Acc3D: Accelerating Single Image to 3D Diffusion Models via Edge Consistency Guided Score Distillation (*Supplementary Material*)

A. Theoratical Investigation of Edge Consistency V.S. Full Consistency Distillation

As mentioned in the manuscript, the consistency model regularizes the consistency of estimated clean samples as

$$\mathcal{L}_c = \left\| \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{X}_t, t) - \mathcal{F}_{\boldsymbol{\theta}^-}(\mathbf{X}_{t-\Delta t}, t-\Delta t) \right\|_F . \tag{S.1}$$

The original consistency models hypothesize that when the consistency training loss $\mathcal{L}_c$ decreases to 0, the error of consistency function follows $\mathcal{O}((\Delta t)^p)$. However, there is a critical issue that such hypnosis is too ideal for large-scale model and dataset training processes. The original loss $\mathcal{L}_c$ cannot be optimized to 0. Thus, there is an inevitable accumulated error, shifting the estimations of the consistency model from the original data manifold.

To further take such optimization error into consideration, we assume that the error, e.g., the difference between two step inference results, follows a uniform distribution, i.e., $\mathbf{u} = \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{X}_t, t) - \mathcal{F}_{\boldsymbol{\theta}^-}(\mathbf{X}_{t-\Delta t}, t-\Delta t)$, and $\mathbf{u} \sim \mathcal{U}(0, \delta)$, where δ is a quite small number.

Notions. We denote the consistency function of empirical PF ODE as $\mathcal{F}_{\boldsymbol{\Phi}}(\cdot,\cdot)$ with $\boldsymbol{\Phi}$ as the parameters of well-trained (ideal) diffusion model, which can be realized by the integral process. We then calculate the error term $\left\| \mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_n}, t_n) - \mathcal{F}_{\boldsymbol{\Phi}}(\boldsymbol{x}_{t_n}, t_n) \right\|_F$. We denote $\mathcal{E}_n$ as the error term with the timestamp t_n.

$$\mathcal{E}_n := \mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_n}, t_n) - \mathcal{F}_{\boldsymbol{\Phi}}(\boldsymbol{x}_{t_n}, t_n). \tag{S.2}$$

Moreover, for the term with timestamp t_{n+1} , we have

$$\begin{aligned}
\mathcal{E}_{n+1} &= \mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{n+1}}, t_{n+1}) - \mathcal{F}_{\boldsymbol{\Phi}}(\boldsymbol{x}_{t_{n+1}}, t_{n+1}); \\
&= \mathcal{F}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}^{\boldsymbol{\phi}}_{t_n}, t_n) + \boldsymbol{u} - \mathcal{F}_{\boldsymbol{\Phi}}(\mathbf{x}_{t_n}, t_n); \\
&= \mathcal{F}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}^{\boldsymbol{\phi}}_{t_n}, t_n) - \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_{t_n}, t_n) + \boldsymbol{u}_n + \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_{t_n}, t_n) - \mathcal{F}_{\boldsymbol{\Phi}}(\boldsymbol{x}_{t_n}, t_n); \\
&= \mathcal{F}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}^{\boldsymbol{\phi}}_{t_n}, t_n) - \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_{t_n}, t_n) + \boldsymbol{u}_n + \mathcal{E}_n.
\end{aligned}$$

Then, we have

$$\begin{aligned}
\|\mathcal{E}_n\|_F &\leq \|\mathcal{E}_1\|_F + \sum_{k=1}^{n} \left(\|\boldsymbol{u}_k\|_F + \|\mathcal{F}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}^{\boldsymbol{\phi}}_{t_k}, t_k) - \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_{t_k}, t_k)\|_F \right); \\
&\overset{①}{=} n \times \mathbb{E}_{\mathbf{x}_{t_n} \sim \mathbf{P}(\mathbf{x}_{t_n})} \left[\|\boldsymbol{u}_n\|_F \right] + \sum_{k=1}^{n} \mathcal{O}\left((t_{k+1} - t_k)^{p+1} \right); \\
&\overset{②}{\leq} n \times \mathbb{E}_{\mathbf{x}_{t_n} \sim \mathbf{P}(\mathbf{x}_{t_n})} \left[\|\boldsymbol{u}_n\|_F \right] + \sum_{k=1}^{n} \mathcal{O}\left((\Delta t)^{p+1} \right); \\
&= n \times \left\{ \mathbb{E}_{\mathbf{x}_{t_n} \sim \mathbf{P}(\mathbf{x}_{t_n})} \left[\|\boldsymbol{u}_n\|_F \right] + \mathcal{O}((\Delta t)^p) \right\},
\end{aligned}$$

