Diffusion Time-step Curriculum for One Image to 3D Generation

Xuanyu Yi[1,4], **Zike Wu**[1], **Qingshan Xu**[1], **Pan Zhou**[3*], **Joo-Hwee Lim**[4], **Hanwang Zhang**[2,1]

[1]Nanyang Technological University, [2]Skywork AI

[3]Singapore Management University, [4]Institute for Infocomm Research

xuanyu001@e.ntu.edu.sg, zike001@e.ntu.edu.sg, qingshan.xu@ntu.edu.sg, panzhou@smu.edu.sg, joohwee@i2r.a-star.edu.sg, hanwangzhang@ntu.edu.sg

Figure 1. DTC123 can generate high-fidelity and multiview-consistent 3D assets from a single arbitrary image, significantly enhancing the original SDS pipeline through the *diffusion time-step curriculum*.

Abstract

Score distillation sampling (SDS) has been widely adopted to overcome the absence of unseen views in reconstructing 3D objects from a ***single*** *image. It leverages pre-trained 2D diffusion models as teacher to guide the reconstruction of student 3D models. Despite their remarkable success, SDS-based methods often encounter geometric artifacts and texture saturation. We find out the crux is the overlooked indiscriminate treatment of diffusion time-steps during optimization: it unreasonably treats the student-teacher knowledge distillation to be equal at all time-steps and thus entangles coarse-grained and fine-grained modeling. Therefore, we propose the Diffusion Time-step Curriculum one-image-to-3D pipeline (DTC123), which involves both the teacher and student models collaborating with the time-step curriculum in a coarse-to-fine manner. Extensive experiments on NeRF4, RealFusion15, GSO and Level50 benchmark demonstrate that DTC123 can produce multi-view consistent, high-quality, and diverse 3D assets. Codes and more generation demos will be released in https://github.com/yxymessi/DTC123.*

*Corresponding author.

1. Introduction

We consider the problem of obtaining a 3D asset from a *single* image. This endeavor holds tremendous industrial promise, notably in realms such as AR/VR content creation from a single snapshot and enhancing robotic navigation through individual captures [9, 42]. However, reconstructing 3D models (*e.g.*, NeRF [2, 28, 29] and mesh [14, 23]) from a single image has remained a formidable challenge due to its severely ill-posed nature, as one image does not contain sufficient unseen views of a 3D scene.

Fortunately, recent advances in large-scale pretrained 2D diffusion models [35, 36, 61] have paved the way for such an ill-posed challenge by synthesizing the unseen views of quality. Pioneered by Score Distillation Sampling (SDS) [30], we can use the 2D models as *teacher* to guide the reconstruction of the 3D models as *student*. The key is rooted in a symbiotic teacher-student cycle: As illustrated in Figure 2 (*Top*), in the early iterations with significantly flawed student output, the teacher provides the rough shape of objects as guidance; as the student gradually improves, it reciprocates with more precise conditions for the teacher, who in turn provides more accurate and fine-grained supervision. Along with the teacher's diffusion time steps, SDS optimizes the student by minimizing the reconstruction error between the student-rendered and the teacher-generated 2D images (see details in Sec. 3.1).

However, 3D models generated by SDS have defects that cannot be ignored. As shown in Figure 1, empirical observations indicate that they often encounter collapsed geometry and limited fidelity. Such issues arise primarily from the confusion of holistic structures and local details as stated in the field of novel view synthesis [7, 34, 51]. We conjecture that this may be attributed to the uniform sampling of diffusion time steps during the calculation of reconstruction errors, that is, SDS treats the student-teacher knowledge distillation to be *equal at all time steps*. This is counter-intuitive, *e.g.*, the teacher shouldn't teach low-level details when the student is still grappling with high-level concepts.

Our key insight is that an optimal SDS should follow a *diffusion time-step curriculum* : larger time steps capture coarse-grained knowledge like geometry formation and smaller time steps focus on enhancing fine-grained details like texture nuance. To this end, we propose the **D**iffusion **T**ime-step **C**urriculum **1** image **to 3**D pipeline dubbed **DTC123**, where both the teacher and the student model collaborate with the annealing time-step, exhibiting a coarse-to-fine generation process. More concretely,

- **Student-wise**: 3D models should progress from low-resolution concepts to high-resolution. We leverage resolution constraints from the hash-encoding band (NeRF) and the tet grid (DMTet) to gradually absorb knowledge, beginning with broader structural elements and eventually focusing on localized textures and complex scene illumination (Sec. 4.2).
- **Teacher-wise**: diffusion models should prioritize on coarse shape to visual details. We employ a pose-aware prior, Zero-1-to-3 [22], to establish a coarse-grained structure that aligns with the reference image. Subsequently, we harness the combined guidance of Zero-1-to-3 and Stable Diffusion, with LLM-augmented prompt and multi-step sampling (Sec. 4.3), to further provide fine-grained texture intricacies.

