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

Weiyu Li [*]
HKUST
LightIllusions
weiyuli.cn@gmail.com

Jiarui Liu [*]
HKUST
LightIllusions
jiaruiliu199509@gmail.com

Rui Chen
HKUST
riorui@foxmail.com

Yixun Liang
HKUST(GZ)
LightIllusions
lyxun2000@gmail.com

Xuelin Chen
Tencent AI Lab
xuelin.chen.3d@gmail.com

Ping Tan
HKUST
LightIllusions
pingtan@ust.hk

Xiaoxiao Long [†]
HKU
xxlong@connect.hku.hk

Abstract

We present a novel generative 3D modeling system, coined CraftsMan, which can generate high-fidelity 3D geometries with highly varied shapes, regular mesh topologies, and 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, irregular mesh topologies, noisy surfaces, and difficulties in accommodating user edits, 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 employ a 3D native diffusion model, which operates on latent space learned from latent set-based 3D representations, to generate coarse geometries with regular mesh topology in seconds. In particular, this process takes as input a text prompt or a reference image, and leverages a powerful multi-view (MV) diffusion model to generates multiple views of the coarse geometry, which are fed into our MV-conditioned 3D diffusion model for generating the 3D geometry, significantly improving robustness and generalizability. Following that, a normal-based geometry refiner is used to significantly enhance the surface details. This refinement can be performed automatically, or interactively with user-supplied edits. Extensive experiments demonstrate that our method achieves high efficacy in producing superior quality 3D assets 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 3D asset creation. However, manually creating these 3D assets is often

[*] Joint first authors

[†] Corresponding author

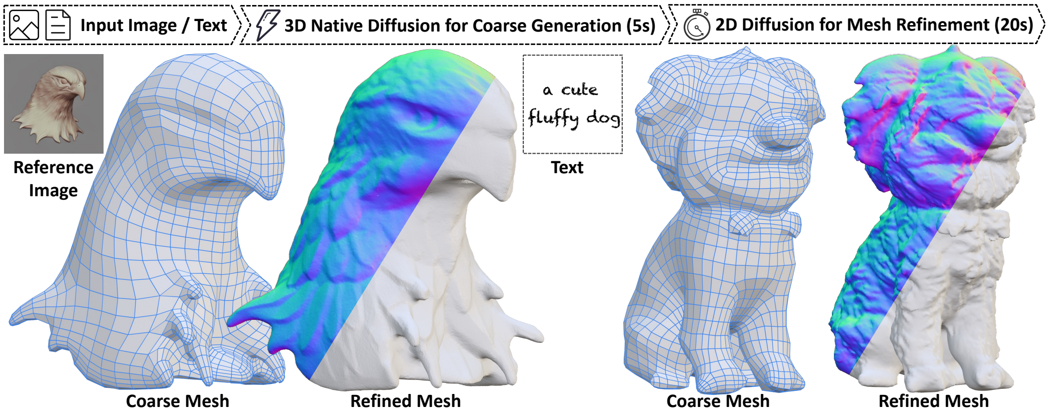


Figure 1: Our method, given a single reference image or text prompt, can generate intricate 3D shapes with high fidelity in just 30 seconds. Drawing inspiration from the typical workflow of craftsman, we start by creating a coarse shape using a 3D native diffusion model. We then enhance the surface details using either an automatic global geometry refiner or, more intriguingly, an interactive geometry refiner that allows for user edits. For more visually compelling results, please refer to the supplementary video.

time-consuming and expensive. Consequently, many methods now leverage generative techniques to simplify 3D generation, requiring only a single image or text prompt as input.

3D generative methods can be broadly categorized into three types: i) Score-Distillation Sampling (SDS) based methods Poole et al. [2022], Lin et al. [2023], Chen et al. [2023] 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 Li et al. [2023], Long et al. [2023]. While these methods significantly improve generation efficiency and robustness, the resulting 3D assets tend to have irregular geometric structures, noisy surfaces, and over-smoothed geometries. iii) 3D native generation methods Nichol et al. [2022], Jun and Nichol [2023], Zhang et al. [2023a] attempt to directly model the probalistic distribution of 3D assets via training on 3D assets. However, these methods are only tested on limited categories, and therefore show poor generalization on unseen cases. More importantly, all of these methods do not support user edits to improve the generated 3D interactively.

In this paper, we present a novel generative 3D modeling 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: i) a native 3D diffusion model, that is conditioned on a set of intermediately generated MV images and directly generates coarse 3D geometries; and ii) a generative geometry refiner that authors intricate details in either an automatic manner or interactive manner.

Specifically, the 3D native diffusion model, trained on 3D data, learns the probabilistic distribution of 3D geometries. This enables the generation of regular topologies and complex geometries, particularly those with high concavity. However, a 3D diffusion model trained on limited 3D datasets inevitably suffers from poor generalization to unseen input images and text prompts, resulting in significantly degraded geometries. We combine the 3D diffusion model with the multi-view diffusion model to address this issue. Concretely, we feed the single reference image or text prompt into the multi-view diffusion model to obtain multi-view images as intermediate conditions, from which our 3D diffusion model learns to generate 3D geometries. The rich geometric and semantic priors learned in the multi-view diffusion model significantly enhance the generalizability and robustness of our coarse generation stage.

On the other hand, our generative geometry refiner includes several key techniques to enhance the details and usability of the meshes derived from the generated coarse geometries. This is achieved with a combination of ControlNet-tile Zhang et al. [2023b] and surface normal map diffusion, preserving the superior quality and generalizability of the 2D diffusion model without the need to train a ControlNet-tile model from scratch. Direct mesh optimization is an efficient method that typically takes 10-20 seconds and maintains the original topology of the shape, making it suitable for downstream tasks such as editing. In particular, we directly optimize the vertices of the mesh, which only takes around 10 seconds and can preserve the original topology of the geometry with ease, rendering its suitability in the user interactive modeling. An automatic global refinement and an interactive local refinement allow users to edit and improve the 3D mesh in a user-friendly and controllable manner.

In summary, our system enables efficient 3D generation featuring high-quality and highly complex geometries, given only a single reference image or a text prompt. In addition, our system allows for user-interactive edits, enhancing the generated coarse geometries to better align with the users' envisioned designs. Extensive experiments demonstrate that our method achieves high efficacy in producing superior quality 3D assets compared to existing methods.

2 Related work

We briefly review the most related literature on 3D generation, with a particular emphasis on learning-based methods with various types of supervision.

