidelity 3D geometries with highly varied shapes, detailed surfaces, and, notably, allows for refining the geometry in an interactive manner. Despite the significant advancements in 3D generation, existing methods still struggle with lengthy optimization processes, self-occlusion, irregular mesh topologies, and difficulties in accommodating user editing, consequently impeding their widespread adoption and implementation in 3D modeling softwares. Our work is inspired by the craftsman, who usually roughs out the holistic figure of the work first and elaborates the surface details subsequently. Specifically, we first introduce a robust data preprocessing pipeline that utilizes visibility check and winding number to maximize the use of existing 3D data. Leveraging this data, we employ a 3D-native DiT model that directly models the distribution of 3D data in latent space, generating coarse geometries in seconds. Subsequently, a normal-based geometry refiner enhances local surface details, which can be applied automatically or interactively with user input. Extensive experiments demonstrate that our method achieves high efficacy in producing superior quality 3D meshes compared to existing methods.

1. Introduction

The rapid development of industries such as video gaming, augmented reality, and film production has led to a surge in demand for automatic 3D asset creation. However, existing methods still struggle to produce results that are ready to use.

3D generative methods can be broadly categorized into three types: i) Score-Distillation Sampling (SDS) based methods [5, 24, 45] typically distill priors in pretrained 2D diffusion models for optimizing a 3D representation, eventually producing 3D assets. However, these methods often suffer from time-consuming processing, unstable optimization, and multi-face geometries. ii) Multi-view (MV) based methods propose generating multi-view consistent images as intermediate representations, from which the final 3D can be reconstructed [20, 28, 30]. While these methods significantly improve generation efficiency and robustness, the resulting 3D assets tend to have artifacts and struggle to generate assets of complex geometric structures. iii) 3D native generation methods [18, 37, 57, 65] attempt to directly model the probalistic distribution of 3D assets via training on 3D assets. However, due to the limited 3D data and high-dimensional 3D representation, existing 3D generative models can not produce high-fidelity details. More importantly, all of these methods do not support user editing to improve the generated 3D interactively.

Challenges of scaling up native 3D generative models largely due to the uniform requirement of training data. Unlike the standardized structures of text and 2D images, 3D assets are from various sources—procedural functions, 3D modeling, or scanning, resulting in diverse mesh topologies such as closed, open, double-sided, non-manifold that require careful handling to maintain geometric integrity, making uniform dataset creation difficult. Point-E [37] pioneers a large-scale model trained on millions of 3D assets to generate 3D point clouds from text prompts. While point clouds reduce data acquisition costs, they lack topological detail, limiting their real-world utility. Implicit distance fields, like signed distance fields (SDF), offer a better alternative due to their continuous, watertight properties, allowing for high-quality 3D mesh extraction. Consequently, existing 3D datasets often require preprocessing to convert meshes into SDFs. Leveraging this, Shape-E [18] improves 3D generation quality, while recent models like CLAY [65] and Direct3D [57] adopt advanced diffusion techniques. However, none of these methods can generate high-fidelity geometric details and limitations in mesh-to-SDF conversions still result in training difficulty.

To tackle problems mentioned above, we first propose an efficient and robust mesh-to-SDF algorithm that maximizes the utilization of existing 3D data [8, 9]. By integrating visibility checks with winding number analysis, we significantly enhance the success rate of the watertight conversion and form a high-quality 3D dataset based on Objaverse [8]. Built on the 3D data, we present a two-stage generative 3D native generation system, coined CraftsMan, which takes as input single images as reference or text prompts and generates high-fidelity 3D geometries featuring highly varied shapes, regular mesh topologies, and detailed surfaces, and, notably, allows for interactively refining the geometry. Drawing inspiration from craftsmen, who typically begin by shaping the overall form of their work before subsequently refining the surface details, our system is comprised of two stages: 1) a native 3D diffusion model, that is conditioned on single image and directly generates coarse 3D geometries; and 2) a robust generative geometry refiner that provides intricate details powered by Poisson Normal Blending and Relative Laplacian Smoothing regularization.

In summary, our main contribution lies in three aspects:

- A robust and efficient data pre-processing pipeline that integrates visibility checks enhanced by the winding number and significantly improves the success rate of watertight mesh conversion.
- A simple yet effective 3D Native DiT model. Extensive experiments demonstrate that our simple structure achieves high efficacy in producing superior quality 3D assets compared to existing methods.
- A novel normal-based interactive mesh refiner which can produce highly enhanced geometries within just 20 seconds and support interactive manipulation, enhancing the generated coarse geometries to better align with the users' envisioned designs.

2. Related work

In this section, we will first provide a brief review of the relevant literature on 3D generation, followed by a discussion of recent works focused on 3D native generative models.

2.1. 3D Generation using 2D Supervision

In recent years, generative models have achieved significant success in producing high-fidelity and diverse 2D images, and we have seen a surge of interest in lifting this powerful 2D prior to 3D generation. Most of these methods generate 3D contents, typically in the form of NeRF [35] or Triplane [2] representations, which are turned into images by a differentiable renderer. Then the multiview images can be compared with either real-world dataset samples or images rendered from 3D models to train a generative model. [3, 11, 38, 50] perform GAN-like [12] structure to synthesize 3D-aware images via adversarial training.

However, these methods are often trained on limited views with specific categories, and therefore shows poor generalization on unseen categories. [45] develop techniques to distill 3D information from a large-scale pretrained 2D text-to-image diffusion models to optimize 3D representation, thus yielding 3D assets. Subsequent works [5, 22–24, 51, 55] are proposed to further enhance the quality of 3D generation. By leveraging existing powerful 2D priors, these per-shape optimization methods take dozens of minutes and require a huge computational cost.

Instead of performing a time-consuming optimization, recent works [20, 26, 28, 30] attempt to generate multi-

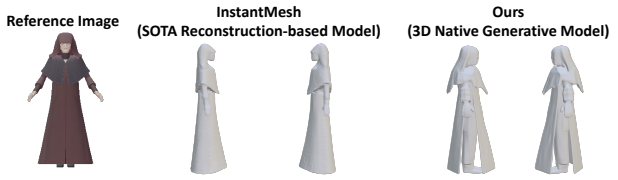


Figure 2. Compared to the SOTA reconstruction-based models, our result produces accurate complex geometric structures, including those that are self-occluded in the input images.

.

view images simultaneously and bring 3D-awareness by finetuning the 2D diffusion models. The generated multi-view images are then used to reconstruct a 3D shape using sparse view reconstruction algorithms or Large Reconstruction Models (LRM). Although these methods achieve high efficiency, the generated results are heavily dependent on the quality of the 2D images. Indirect modeling of 3D probability distributions is insufficient for faithfully recovering geometric information. Self-occlusion, complex lighting conditions, and multi-view inconsistency are still challenging, usually result in degraded final generation quality, which can be validate in Figure 2. In contrast, our approach modeling the distribution of 3D data, enabling high-quality mesh generation even with complex inputs.

2.2. 3D Native Generative Models

Unlike approaches that rely on 2D supervision, many works adopt various 3D representations such as point clouds [19, 59, 67], meshes [29, 36], and implicit functions [6, 40, 53] to train native 3D generative models. Building up on recently advanced diffusion models [13], a series of works began to conduct 3D diffusion models with the representation of point cloud [31, 67], meshes [29] and implicit fields [7, 52, 60]. However, training these 3D generative models directly on 3D data is quite challenging, due to the high memory footprint and computational complexity. To tackle these challenges, inspired by the success of latent diffusion [49], recent studies [16, 66] first compress 3D shapes into compact latent space, and then perform diffusion process in the latent space. For instance, [62] and [63] propose a method to encode occupancy fields using a set of either structured or unstructured latent vectors. Neural Wavelet [16] advocates a voxel grid structure containing wavelet coefficients of a Truncated Signed Distance Function (TSDF). One-2-3-45++ [25] and XCube [48] focus on explicit dense grid volume. The most recent works, Michelangelo [66] and CLAY [65], train a diffusion model on latent set representations, and Direct3D [57] explores a triplane representation to enhance training scalability. However, these works often suffer from lacking geometric details, over-smoothing surfaces, and unstable training processes. Our work harnesses the feed-forward nature of 3D diffusion models while enhancing its generalization capability by leveraging the prior from pre-trained multi-view 2D diffusion as the condition. This approach significantly facilitates zero-shot ability and robust generation.