In addition, we introduce several geometric regularization to alleviate the Janus Face [12, 27] and high-frequency surface artifacts (Sec. 4.4). As illustrated in Figure 1, by integrating the time-step curriculum, which includes three coherent parts — time-step schedule, progressive student representation and teacher guidance with the aforementioned regularization techniques, DTC123 significantly enhances the geometry quality and texture fidelity of the SDS-based pipeline. We demonstrate the superiority of DTC123 on NeRF4 [28], RealFusion15 [26], GSO [5] and our benchmark Level50. Comprehensive quantitative and qualitative evaluations clearly show that DTC123 can efficiently generate multi-view consistent, high-fidelity, and diverse 3D assets, continually outperforming other state-of-the-art methods. Furthermore, DTC123 enables wide range of applications, *e.g.*, multi-instance generation and mesh refinement. In summary, we make three-fold contributions:

- We develop an end-end one image-3D pipeline DTC123, boosting the efficiency, diversity and fidelity of SDS-based methods in real-world and synthetic scenarios.
- *Diffusion time-step curriculum* is a plug-and-play training principle that could further unleash the potential of SDS-based teacher-student models.
- We systematically and theoretically validate the proposed diffusion time-step curriculum.

2. Related Work

Text-3D Generation focuses on generating 3D assets from a given text prompt. The core mechanism of such approaches is the score distillation sampling (SDS) proposed by [30], where the diffusion priors are used as score functions to supervise the optimization of a 3D representation. Recent advancements aim to enhance the training stability and generation fidelity via advanced shape guidance [19, 39, 62], disentangled 3D representation [3, 52, 57] and loss design [15, 49, 54, 64]. Note that some concurrent works [13, 44, 64] also leverage annealed time-step schedule for efficient training, but they fail to combine such time-step sampling schedule with the teacher-student knowledge transfer, *i.e.*, thus have not fully unleashed the potential of their symbiotic cycle.

Image-3D Generation focuses on generating 3D assets from a given reference image, which can also be considered as an ill-posed single-view reconstruction problem [4]. The above text-to-3D methods can be directly adapted for image-to-3D generation with image captioning, *e.g.*, Realfusion [26] and NeRDi [4] directly adapted SDS [30] into Image-to-3D with textual inversion of the given image. Magic123 [31] and Consistent123 [50] combine Zero-1-to-3 [22] with RealFusion [26] to further improve the generation quality. One-2-3-45 [22] and Dreamgaussian [44]

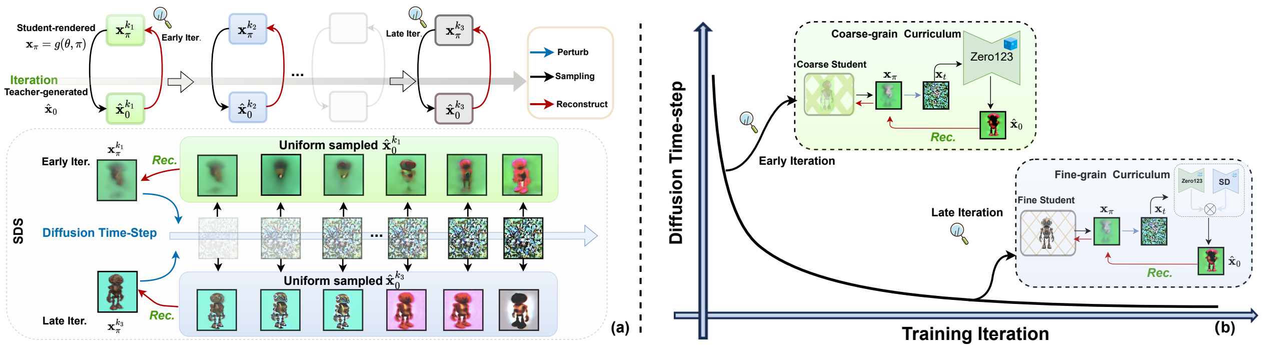

Figure 2. **(a)** SDS embraces an symbiotic teacher-student cycle with the training iteration progresses (*Top*). However, it entangles coarse-grained and fine-grained modeling with uniform sampling of time steps (*Bottom*) and equal treatment of student and teacher, where $k_1 \dots k_3$ denotes the training iteration from early to late. **(b)** Our DTC123 follows the *diffusion time-step curriculum*, where larger time steps capture coarse-grained concept and smaller time steps focus on fine-grained details.

adopted NeuS [47, 48] and Generative Gaussian Splatting [16] as the 3D representation respectively, significantly reducing the generation time at the expense of the generation quality. We build upon these optimization-based approaches by implementing a coarse-to-fine optimization strategy that explores the *diffusion time-step curriculum* they overlooked, which enables efficient generation of high-fidelity, multi-view consistent 3D assets from one image.

3. Preliminary

Given a single image of an object, our goal is to optimize a coherent 3D model (*e.g.*, NeRF, mesh) so that it can restore the given image from the reference view π_{ref} and generate a highly plausible image from any unseen view π with the supervision of the diffusion model. Here, we mainly introduce Score Distillation Sampling (SDS) [30] with *student* 3D models, *teacher* diffusion models (see *Appendix*) that will help to build up our approach.