3D Native Generative Models Many works adopt various 3D representations such as point clouds Li et al. [2018], Zhou et al. [2021], Yang et al. [2019], meshes Nash et al. [2020], Liu et al. [2023c], and implicit functions Chen and Zhang [2019], Park et al. [2019] to train native 3D generative models. The early efforts in the field Wu et al. [2016], Chen and Zhang [2019], Ibing et al. [2021] primarily concentrated on Generative Adversarial Networks (GANs) Goodfellow et al. [2014]. Autoregressive models Sun et al. [2020], Nash et al. [2020], Mittal et al. [2022] also draw great attention in 3D generation and have been extensively explored. Recently, diffusion models Ho et al. [2020] show great potentials in 2D image generation but they are not fully explored in the 3D domain. Some recent works extended diffusion models to 3D with the representation of point cloud Luo and Hu [2021], Zhou et al. [2021], meshes Liu et al. [2023c] and inplicit fields Chou et al. [2023], Shue et al. [2023]. 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, recent works propose to first compress 3D shapes into compact latent space, and then perform diffusion process in the latent space. Zhang et al. [2022] and Zhang et al. [2023a] propose a method to encode occupancy fields using a set of either structured or unstructured latent vectors. Neural Wavelet Hui et al. [2022] advocates a voxel grid structure containing wavelet coefficients of a Truncated Signed Distance Function (TSDF). Mosaic-SDF Yariv et al. [2023] approximates the SDF of a given shape by using a set of local grids spread near the shape's boundary. These works often suffer from lacking geometric details and over-smoothing surfaces. Most of the studies have only been tested on limited datasets Chang et al. [2015], Wu et al. [2015] with specific categories, such as chairs, faces, etc. Although recent 3D datasets, such as Objaverse Deitke et al. [2022], have dramatically enriched the state-of-the-art of 3D datasets, scaling up with larger datasets remains a challenge and the generalization capability of these models is still under-explored. Our work harnesses the feed-forward nature of 3D diffusion models while enhancing their generalization capability by leveraging the pre-trained multi-view 2D diffusion prior as the condition. This approach significantly facilitates zero-shot ability.

3D Generation using 2D Supervision In contrast to the often elusive 3D supervision, 2D supervision is more readily available and has been extensively utilized in 3D generation tasks. 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 Mildenhall et al. [2020] or Triplane Chan et al. [2021a] 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. Schwarz et al. [2020], Niemeyer and

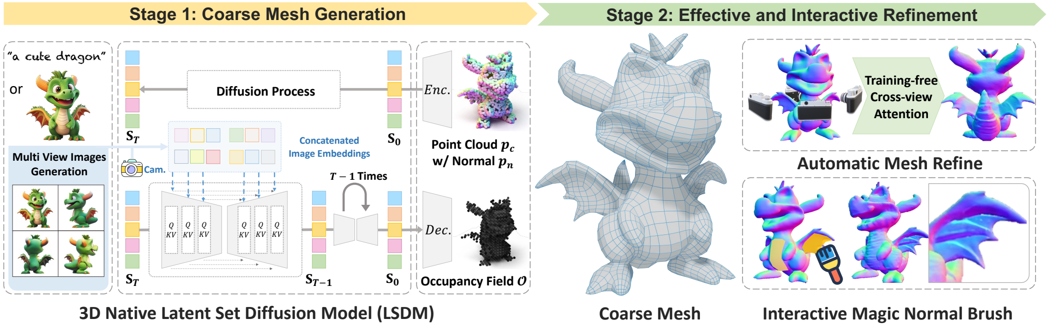

Stage 1: Coarse Mesh Generation
Stage 2: Effective and Interactive Refinement
"a cute dragon"
or
Multi View Images Generation
Cam.
Diffusion Process
S_T
S_0
Concatenated Image Embeddings
$T-1$ Times
Q KV
S_{T-1}
Enc.
Dec.
Point Cloud p_c w/ Normal p_n
Occupancy Field $\mathcal{O}$
3D Native Latent Set Diffusion Model (LSDM)
Coarse Mesh
Training-free Cross-view Attention
Automatic Mesh Refine
Interactive Magic Normal Brush

Figure 2: Overview of CraftsMan. Our method first transforms the input single image or text prompt into multi-view images using a multi-view diffusion model. These generated multi-view images are then fed into a native 3D diffusion model as conditions to produce a coarse mesh with regular topology. Finally, a surface normal-based refinement is employed to improve or edit the coarse geometry, enhancing it with intricate details. The refinement process features two key tools: an automatic global refinement and an interactive Magic Brush, which together enable efficient and controllable 3D modeling.

Geiger [2021], Chan et al. [2021b], Gao et al. [2022] perform GAN-like Goodfellow et al. [2014] structure to synthesize 3D-aware images via adversarial training.

Due to the high costs associated with high-resolution 2D supervision, many of these approaches focus on efficient training, like using efficient 3D representation Chan et al. [2021a], Zhao et al. [2022], patch-based discriminator Schwarz et al. [2020], progressive growing resolution using 2D upsampler Gu et al. [2022], etc. However, these methods are often trained on limited data with specific categories, and therefore shows poor generalization on unseen categories. Poole et al. [2022] 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 Wang et al. [2023], Liang et al. [2023], Chen et al. [2023], Lin et al. [2023], Shi et al. [2023], Li et al. [2024] 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 usually require a huge computational cost.

Instead of performing a time-consuming optimization, some recent works Long et al. [2023], Li et al. [2023], Liu et al. [2024, 2023b] attempt to generate multi-view images simultaneously and bring 3D-awareness by finetuning the 2D diffusion. The generated multi-view images are then used to reconstruct a 3D shape using sparse view reconstruction algorithms. Although these methods achieve high efficiency, the generated results are heavily dependent on the quality of the 2D images. Complex lighting conditions, occlusion, and multi-view inconsistency are still challenging and usually result in low-quality geometric structures. Indirect modeling of 3D probability distributions typically results in degraded final generation quality and cannot produce satisfactory geometry. In contrast, our approach mimics modern modeling workflows by first generating a coarse 3D shape using a feed-forward 3D native generative model, followed by refinement using detailed 2D priors.

3 Method

Figure 2 provides an overview of our generative 3D modeling workflow, which can synthesize 3D assets with rich details using text or a single image as input. Our framework is designed to mimic the artist's workflow of 3D modeling by incorporating a coarse geometry modeling process followed by a refinement process. These steps allow for the generation of high-quality 3D shapes with regular topology and detailed geometry.