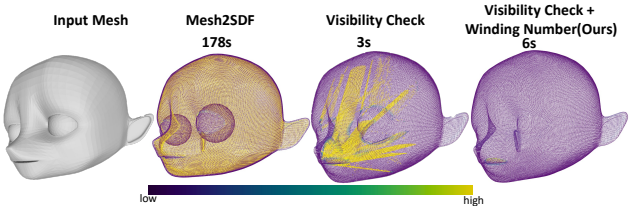


Figure 3. Error maps of different mesh-to-sdf methods. We sample surface points from the processed meshes for each method and show the differences compared to the ground truth mesh.

3. Method

Our 3D generation framework mirrors the 3D artist's workflow, which begins typically with the creation of a rough geometry that is then refined in the subsequent stage. Figure 4 illustrates our generative 3D modeling workflow, that is capable of producing high-quality, detailed 3D assets.

In this section, We begin by introducing our data preprocessing (Sec.3.1), which significantly improves the success rate of watertight conversion and maximizes the utilization of existing 3D data. Following this, we train a Variational Auto-Encoder (VAE) on the watertight meshes to learn latent set-based representations[63] and output a TSDF field. Next, we train a dedicated DiT-based denoising network that operates on these learned latent representations, using the intermediate multi-view image as conditioning (Sec.3.2). Finally, our framework features a normal map-based geometry refinement scheme (Sec.3.3).

3.1. Data Preprocessing.

Standardizing the geometric data is essential for effectively training a 3D generative model. Due to the significant noise in the geometry and appearance, we first filter out low-quality meshes, including those with point clouds, thin structures, holes, and textureless surfaces to form our initial subset. Ensuring that the mesh is watertight is also essential for extracting the SDF (Signed Distance Function) field from the processed meshes as supervision [65] when training a Shape VAE [63, 66]. Although the dataset proposed in [8, 9] claims to have nearly ten million objects, the vast majority of it is non-watertight, such as scanned point clouds and planes, resulting in less than 1% of the data can be directed used. Therefore, we propose an efficient and effective method for converting mesh into a watertight one.

Winding Number-Enhanced Watertight Conversion. Dual Octree Graph Networks (DOGN) [54] proposed a "mesh-to-SDF" approach, which requires a significant amount of time. CLAY [65] introduced a "visibility check"

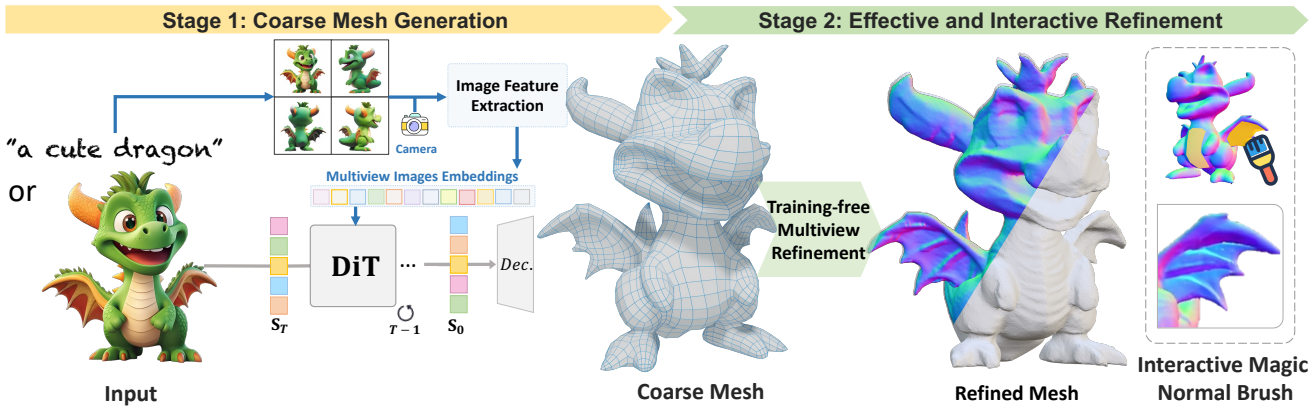


Figure 4. Overview of CraftsMan3D. We first using a multi-view diffusion model to generate a multi-view image from the input single image or text prompt. The generated multi-view image is then fed into our Latent Set-based DiT model as conditioning to produce a coarse mesh. Finally, a dedicated refinement module is employed to improve or edit the surface normals of the coarse geometry, enhancing with intricate details. In particular, this refinement module features two key usages, namely the automatic global refinement and interactive magic brush, that contribute to efficient and controllable 3D modeling of high-quality meshes.

method for remeshing, which maximizes positive volume while faithfully preserving geometric features. However, as shown in Figure 3, for the non-manifold objects with holes, it is easy to encounter floaters inside the converted mesh. To tackle these challenges, we enhance the visibility check by incorporating the concept of the winding number [32], which is an effective tool for determining whether points are inside or outside a shape. When the input point cloud has well-defined normals, the winding number can reliably differentiate between the inside and outside in a global manner. Specifically, we first randomly choose 50 cameras on a sphere and use a dense grid with a resolution of 256 or 512 for visibility check. If the center of a grid cell is not visible to all the cameras, we further check the winding number of it. Once the value of the winding number indicator function is greater than a threshold, which we set to 0.75 by default, we treat that point as being inside the object. Thus, we can get robust inside-outside test results. This approach statistically improves our watertight conversion success rate from 60% to 80% on [8]. Please refer to the supplementary for more details.

3.2. Multi-view guided 3D generation model

3D Shape VAE. Following [66], we adopt a Perceiver-based [17] shape VAE to encode the 3D shape into a set of latent vectors $\mathbf{S}$ and then decode them to reconstruct the neural field of the original 3D shape. Figure 5(a) shows the network architecture. Specifically, for each 3D shape, we first sample on the 3D surface to obtain a set of points $\mathbf{P}_c \in \mathbb{R}^{N\times 3}$, as well as a set of surface normal vectors $\mathbf{P}_n \in \mathbb{R}^{N\times 3}$ at these point positions. The encoder is trained to map points $\mathbf{P}_c$ and $\mathbf{P}_n$ into a latent vector set $\mathbf{Z}$, which a decoder then translates into an implicit field representation. Notably, we replace the original occupancy field with a TSDF field using a threshold of $1/256$ for stable optimization and better performance.

Multi-view Guided 3D Diffusion Model. Instead of directly using the input single image or text prompt as conditioning, our DiT-based diffusion model is conditioned on the multi-view (MV) images that capture the target 3D asset. During inference the pre-trained text-based [51] or image-based [21, 30, 56] MV diffusion models are used to generate the corresponding MV image from the input single image or text prompt accordingly. MV images generated by recent MV diffusion models offer richer geometric and contextual information compared to using a single image or text alone. As a result, the multi-view conditioned DiT model enables improved generation of various 3D shapes, particularly on unobserved regions from the single input image.

