Controllable Text-to-3D Generation via Surface-Aligned Gaussian Splatting

Supplementary Material

A. Introduction

In this supplementary material, we offer additional details regarding our experimental setup and implementation. Subsequently, we present more qualitative results showcasing the performance and diversity of our method with various types of condition images as input.

B. Implementation Detail

B.1. Training Data

Multi-view Images Dataset. We employ the multi-view renderings from the publicly available large 3D dataset, Objaverse [2], to train our MVControl. Initially, we preprocess the dataset by removing all samples with a CLIP-score lower than 22, based on the labeling criteria from [12]. This filtering results in approximately 400k remaining samples. For each retained sample, we first normalize its scene bounding box to a unit cube centered at the world origin. Subsequently, we sample a random camera setting by uniformly selecting the camera distance between 1.4 and 1.6, the angle of Field-of-View (FoV) between 40 and 60 degrees, the degree of elevation between 0 and 30 degrees, and the starting azimuth angle between 0 and 360 degrees. Under the random camera setting, multi-view images are rendered at a resolution of 256×256 under 4 canonical views at the same elevation starting from the sampled azimuth. We repeat this procedure three times for each object. During training, one of these views is chosen as the reference view corresponding to the condition image. Instead of utilizing the names and tags of the 3D assets, we employ the captions from [6] as text descriptions for our retained objects.

Canny Edges. We apply the Canny edge detector [1] with random thresholds to all rendered images to obtain the Canny edge conditions. The lower threshold is randomly selected from the range [50, 125], while the upper threshold is chosen from the range [175, 250].

Depth Maps. We use the pre-trained depth estimator, Midas [7], to estimate the depth maps of rendered images.

Normal Maps. We compute normal map estimations of all rendered images by computing normal-from-distance on the depth values predicted by Midas.

User Scribble. We synthesize human scribbles from rendered images by employing an HED boundary detector [13] followed by a set of strong data augmentations, similar to those described in [14].

B.2. Training Details of MVControl

While our base model, MVDream [10], is fine-tuned from Stable Diffusion v2.1 [8], we train our multi-view ControlNet models from publicly available 2D ControlNet checkpoints[1] adapted to Stable Diffusion v2.1 for consistency. The models are trained on an 8×A100 node, where we have 160 (40×4) images on each GPU. With a gradient accumulation of 2 steps, we achieve a total batch size of 2560 images. The model undergoes 50000 steps of training under a constant learning rate of 4×10^{-5} with 1000 steps of warm-up. Similar to the approach in [14], we randomly drop the text prompt as empty with a 50% chance during training to facilitate classifier-free learning and enhance the model's understanding of input condition images. Moreover, we also employ 2D-3D joint training following [10]. Specifically, we randomly sample images from the AES v2 subset of LAION [9] with a 30% probability during training to ensure the network retains its learned 2D image priors.

B.3. Implementation Details of 3D Generation

Multi-view Image Generation . In our coarse Gaussian generation stage, the multi-view images are generated with our MVControl attached to MVDream using a 30-step DDIM sampler [11] with a guidance scale of 9 and a negative prompt "ugly, blurry, pixelated obscure, unnatural colors, poor lighting, dull, unclear, cropped, lowres, low quality, artifacts, duplicate".

Gaussian Optimization Stage . This stage comprises a total of 3000 steps. During the initial 1500 steps, we perform simple 3D Gaussian optimization with split and prune every 300 steps. After step 1500, we cease densification and prune, and instead introduce SuGaR [3] regularization terms to refine the scene. In the end of the stage, we prune all Gaussians with opacity below $\bar{\sigma} = 0.5$. The 3D SDS ($\nabla_\theta \mathcal{L}^{3D}_{SDS}$) is computed with a guidance scale of 50 using the CFG rescale trick [5], and $\nabla_\theta \mathcal{L}^{2D}_{SDS}$ is computed with a guidance scale of 20. We use $\lambda_{2D} = 0.1$ and $\lambda_{3D} = 0.01$ for 2D and 3D diffusion guidance, with resolutions of 512×512 and 256×256 respectively.

SuGaR Refinement Stage . This stage has totally 5000 steps of optimiztion. The $\nabla_\theta \mathcal{L}_{VSD}$ in the SuGaR refinement stage is computed with a guidance scale of 7.5. The training is under resolution 512×512 for rendering.

All score distillation terms also incorporate the aforementioned negative prompt. Our implementation is based

[1] https://huggingface.co/thibaud

on the threestudio project [4]. All testing images for condition image extraction are downloaded from *civitai.com*.

C. Additional Qualitative Results

C.1. Diversity of MVControl

Similar to 2D ControlNet [14], our MVControl can generate diverse multi-view images with the same condition image and prompt. Please refer to our project page for some of the results.

C.2. Textured Meshes

We also provide additional generated textured mesh. Please refer to our project page for video and interactive mesh results.

D. Textual Prompts for 3D Comparison

Here we provide the missing textual prompts in Fig. 4 of our main paper as below:

1. "RAW photo of A charming long brown coat dog, border collie, head of the dog, upper body, dark brown fur on the back,shelti,light brown fur on the chest,ultra detailed, brown eye"
2. "Wild bear in a sheepskin coat and boots, open-armed, dancing, boots, patterned cotton clothes, cinematic, best quality"
3. "Skull, masterpiece, a human skull made of broccoli"
4. "A cute penguin wearing smoking is riding skateboard, Adorable Character, extremely detailed"
5. "Masterpiece, batman, portrait, upper body, superhero, cape, mask"
6. "Ral-chrome, fox, with brown orange and white fur, seated, full body, adorable"
7. "Spiderman, mask, wearing black leather jacket, punk, absurdres, comic book"
8. "Marvel iron man, heavy armor suit, futuristic, very cool, slightly sideways, portrait, upper body"

References

[1] Canny, J.: A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence (6), 679–698 (1986) 1

[2] Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13142–13153 (2023) 1

[3] Guédon, A., Lepetit, V.: Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. arXiv preprint arXiv:2311.12775 (2023) 1

[4] Guo, Y.C., Liu, Y.T., Shao, R., Laforte, C., Voleti, V., Luo, G., Chen, C.H., Zou, Z.X., Wang, C., Cao, Y.P., Zhang, S.H.: threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio (2023) 2

[5] Lin, S., Liu, B., Li, J., Yang, X.: Common diffusion noise schedules and sample steps are flawed. arXiv preprint arXiv:2305.08891 (2023) 1

[6] Luo, T., Rockwell, C., Lee, H., Johnson, J.: Scalable 3d captioning with pretrained models. arXiv preprint arXiv:2306.07279 (2023) 1

[7] Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021) 1

[8] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 1

[9] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022) 1

[10] Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023) 1

[11] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 1

[12] Sun, Q., Li, Y., Liu, Z., Huang, X., Liu, F., Liu, X., Ouyang, W., Shao, J.: Unig3d: A unified 3d object generation dataset. arXiv preprint arXiv:2306.10730 (2023) 1

[13] Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE international conference on computer vision. pp. 1395–1403 (2015) 1

[14] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 1, 2