Specifically, we first compress 3D assets into a compact latent space via a encoder-decoder structure, which enables efficient 3D shape latent diffusion (see Sec. 3.1). Secondly, we transform the input single image or text prompt into multi-view images via MV diffusion model. The generated MV images are used as conditions for the 3D latent diffusion model, which leads to robust and high-quality 3D generation (see Sec. 3.2). Finally, we introduce a surface normal based refinement scheme that

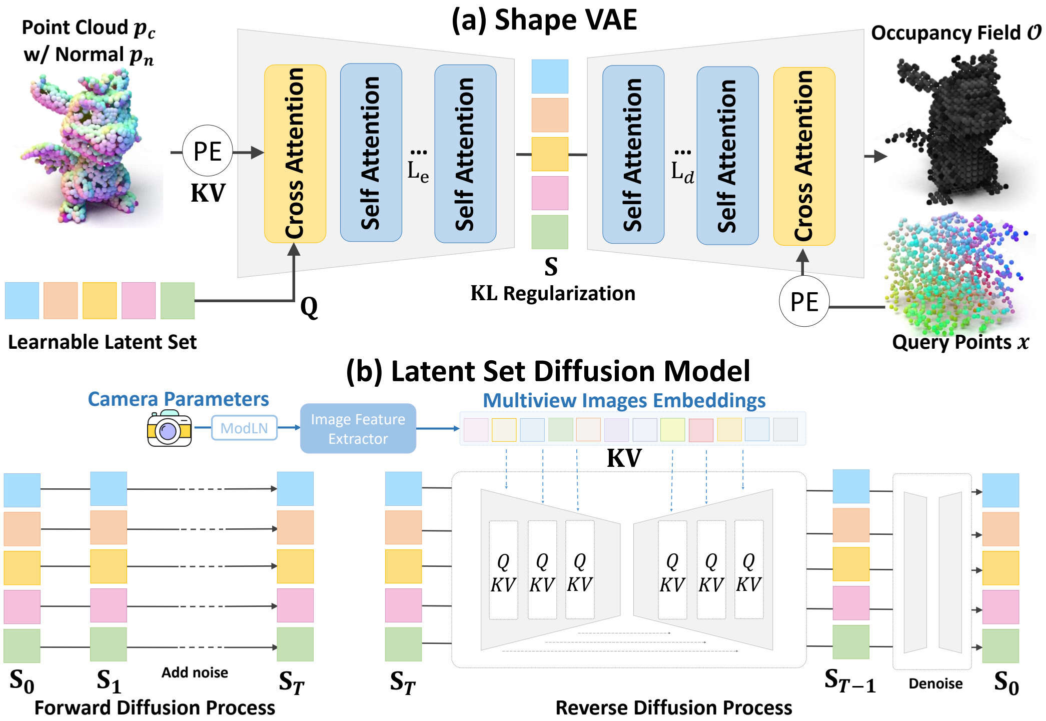


Figure 3: The illustration of 3D generation. We first train a 3D Variational Autoencoder (VAE) to compress 3D assets into a latent space, where the VAE takes point clouds with normals as input and outputs occupancy fields. With the learned latent space, we further train a 3D Latent Set Diffusion Model to produce 3D shapes, using multi-view images as conditions.

further improves or edits the generated coarse geometry in an effient and controllable manner (see Sec. 3.3).

3.1 3D Latent Set Representation

We propose to directly learn the distribution of 3D shapes by training a native 3D diffusion model on 3D datasets, enabling the generation of high-quality 3D shapes with complex and regular geometry. Initially, we encode the 3D assets into a latent space using a Variational Autoencoder (VAE), and then we train a diffusion model in this latent space to exploit its compactness and efficiency.

Shape Representation The success of the Latent Diffusion Model (LDM) Rombach et al. [2022] proves that a compact, efficient, and expressive representation is essential for training a diffusion model. Therefore, we first encode 3D shapes into a latent space and then train a 3D latent diffusion model for 3D generation. Similar to the methods in Moreno et al. [2022], Zhang et al. [2023a], each 3D asset is encoded into a compact one-dimensional latent set $\mathbf{S} = \left\{\mathbf{s}_i \in \mathbb{R}^C\right\}_{i=1}^D$, where D is the number of the latent sets and C is the feature dimension.

Shape Encoding. We adopt an auto-encoder structure to encode the 3D shapes into the latent sets and then decode them to reconstruct a neural field. Specifically, for each 3D shape, we first sample a set of points clouds $\mathbf{P}_c \in \mathbb{R}^{N\times 3}$ from the surface, along with its surface normal vectors $\mathbf{P}_n$. Next, we implement a cross-attention layer to integrate the information of the concatenated Fourier positional encodings with their respective normals into the shape encoder. Following Zhao et al. [2023], we leverage the Perceiver Jaegle et al. [2021]-based shape encoder that effectively captures the geometric characteristics of the 3D shape to learn a set of latent vectors $\mathbf{S}$.

$$\begin{aligned}\hat{\mathbf{P}} &= \mathrm{Concat}(\mathrm{PE}(\mathbf{P}_c), \mathbf{P}_n),\\ \mathrm{Enc}(\mathbf{P}_c, \mathbf{P}_n) &= \mathrm{SelfAttn}^{(i)}(\mathrm{CrossAttn}(\mathbf{S}, \hat{\mathbf{P}})), for\ i = 1, 2, \ldots, L_e,\end{aligned} \tag{1}$$

where $\hat{\mathbf{P}}$ is the concatenated point feature, L_e is the number of Self Attention layers in the shape encoder and PE denotes the column-wise Fourier positional encoding function.

Shape Decoding. We use a similar perceiver-based decoder while moving all self-attention layers before cross-attention layers that decode the latent set $\mathbf{S}$ to neural fields. Given a query 3D point $x \in \mathbb{R}^3$ in space and a learned shape latent embeddings $\mathbf{S}$, the decoder predicts its occupancy value.