where ① is for the importance sampling of $\boldsymbol{u}$, $\|\mathcal{E}_1\|_F = 0$ (the boundary condition) and Taylor expansion of $\mathcal{F}_{\boldsymbol{\theta}}$; in ②, $\Delta t = \max(t_{n+1} - t_n)$ is selected as the largest time interval. From the results, we can derive that instead of the inherent Tylor high order residues $\mathcal{O}((\Delta t)^p)$, there is further training residual error term of $N \times \mathbb{E}\|\boldsymbol{u}\|_F$. Moreover, both of them are scaled by the number of intervals N, which indicates the accuracy of consistency is negatively associated with the length of the interval we want to regularize, i.e., the full consistency resulting in the largest potential error. Here we denote the aforementioned upper error boundary $\mathbb{E}_{\mathbf{x}_{t_n} \sim \mathbf{P}(\mathbf{x_{t_n}})}\|\boldsymbol{u}_n\|_F + \mathcal{O}((\Delta t)^p)$ by $\|\mathcal{E}_r\|_F$. We then have $\|\mathcal{E}_n\|_F \leq n \times \|\mathcal{E}_r\|_F$.
Moreover, considering the distillation process, we feed the one-step generation result into the edge consistency region via interpolating the latent at timestamp t as

$$\mathbf{x}_{t|0,T} = \alpha_t \mathbf{x}_{0|T} + \sigma_t \boldsymbol{\epsilon}.$$

Since such one-step generation is coarse due to the error of the score function in the pure noise state, note that such interpolation is a contraction mapping for our estimation term $\mathbf{x}_{0|T}$ ($\alpha_t \leq 1$). By assuming the consistency function has $L-$Lipschitz continuity, we have

$$\|\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_t^*, t) - \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_{t|0,T}, t)\|_F \leq L\|\mathbf{x}_t^* - \mathbf{x}_{t|0,T}\|_F, = L\alpha_t\|\mathbf{x}_0^* - \mathbf{x}_{t|0,T}\|_F,$$

which indicates the error of our distillation target is positively associated with data ratio α_t. However, note that such errors are estimated over the trained consistency model parameterized by $\boldsymbol{\theta}$, whose error is shown as $\|\mathcal{E}_n\|_F$. Thus, taking both consistency model training error with inherent noised latent interpolation error into consideration, we have

$$\begin{aligned}\mathcal{E}_A(\mathbf{x}_{0|t}) &\leq L\alpha_t\|\mathbf{x}_0^* - \mathbf{x}_{t|0,T}\|_F + n \times \|\mathcal{E}_r\|_F, \\ &\overset{①}{=} L\alpha_t\|\mathbf{x}_0^* - \mathbf{x}_{t|0,T}\|_F + t \times \|\mathcal{E}_r\|_F,\end{aligned}$$

where ① is derived by setting the distillation step t to be the same as the largest edge consistency training timestamp n. Then, it's natural to derive a lower error upper bound as setting the t^* to ensure $\frac{d\alpha_t}{dt}|_{t=t^*} = -\frac{L\alpha_t\|\mathbf{x}_0^* - \mathbf{x}_{t|0,T}\|_F}{\|\mathcal{E}_r\|_F}$. It indicates that we need to train and utilize the consistency in the region of $[t^*, 0]$, which is exactly the proposed edge consistency. Moreover, we utilize empirical experiments to validate the choice of t^*, as shown in our ablation studies.

