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

Weiyu Li [*] Jiarui Liu [*] Rui Chen Yixun Liang

Xuelin Chen Ping Tan Xiaoxiao Long [†]

1 Appendix

In this Appendix, we describe a more detailed implementation in Sec. 1.1, including the data preprocessing and method training details. Subsequently, we delve into a more comprehensive discussion highlighting the advantages of 3D generative models in comparison to reconstruction models. Furthermore, we present the outcomes from various configurations aimed at enhancing normal maps.

1.1 More Implementation Details

We present more implementation details of data preparation for each component in our method.

1.1.1 Data Preparation

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]. 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.

1.1.2 Model Training

Following the approach in Zhao et al. [2023], 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 Adam 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.

[*] Joint first authors

[†] Corresponding author

For the normal-adapted diffusion model, which is derived from SD1.5, we opt for convenience to fine-tune the model introduced in Huang et al. [2024]. 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.

1.2 Different Settings of Normal Enhancement

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

The Effective of Different CFG Scale We demonstrate the results of different classifier-free guidance weights in Fig. 1. As this variable becomes larger, the refinement process produce results that are more consistent with the text description. If this value is too large, the quality of the generated image will decrease. We set this value to 20 by default in order to strike a balance between effect and quality.

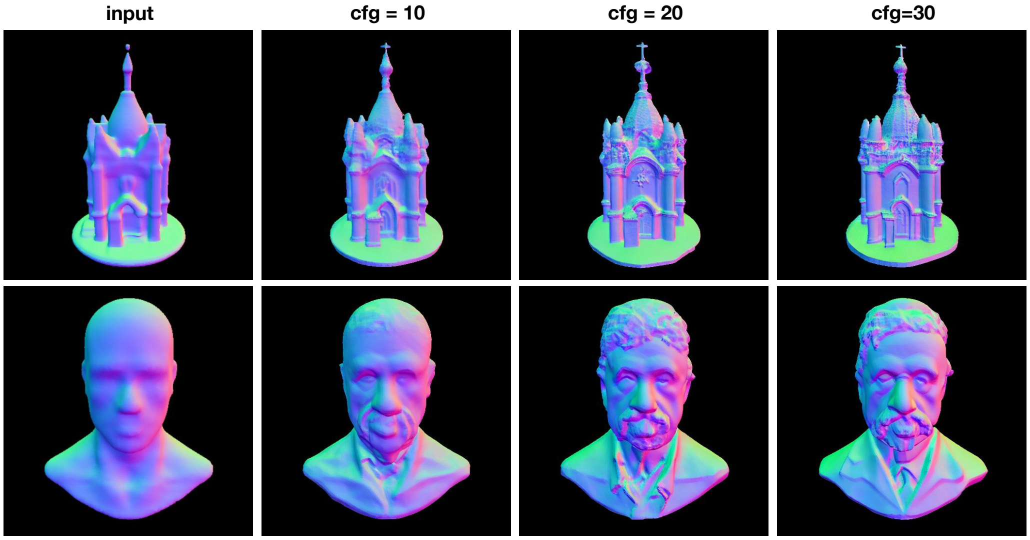

Figure 1: Normal refinement results with different CFG settings

The Effective of Control Scale for Tile Model The control scale defines how much the refinement process will refer to the control image, which is the normal map rendered from coarse mesh. As shown in the fig.2, a larger control scale results in less structural diversity and the refined normal maps are more likely to align with the 3D shape. We set this value to 0.8 by default to enhance details while preserving the overall shape of the coarse mesh.

References

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.

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.

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.

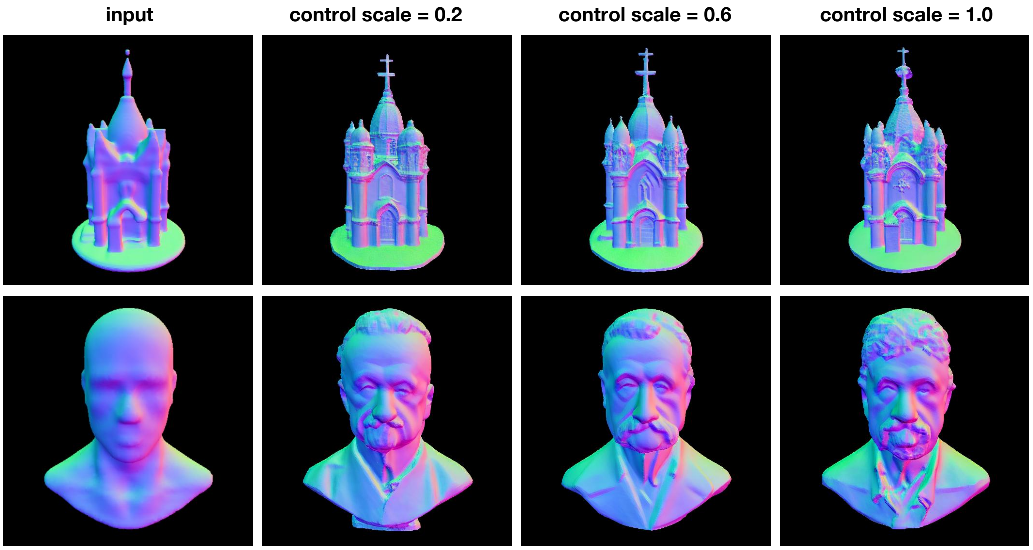

Figure 2: Normal refinement results with different control-scale settings