The training objective is as:

$$\begin{gathered} \mathbf{S} \leftarrow \mathrm{SelfAttn}^{(i)}(\mathbf{S}), \quad \forall i = 1, 2, \ldots, L_d, \\ \mathcal{L}_{vae} = \mathbb{E}_{x \in \mathbb{R}^3}\left[\mathrm{BCE}\left(\hat{\mathcal{O}}(x), \mathcal{O}(\mathrm{CrossAttn}(\mathrm{PE}(x), \mathbf{S}))\right)\right] + \lambda_{kl}\mathcal{L}_{kl}, \end{gathered} \tag{2}$$

where L_d is the number of Self Attention layers in the shape decoder, $\hat{\mathcal{O}}(x)$ is the ground truth occupancy value of x and $\mathcal{O}$ is the occupancy prediction function using a single MLP layer. The KL divergence loss $\mathcal{L}_{kl}$ is used to regularize the latent space distribution to a standard Gaussian distribution. Subsequently, we sample query points in a regular grid to reconstruct the final surface using Marching Cubes Lorensen and Cline [1998]. Please refer to Zhang et al. [2023a], Zhao et al. [2023] for more details.

3.2 3D Native Diffusion Model

Instead of directly using a single image or text prompt as conditions, we leverage recent multi-view diffusion models to produce multi-view images from the input single image or text prompt. The multi-view conditioned 3D diffusion model enables robust and accurate 3D generation of various shapes.

Generating MV Images as Intermediate Conditions. Multi-view (MV) images generated by recent MV diffusion models offer richer geometric and contextual priors compared to using a single image or text alone. Importantly, adopting MV images leads to a unified scheme that avoids the need for separately training text-conditioned and single-image-conditioned models. As a result, the recent text-to-MV Shi et al. [2023], Li et al. [2023] and image-to-MV Wang et al. [2024], Long et al. [2023] methods are leveraged to generate multi-view images. Formally, we can express the process as,

$$\hat{\mathbf{y}} = \{f(y, \pi_i)\}_{i=1}^{K}, \tag{3}$$

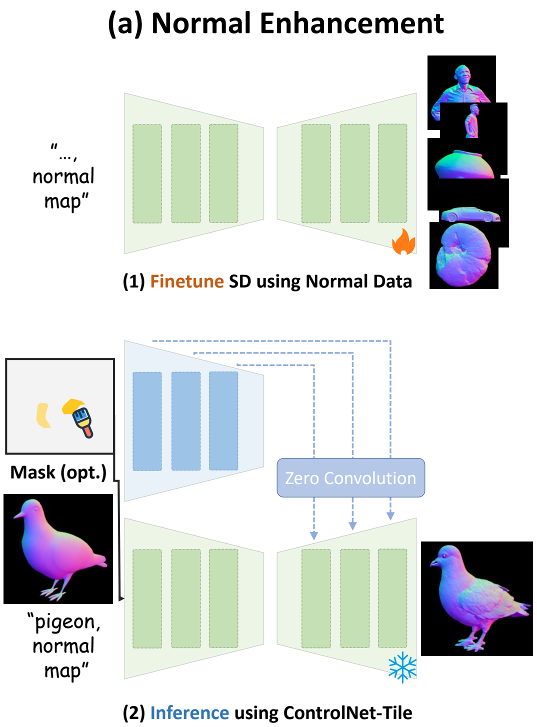
where f is the multi-view diffusion model, y is the input text or image condition, π_i is the given camera parameters, and $\hat{\mathbf{y}}$ is the generated K images.

MV-conditioned 3D Latent Set Diffusion Once obtaining the encoded shape latent set $\mathbf{S}$ and corresponding multi-view images $\hat{\mathbf{y}}$, we can train a conditional 3D diffusion model on the latent space to generate 3D shapes. The multi-view images are first fed into a large pre-trained image feature extractor τ_θ like Radford et al. [2021], Caron et al. [2021] to acquire their corresponding embeddings. Notably, to distinguish between multi-view images, some recent works Li et al. [2023], Long et al. [2023], Shi et al. [2023] modulate the image feature extractor using camera parameters π. We follow the method in Li et al. [2023] by employing an adaptive layer normalization (adaLN) Perez et al. [2018]. Such modulation (ModLN) is applied to each attention sub-layer in the image feature extractor τ_θ, and the modulation layers are optimized during training to make image embeddings be aware of the camera position.

Thus, we can learn the conditional Latent Set Diffusion Model (LSDM) via:

$$\mathcal{L}_{LSDM} := \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_\theta\left(\mathbf{S}_t, t, \tau_\theta(\hat{\mathbf{y}}, \mathrm{ModLN}(\pi))\right)\|_2^2\right], \tag{4}$$

where ϵ_θ is build on a UNet-like transformer Ronneberger et al. [2015], Vaswani et al. [2017], t is uniformaly samppled from $\{1, \ldots, T\}$ and $\mathbf{S}_t$ is a noisy version of $\mathbf{S}_0$. Furthermore, we also adapt the classifier-free guidance (CFG) training strategy Rombach et al. [2022] that randomly drops the conditions during training to improve the fidelity and diversity of the generated shapes. As a result, with text-based or image-based conditions, the 3D native diffusion model could produce a corresponding occupancy field, and a subsequent marching cube algorithm can be used to extract explicit mesh from the occupancy field. To better visualize the smooth geometry of our generated meshes, we have also utilized a remeshing tool Maxime [2024] to convert the triangular meshes into quadrilateral meshes.



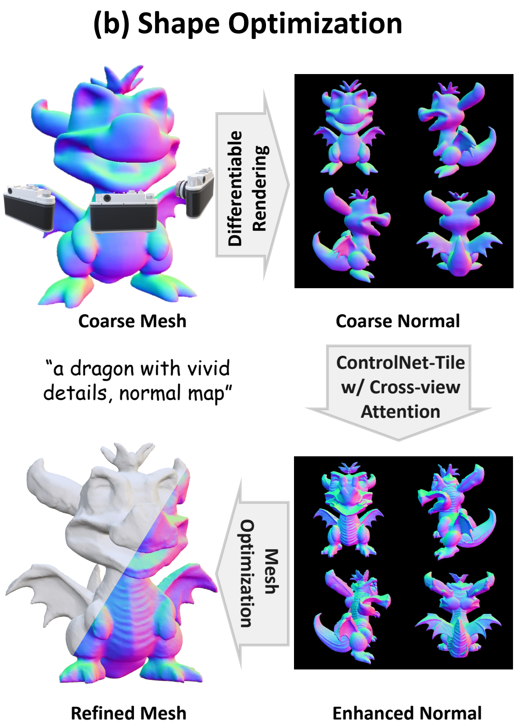


Figure 4: The illustration of surface normal-based geometry refinement. (a.1) We first finetune the villina stable diffusion model using normal maps to adapt to normal data. (a.2) The normal-adapted diffusion model is combined with ControlNet-Tile to enhance a normal with intricate details. (b) The automatic mesh refinement process via training-free cross-view attention.

