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

Supplementary Material

1. More Implementation Details

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

1.1. 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 [1] 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.

1.2. Model Training

Following the approach in [9], 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 [2]. 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.

2. 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 the outcomes from various configurations aimed at enhancing normal maps.

2.1. More Qualitative Results

We provide more qualitative results for both mesh generate stage and mesh refinement stage in Fig. 1 and Fig. 2. 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 [8]. Due to the unavailability of the original mesh displayed in [8], we copyed the rendered images from their paper and present our results alongside for visual comparison. As is shown in 3, our method generate meshes align better with input images. Considering Clay [8] 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.

2.2. Ablation study

We provide qualitative ablation results in Fig. 4 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

Figure 1. Raw coarse meshes generated by our proposed method using a single image as a reference or a text prompt.

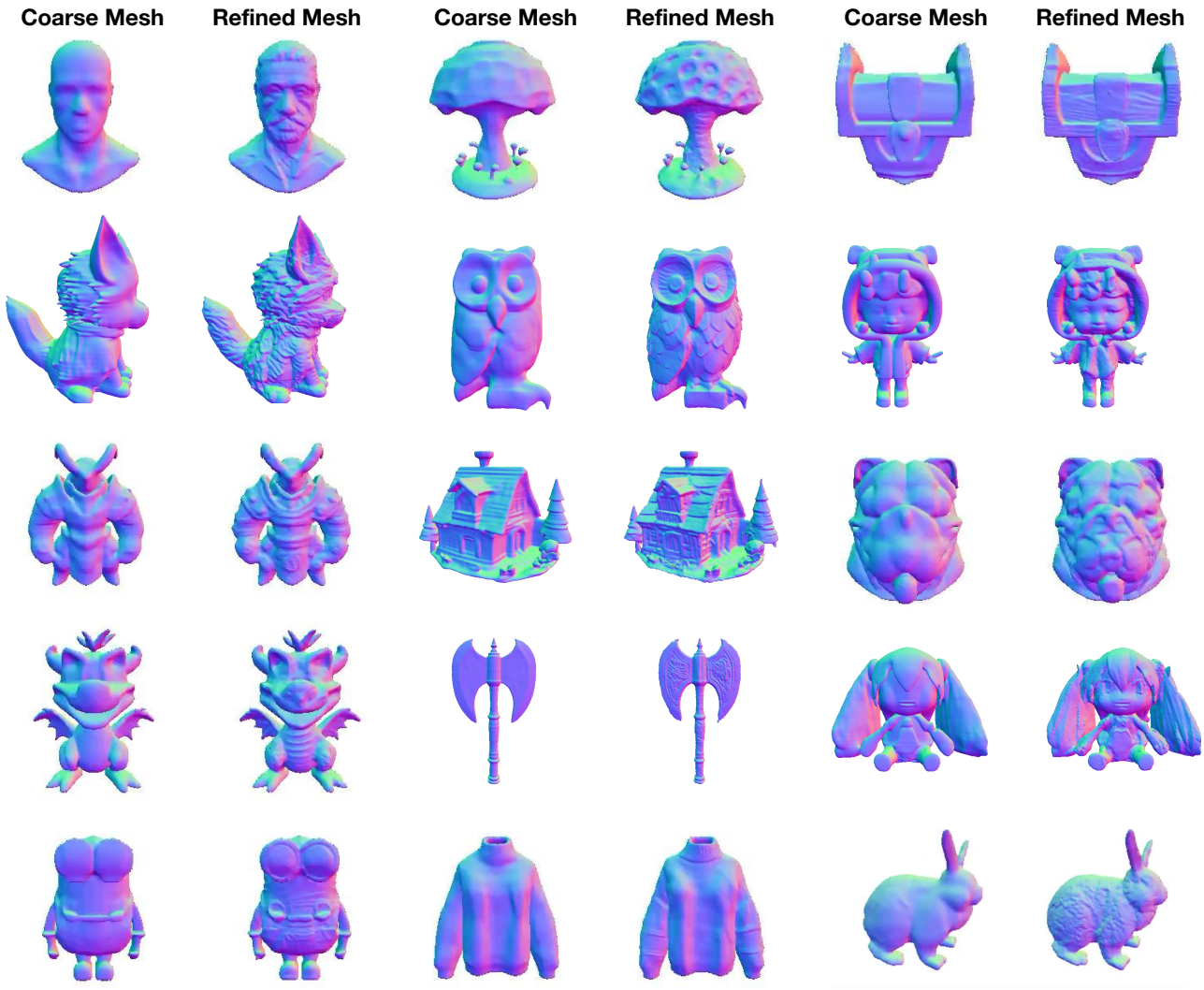

Figure 2. Results of the Automatic Mesh Refinement. We visualize the normal images of the coarse mesh and the refined mesh. It can be seen from the images that our proposed automatic normal refinement module significantly enhances the geometric details, such as the fur of the dog and the details of the face.

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.

2.3. 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

Figure 3. Comparison of image to 3d generation between ours coarse output and Clay [8]

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.

3. Application

3.1. Magic Normal Brush

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

Similar to ZBrush [6] 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 [3–5]. 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 7(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 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.

3.2. Image as Prompt for Mesh Refinement

As presented in Fig. 7(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 [7] 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 im-