B. Details of the Adversarial Architecture

We extract the features from the last three layers of the diffusion model. The feature channels are processed to attain a uniform channel value of 640, achieved through the use of a 1×1 convolution. Next, average pooling is used to ensure consistent sizing across the features. The three layers of features are then concatenated along the channel dimension and fed into the prediction head. Inspired by DMD [7], the prediction head is composed of a series of 4x4 convolutions with a stride of 2, group normalization, and SiLU activations. All feature maps are downscaled to a 4×4 resolution, which is then followed by a singular convolutional layer with a kernel size and stride of 4. This layer aggregates the feature maps into a single vector, which is subsequently fed into a linear projection layer to predict the classification score.

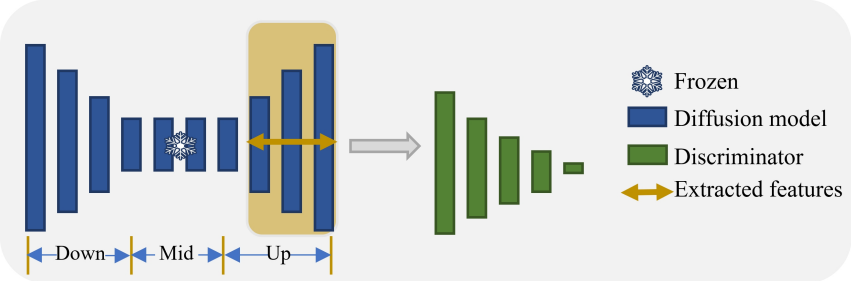

Figure 1. Illustration of the adversarial architecture, where we utilize the features from the last three layers of the feature extractor (the features in the yellow box).

C. Implementation Details

The training phase of the Generative Adversarial Network (GAN) is unstable and susceptible to influence from the initial phase. Hence, we warm up the single-step diffusion model using a typical distillation framework to distill knowledge from the multi-step diffusion model to the single-step one, to avoid the failure of single-step prediction. Specifically, the multi-step model executes inference in 6 steps and the warm-up phase trains approximately 6,000 iterations.

D. More Visual Results

In this section, We present additional visual results generated by our accelerated model to further highlight its capabilities and performance. These examples include a range of typical demonstrations for Image-to-3D tasks, which are commonly used benchmarks in multi-view image diffusion models [2–4]. The results, shown in Figure 2, demonstrate the model's ability to produce high-quality, multiview-consistent 3D reconstructions with remarkable efficiency. Additionally, we present further outputs produced by the Text-to-Image (T2I) model, Flux [6], which are visualized in Figure 3. In the main paper, we conducted a comprehensive evaluation of our accelerated model on the GSO [1] and DTC [5] datasets, which are well-established benchmarks for assessing 3D generation tasks. To provide deeper insights and reinforce the superior performance of our approach, we include additional visual results in Figure 4 and Figure 5. These results further demonstrate the model's ability to achieve high-quality, multiview-consistent 3D reconstructions with significantly fewer inference steps, setting a new benchmark for diffusion-based 3D generation methods.

E. More Visual Results of Ablation Studies

In this section, we provide a detailed analysis of Table 3 from the main paper and present some intuitive visualization results.

E.1. Risk of Mode Collapse

As shown in Fig. 6 (1) and (3), in the absence of distillation, adversarial learning is prone to mode collapse, with all results skewing towards an unusual pattern. This is particularly evident with the 2^{nd} and 3^{rd} samples in (1). Fig. 6 (2) showcases the outcomes of fully guided score distillation. As shown in 1^{st} and 3^{rd} samples in (2), fully guided distillation potentially amplifies the learning burden, subsequently lowering generative performance.

Figure 2. The qualitative results generated by Acc3D on typical demonstrations in Image-to-3D. Acc3D is capable of producing outstanding multi-view outputs in just two steps. Zoom in for details.

