CAD: Photorealistic 3D Generation via Adversarial Distillation
Supplementary Material

A. Overview

In this supplemental material, additional implementation details and experimental results are provided, including:

- Results of the user study (Section B);
- More details about the whole distillation pipeline (Section C);
- More analysis regarding the pose pruning, distribution refinement, diversity and consistency training (Section D);
- Video demo, which shows the photorealistic generation of our method and qualitative video comparisons. Please refer to *video demo* in the project website.

B. User Study

To better evaluate the subjective quality and diversity of 3D generation, we conducted a user study to compare our method with existing baselines. Specifically, we created a test set comprising 10 images and prompts. For each case, we utilized Zero-1-to-3 [4] and Magic123 [6] for generating one 3D object, and Dreamfusion [5], ProlificDreamer [8], along with our method for synthesizing multiple 3D objects. Participants were then asked to rank the five groups from highest to lowest, based on a comprehensive inspection of various aspects, including rendering quality, photorealism, and generation diversity. We collected surveys from 22 participants and calculated the percentages of each method being selected as the top 1, 2, and 3, with the statistics shown in Figure 1. Our method demonstrates clear superiority over other methods, with more than 92% of users selecting it as the best result, significantly surpassing the second-highest percentage of 3.6% for Magic123 [6].

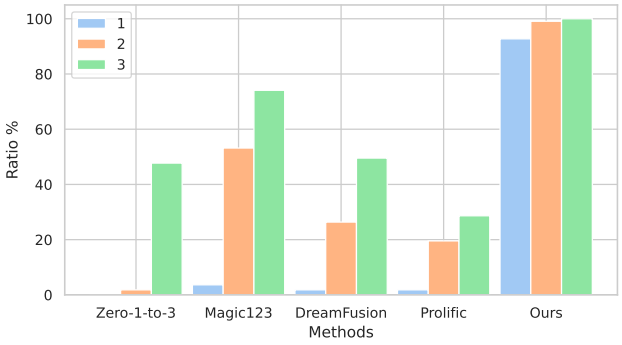


Figure 1. **Distribution of User Preferences**. This graph illustrates the percentages of participatns that ranked each method as their top three choices, indicated by 1, 2, and 3, respectively.

C. Implementation Details

The 3D generator architecture implemented in our paper is an adaptation of EG3D [2], with several critical modifications designed for our 3D adversarial distillation framework: 1) We have omitted the generator pose conditioning since our model targets general 3D objects, not specifically pose-correlated faces. 2) To achieve increased compactness and accelerated optimization, the maximum resolution of the triplanes is reduced to 128^2. 3) The raw (volumetric) rendering resolution of the triplanes is fixed at 64^2 throughout the distillation process. We rely on 3D upsampling for maintaining multiview consistency. The generated triplanes are confined within a $[-0.7, 0.7]^3$ box, and we utilize a field of view (FOV) of $49.1°$, following a pinhole camera model. We incorporate the absolute camera pose into the discriminator to facilitate the generator in learning the correct 3D prior. For the reference viewpoint, we set the azimuth angle (ϕ) to $180°$, the distance from the center $r = 2$, and determine the polar angle (θ) using the pose estimation module from One-2-3-45 [3]. The pose for any other viewpoint is then derived through relative transformation.

To ensure the efficiency of the adversarial distillation, we do not adopt the on-the-fly way to sample $I \sim p_0$ from a frozen diffusion model, instead we cache the sampling in advance so that during the distillation we could directly fetch the prior from the memory. Although sampling more data prior will can lead to improved quality and diversity, considering the efficiency we only restrict the pre-sampling number to be 10K and enable the adaptive discriminator augmentation (ADA) to stabilize the adversarial training. Each sample is associated with a unique pose, uniformly distributed across a sphere with a radius of 2.

In the process of distribution refinement, we apply different noise strengths to different diffusion priors. Specifically, a noise strength of 0.8 is utilized for the depth-conditioned Control-Net. In the case of DeepFloyd, noise strengths of 0.3 and 0.7 are employed for its low-resolution and high-resolution branches, respectively. Additionally, since during refinement we are aware of the camera pose, thus we also leverage the view-dependent prompting to effectively prevent incorrect face generation when dealing with asymmetric objects. Please noted in the main submission, for certain data we use Control-Net for the refinement, while for the rest we leverage DeepFloyd.

We adopt a two-stage approach for training the consistent 3D generator. During the first stage, image-space convolutions are utilized to upscale the raw rendering from 128^2 triplanes to calculate the adversarial loss. In the second stage, a compact 3D upsampler is employed to lift the triplanes to a 256^2 resolution. This stage also incorporates patch-level

Figure 2. **Effects of the noise strength.** We show the refinement results by using different diffusion models and various noise strengths.