3.3 Normal-based Geometry Refinement

To further enhance the coarse mesh with rich details, instead of directly manipulating the mesh vertices, we propose to improve and edit the initial mesh using normal maps as an intermediate representation. We first render the coarse mesh into normal maps and then leverage normal-based diffusion model 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, we design two processes with practical applicability, namely automatic global refinement and interactive local refinement, which respectively cater to the needs of artists for automatic high-quality mesh refinement and interactive mesh editing.

Coarse Normal Enhancement We adopt ControlNet-Tile Zhang et al. [2023b] to enhance the rendered normals with details. Instead of directly training a normal-adapted ControlNet model, we propose a more lightweight adaptation that inherits the priors in the 2D diffusion model in the RGB domain. Specifically, we only finetune a diffusion model Rombach et al. [2022] using normal images, and then directly a ControlNet-tile network φ that pretrained on the RGB domain to generate the refiner results. Formally, for the i_{th} view with a rendered normal map n_i, the output from by the ControlNet-tile network φ can be succinctly represented as: $\hat{\mathbf{n}}_i = \varphi(n_i, y_{text})$, where y_{text} is the input text condition and can be set to empty if we use the guess mode in Zhang et al. [2023b]. The only model we finetuned is the text-to-image diffusion model, the lightweight adaptation inherits the powerful zero-shot generalization ability and shows superior details in generated images.

Shape Optimization via Differentiable Rendering We advocate for direct vertex optimization through continuous remeshing Palfinger [2022], 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

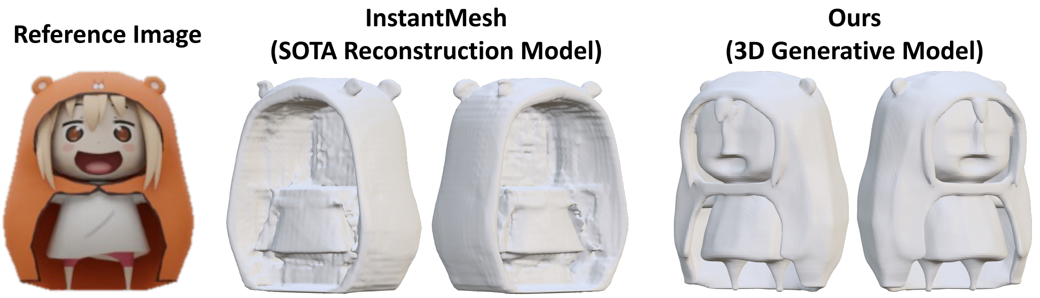


Figure 5: Compared to the strong baseline InstantMesh, our result maintains accurate complex geometric structures and avoid shape ambiguity.

differences between the rendered normals and the refined normals via:

$$\mathcal{L}_{remeshing} = \sum_i \|\hat{n}_i - \mathcal{R}_n(V, F, \pi_i)\|_1^1, \tag{5}$$

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. Then, a remeshing operation is executed to adeptly split, merge, and flip edges as proposed in Palfinger [2022].

To stabilize the mesh optimization, we introduce a relative Laplacian smoothing term, whose vertex updating process can be formulated as:

$$x \leftarrow x_{init} + \lambda \cdot \mathbf{v} \cdot \mathbf{W}\left(x - x_{init}\right), \tag{6}$$

where x_{init} is the initial position for each vertex, λ is a smoothing hyperparameter, $\mathbf{v}$ is the relative speed as mentioned in Palfinger [2022], and $\mathbf{W}$ is the combinatorial Laplacian matrix.

Automatic Mesh Refinement We can simultaneously refine the normal maps across different perspectives and subsequently optimize the mesh using enhanced multi-view normals. This operation is named as *Auto Normal Brush*. A pivotal challenge arises from the inconsistencies observed in the normal images generated by diffusion models across different views. Recent advancements, as detailed in Shi et al. [2023], Long et al. [2023], address this issue by employing a cross-view attention mechanism. The cross-view attention facilitates the propagation of information across perspectives by interlinking keys and values, enabling the perception of correlations between multiple views.

However, methods trained on synthesized datasets are prone to overfitting, leading to unrealistic and over-smoothed results. 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. The refined normals are then used to optimize the mesh, resulting in a high-quality 3D model with rich details.

Interactive Local Refinement We also offer an interactive editing tool, dubbed as *Magic Normal Brush*, that enables precise adjustments to specific local regions of the mesh in a controllable manner. Users can select the areas to be edited using a painting brush, creating a binary mask that indicates the regions to be updated. We then render a normal map of the current mesh via differentiable rendering. The binary mask and the rendered normal map are combined to produce a masked normal map with empty pixels that need updating. This masked normal map is fed into a normal diffusion model for inpainting. The masked normal map is updated with a user input text prompt. Finally, we optimize the mesh shape using the updated normal map as supervision, resulting in the final edited mesh.

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

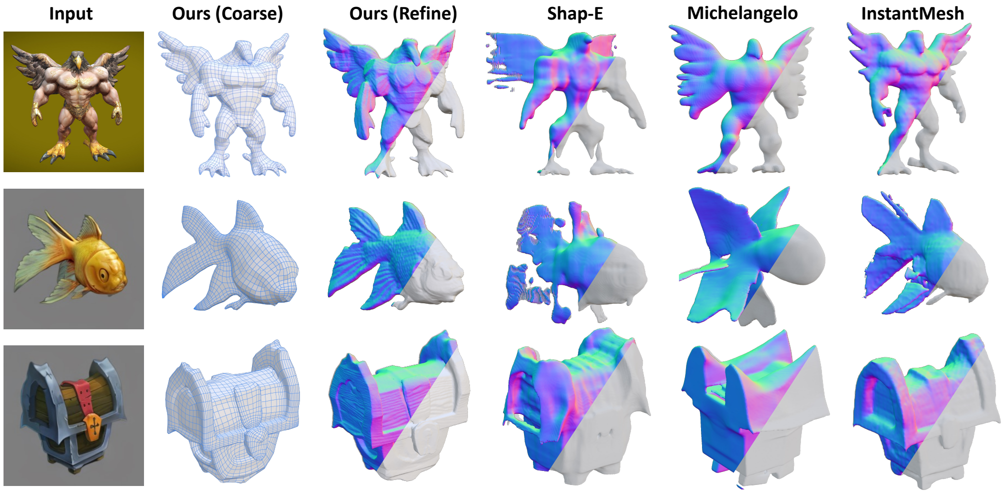


Figure 6: Qualitative comparisons with baseline methods for the task of single-view reconstruction. The coarse geometries are remeshed to quadrilateral meshes using Maxime [2024] for better visualization of the smooth geometry of our generated results.

our method as described in Section 4.3 and Table 1, 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. Moreover, we demonstrate the versatility of our method by supporting downstream applications, as shown in Fig. 10. More intriguing visual results can be found in our accompanying video and supplementary.