Figure 3. The qualitative results produced by Acc3D, utilizing **less than four** inference steps on a variety of image styles generated by the Text-to-Image model Flux [6]. Zoom in for details.

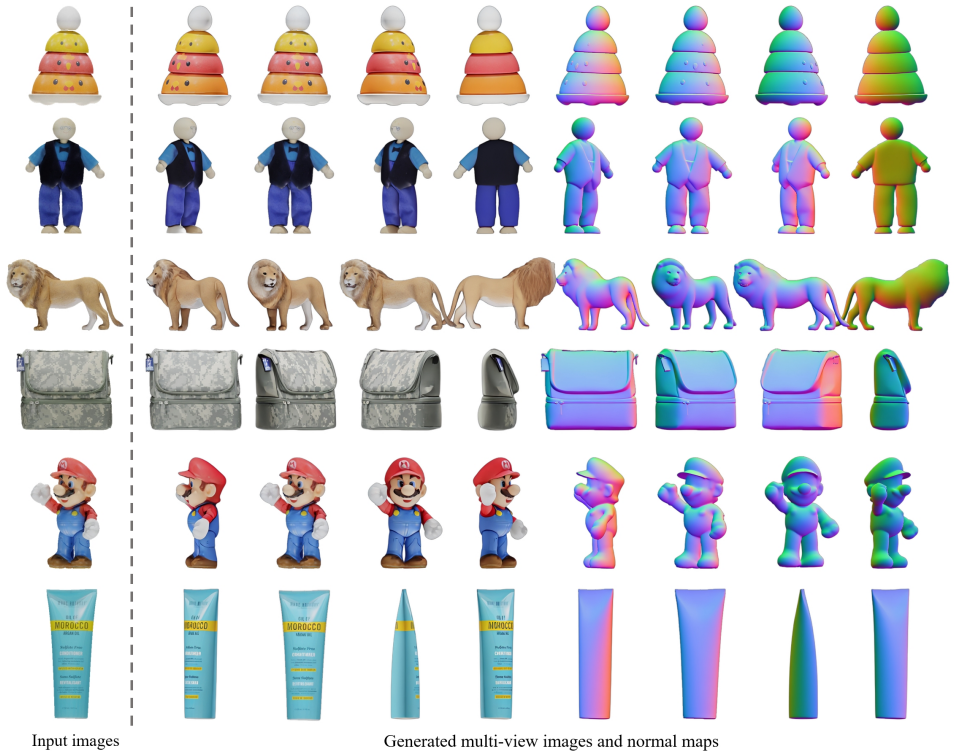

Figure 4. The qualitative results on GSO dataset [1]. Zoom in for details.