With the latent set representation $\mathbf{S}$ of a shape and its corresponding multi-view images $\hat{\mathbf{y}}$, we now train a MV-conditioned DiT model. To make image embeddings be aware of the camera position, we follow the method [20] to modulate the camera parameters π during the feature extraction, by employing an adaptive layer normalization (adaLN) [42]. Formally, the conditioned embeddings c can be represented by:

$$c = \varphi_{clip}(\hat{\mathbf{y}}, ModLN(\pi)) + \varphi_{mlp}(\varphi_{dino-v2}(\hat{\mathbf{y}}, ModLN(\pi))) \quad (1)$$

where φ_{clip} and $\varphi_{dino-v2}$ are pretrained CLIP [47] and DINO-v2 [1] and φ_{mlp} is a small MLP that aligns DINO features with CLIP features. Then, we can learn the Multi-view guided Latent Set Diffusion Model (LSDM) via:

$$\mathcal{L}_{LSDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{S}_t, t, c)\|_2^2\right], \quad (2)$$

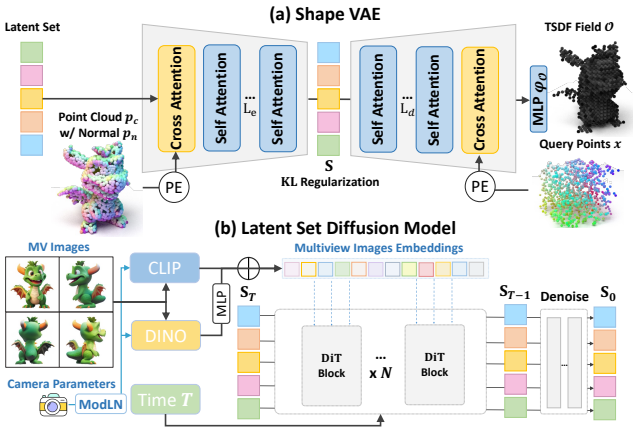


Figure 5. The illustration of 3D generation. (a.) We first train a 3D Variational Autoencoder (VAE) to compress 3D shape into a latent space, which takes point clouds with normals as input and outputs TSDF fields. (b.) With the learned latent space, we train a 3D Latent Set DiT Model that using multi-view images as conditions.

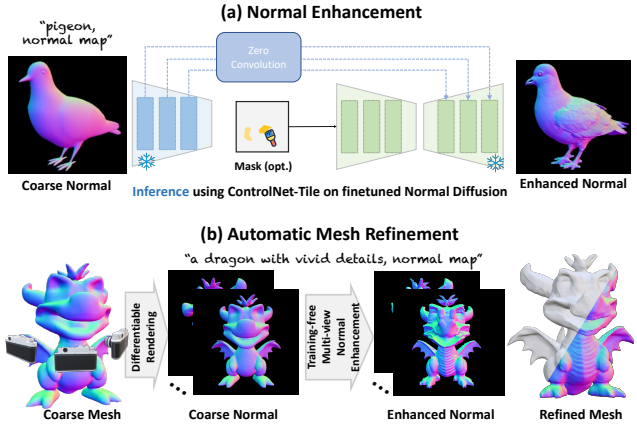


Figure 6. The illustration of surface normal-based geometry refinement. (a) The normal-adapted diffusion model is combined with ControlNet-Tile to enhance a normal with intricate details. (b) The automatic mesh refinement process via training-free cross-view attention.

where ϵ_θ is build on a DiT [41] model, t is time step and $\mathbf{S}_t$ is a noisy version of $\mathbf{S}_0$. To reduce the number of parameters and computational cost, we employ adaLN-single [4] in each DiT block.

3.3. Normal-based Geometry Refinement

To further enhance the coarse mesh, we propose to improve the initial mesh using normal maps as an intermediate representation. We first render the normal maps of coarse mesh and then leverage normal-based diffusion to enhance the rendered normals with intricate details. Subsequently, the refined normals serve as supervision to optimize the mesh, thus yielding a refined mesh with rich details. Moreover, this process also can be performed in an interactive way. Users can select the areas to be edited using a painting brush, creating a binary mask that indicates the regions to be updated. Please refer to the supplementary video for more visual results.

Intermediate Normal Guidance Generation We adopt ControlNet-Tile [64] that is finetuned on a normal dataset [8, 15] to enhance the rendered normals with details. A pivotal challenge arises from the inconsistencies observed in the normal images generated by diffusion models across different views. Recent advancements, as detailed in [30, 51], address this issue by employing a cross-view attention mechanism. Interestingly, we have observed that the cross-view attention mechanism can be directly applied to our task in a training-free manner. This is partially attributable to the inherent constraints of the coarse normal maps and the design of ControlNet-Tile, which hallucinates new details without significantly altering the original input conditions. Formally, during the diffuse process, for the i_{th} view with a rendered normal map n_i, we replace the K and V in the original attention layer with:

$$K = W^K \left[z^0, \cdots, z^K\right], V = W^V \left[z^0, \cdots, z^K\right], \quad (3)$$

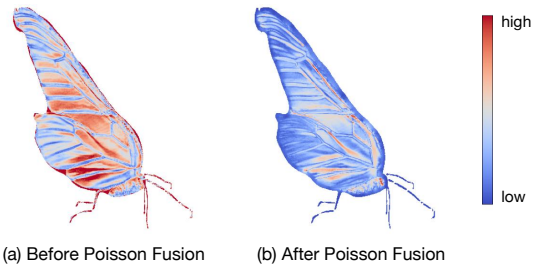
Here, the key K and value V are globally shared for all input views.

Shape Editing via Normal-based Optimization We advocate for direct vertex optimization through continuous remeshing [39], which is favored for its computational efficiency and explicit control over the optimization process. Given a mesh with vertices V and faces F, we optimize the mesh details by directly manipulating the triangle vertices and edges, with the supervision of the refined normal maps $\hat{n}_i$. Specifically, in each optimization step, we render normal maps from the current mesh via differentiable rendering, denoted as $\mathcal{R}_n(V, F, \pi_i)$. Then, we minimize the L1 differences between the rendered normals and the refined normals via:

$$\mathcal{L}_{remeshing} = \sum_i \|\hat{n}_i - \mathcal{R}_n(V, F, \pi_i)\|_1^1, \quad (4)$$

where $\mathcal{R}_n$ denotes the differentiable normal rendering function and π_i is the camera information of i_{th} rendering camera. In each step, an update operation is executed to update the position for each vertex according to the gradient computed in the loss backward process.

Poisson Normal Blending Diffusion models generate normals maps by regarding them as a specific domain of images. We found that normal maps generated this way sometimes are inaccurate, which results in unstable optimization.



(a) Before Poisson Fusion (b) After Poisson Fusion

Figure 7. Distance map with coarse normal. Normal maps enhanced by stable diffusion contain low-frequency changes from original normal map(shown in red in (a)), which will result in global distortion of input shapes. Applying Poisson Fusion eliminate global distortions, resulting in the preservation of global shapes and the enhancement of high details(b).

Figure 7(a) shows the pixel-wise L2 distance between the normal map rendered from coarse mesh and the normal map enhanced by the normal stable diffusion, which shows significant changes in the initial shape and leads to stretched shape during the optimization. To address this, we try to eliminate the influence of those low frequency changes and only take use of the local details contained in the enhanced normal map. We accomplish this by employing the efficient and traditional Poisson Blending algorithm [43]:

$$n_{fused} = \Gamma\left(\hat{n}, \mathcal{R}_n(V, F, \pi), m\right). \tag{5}$$

we denote Γ as the Poisson Blending algorithm, $\mathcal{R}_n(V, F, \pi)$ and $\hat{n}$ are the rendered normal map and enhanced normal map respectively. m denotes the mask rendered from coarse mesh, which will be used to label the target region to be fused.

Relative Laplacian Smoothing Previous methods [39] often achieve stable optimization by introducing Laplace regularization term. However, this term avoids undesirable results by forcing each vertex close to the coordinate origin in a local Laplace coordinate, which inevitably cause the shrink of the shape. Fortunately, in our detail enhancement task, our initial coarse mesh contains a good prior, thus we do not need to constrain the smoothness by enforcing the Laplace coordinate to zero, but punishing the change of the Laplace coordinate comparing to the initial shape, which is called relative Laplacian smoothing term. Given a coarse shape with vertices $\mathbf{x}$, we compute the initial Laplace coordinate by $V_{init}^{\mathbf{W}} = \mathbf{W_{init}} V_{init}$, here V_{init} is the initial vertex coordinate of coarse mesh, $\mathbf{W_{init}}$ is the corresponding Laplacian matrix. Then in every optimization step, we regularize the deformation process by

$$x \leftarrow x + \lambda v(\mathbf{W}V - V_{init}^{\mathbf{W}}), \tag{6}$$

Table 1. Quantitative comparison with baseline methods on the GSO dataset [10]. We follow [28, 56] and randomly choose 30 shapes from GSO for comparison. Each shape in aligned by conducting an ICP register to calculate the metrics[33].