4.1 Dataset Preparation

We use a public 3D dataset–Objaverse Deitke et al. [2022], which contains around 800k models created by artists, as our training data. Due to the presence of considerable noise in the geometry and appearance, we filter out meshes of inferior quality, such as those with point clouds, thin structures, holes, and texture-less surfaces, from our training data to ensure its high quality, resulting in a curated dataset of approximately 170k objects. For each mesh, we first normalize the object to fit within a unit cube and then convert it into a water-tight mesh as in Mescheder et al. [2019]. Then, we render 4-orthogonal views with a random rotation of each object to serve as conditions for the Latent Set Diffusion Model. Additionally, we render normal maps for each object to finetune a 2D normal diffusion model. Please refer to the supplementary for more details

4.2 Implementation Details

We follow the same architecture as in Zhao et al. [2023] 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 UNet-like transformer, comprising 104 million parameters. Additionally, we utilize pre-train image encoders in the CLIP Radford et al. [2021] (ViT-L-14) as the image feature extractor and freeze it during training and sampling.

Please refer to the Zhao et al. [2023] and the supplementary for more details.

4.3 High-quality Mesh Generation

We present the results of our proposed method in Figure 1 and the supplementary. It is crucial to note that the evaluation of geometry quality, especially aspects like smoothness and intricate details of the 3D generation results is inherently challenging Wu et al. [2024]. Additionally, our model is generative rather than reconstructive in nature, making it inequitable to directly compare our outputs with the ground truth, as is typically done in reconstruction models. In this evaluation, our primary

Table 1: Quantitative comparison with baseline methods on the GSO dataset Downs et al. [2022]. Notably, as a generative model, our method inherently differs from reconstruction methods, making a direct numerical comparison infeasible. Nonetheless, our method achieves comparable performance.

Type	Method	Chamfer Dist↓	Volume IoU↑	Inference Time↓
Sparse View Recontruction	One-2-3-45 Liu et al. [2024]	0.0629	0.4086	~45s
	zero123 Liu et al. [2023a]	0.0339	0.5035	~10min
	InstantMesh Xu et al. [2024]	**0.0187**	**0.6353**	~10s
2D Distillation	Realfusion Melas-Kyriazi et al. [2023]	0.0819	0.2741	~90min
	Magic123 Qian et al. [2024]	0.0516	0.4528	~60min
3D Generative Model	Point-E Nichol et al. [2022]	0.0426	0.2875	~40s
	Shap-E Jun and Nichol [2023]	0.0436	0.3584	~10s
	Michelangelo Jun and Nichol [2023]	0.0404	0.4002	~3s
	Ours	**0.0355**	**0.5092**	~5s

focus is on the qualitative quality of the 3D generation results through a variety of results and we also present quantitative data for reference in Table 1. As depicted in Fig. 6 and 7, our method is adept at generating 3D coarse meshes with smooth topology, which are further enhanced to exhibit finer details following the refinement process.

Qualitative Evaluation. We incorporate images with varied styles, obtained from the internet, in our evaluation to gauge the generalization capacity of our model. To extensively evaluate the performance of our method, we compare our model with the 3D generative models Jun and Nichol [2023], Zhao et al. [2023] and state-of-the-art reconstruction models Xu et al. [2024]. As shown in Fig. 6 and supplementary, 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 Fish in Fig. 6. The strong baseline, InstantMesh, could produce accurate geometries but still lacks geometric details.

Quantitative Evaluation. Following the prior works, we employed the Google Scanned Object dataset—a rich collection of common everyday objects—to evaluate the performance of our 3D Diffusion Model in generating 3D models from single images. We present the quantitative evaluation of the quality of our image-to-3D generation in Table 1. For each object in the evaluation set, we use the front view image with a resolution of 256x256 as input. To quantitatively evaluate our method, we adopt two widely-used metrics, namely Chamfer Distances (CD) and Volume Intersection over Union (IoU), between the ground-truth shapes and generated ones.

Our approach exhibits superior performance when compared to the 3D generative models, as illustrated in Table 1. To comprehensively evaluate the overall quality and detail richess of our method, we futher conduct user study that compares our method with three other methods Jun and Nichol [2023], Zhao et al. [2023], Xu et al. [2024]. The study containing valid 40 samples shows that 86% users vote that our mesh quality achieves the best, and 98% believe that our results keep more geometry details.

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. Fig. 8 illustrates the ablation results of each component.

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, which is beneficial for the generation of unseen parts of 3D meshes. 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