3.1. Score Distillation Sampling (SDS)

SDS [30] distills 2D priors from a pre-trained conditional diffusion *teacher* model $\boldsymbol{\epsilon}_\phi(\cdot)$ into differentiable 3D *student* representations θ. In particular, given a certain camera parameter π, we **randomly** select a diffusion time-step t and perturb the student-rendered $\mathbf{x}_\pi \xrightarrow{+\sigma_t\boldsymbol{\epsilon}} \mathbf{x}_t$ by adding a Gaussian noise $\boldsymbol{\epsilon}$, and reformulate SDS from the perspective of reconstruction by calculating:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{SDS}}(\boldsymbol{\theta}, t) =& \mathbb{E}_{\boldsymbol{\epsilon}}\left[\omega(t)\left(\boldsymbol{\epsilon}_\phi\left(\mathbf{x}_t; t, \boldsymbol{y}\right) - \boldsymbol{\epsilon}\right)\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}\right] \\ =& \mathbb{E}_{\boldsymbol{\epsilon}}\left[\bar{\omega}(t)\left(\mathbf{x}_\pi - \hat{\mathbf{x}}_0\right)\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}\right], \end{aligned} \tag{1}$$

where $\hat{\mathbf{x}}_0 = \mathbf{x}_t - \sigma_t\boldsymbol{\epsilon}_\phi\left(\mathbf{x}_t; \boldsymbol{y}, t\right)$, can be considered as single-step de-nosing output with starting point $\mathbf{x}_t$; $\boldsymbol{y}$ is the condition (*e.g.*, text, camera pose) and depends on the types of teacher diffusion models; $\bar{\omega}(t) = \omega(t)/\sigma_t$, denotes the weight function. Thus, we reveal that the crux of this teacher-student optimization process is directly determined by such perturbed-and-denoised output, *i.e.*, the quality of the teacher-generated $\hat{\mathbf{x}}_0$. As illustrated in Figure 2(a), intuitively, not all time-steps t can provide useful and valid guidance $\hat{\mathbf{x}}_0$, which motivates the following proposed diffusion time-step curriculum.

4. DTC123

In this section, we introduce our *diffusion time-step curriculum* one-image-to-3D pipeline, called "DTC123", with a theoretical justification in *Appendix*. As shown in Figure 3(a): We start by taking a reference image and extracting its geometric estimation and text description. Then the optimization procedure can be categorized into unseen view guidance and reference restoration:

- For the **unseen view**, we employ our *diffusion time-step curriculum*, where larger time-steps capture coarse-grained concepts, and smaller time-steps focus on fine-grained details. Specifically, we implement such a curriculum with an annealed time-step sampling schedule (Sec. 4.1), progressive *student* representation (Sec. 4.2), and coarse-to-fine *teacher* guidance (Sec. 4.3).
- For the **reference view**, we basically penalize the 3D model to align with the given image, employing traditional reconstruction constraints.

Moreover, DTC123 incorporates several techniques (Sec. 4.4) to enhance generation efficiency, geometric robustness, and alleviate the Janus Face problem.

4.1. Annealed Time-Step Schedule

We develop a time-step annealed sampling schedule for our time-step curriculum, thereby facilitating teacher-student knowledge transfer in a progressive, coarse-to-fine manner.

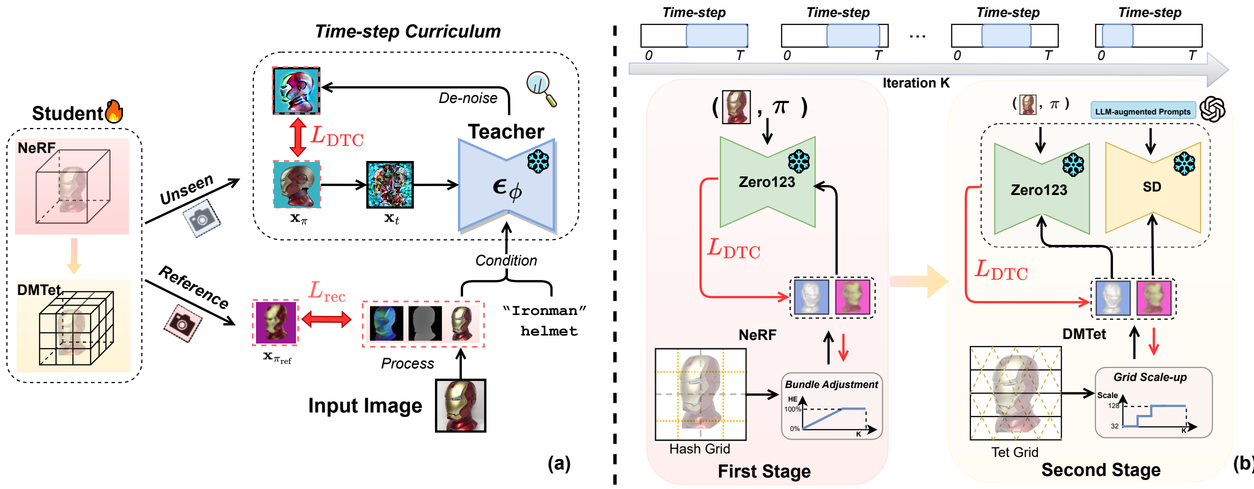

Figure 3. **(a)** Overall pipeline of DTC123, which have two optimization stages and includes the reference view reconstruction and unseen view imagination. **(b)** The zoom-in diagram of unseen view imagination with the proposed *diffusion time-step curriculum*.