Type	Method	CD↓	IoU↑	Time↓
Recon.-based Model	One-2-3-45 [26]	0.0629	0.4086	~45s
	zero123 [27]	0.0339	0.5035	~10min
	InstantMesh [58]	**0.0187**	**0.6353**	~10s
SDS-based Model	Realfusion [33]	0.0819	0.2741	~90min
	Magic123 [46]	0.0516	0.4528	~60min
3D Generative Model	Point-E [37]	0.0426	0.2875	~40s
	Shap-E [18]	0.0436	0.3584	~10s
	Michelangelo [66]	0.0404	0.4002	~3s
	One2345++ [25]	0.0437	0.3386	~20s
	Ours	**0.0291**	**0.5347**	~5s

Table 2. Quantitative comparison on subset which contained self-occlusion in the input images. Our 3D generative model demonstrated a significant performance.

Method	CD↓	IoU↑
InstantMesh [58]	0.04909	0.50151
Ours	**0.03943**	**0.53215**

where x_{init} is the initial vertex position, λ is a smoothing hyperparameter. Please refer to the [39] for more details.

4. Experiments

To validate the effectiveness of our proposed workflow, we extensively evaluate our proposed framework using a rich variety of inputs. We present the qualitative and quantitative evaluation of our method as described in Section 4.2 and Section 3.3, as well as comparison results against other baseline methods, showing the effectiveness and efficiency compared to other generation methods. We also conduct ablation studies to validate the effectiveness of each component in our framework, as described in Section 4.4. More intriguing visual results can be found in our accompanying video and supplementary.

4.1. Implementation Details

We follow the same architecture as in [66] for our shape auto-encoder, with the exception of the layer dedicated to contrastive learning, and for our latent set diffusion model. The shape auto-encoder is based on a perceiver-based transformer architecture with 185M parameters, while the latent set diffusion model is based on a DiT, comprising 500 million parameters. We train the diffusion model on 32 A800 GPUs using ground truth multi-view images, which share common approaches in related works in this area like [14, 25], etc. Additional details, including dataset, training settings can be found in our supplementary.

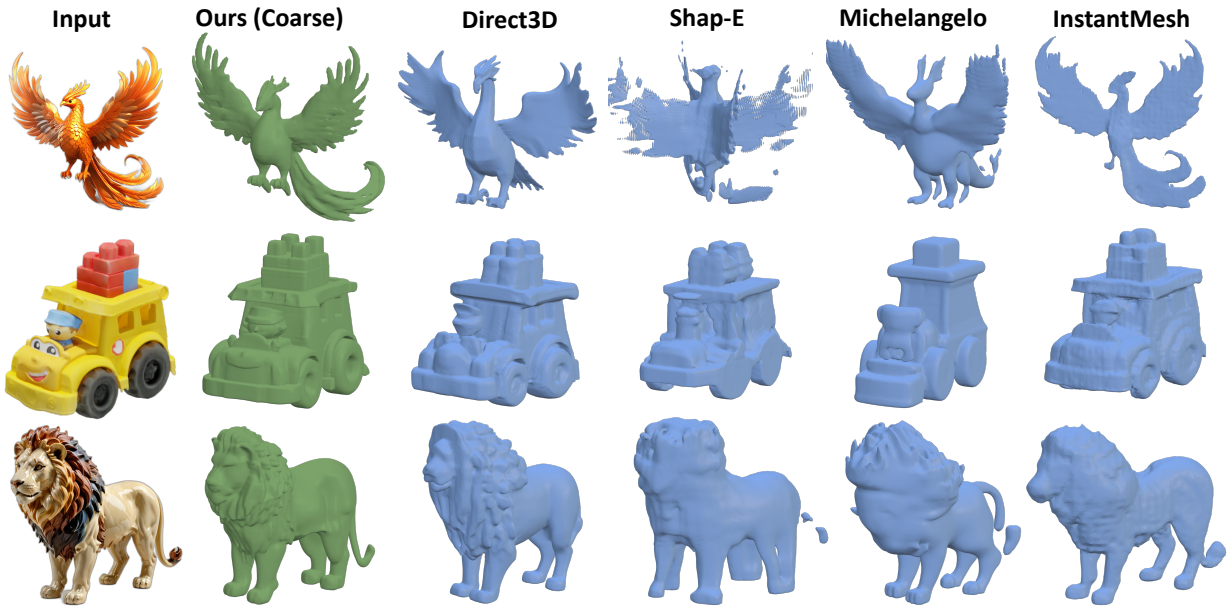


Figure 8. Qualitative comparisons with baseline methods for the task of single-view generation.

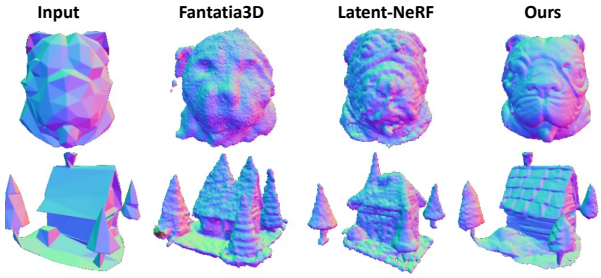


Figure 9. Qualitative comparisons with baseline methods for mesh refinement. To better showcase the effects of our mesh refinement, we performed a decimation operation on the input mesh.

4.2. Evaluation of Mesh Generation

In this evaluation, we focus on presenting the quality of our 3D generation model through a variety of results, and also present quantitative data for reference. We compare our model with several 3D generative models [18, 57, 66] and the state-of-the-art large reconstruction model(LRM) [58]. Given that CLAY [65] is not publicly available and our request to obtain their results haven't been responded, we only present the visual results in supplementary.

As shown in Fig. 8, our 3D native diffusion model produces coarse geometry with regular topology, and the coarse meshes are further enhanced with more intricate details. On the contrary, the 3D native counterpart Shap-E tends to produce noisy surfaces and incomplete shapes, while Michelangelo produces over-smoothed geometries and also suffers from shape ambiguity, like the second example in Fig. 8. Although InstantMesh produces accurate geometries, it can not handle complex geometry structures which results in adhesive geometry and lacks geometric details, take the phoenix in the first line for an example. Compared with Direct3D, our method achieves better consistency between the input image and the generated mesh.

Following the prior works [28, 30, 58], we also employ the Google Scanned Object dataset [10]—a rich collection of common everyday objects—to evaluate the performance of our 3D Diffusion Model in generating 3D models from single images. We adopt widely-used Chamfer Distances (CD) and Volume Intersection over Union(IoU) as the metrics. For each object in the evaluation set, we use the front view image as input. To align the input for a fair comparison, we first generate multi-view images from input image using existing multi-view diffusion models [30, 56]. The quantitative evaluation of the quality of our image-to-3D generation is shown in Table 1. Our method surpasses all the generation based methods and displays comparable results in a shorter time compared to the reconstruction based method InstantMesh [58]. We notice that the distribution of the GSO dataset is kind of monotonous,lacking mesh with complex structures and self occlusion, which is exactly where our model excels. To fully demonstrate the superiority of our method, we randomly choose a subset from the Objaverse dataset for further evaluation. As shown in Table 2, in this dataset with more complex geometries, the performance of our method is superior to InstantMesh. We also report the time consumption of different methods. In contrast to the SDS-based methods that usually require hours to optimize, our method obtains the resulting mesh in just a few seconds.

