IT3D: Improved Text-to-3D Generation with Explicit View Synthesis

Yiwen Chen[*1,2], **Chi Zhang**[*3], **Xiaofeng Yang**[2],
Zhongang Cai[4], **Gang Yu**[3], **Lei Yang**[4], **Guosheng Lin**[†1,2]

[1]S-Lab, Nanyang Technological University
[2]School of Computer Science and Engineering, Nanyang Technological University
[3]Tencent PCG, China
[4]SenseTime Research

Abstract

Recent strides in Text-to-3D techniques have been propelled by distilling knowledge from powerful large text-to-image diffusion models (LDMs). Nonetheless, existing Text-to-3D approaches often grapple with challenges such as over-saturation, inadequate detailing, and unrealistic outputs. This study presents a novel strategy that leverages explicitly synthesized multi-view images to address these issues. Our approach involves the utilization of image-to-image pipelines, empowered by LDMs, to generate posed high-quality images based on the renderings of coarse 3D models. Although the generated images mostly alleviate the aforementioned issues, challenges such as view inconsistency and significant content variance persist due to the inherent generative nature of large diffusion models, posing extensive difficulties in leveraging these images effectively. To overcome this hurdle, we advocate integrating a discriminator alongside a novel Diffusion-GAN dual training strategy to guide the training of 3D models. For the incorporated discriminator, the synthesized multi-view images are considered real data, while the renderings of the optimized 3D models function as fake data. We conduct a comprehensive set of experiments that demonstrate the effectiveness of our method over baseline approaches.

Introduction

In recent years, the field of text-to-image has witnessed remarkable progress, sparking a surge of interest within the research community to extend this advancement to the domain of 3D generation. This enthusiasm can largely be attributed to the emergence of methods capitalizing on pre-trained 2D text-to-image diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Song et al. 2021). By harnessing the immense volumes of training image data, these models acquire a comprehensive understanding of object appearance and geometry. Such advances have paved the way for text-to-3D generation without the explicit need for 3D datasets, showcasing an unparalleled combination of flexibility, diversity, and potential for cost and time savings.

A cornerstone in this domain is the innovative work of Dreamfusion (Poole et al. 2022). Their introduction of the Score Distillation Sampling (SDS) algorithm has been a game-changer, primarily due to its capability to generate diverse 3D objects using mere textual prompts (Poole et al. 2022). Despite its revolutionary approach, it comes with its set of challenges. A significant limitation is its control over the geometry and texture of the generated models, often leading to issues like over-saturation and the multi-face appearance of models (as depicted by coarse models in Fig. 1). Furthermore, there is an observed inefficacy in improving the model quality by merely amplifying the textual prompts (Poole et al. 2022; Lin et al. 2022; Wang et al. 2023).

To address these challenges, our research seeks to present an enhanced methodology for 3D generation. At its core, our method emphasizes the explicit synthesis of multi-view images of the intended 3D model, subsequently employing these images for 3D object reconstruction. This process begins by utilizing an existing text-to-3D generation model, such as DreamFusion (Poole et al. 2022), to craft a rudimentary depiction of the object. By rendering these foundational models, we gain an elemental representation of the object's geometry and spatial arrangement. Building upon these initial renderings, our method refines the view images using an image-to-image (I2I) generation pipeline (Zhang and Agrawala 2023; Brooks, Holynski, and Efros 2023; Rombach et al. 2022). By focusing on generating 2D images, a task that is simpler and more extensively studied, we can employ a range of techniques to enhance the image quality, resulting in the creation of more realistic and detailed images from different views of the 3D objects.

However, our approach is not without its unique set of challenges. The independent view generation can lead to multi-view inconsistencies, such as differences in textures and geometries. These inconsistencies pose a significant barrier to traditional 3D reconstruction methods and hinder them from producing satisfactory outcomes. To circumvent this obstacle, we introduce a discriminator paired with an adversarial loss to guide the learning of 3D models (Goodfellow et al. 2014; Chan et al. 2022). Here, we treat the enhanced multi-view images generated by I2I models as real data and the renderings of the optimized 3D models as fake data. By leveraging adversarial losses, our model effectively addresses these multi-view discrepancies, significantly enhancing the stability of training and improving the overall

[*]These authors contributed equally.
[†]Corresponding Authors.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