Please refers to *Appendix* for its theoretical justification. In particular, we use an annealed interval $[t_{\text{mid}}-\Delta, t_{\text{mid}}+\Delta]$ to randomly select the time-step t, with the interval midpoint t_{mid} decreasing monotonically. The interval radius Δ narrows progressively during optimization, and the midpoint t_{mid} is determined by the current iteration k:

$$t_{\text{mid}} := t_{\max} - (t_{\max} - t_{\min}) \cdot \log\left(1 + \frac{\lfloor k/l \rfloor \cdot l}{N}\right), \quad (2)$$

where $t_{\max}$ and $t_{\min}$ refers to the entire diffusion time-step range; l and N denotes the step length and total training iteration, respectively. This schedule not only introduces *local randomness* to preserve the vibrancy of the model's coloration but also allocates more iterations to smaller time-steps for in-depth detail exploration.

4.2. Student: Progressive 3D Representation

The student 3D model adheres to the *diffusion time-step curriculum* by initially representing coarse-grained features at larger time-steps and subsequently fine-grained details at smaller time-steps. Due to computational memory constraints [20, 31], we leverage NeRF [29] for low-resolution scene modeling in the *first* stage, and then adopt DMTet [38] for high-resolution mesh fine-tuning in the *second* stage. Recall that grid-based 3D models embrace the inherent multi-resolution representation [41, 48, 60], where the lower-level spatial grids (*e.g.*, hash grid, tetrahedral grid) store the general contours while the higher-level counterparts store the finer textures and scene illumination. Accordingly, we employ progressive resolution constraints in **both** stages, which allows the student to initially assimilate the correct geometry structure information with a smooth coarse-grained signal and later shifts to learn a high-fidelity scene representation. More concretely:

Progressive bundle for NeRF. The design of hash encoding captures both low and high-resolution features, which might inadequately represent basic geometry at large time-step as inaccurate and diverse finer details could overshadow essential structural information. To mitigate this issue, given the hash-grid encoding (HE), we gradually regularize the visible resolution bands by applying a dynamic soft feature mask:

$$m_i(k, N, L) = \begin{cases} 1, & \text{if } i \leq 4 + \min\left(\left\lfloor \frac{10k}{N} \right\rfloor, L-4\right) \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where i denotes the HE feature level from coarsest to finest and L denotes the total number of feature levels.

Progressive Tetrahedral Grid for DMTet. As shown in Figure 3(b), we initially convert [43] the neural density field to a signed distance function (SDF) field and re-use the neural color field for texture representation, progressing to higher resolutions *e.g.*, $64 \times 64 \rightarrow 512 \times 512$ for detailed surface rendering. To circumvent issues like mesh distortion or topology errors from abrupt resolution changes, we employ a graded approach, scaling up the tetrahedral grid size, *i.e.* $32 \rightarrow 64 \rightarrow 128$, and the rasterizing resolution progressively with the reducing time-step. This strategy ensures a seamless transition from initialized structural capture to rendering sophisticated surface nuances.

4.3. Teacher: Coarse-to-fine Diffusion Prior

The teacher diffusion model should follow the *diffusion time-step curriculum* by initially offering a rough silhouette of the object in the desired pose in larger time steps, subsequently prioritizing intricate texture refinement in small time steps. A natural question follows : *What is the suitable teacher for the time-step curriculum?* Empirical evidence[1] suggests: the view-conditioned Zero-1-to-3 serves as a coarse-grained teacher by providing a more accurate geometry structure at large time-step. Conversely, the Stable Diffusion is suitable for a fine-grained teacher as it yields realistic texture details at smaller time-step.

Such evaluation drives our search for the following utilization of teacher guidance that makes the best use of both diffusion priors [58]. As illustrated in Figure 3(b), in the first stage, we use Zero-1-to-3 solely for efficient coarse-grained contour and structural guidance. Subsequently, in the second stage, a collaborative diffusion guidance with dynamic prior reweighting is leveraged, wherein Zero-1-to-3 guides SDF field geometry optimization supervised by textureless shading; Stable Diffusion aids in the optimization of color field respectively at smaller time-steps. For clarity, we denote the Zero-1-to-3 guidance as $\mathcal{L}^{\text{geo}}_{\text{SDS}}(\boldsymbol{\theta}, t)$ and the Stable Diffusion guidance as $\mathcal{L}^{\text{tex}}_{\text{SDS}}(\boldsymbol{\theta}, t)$ [2]. The diffusion objective for the time-step curriculum can then be formulated as:

$$\nabla_\theta \mathcal{L}_{\text{DTC}}(\theta, t) = \nabla_\theta \mathcal{L}^{\text{geo}}_{\text{SDS}}(\boldsymbol{\theta}, t) + \nabla_\theta \lambda \mathcal{L}^{\text{tex}}_{\text{SDS}}(\boldsymbol{\theta}, t), \quad (4)$$

where the time-step t follows the sampling schedule in Sec. 4.1; λ is the trade-off hyper-parameter, set as 0 in the first stage and gradually increase with the reducing time-step in the second stage. We apply DDIM sampling [6, 40] for multi-step de-nosing process, *i.e.*, $\mathbf{x}_t \to \mathbf{x}_{t-r} \ldots \to \hat{\mathbf{x}}_0$ instead of the imprecise single-step de-noising in the last few iterations for both efficiency and performance concerns. Empirically, it equips the teacher-generated $\hat{\mathbf{x}}_0$ with more decent details and mitigates issues like texture flickering or overly saturated color blocks at smaller time-steps. In addition, we propose the following approaches to further improve the fine-grained supervision quality in scenarios with sophisticated textures by clarifying the ambiguous inferred text description and alleviating the Janus face problem.