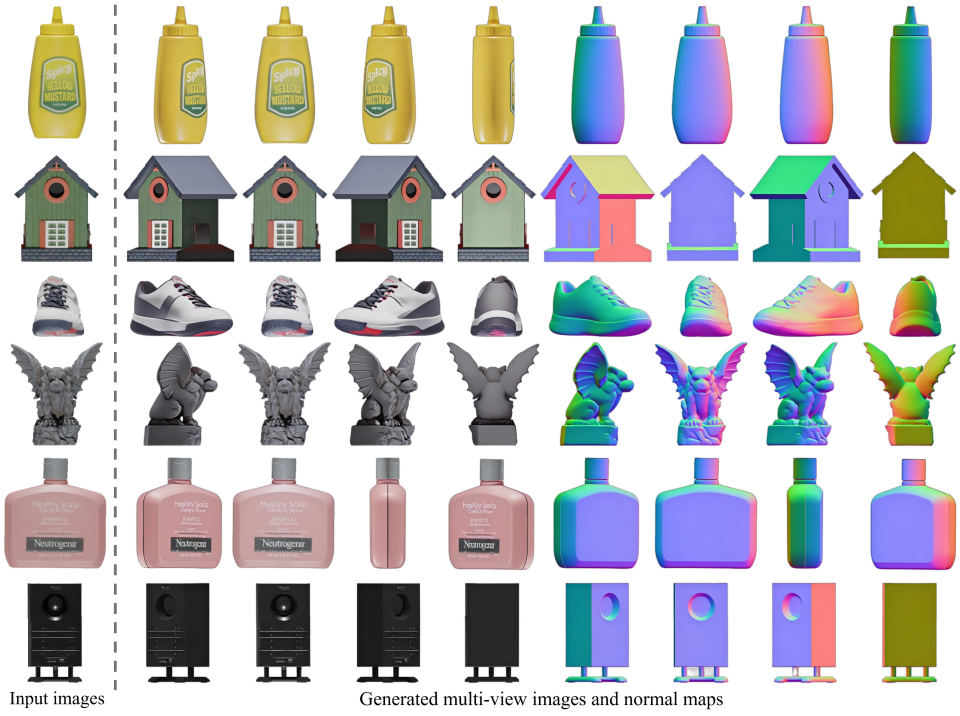

Figure 5. The qualitative results on DTC dataset [5]. Zoom in for details.

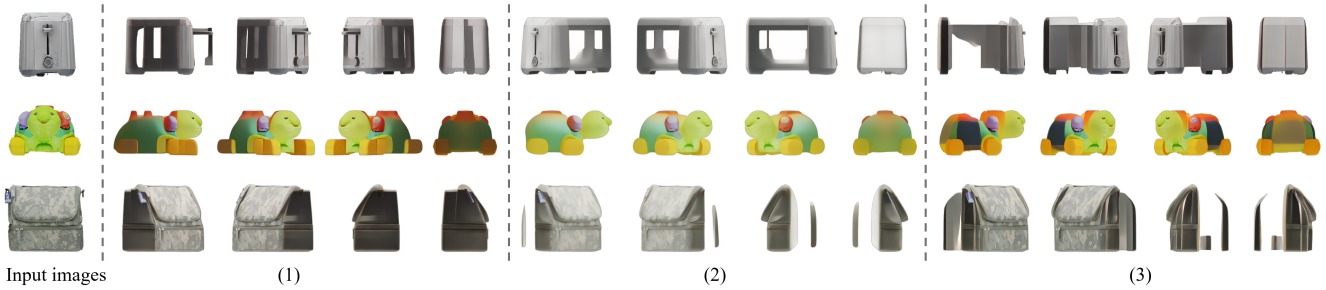

Figure 6. Visualizations of some collapse under various experiment settings. All results are generated in two steps. The results (1), (2), and (3) correspond to the experimental configurations (b), (c), and (e) outlined in Table 3 from the main paper, respectively.

E.2. Negative Effect of Single Discriminator

As depicted in Fig. 7, the fusing of these two modalities adversely affect the discriminator/model's performance, shown obviously for the normal maps of 2^{nd} and 3^{rd} samples. Independent learning of different modalities in the discriminator can enhance the stability of adversarial learning and result in superior quantitative outcomes.

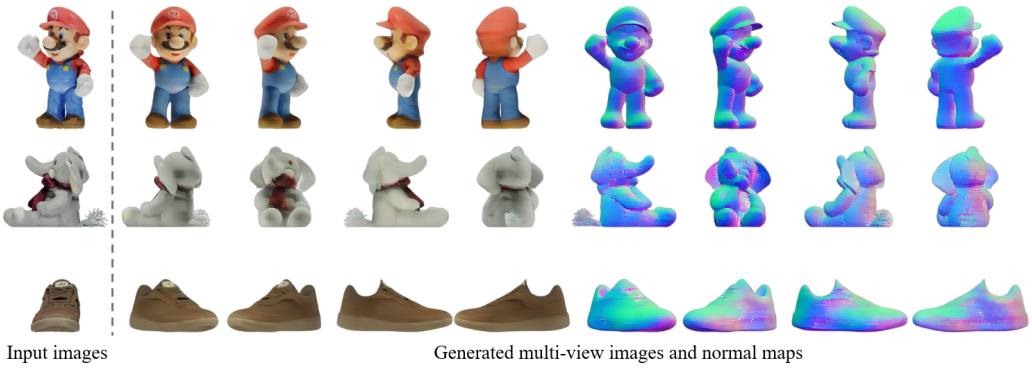

Figure 7. Visualizations of single discriminator processing both colors and normal maps simultaneously.