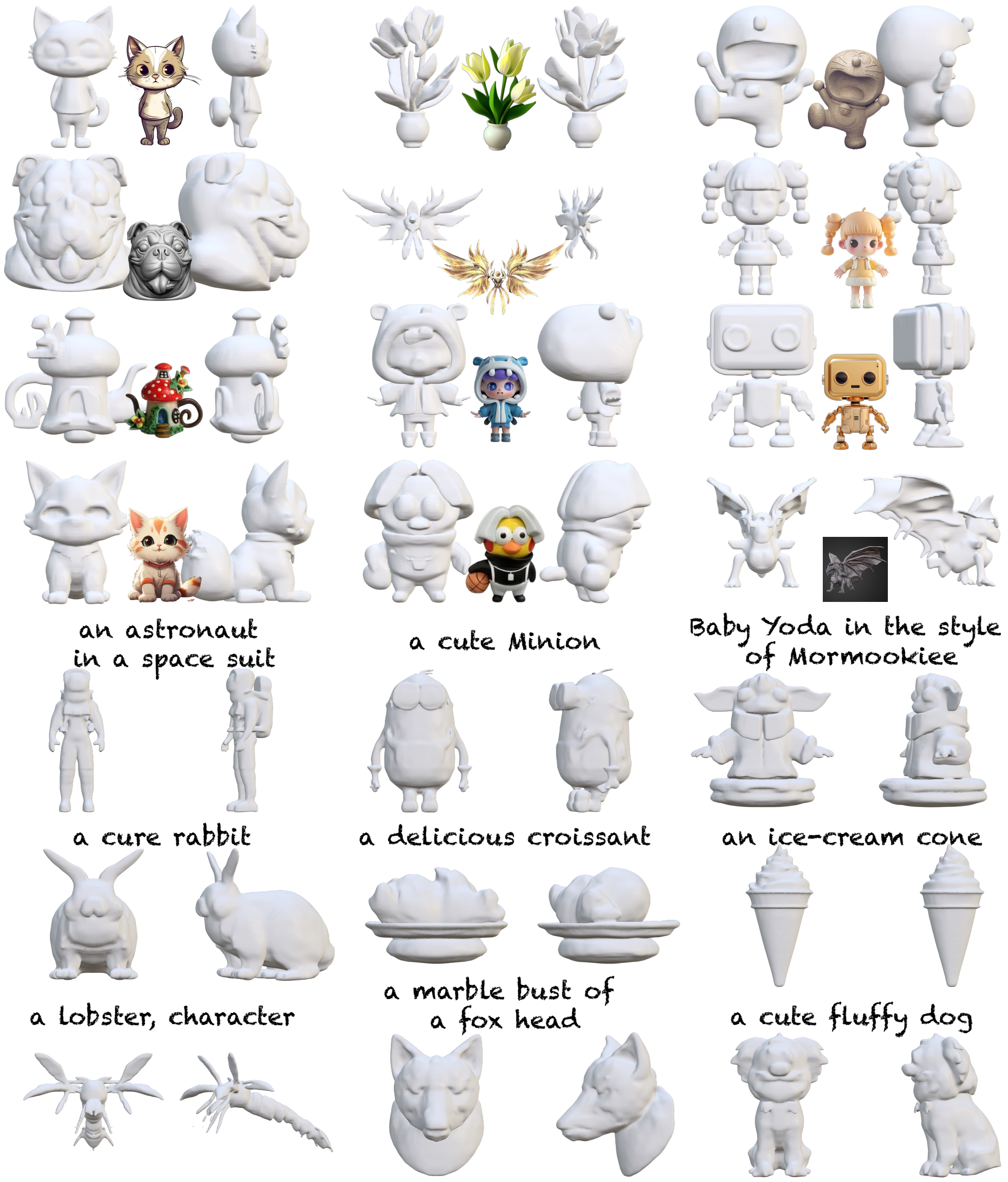


Figure 7: Raw coarse meshes (w/o quad remeshing) generated by our proposed method using a single image as a reference or a text prompt.

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.

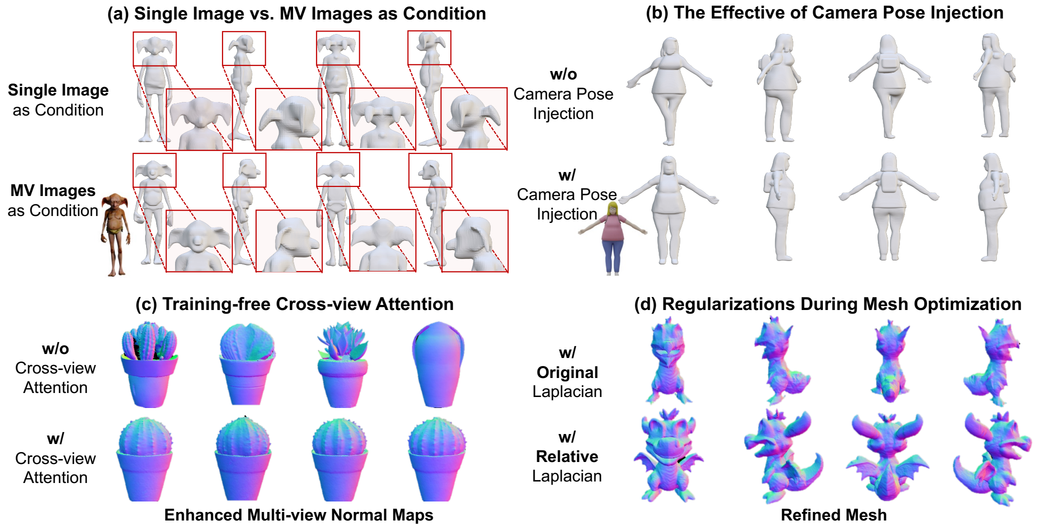

(a) Single Image vs. MV Images as Condition
Single Image as Condition
MV Images as Condition
(b) The Effective of Camera Pose Injection
w/o Camera Pose Injection
w/ Camera Pose Injection
(c) Training-free Cross-view Attention
w/o Cross-view Attention
w/ Cross-view Attention
Enhanced Multi-view Normal Maps
(d) Regularizations During Mesh Optimization
w/ Original Laplacian
w/ Relative Laplacian
Refined Mesh

Figure 8: 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.

4.5 Image as Prompt for Mesh Refinement

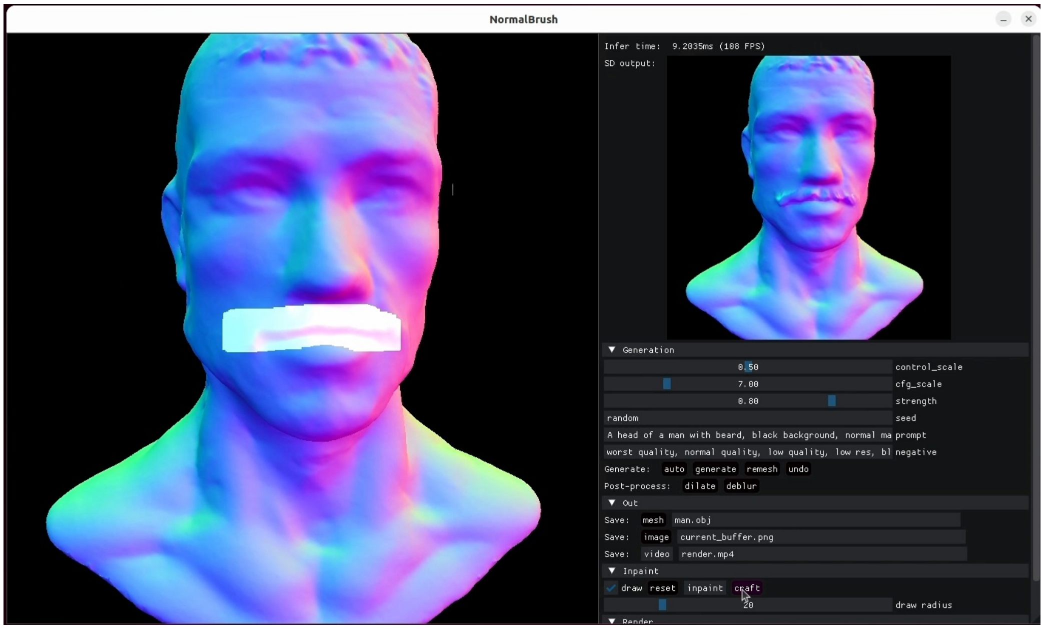
Our refinement module is designed to be versatile and can be applied to a variety of real-world modeling applications. As presented in Fig. 10, in addition to using text prompts as conditional for normal refinement, our model is also capable of incorporating images as conditions, thanks to the advancements in the 2D diffusion community. Specifically, we leverage the IP-Adapter Ye et al. [2023] 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.