LLM-augmented Prompts. Previous methods [20, 27, 45] directly utilize ambiguous view-dependent prompts within {`front`, `back`, `side`} as the condition for Stable Diffusion. Paralleled to [46], we notice that more specific language prompts force the image distribution of Stable Diffusion to be narrower and more beneficial to the mode-seeking SDS algorithm. To this end, we instead leverage instruction-tuned large language models (LLMs) [33] with carefully designed task prompts to specify unseen view descriptions from the original text description. The core objective of LLM-augmented prompts is to intricately enhance multi-view descriptions by integrating additional details, *e.g.*, *'**The back of** Ironman helmet, **with metallic sheen**'*, while meticulously avoiding description conflict among different views. Detailed instruct-LLM design is in *Appendix*.

Camera Pose De-biasing. When the given object is human / animal-like, the intrinsic pre-trained bias [1, 12, 22] of diffusion models always causes the Janus face problems. Empirical observations suggest that Zero-1-to-3 tends to generate a Janus face on the backside, tends to 'copy' the conditioned front face to the back view with symmetric contour, whereas Stable Diffusion is more prone to this issue on the sides. To mitigate these anomalies in the second stage, we dynamically employ gradient clipping and randomized dropout for $\nabla_\theta \mathcal{L}^{c}_{\text{SDS}}(\boldsymbol{\theta}, t)$ within the azimuth range $\left[\frac{11\pi}{12}, \frac{13\pi}{12}\right]$ and $\nabla_\theta \mathcal{L}^{f}_{\text{SDS}}(\boldsymbol{\theta}, t)$ within the azimuth range $\pm\left[\frac{\pi}{6}, \frac{\pi}{4}\right]$, respectively.

4.4. Advanced Regularization Techniques

Due to the ambiguity and inconsistency guidance of diffusion-based 3D generation [8, 44], high-frequency artifacts often appear on the crisped surface of the student renderings. To counteract this, following [43, 63], we basically improve the smoothness of the normal map by calculating the normal vector using finite depth differences and incorporate additional continuous connection from the input coordinate for hash-based NeRF regularization. As for DMTet and mesh exportation, we apply Laplacian-based regularization [59] to achieve mesh smoothness, capitalizing on the uniform Laplacian matrix derived from the mesh vertices and adjacent faces, as well as removal and calibration of unreferenced vertices and faces. We denote the above smoothness regularization as $\mathcal{L}_{\text{reg}}$ in both stages.

4.5. Training Objective

Our final objective then comprises three key terms:

$$\nabla_\theta \mathcal{L} = \nabla \mathcal{L}_{\text{DTC}} + \lambda_{\text{reg}} \cdot \nabla \mathcal{L}_{\text{reg}} + \lambda_{\text{rec}} \cdot \nabla \mathcal{L}_{\text{rec}}\,, \quad (5)$$

where $\mathcal{L}_{\text{DTC}}$ represents the time-step curriculum diffusion prior objective for unseen views; $\mathcal{L}_{\text{rec}}$ denotes the traditional reconstruction objective for reference views, which aligns the given image in depth space with Pearson correlation [37] and the RGB, mask space through mean squared error (MSE) [25]; $\mathcal{L}_{\text{reg}}$ is the regularization term ensuring the geometrical smoothness. By synergistically integrating these objectives, our DTC123 pipeline demonstrably achieves a high degree of geometric robustness and superior texture quality.

[1]To answer this question, we quantitatively analyze two most popular teacher diffusion models by investigating the contour exploration consistency via MaskIoU and the perceptual generation quality computed by CLIP-R [21] in the OmniObject3D [53] dataset. See detailed experimental results in *Appendix*.

[2]Note that the weight function $\bar{\omega}(t)$ is calibrated relative to the spherical distance from the reference view, as 3D models tend to reconstruct more accurately near the reference view, requiring less teacher-guided imagination.

5. Experiments

5.1. Implementation Details

Pipeline Settings. We consistently applied the *same* set of hyper-parameters across all experiments. DTC123 was implemented in PyTorch with a single NVIDIA A100 GPU. We trained both the first and second stages for 3,000 iterations using the Adan [55] optimizer with 1e-3 learning rate and 2e-5 weight decay, which cost approximately *20-25* minutes for the entire pipeline. In the first stage, we adopted Zero123-XL [22] (cfg=5) as the only teacher model to supervise an Instant-NGP [29] with three MLP layers and hash-encoder. Then Stable Diffusion v2.0 [35](cfg=25) was integrated with zero123-XL, jointly enhancing the geometric robustness and texture refinement of a DMTet, initialized from the prior stage. The reference view was sampled with a 25% probability, and other views at 75%, in both stages. Final rendering resolutions were set at 64×64 for the first stage and 512×512 for the second stage, respectively.