F. Comparison with Base Model Era3D

We visualize the results of our method and the base model Era3D in Fig. 8. Our method produces clearer and visually superior results at the same resolution (512), demonstrating finer details and better structural consistency. Additionally, our approach requires significantly fewer steps to achieve high-quality novel view synthesis, making it more efficient while maintaining superior visual fidelity. This highlights the effectiveness of our framework in generating high-resolution, high-quality 3D-aware images with reduced computational cost.

G. Diversity of Sampling

Diversity is an important evaluation indicator for single image-to-3D model. Our model demonstrates the capability to produce an array of superior-quality examples, as depicted in Fig 9. In Table 1, we present the variance observed in multiple samplings relative to the direct regression from the real images. Directly mapping the diffusion model's outputs with real images results in a diminished diversity.

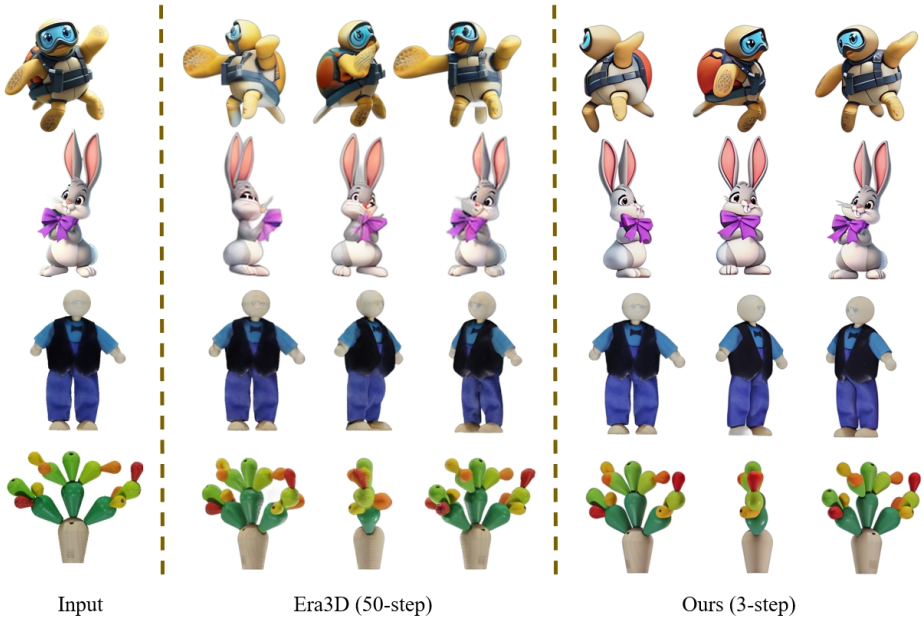

Figure 8. Visual comparisons with baseline model Era3D.

Table 1. Variance comparisons with the regression strategy on sampling diversity. "Regression" signifies the direct regression of the diffusion model's outputs with real images.

	Camera1	Camera2	Pencilcase	Lunchbox1	Lunchbox1	Average Variance ↑
Ours	1662.66	399.56	902.21	618.93	1281.10	972.89
Regression	342.09	125.43	211.42	91.71	147.02	183.53

Figure 9. Diversity in the synthesis of novel views under different seeds. The diverse results highlight a broad range of diversity, capturing both the geometrical and visual characteristics that are not present in the input view.

References

[1] 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. 3, 6

[2] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *arXiv preprint arXiv:2405.11616*, 2024. 3

[3] 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.

[4] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 3

[5] Meta. Dtc object dataset, 2024. 3, 6

[6] StabilityAI. Flux.1 model on huggingface, 2023. 3, 5

[7] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024. 2