Figure 3. **Diffusion refinement to ensure the sampled prior from target distribution is high-quality.**

GAN loss and LPIPS for supervision. Notably, we observe a significant quality degradation when training the patch discriminator with adaptive discriminator augmentation (ADA), since the combination of random patch-level augmentation will result in a larger gap compared with original rendering. Hence, we disable ADA in the second stage of training to maintain quality.

To obtain a 3D distribution based on a given reference image and prompt, the entire distillation process takes around three days at a 256^2 image resolution, which includes 1.5 days for training the 2D upsampler branch, followed by another 1.5 days for finetuning together with the 3D upsampler, using 4 NVIDIA Tesla V100 GPUs. It is worth noting that GAN training can sometimes be unstable, but the integration of our proposed pose pruning strategy could effectively mitigate the need for dense parameter tuning in R1 regularization, for which we consistently set the weight $\gamma = 3$ across all experiments.

Figure 4. **Visualizations of pose pruning.** We show the frequently appeared errors while leveraging the view-dependent diffusion model. By simultaneously considering the parallel sampled results for the same viewpoint, our pruning strategy could effectively filter out the wrong generations.

D. Additional Analysis

Importance of distribution refinement. The impact of the distribution refinement step on the distillation quality has been demonstrated in the main paper. To further illustrate its importance, we present a visualization in Figure 3. This figure compares directly sampled results from Zero-1-to-3 with those refined by a 2D diffusion model. The refinement process greatly enhances both the quality and diversity of the

2D upsampling 3D upsampling

Figure 5. **Comparisons between 3D generation without consistency training and using our full model.** By directly upsampling the triplanes, the rendering from our model is naturally consistent.

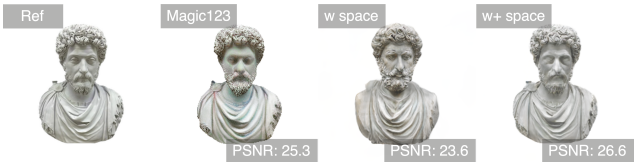
target distribution, which makes the GAN training process more effective.

Controlling Generation Diversity. We observed the most significant factor that influences the generation diversity is the noise strength used when sampling the prior. To better demonstrate this observation, we show the refinement results of different diffusion models with continuous noise strengths in Figure. 2. Increasing noise strength results in larger differences from the non-refined input, as it removes more contents from the input. This makes noise strength an effective tool for controlling diversity. However, when using solely text-conditioned diffusion models like DeepFloyd [1], a higher noise strength (e.g., $s \geq 0.7$) can also induce pose changes. In contrast, the depth-conditioned Control-Net can effectively preserve pose-image correlation due to the geometric constraints from depth. In our main paper, to demonstrate the generality of our adversarial distillation pipeline, we concurrently try the two diffusion models with fixed noise strength. Consequently, the renderings of some models closely resemble the reference image, while others show larger differences yet retain similar semantics.

Visualization of pose pruning. As shown in Figure. 4, incorporating view-conditioned diffusion models to address sampling bias can result in errors such as mismatched semantics with input views (first row), incorrect camera poses (second row), and wrong 3D structures (third row), all of which will have negative influences on the adversarial training. By considering two aspects including the geometry consistency and semantic consistency, our pruning strategy could effectively filter out these bad samples to make the 3D GAN training stable again even with limited data.

Importance of consistency training. Limited by the efficiency of volumetric rendering, most 3D GANs [2, 7] rely on image-space convolution for upsampling raw renderings to higher resolutions. While this approach may not result in very visible inconsistencies for face modeling, it is insufficient for synthesizing general 3D objects with a 360° azimuth angle. In Figure 5, we compare 2D and 3D upsampling methods qualitatively. Although 2D upsampling can ensure high rendering quality for each view, the appearance differences become dramatically high when changing camera poses. Our distillation method, however, bakes the abilities of the 2D branch into the 3D branch, which could effectively render multi-view consistent and photorealistic images.

Inconsistent reconstruction with the reference. In the main paper we only leverage w space for inversion to showcase the potential application of CAD, which is not powerful enough to faithfully recover the input, as pointed out by the reviewer. To address this, we try the $w+$ space and note that the reconstruction quality is improved. We believe integrating more advanced inversion techniques such as pivot tuning could further improve the SVR consistency.



References

[1] Deepfloyd. https://github.com/deep-floyd/IF, 2023. 2, 3

[2] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 1, 3

[3] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 1

[4] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 1, 2

[5] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

[6] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 1

[7] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20991–21002, 2023. 3

[8] 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. 2023. 1

[9] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 2