Image Pre-processing Details. Given an arbitrary reference image, DTC123 pipeline systematically processed it for the follow-up 3D generation. The first step employed the state-of-the-art segmentation model, SAM [17], to meticulously distinguish foreground objects from their background context. Subsequently, the Dense Prediction Transformer [32] was harnessed to estimate both depth and normal maps, ensuring that rich geometric information has been captured. Then, BLIP2 [18] crafted a descriptive caption on the segmented object for text-conditioned guidance. The culmination of this workflow is a trio of outputs: a sharply segmented image, its corresponding depth and normal map, and its semantic description. Note that unlike [31], [26], textual inversion with image-level augmentations, which is time consuming (more than an hour), was not leveraged in ours.

Figure 4. Multi-instance generation by customized prompts.

5.2. Experimental Protocol

An exceptional 3D model should not only mirror the reference image but also maintain a consistent correlation with the reference and plausible results when observed from other poses. Following [31], we compared our DTC123 in RealFusion15 [26], NeRF4 benchmark for quantitative comparison, with PSNR and LPIPS metrics to measure reconstruction quality and CLIP-similarity to evaluate appearance similarity. For qualitative results, we manually collected 50 reference images from the Internet out of the range of Objaverse, covering a wider range of difficult items.

We adopted Zero123-SDS [22], RealFusion [26], NeuralLift [56] and Magic123 [31] as baseline SDS-based methods with their default experimental settings. We also compared with the state-of-the-art methods One-2-3-45 [21] and SyncDreamer [24], which perform Image-3D generation in a feed-forward manner instead of SDS optimization. For Zero123-SDS and Magic123, we adopted the implementation from threestudio [10] and leveraged Zero123-XL as ours for fair comparison, while others from their official codebase. We are quite confident that the baselines presented here are the finest re-implementations we have come across.

5.3. Image-3D Generation

Qualitative Results. Figure 5 demonstrates that DTC123 maintains high fidelity and plausible generation in complex scenarios. In contrast, generation results from most baselines, even those utilizing the more advanced Zero123-XL, are plagued by multi-view inconsistency, geometric distortion, and texture conflict. For example, in the yellow robot case in the *fourth* line of Figure 5, competing textures cause blurring and exhibit unreasonable features, such as an extra eye on the back. In comparison, DTC123 generates high-fidelity novel views with realistic metallic textures.

Quantitative Results. Table 1 shows that DTC123 consistently outperforms other methods across all metrics, demonstrating its superior reconstruction (PSNR, LPIPS) and 3D consistency (CLIP-similarity) capabilities. Specifically, in the reference view reconstruction, DTC123 is on par with Magic123 and significantly exceeds RealFusion and NeuralLift. In terms of view consistency, as indicated by CLIP-similarity, DTC123 exceeds Magic123 by a large margin. The primary inconsistency in Magic123 stems from the boundary disparities between occluded and non-visible regions, resulting in pronounced seams. More results on GSO dataset is included in *Appendix*.

5.4. Multi-instance Generation

While other Image-to-3D approaches [22, 31] can only generate plausible instances with limited diversity by different random seeds for initialization, DTC123 facilitates amazing multi-instance generation with different localized details by

Figure 5. **Qualitative comparisons on image-to-3D generation**. We randomly sample several new views to present, while other views and methods are included in *Appendix*. Our DTC123 consistently outperforms other state-of-the-art methods by generating multi-view consistent and high-fidelity results.

Table 1. **Quantitative results.** We show quantitative results in terms of CLIP-Similarity↑ / PSNR↑ / LPIPS↓. The results are shown on the NeRF4 and RealFusion datasets.

Dataset	Metrics\Methods	NeuralLift [56]	RealFusion [26]	Magic123 [31]	**DTC123**
NeRF4	CLIP-Similarity↑	0.52	0.38	0.80	0.84
	PSNR↑	12.55	15.37	24.62	25.14
	LPIPS↓	0.50	0.20	0.03	0.02
RF15	CLIP-Similarity↑	0.65	0.67	0.82	0.87
	PSNR↑	11.08	0.67	19.50	21.42
	LPIPS↓	0.53	0.14	0.10	0.08

specific user prompt. For example, when the reference image presents the side view of a yellow car, users can specify their desired object, such as a 'BMW' or a 'Porsche Panamera'. DTC123 proficiently generates realistic and coherent 3D contents based on the given instructions, thanks to the appropriate time-step curriculum, where the student model captures the car's general shape initially, followed by brand-specific details. Please check Figure 4 for additional results, which vividly demonstrates the potential of DTC123 for user-guided controllable generation and 3D editing.

6. Ablation and Discussion

Q1: ***What impacts performance of DTC123 in terms of component-wise contributions?*** We discarded each core component of DTC123 to validate its component-wise effectiveness. The results are depicted in figure 6.