Table 3. Quantitative comparison for mesh refinement

Method	CLIP similarity ↑	Time ↓
Fantasia3D [5]	0.2567	~15min
Latent-NeRF [34]	0.2725	~1h
Ours	**0.2821**	~20s

Table 4. Ablation study of multi-view guided 3D diffusion model

Method	CD↓	IoU↑
w/o MV condition	0.0317	0.5892
w/o Camera Injection	0.0249	0.6561
ours-Cost Volume	0.0223	0.6583
ours	**0.0188**	**0.7059**

4.3. Evaluation of Mesh Refinement

To further assess the efficiency of our mesh refinement technique, we compare our method with recent approaches, specifically Fantasia3D [5] and Latent-NeRF [34]. To reduce the influence of the initial mesh and validate the strong detail enhancement power of our refinement, we reduce the number of face of initial shapes to 500. For the comparison with Fantasia3D, we employ the coarse mesh for initialization and only conduct the geometry modeling stage. In the case of Latent-NeRF, we use the input mesh as Sketch Shapes and train the NeRF in Sketch-Shape mode. All comparative experiments were conducted under their default settings. The visual results presented in Figure 9 demonstrate that our mesh refinement technique outperforms previous methods, producing not only clear and coherent outcomes but also effectively integrating high-quality details without compromising the overall structural integrity of the original mesh. Additionally, we provide a quantitative evaluation of our mesh refinement. We selected 20 objects from the Objaverse dataset and employed the same text descriptions as guidance. Table 3 presents the CLIP [47] similarity scores and the corresponding running times for each method. Our mesh refinement achieved a higher CLIP similarity compared to previous methods, while also demonstrating faster refinement speeds.

4.4. Ablation Study

We conduct comprehensive ablation studies to substantiate the effectiveness of each design element within our workflow, showing the importance of each component in the generation of high-quality 3D meshes.

Mutil-view images condition. In comparison to the single-image condition, the multi-view images generated by the 2D diffusion model offer enhanced information about the object, which is advantageous for generating unseen parts of 3D meshes. By incorporating camera poses into the image feature extractor, our model can better differentiate embeddings from various views of the object, ultimately leading to more accurate 3D shape generation. In the absence of camera pose information, the model is prone to producing 3D geometries with incorrect orientations. Unlike CLAY [65], which employs a cost volume that integrates camera pose information before feeding it into their diffusion model, which requires precise camera poses for accurate back projection. To demonstrate the superiority of our design in the context of multi-view images with camera pose injection, we conducted a comparison on our selected subset, which evaluated by the metrics of Chamfer Distance (CD) and Intersection over Union (IoU). As shown in Table 4, our approach achieved the best performance.

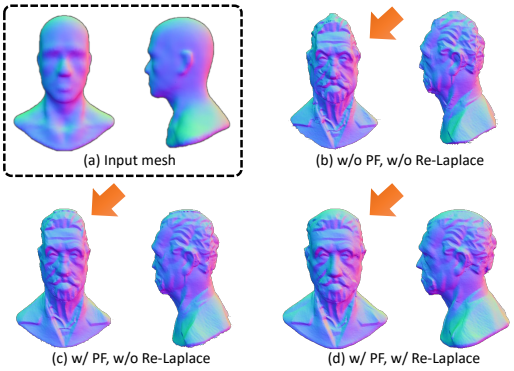


Figure 10. Ablation Study of the normal-based geometry refinement. We demonstrate the enhancement of our Poisson Fusion(PF) and Relative Laplace(Re-Laplace) module.

Regularizations During Mesh Optimization. Our proposed regularization terms eliminate the global distortions introduced in the detail enhancement process by normal stable diffusion, constraint the vertices towards the proximity of the coarse mesh, avoiding the mesh shrink introduced by the shape independent local smoothness thereby enabling a robust optimization process. As shown in Figure 14, directly refining the mesh without Poisson Fusion (PF) and Relative Laplace regularization (R-Laplace) results in an oddly sharp head due to global bias from normal stable diffusion. Although Poisson Fusion corrects this bias, the shape still shrinks. Replacing Original Laplace regularization with R-Laplace leading to a more reasonable shape.

5. Conclusion and Discussion

We present *CraftsMan3D*, a pioneering framework for the creation of high-fidelity 3D meshes that mimics the modeling process of a craftsman, all within a mere 30 seconds. Our approach begins with the generation of a coarse geometry, followed by a refinement phase that enhances surface details. Despite our method's capability to produce high-quality 3D meshes, the controllability of the Latent Set Diffusion model warrants further investigation, and the generation of texture for 3D meshes presents a promising avenue for future research.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the International Conference on Computer Vision (ICCV), 2021. 4

[2] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In arXiv, 2021. 2

[3] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 5799–5809, 2021. 2

[4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 5

[5] 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 (ICCV), 2023. 2, 8

[6] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 5939–5948, 2019. 3

[7] Gene Chou, Yuval Bahat, and Felix Heide. Diffusion-sdf: Conditional generative modeling of signed distance functions. In International Conference on Computer Vision (ICCV), pages 2262–2272, 2023. 3

[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. arXiv preprint arXiv:2212.08051, 2022. 2, 3, 4, 5, 11

[9] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. arXiv preprint arXiv:2307.05663, 2023. 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. 6, 7

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

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014. 2

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. arXiv preprint arxiv:2006.11239, 2020. 3

[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, 2024. 6

[15] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation, 2024. 5, 11

[16] Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. Neural wavelet-domain diffusion for 3d shape generation. 2022. 3

[17] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In International Conference on Machine Learning (ICML), pages 4651–4664. PMLR, 2021. 4

[18] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463, 2023. 2, 6, 7

[19] Chun-Liang Li, Manzil Zaheer, Yang Zhang, Barnabas Poczos, and Ruslan Salakhutdinov. Point cloud gan. arXiv preprint arXiv:1810.05795, 2018. 3

[20] 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. arXiv preprint arXiv:2311.06214, 2023. 2, 4, 16

[21] 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

[22] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. International Conference on Learning Representations (ICLR), 2024. 2

[23] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. arXiv preprint arXiv:2311.11284, 2023.

[24] 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 2

[25] 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, 6

[26] 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. 2024. 2, 6

[27] 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, 2023. 6

[28] 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, 6, 7, 16

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

[30] 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, 4, 5, 7, 16

[31] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 3

[32] A.L.F. Meister. Generalia de genesi figurarum planarum et inde pendentibus earum affectionibus. 1769. 4

[33] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360 reconstruction of any object from a single image. In Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 6

[34] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. arXiv preprint arXiv:2211.07600, 2022. 8

[35] 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. In European Conference on Computer Vision (ECCV), 2020. 2

[36] Charlie Nash, Yaroslav Ganin, SM Ali Eslami, and Peter Battaglia. Polygen: An autoregressive generative model of 3d meshes. In International conference on machine learning, pages 7220–7229. PMLR, 2020. 3

[37] 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, 6

[38] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 11453–11464, 2021. 2

[39] Werner Palfinger. Continuous remeshing for inverse rendering. Computer Animation and Virtual Worlds, 33(5):e2101, 2022. 5, 6

[40] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 165–174, 2019. 3

[41] William Peebles and Saining Xie. Scalable diffusion models with transformers. In International Conference on Computer Vision (ICCV), pages 4195–4205, 2023. 5

[42] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In Association for the Advancement of Artificial Intelligence(AAAI), 2018. 4

[43] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pages 577–582. 2023. 6

[44] ZBrush: 3D Sculpting Software. Pixologic, USA, 2023 edition, 2023. https://pixologic.com/. 16

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

[46] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In International Conference on Learning Representations (ICLR), 2024. 6

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (ICML), pages 8748–8763. PMLR, 2021. 4, 8

[48] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 3

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 10684–10695, 2022. 3

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

[51] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. arXiv:2308.16512, 2023. 2, 4, 5

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

[53] Jia-Mu Sun, Tong Wu, and Lin Gao. Recent advances in implicit representation-based 3d shape generation. Visual Intelligence, 2(1):9, 2024. 3

[54] Peng-Shuai Wang, Yang Liu, and Xin Tong. Dual octree graph networks for learning adaptive volumetric shape rep-

resentations. ACM Transactions on Graphics (SIGGRAPH), 41(4), 2022. 3

[55] 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. arXiv preprint arXiv:2305.16213, 2023. 2

[56] 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. 4, 6, 7

[57] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. arXiv:2405.14832, 2024. 2, 3, 7

[58] 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. 6, 7

[59] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. arXiv, 2019. 3

[60] Lior Yariv, Omri Puny, Natalia Neverova, Oran Gafni, and Yaron Lipman. Mosaic-sdf for 3d generative models. arXiv, 2023. 3

[61] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023. 17

[62] Biao Zhang, Matthias Nießner, and Peter Wonka. 3DILG: Irregular latent grids for 3d generative modeling. In Advances in Neural Information Processing Systems (NeurIPS), 2022. 3

[63] 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 (SIGGRAPH), 42(4), 2023. 3

[64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 5

[65] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. arXiv preprint arXiv:2406.13897, 2024. 2, 3, 7, 8, 13, 14, 17, 18

[66] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, BIN FU, Tao Chen, Gang YU, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. In Advances in Neural Information Processing Systems (NeurIPS), 2023. 3, 4, 6, 7, 11

[67] 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. 3

6. More Implementation Details

In this section, we describe a more detailed implementation, including the data preparation and model training details.