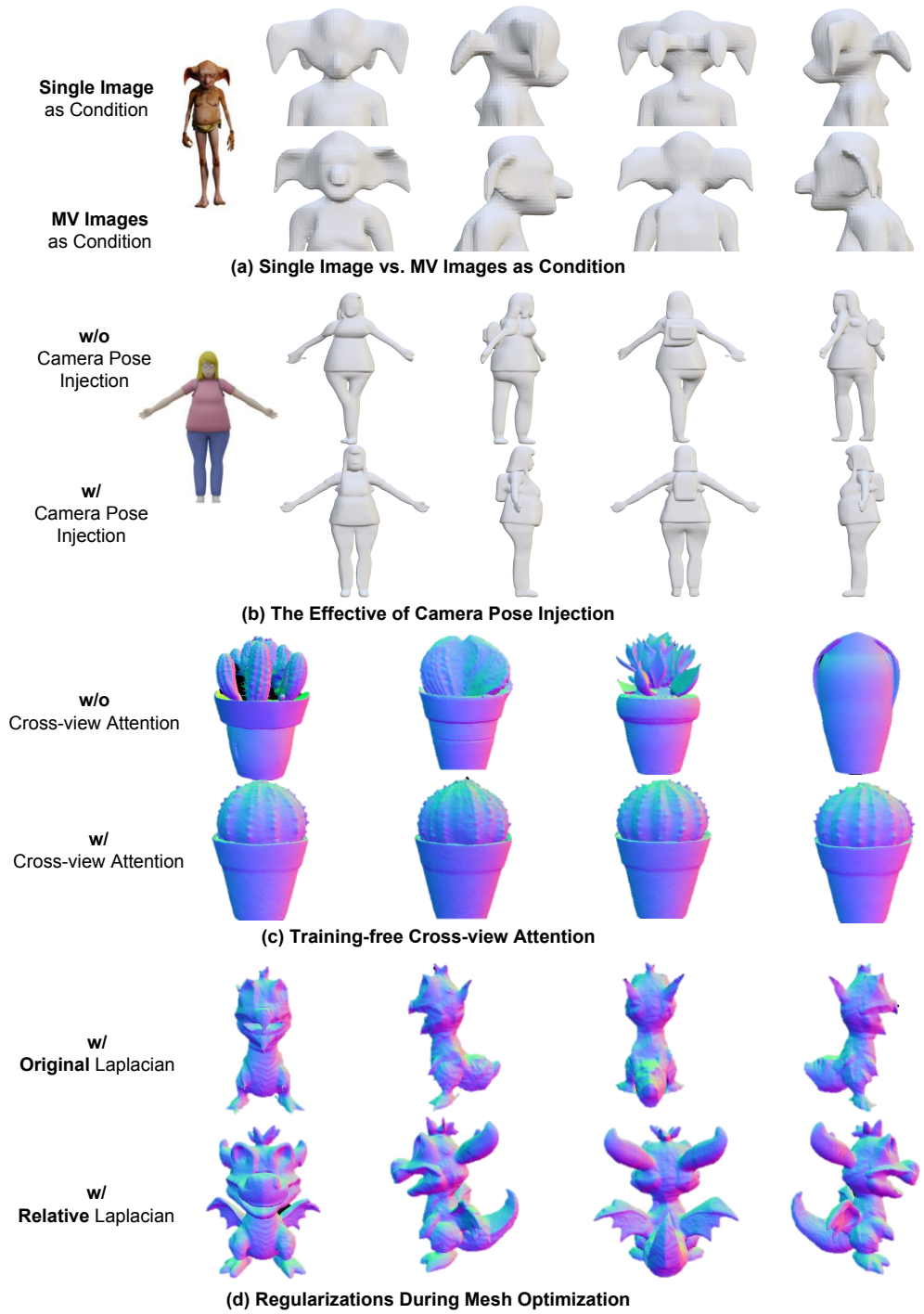

Figure 4. Ablation Study. (a) When only using a single image as a reference, the absence of information for the occluded parts can result in erroneous interpretations, as exemplified by the four ears of the goblin. (b) Incorporating the camera pose significantly enhances the diffusion model to comprehend 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.