A1: In a multi-component pipeline, we observed that the exclusion of any component from DTC123 resulted in a significant degradation in performance. In particular, when the annealed sampling schedule is replaced with a random sampling schedule, it leads to geometric irregularities and excessive texture details. The similar situation occurs if

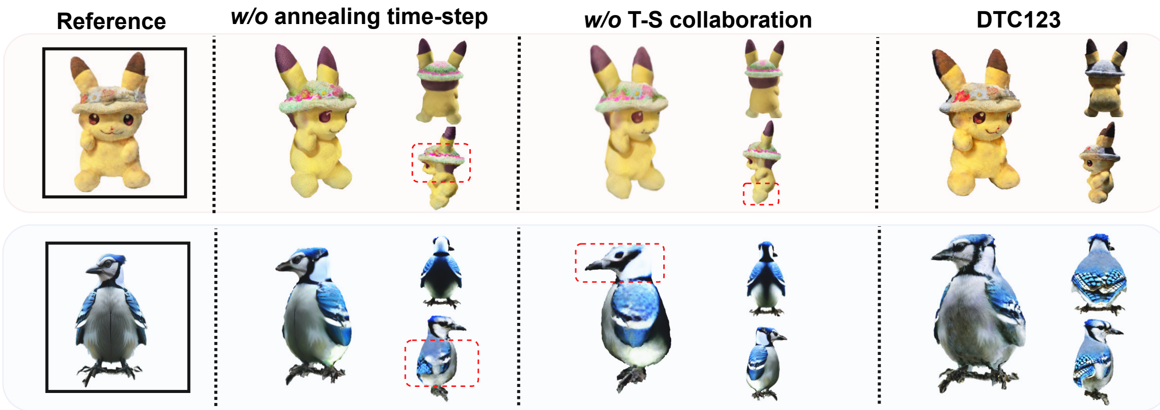

Figure 6. Ablation study on the component-wise contribution of DTC123. **T-S** denotes the **T**eacher-**S**tudent collaboration.

teacher and student models don't collaborate with the annealed time-step. In contrast, when both teacher and student follow the time-step curriculum, it effectively generates results with sophisticated texture and illumination while ensuring geometric stability.

Q2: ***How about the robustness of DTC123?*** To better diagnose the robustness of DTC123, we meticulously analyzed and quantified the occurrences of generation failure (*e.g*., Janus face, geometry distortion, extremely abnormal coloration) on Level50 with different random seeds and under different difficulty levels.

A2: As depicted in Figure 7, DTC123 consistently exhibits a lower failure rate compared to other methods across various initialization (random seeds) and at different difficulty levels by a large margin. Such robustness should be attributed to the proposed time-step curriculum, where the student 3D model initially captures coarse features, significantly reducing the instability typically associated with randomized NeRF initialization.

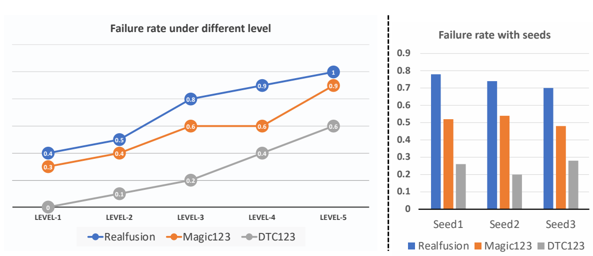

Figure 7. Failure analysis with different initialization and levels.

Q3: ***What's the difference between diffusion time-step curriculum and annealing time-step?***

A3: We clarify that annealing time-step is **NOT** equivalent to, but a part of time-step curriculum. Our DTC includes three coherent parts : time-step schedule, progressive student representation and teacher guidance. Different from the viewpoint of DDPM sampling from Dreamtime [13], we consider that SDS leverages teacher-generated $\hat{\mathbf{x}}_0$, which is a single-step de-nosing output with starting point $\mathbf{x}_t$, to optimize student-rendered $\mathbf{x}_\pi$. Modeling 3D generation as a data corruption reduction process, we reveal that there exists a time-step lower bound in SDS for *teacher* diffusion model to well estimate the desired score function and thus provide quality guidance $\hat{\mathbf{x}}_0$. In order to alleviate the marked divergence of the teacher-generated $\hat{\mathbf{x}}_0$ [11] at large time-steps, we further design the progressive student representation and teacher guidance to cooperate with the annealing time-step for a more stable coarse-to-fine generation. Please refers to *Appendix* for more details.

7. Conclusion

We revisit Score Distillation Sampling (SDS) and point out that the crux of its enhancement lies in the proposed diffusion time-step curriculum. We then design an improved coarse-to-fine Image-3D pipeline (DTC123) which collaborates the teacher diffusion model and 3D student model with the time-step curriculum. Through qualitative comparisons and quantitative evaluations, we show that our DTC123 significantly improves the photo-realism and multi-view consistency of Image-to-3D generation. In the future, we will focus on exploring the potential of diffusion time-step curriculum with advanced teacher models and diverse student models to further improve the efficiency and quality.