6.1. Data Preparation

We present more implementation details of data preparation for each component in our method. We obtained a collection of 190K objects refined from [8] for training. For each mesh, we first normalize the object to fit within a unit cube and then convert it into a water-tight mesh. To facilitate the training of the shape auto-encoder, we uniformly sample 500k points on the surface as input for the shape encoder. Furthermore, we sample 500k points in the volume and another 500k points near the surface of each mesh and then compute the occupancy value through SDF as the target for the shape decoder. For the 3D latent set diffusion model, we render 4-orthogonal views of each object as multi-view guidance, with a random rotation of azimuth in the range of $[-45, 45]$ and elevation angles in the range of $[-10, 30]$ for 5 times, resulting in a total of 25 images for each object. We also render 20 images for each object with random camera poses to generate the normal map for finetuning the 2D normal diffusion model.

6.2. Model Training

Following the approach in [66], we use the following architecture for the shape auto-encoder: the number of self-attention layers L_e and L_d are set to 8 and 16 respectively, while the number of the latent sets D and feature dimension C are set to 256 and 768 respectively. It is trained on the Adam optimizer with a learning rate of 5e-5 and a total batch size of 1024 using 8x A100 GPUs for 3 days. For the conditional Latent Set Diffusion Model (LSDM), we implement ϵ_θ with an Unet-like transformer consisting of 13 self-attention blocks. Each block contains 12 heads with 64 dimensions. We train ϵ_θ on the Adma optimizer with a learning rate of 5e-5 and a total batch size of 1024 using 32x A800 GPUs for around 7 days.

For inference, we use DDIM sampling scheduler with 50 steps, which generates a 3D mesh within 10 seconds. For the normal-adapted diffusion model, which is derived from SD1.5, we opt for convenience to fine-tune the model introduced in [15]. This model was originally fine-tuned on high-quality human normals and is further refined using our rendered normal images, trained using 8 A100 GPUs for one day.

7. More Results

In this section, we firstly present more results for a more intuitive perception of the effectiveness of our method, then delve into a more comprehensive discussion highlighting the advantages of each part of our method. Then, we present

Figure 11. Raw coarse meshes generated by our proposed method using a single image as a reference or a text prompt.

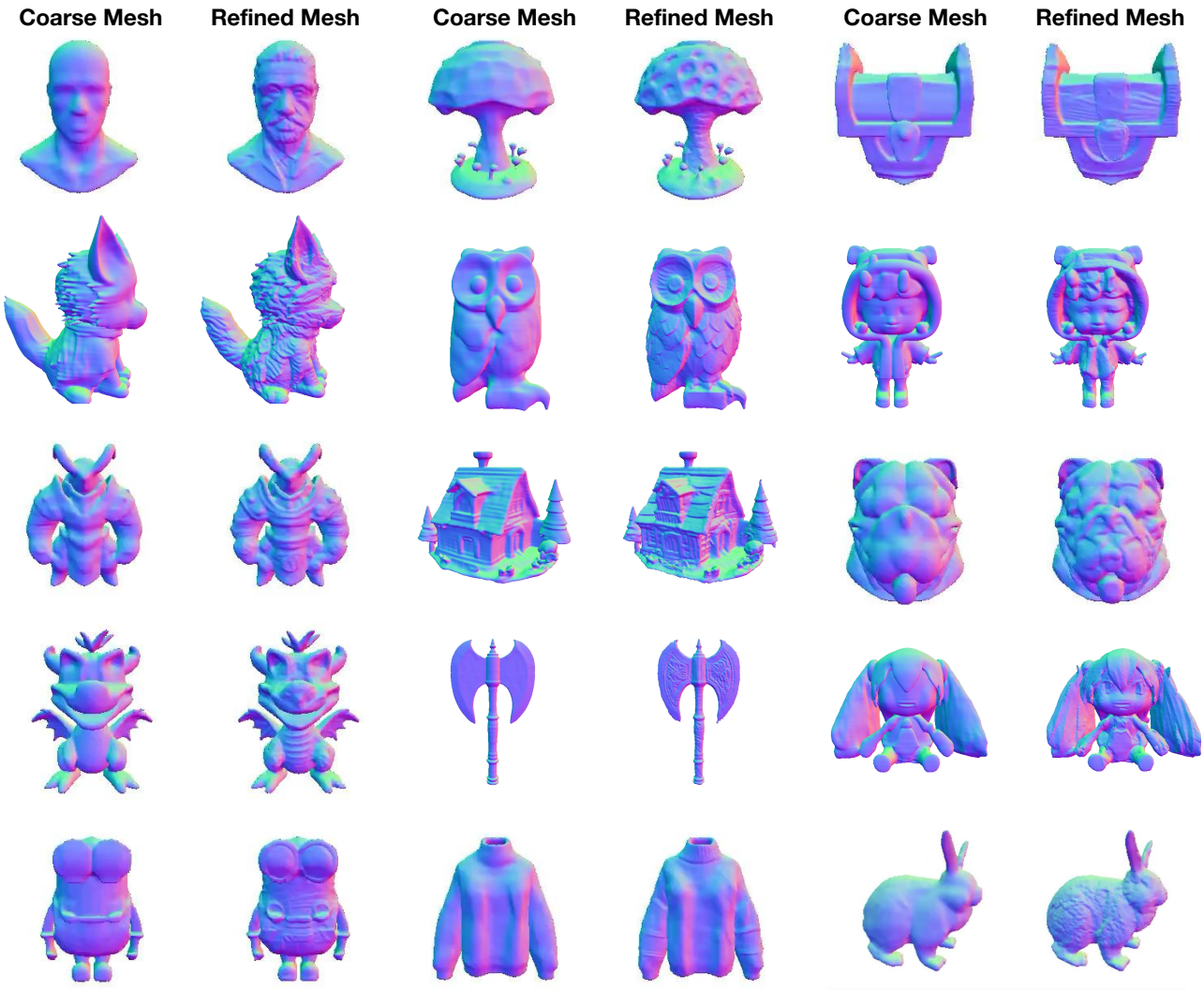


Figure 12. Results of the Automatic Mesh Refinement. We visualize the normal images of the coarse mesh and the refined mesh. It can be seen from the images that our proposed automatic normal refinement module significantly enhances the geometric details, such as the fur of the dog and the details of the face.

the outcomes from various configurations aimed at enhancing normal maps.

7.1. More Qualitative Results

We provide more qualitative results for both mesh generate stage and mesh refinement stage in Fig. 11 and Fig. 12. Refined meshes are rendered as normal maps to highlight the enhanced local detail.

We incorporate images with varied styles, obtained from the internet, in our evaluation to gauge the generalization capacity of our model. We also gathered several text prompts for the evaluation of the text-to-3D generation capability. Our 3D native diffusion model produces coarse geometry with neat shape and regular topology. Our mesh refinement stage further enhanced the generated mesh with more intricate details, such as wood grain on the box, human hair and wrinkles, wrinkles on clothes.

We further provide comparison results with Clay [65]. Due to the unavailability of the original mesh displayed in [65], we copyed the rendered images from their paper and present our results alongside for visual comparison. As is shown in 13, our method generate meshes align better with input images. Considering Clay [65] trained their model on 527K objects, we believe that our results have demonstrated the strong generation capability of our model, which is trained with barely 190K objects.