4.6 Magic Normal Brush

Our proposed *Magic Normal Brush* supports meshes produced by various approaches, including manual crafting and other 3D generation methods Long et al. [2023], Liu et al. [2023b], Li et al. [2023]. 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 9, this tool enables users to efficiently add whiskers to a man's face via simply drawing and typing text.

5 Conclusion and Discussion

We present *CraftsMan*, 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. To achieve this, we utilize a diffusion model that is directly trained on 3D geometries. Specifically, we employ a robust multi-view diffusion model to generate a series of views, which serve as conditions for the 3D diffusion model. This strategy overcomes the scarcity of 3D datasets, significantly enhancing the robustness and generalizability of our approach. Subsequently, we harness the power of the 2D diffusion model to refine the normal map rendered from the coarse geometry. This refined map is then utilized as a guide for detailed surface enhancement. Despite our method's



```
Infer time:  9.2035ms (108 FPS)
SD output:
▼ Generation
                0.50                          control_scale
                7.00                          cfg_scale
                0.80                          strength
random                                        seed
A head of a man with beard, black background, normal ma prompt
worst quality, normal quality, low quality, low res, bl negative
Generate:  auto  generate  remesh  undo
Post-process:  dilate  deblur
▼ Out
Save:  mesh  man.obj
Save:  image  current_buffer.png
Save:  video  render.mp4
▼ Inpaint
☑ draw  reset  inpaint  craft
                20                            draw radius
▼ Render
```

Figure 9: 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.

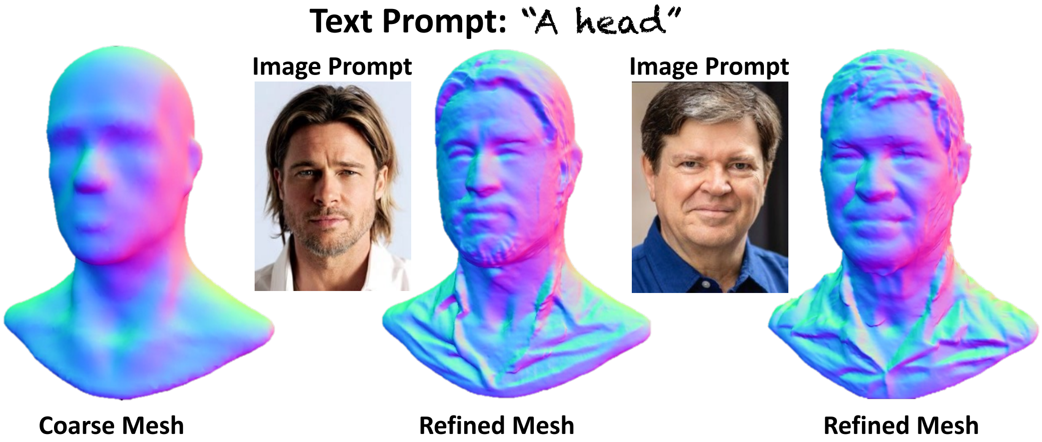


Figure 10: Our mesh refinement module is capable of accepting an image as the prompt. By incorporating a facial image to guide the normal mapping enhancement, we can refine the mesh according to the identity in the image.

capability to produce high-quality 3D meshes with regular mesh topology, there remains ample room for future exploration. 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

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.

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*, 2021a.

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, 2021b.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

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)*, October 2023.

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.

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.

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.

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.

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.

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.

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=iUuzzTMUw9K`.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

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

Moritz Ibing, Isaak Lim, and Leif P Kobbelt. 3d shape generation with grid-based implicit functions. in 2021 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 2021.

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.

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

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

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.

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.

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, 2023.

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.

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.

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, 2023a.

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*, 2023b.

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*, 2023c. URL `https://openreview.net/forum?id=0cpM2ApF9p6`.

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.

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

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)*, June 2021.

Maxime. *Quad Remesher*. Exoside, 2024.

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. URL `https://arxiv.org/abs/2302.10663`.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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.

Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. AutoSDF: Shape priors for 3d completion, reconstruction and generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Pol Moreno, Adam R. Kosiorek, Heiko Strathmann, Daniel Zoran, Rosalia G. Schneider, Björn Winckler, Larisa Markeeva, Théophane Weber, and Danilo J. Rezende. Laser: Latent Set Representations for 3D Generative Modeling. *arXiv*, 2022. URL `https://laser-nv-paper.github.io/`.

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.

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.

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.

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

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.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. volume 32, 2018.

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

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. URL `https://openreview.net/forum?id=0jHkUDyEO9`.

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.

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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

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.

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

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.

Yongbin Sun, Yue Wang, Ziwei Liu, Joshua E Siegel, and Sanjay E Sarma. Pointgrow: Autoregressively learned point cloud generation with self-attention. In *Winter Conference on Applications of Computer Vision*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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.

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.

Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016.

Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. *arXiv preprint arXiv:2401.04092*, 2024.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

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.

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.

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

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.

Biao Zhang, Matthias Nießner, and Peter Wonka. 3DILG: Irregular latent grids for 3d generative modeling. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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), jul 2023a. ISSN 0730-0301. doi: 10.1145/3592442. URL `https://doi.org/10.1145/3592442`.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023b.

Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *European Conference on Computer Vision*, pages 18–35. Springer, 2022.

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. URL `https://openreview.net/forum?id=xmxgMij3LY`.

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.