Acknowledgments

DTC123 is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-022) and the Agency for Science, Technology AND Research. Pan Zhou is supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

References

[1] Mohammadreza Armandpour, Huangjie Zheng, Ali Sadeghian, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*, 2023. 5

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 1

[3] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 2

[4] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20637–20647, 2023. 2

[5] 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. 2

[6] Prasad Gabbur. Improved ddim sampling with moment matching gaussian mixtures. *arXiv preprint arXiv:2311.04938*, 2023. 5

[7] Kyle Gao, Yina Gao, Hongjie He, Dening Lu, Linlin Xu, and Jonathan Li. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379*, 2022. 2

[8] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 35:14715–14728, 2022. 5

[9] Faiza Gul, Wan Rahiman, and Syed Sahal Nazli Alhady. A comprehensive study for robot navigation techniques. *Cogent Engineering*, 6(1):1632046, 2019. 1

[10] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. `https://github.com/threestudio-project/threestudio`, 2023. 6

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 8

[12] Susung Hong, Donghoon Ahn, and Seungryong Kim. Debiasing scores and prompts of 2d diffusion for robust text-to-3d generation. *arXiv preprint arXiv:2303.15413*, 2023. 2, 5

[13] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023. 2, 8

[14] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018. 1

[15] Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. Noise-free score distillation. *arXiv preprint arXiv:2310.17590*, 2023. 2

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 3

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 6

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 6

[19] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arXiv preprint arXiv:2310.02596*, 2023. 2

[20] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 4, 5

[21] 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. 5, 6

[22] 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. 2, 5, 6

[23] Shichen Liu, Weikai Chen, Tianye Li, and Hao Li. Soft rasterizer: Differentiable rendering for unsupervised single-view mesh reconstruction. *arXiv preprint arXiv:1901.05567*, 2019. 1

[24] 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. 6

[25] Hans Marmolin. Subjective mse measures. *IEEE transactions on systems, man, and cybernetics*, 16(3):486–489, 1986. 5

[26] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2023. 2, 6, 7

[27] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673, 2023. 2, 5

[28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[29] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 1, 4, 6

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

[31] 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. 2, 4, 6, 7

[32] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 6

[33] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 2023. 5

[34] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12216–12225, 2021. 2

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 6

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[37] Philip Sedgwick. Pearson's correlation coefficient. *Bmj*, 345, 2012. 5

[38] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021. 4

[39] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[41] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 4

[42] Tianyu Sun, Guodong Zhang, Wenming Yang, Jing-Hao Xue, and Guijin Wang. Trosd: A new rgb-d dataset for transparent and reflective object segmentation in practice. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1

[43] Jiaxiang Tang. Stable-dreamfusion: Text-to-3d with stable-diffusion, 2022. https://github.com/ashawkey/stable-dreamfusion. 4, 5

[44] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2, 5

[45] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023. 5

[46] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 5

[47] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 3

[48] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3295–3306, 2023. 3, 4

[49] 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. *arXiv preprint arXiv:2305.16213*, 2023. 2

[50] Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, CL Chen, and Lei Zhang. Consistent123: Improve consistency for one image to 3d object synthesis. *arXiv preprint arXiv:2310.08092*, 2023. 2

[51] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020. 2

[52] Jinbo Wu, Xiaobo Gao, Xing Liu, Zhengyang Shen, Chen Zhao, Haocheng Feng, Jingtuo Liu, and Errui Ding. Hd-fusion: Detailed text-to-3d generation leveraging multiple noise estimation. *arXiv preprint arXiv:2307.16183*, 2023. 2

[53] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 5

[54] Zike Wu, Pan Zhou, Xuanyu Yi, Xiaoding Yuan, and Hanwang Zhang. Consistent3d: Towards consistent high-fidelity text-to-3d generation with deterministic sampling prior. *arXiv preprint arXiv:2401.09050*, 2024. 2

[55] Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *arXiv preprint arXiv:2208.06677*, 2022. 6

[56] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4479–4489, 2023. 6, 7

[57] Xudong Xu, Zhaoyang Lyu, Xingang Pan, and Bo Dai. Matlaber: Material-aware text-to-3d via latent brdf auto-encoder. *arXiv preprint arXiv:2308.09278*, 2023. 2

[58] Xuanyu Yi, Jiajun Deng, Qianru Sun, Xian-Sheng Hua, Joo-Hwee Lim, and Hanwang Zhang. Invariant training 2d-3d joint hard samples for few-shot point cloud recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14463–14474, 2023. 5

[59] Ming Yin, Junbin Gao, and Zhouchen Lin. Laplacian regularized low-rank representation and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):504–517, 2015. 5

[60] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 4

[61] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023. 2

[62] Minda Zhao, Chaoyi Zhao, Xinyue Liang, Lincheng Li, Zeng Zhao, Zhipeng Hu, Changjie Fan, and Xin Yu. Efficientdreamer: High-fidelity and robust 3d creation via orthogonal-view diffusion prior. *arXiv preprint arXiv:2308.13223*, 2023. 2

[63] Hao Zhu, Fengyi Liu, Qi Zhang, Xun Cao, and Zhan Ma. Rhino: Regularizing the hash-based implicit neural representation. *arXiv preprint arXiv:2309.12642*, 2023. 5

[64] Joseph Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance. *arXiv preprint arXiv:2305.18766*, 2023. 2