Figure 13. Comparison of image to 3d generation between ours coarse output and Clay [65]

7.2. Ablation study

We provide qualitative ablation results in Fig. 14 to show the importance of each component in our method.

Single Image vs. Multi-view Images Condition. Compared to the single-image condition, the multi-view images generated by the 2D diffusion model provide more information regarding the object. The generated shapes are prone to have anomalous deformation in the single-image condition, whereas the multi-view condition generates a more comprehensive 3D mesh.

Camera Pose Injection. Incorporating camera poses in the image feature extractor helps the model to distinguish embeddings from different views of the object, ultimately leading to more precise 3D shape generation. Without camera pose injection, the model tends to generate a 3D geometry with an incorrect orientation.

Training-free Cross-view Attention. Cross-view attention enables the propagation of information across disparate viewpoints, thereby enhancing the consistency of generated images. Although without fine-tuning on multi-view datasets, this mechanism substantially bolsters the multi-view consistency of images.

Regularizations During Mesh Optimization. Our proposed relative Laplacian constraint the vertices towards the proximity of the coarse mesh, avoiding the mesh collapse introduced by the self-consistent local smoothness, thereby enabling a robust optimization process.

7.3. Different Settings of Normal Enhancement

We demonstrate the flexibility of our framework through experiments with different settings.

The Effective of Different CFG Scale We demonstrate the results with different classifier-free guidance weights. As this variable becomes larger, the refinement process takes more prompt information as guidance, and produces results that are more consistent with text descriptions. The quality of the generated image will be reduced if this value is too large, we balance between the effect and quality by setting this value to 20 by default.

The Effective of Control Scale for Tile Model The control scale defines how much the refine process will refer to the control image, which is the normal map renderer from coarse mesh in our situation. A larger control scale results in less structural diversity and the refined normal maps are more likely to align with the 3D shape. We default set this value as 0.8, for the purpose of enhancing details while preserving the overall shape of the coarse mesh.

8. Application

8.1. Magic Normal Brush

Our refinement module is designed to be versatile and can be applied to a variety of real-world modeling applications.

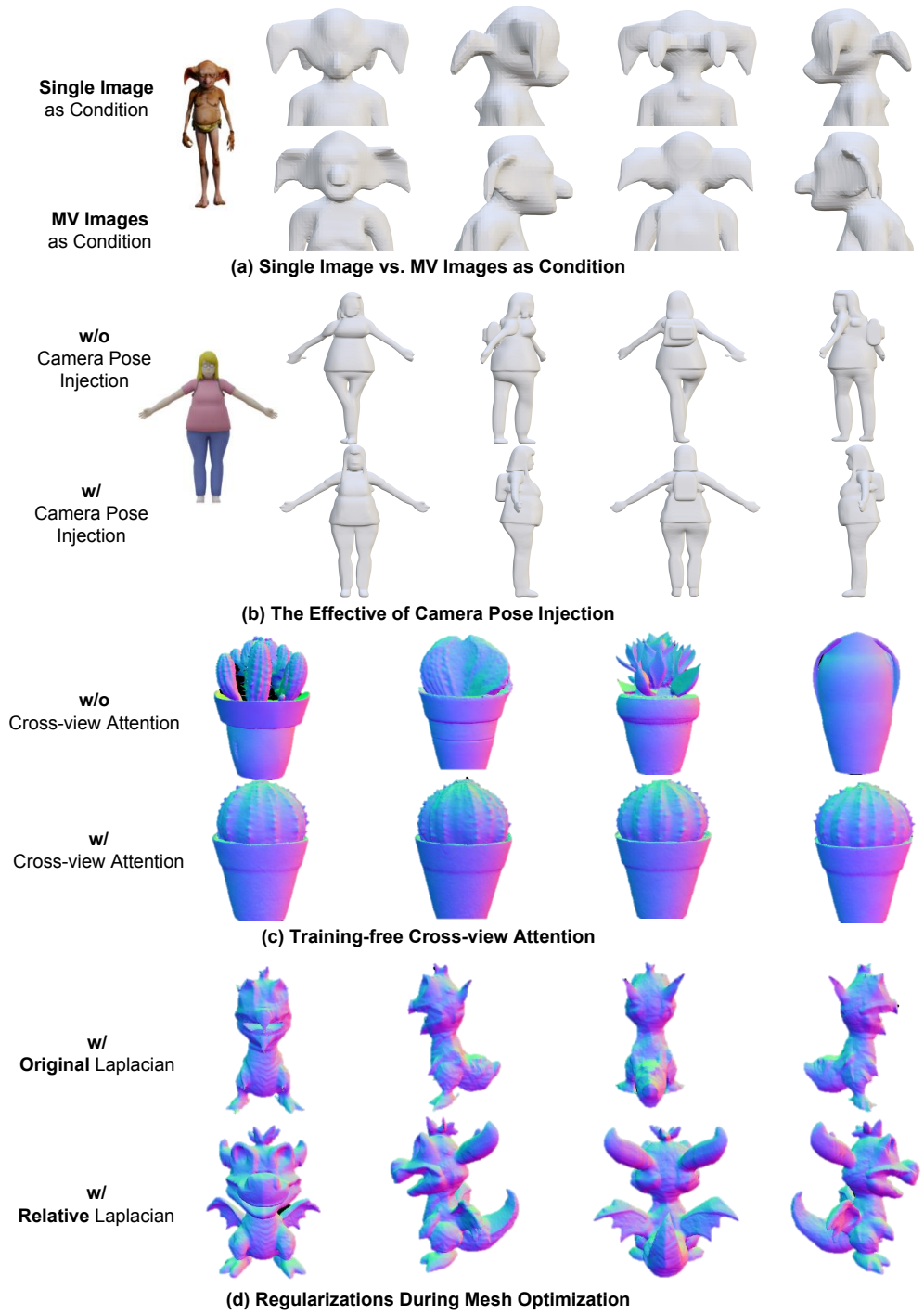


Figure 14. Ablation Study. (a) When only using a single image as a reference, the absence of information for the occluded parts can result in erroneous interpretations, as exemplified by the four ears of the goblin. (b) Incorporating the camera pose significantly enhances the diffusion model to comprehend spatial information. Without this, the model may inaccurately predict the geometry, potentially leading to distorted geometry, such as the unnaturally twisted body. (c) Introducing Cross-view attention significantly increases the multi-view consistency of normal prediction, especially for round objects. (d) Employing relative Laplacian constraints addresses the issue of thin mesh diminishing due to the local smoothness criteria in the standard Laplacian regularization term.