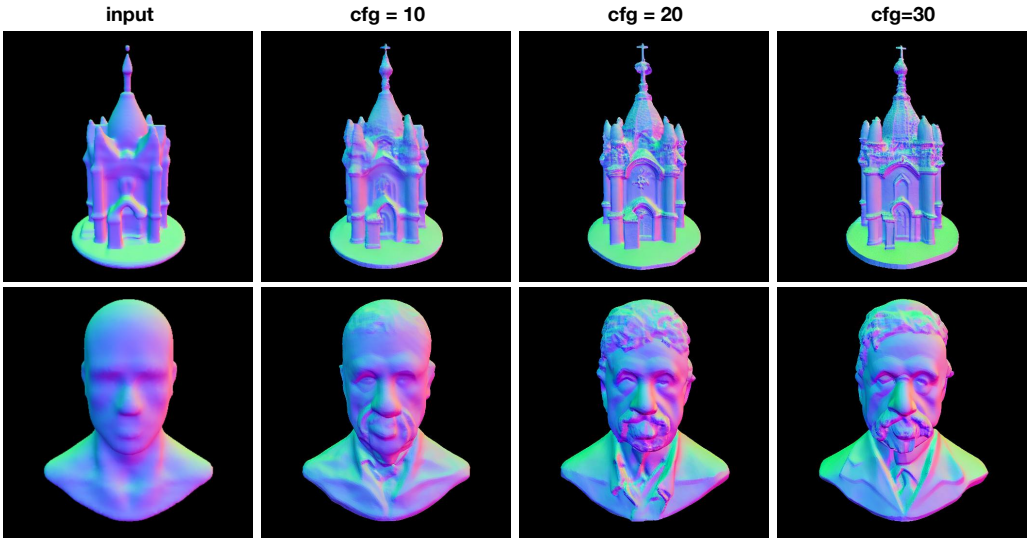

Figure 5. Normal refinement results with different CFG settings

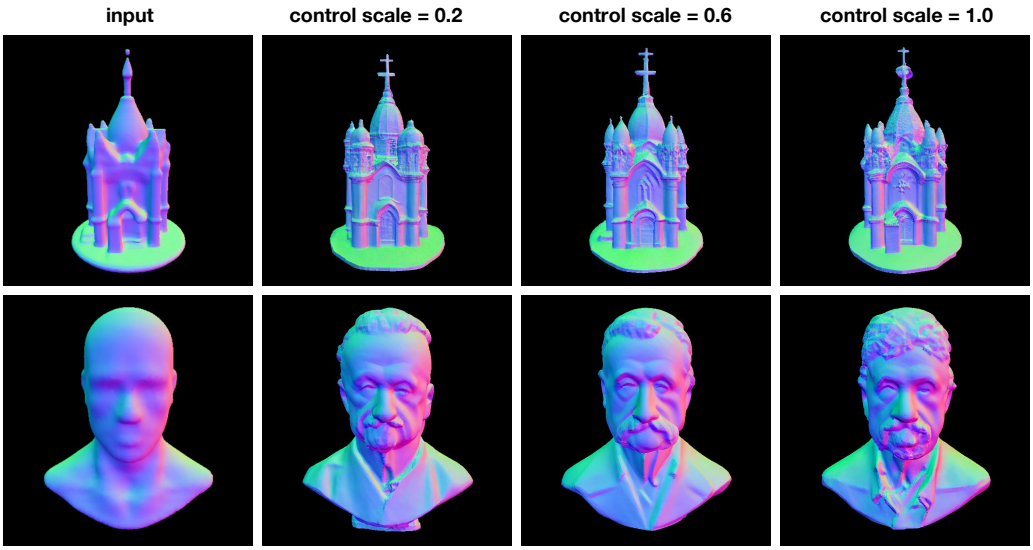

Figure 6. Normal refinement results with different control-scale settings

age, such as the facial features of an individual, to produce a mesh that maintains the same identity-preserving attribute.

3.3. 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 [8]. As is demonstrated in Fig. 9, 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. 10.

4. 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 8 and will add

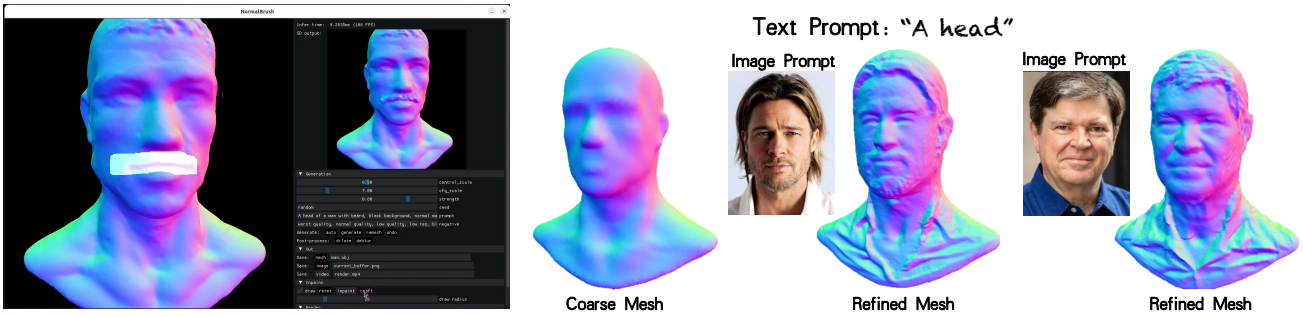

Figure 7. (a) With the *Magic Normal Brush*, it's convenient to edit a mesh via simple drawing and typing text. Whiskers are easily added to the mesh. (b) Our mesh refinement module is capable of accepting an image as the prompt. By incorporating a facial image to guide the normal mapping enhancement, we can refine the mesh according to the identity in the image.

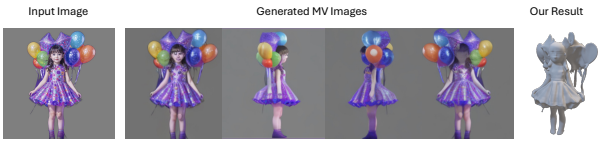

Figure 8. Failure cases due to the poor multi-view prediction and intricate structure.

Figure 9. Our mesh with texture. We implement texture generation using similiar methods as [8].

analysis in the revision.

5. 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.

References

[1] 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. 1

[2] 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. 1

[3] 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. 4

[4] 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.

[5] 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. 4

[6] *ZBrush: 3D Sculpting Software*. Pixologic, USA, 2023 edition, 2023. https://pixologic.com/. 4

[7] 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. 4

[8] 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. 1, 4, 6, 7, 8

[9] 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. 1

Figure 10. Our mesh with texture. We implement texture generation using similiar methods as [8].