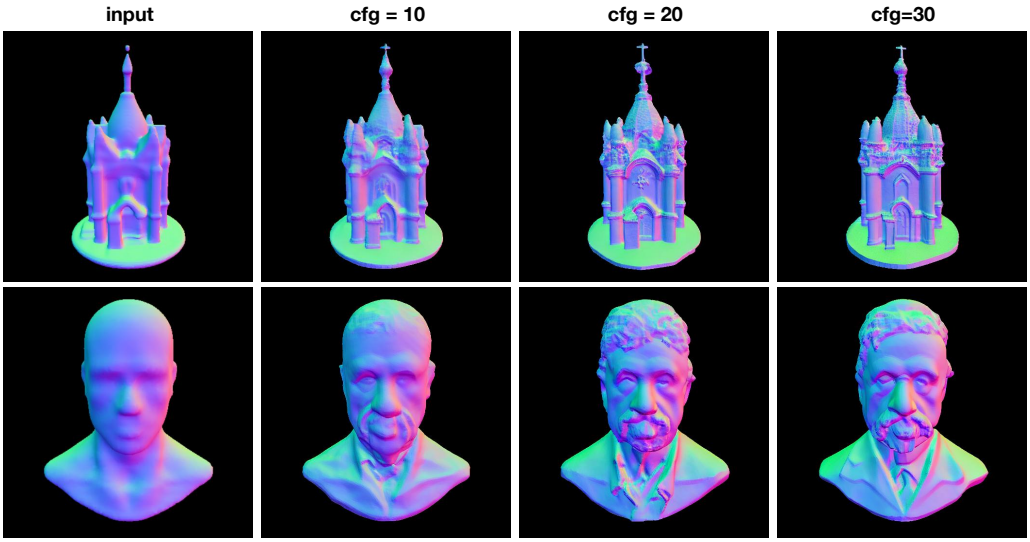


Figure 15. Normal refinement results with different CFG settings

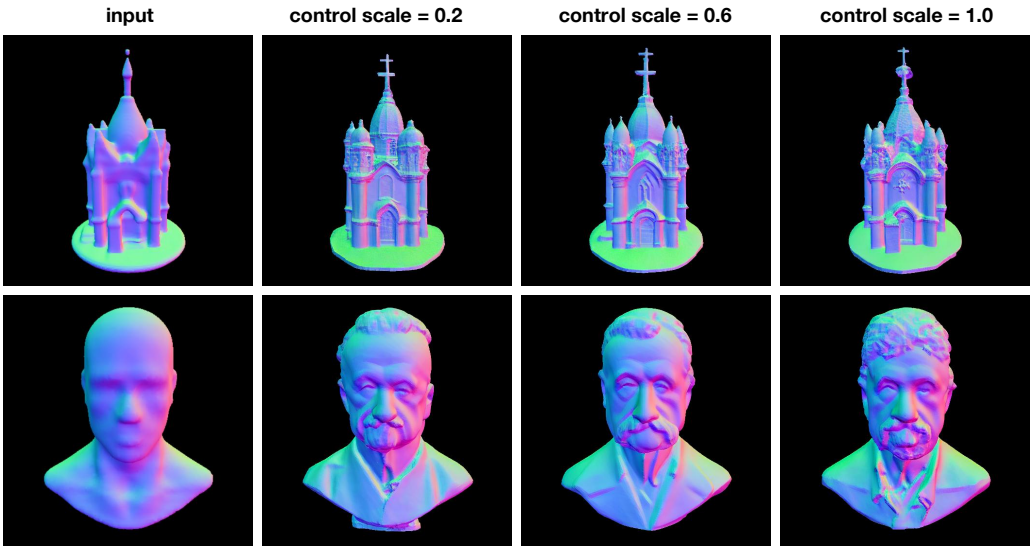


Figure 16. Normal refinement results with different control-scale settings

Similar to ZBrush [44] software, we incorporate a brush tool enabling users to interactively refine the normal map of the mesh with the generative capabilities of the normal diffusion model. Our proposed *Magic Normal Brush* supports meshes produced by various approaches, including manual crafting and other 3D generation methods [20, 28, 30]. Users are required to first select the regions to be updated and then type text prompts to edit the selected areas. As illustrated in Figure 17(a), this tool enables users to efficiently add whiskers to a man's face via simply drawing and typing text.

To fully preserve the high-frequency detail obtained through early refinement, we involve 3D mask into calculation in each mesh optimizing step. Specifically, given a 2D mask I_v^{draw} specified by user under drawing view p_v, we first adopt normal diffusion model into inpainting task to achieve the local editing of the guiding normal map. Then we adopt the mesh optimizer into a 3D mask version by optimizing a 3D mask defined on each vertex. This step is necessary even if the guiding normal map is totally unchanged

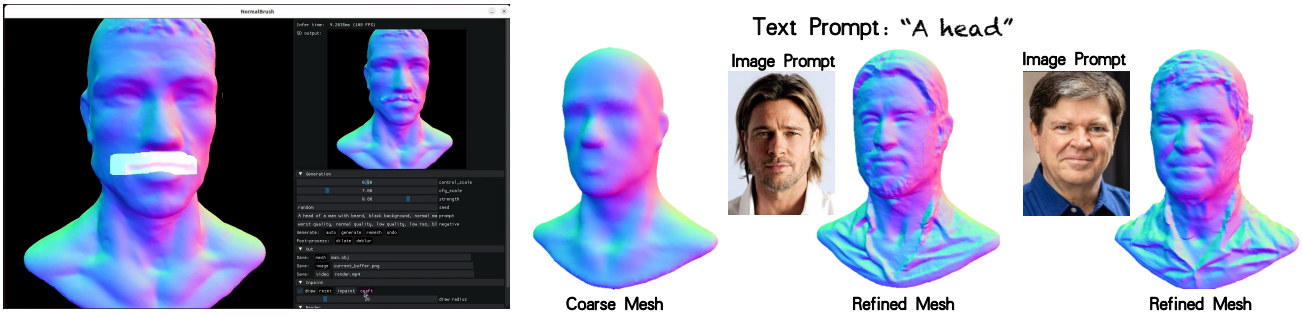


Figure 17. (a) With the *Magic Normal Brush*, it's convenient to edit a mesh via simple drawing and typing text. Whiskers are easily added to the mesh. (b) Our mesh refinement module is capable of accepting an image as the prompt. By incorporating a facial image to guide the normal mapping enhancement, we can refine the mesh according to the identity in the image.

outside the 2D mask, for the remeshing step and regularization term can do harm to the geometry details on any region on the editable 3D surface. We obtain the 3D mask by optimizing a single value $b \in B$ for each vertex, rendering the variables with differentiable renderer under mask drawing views and minimizing the $L1$ loss between the rendererd images and 2D masks. Thanks to our explicit mesh optimizing option, this results in a complete preservation for all the vertices and edges outside the 3D mask, while preserving the local continuity between the edited and preserved meshes.

8.2. Image as Prompt for Mesh Refinement

As presented in Fig. 17(b), in addition to using text prompts as conditional for normal refinement, our model is also capable of incorporating images as conditions, thanks to the advancements in the 2D diffusion community. Specifically, we leverage the IP-Adapter [61] face model to utilize an image as prompt for normal refinement. Consequently, we are able to refine the coarse meshing based on the input IP image, such as the facial features of an individual, to produce a mesh that maintains the same identity-preserving attribute.

8.3. Texture Generation

Our refined meth contains more high frequency details, thus is more suitable for geometry guided texture generation. We trained a multi-view normal map based control-net to generate multi-view aligned texture map, and inject text and image conditions by embedding them as clip features, as is done in [65]. As is demonstrated in Fig. 19, refined meshes offer more precise control in the texture generation process, ensuring a high-quality and richly detailed texture. We present more colored results in Fig. 20.

9. Failure Cases

When the input images are overly intricate or are captured from extreme viewpoints, it may affect the results of the MV Diffusion model, thereby impacting the final geometry generation. We show the result in Figure 18 and will add analysis in the revision.

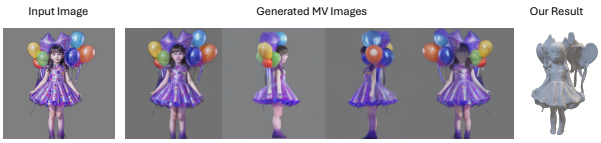


Figure 18. Failure cases due to the poor multi-view prediction and intricate structure.

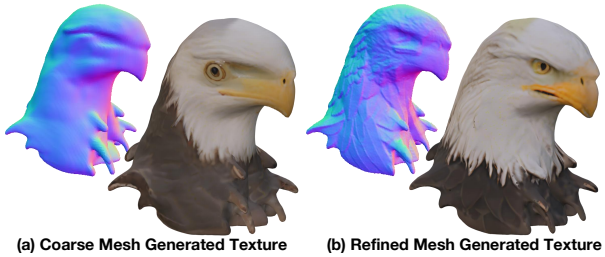


Figure 19. Our mesh with texture. We implement texture generation using similiar methods as [65].

10. Societal Impact

The societal impact of 3D generation technology is overwhelmingly positive in several fields, such as healthcare, education, architecture and manufacturing. 3D generation streamlines processes and promotes creativity, leading to more efficient and innovative solutions without any notable negative effects. The authors believe that this work has small potential negative impacts.

Figure 20. Our mesh with texture. We implement texture generation using similiar methods as [65].