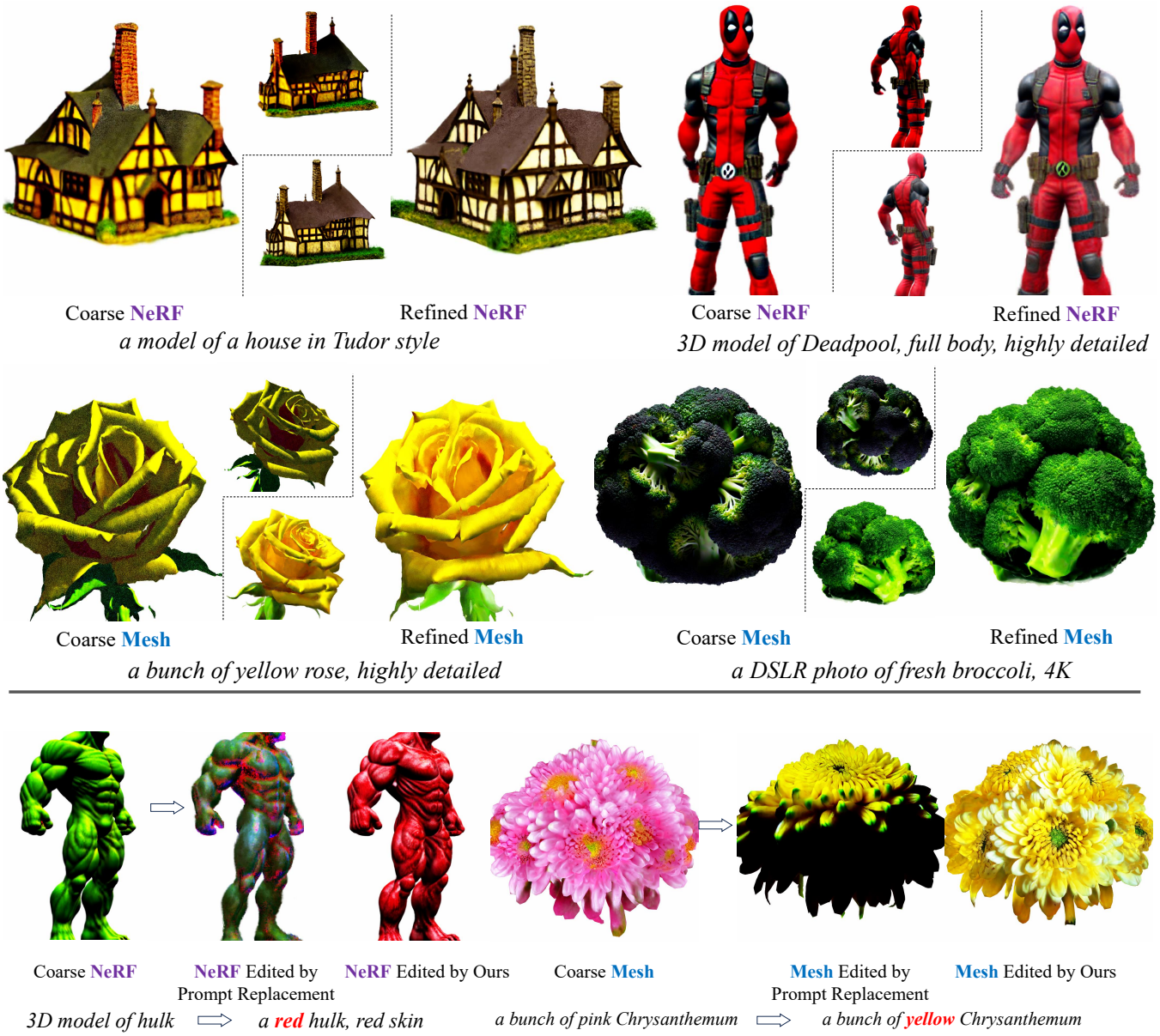


Figure 1: Results and applications of IT3D. Top: high-resolution text-to-3D Refinement. IT3D can refine the texture and geometry of 3D models generated by diffusion prior. Bottom: high-resolution prompt-based editing. Our method can edit 3D models given a prompt. Naively editing 3D models by prompt replacement would potentially cause an editing failure.

quality of the generated 3D objects.

Our method, termed IT3D, presents itself as a versatile plug-and-play solution, seamlessly integrating with any text-to-3D methodology that draws upon 2D diffusion priors. Our method is versatile, supporting a range of 3D output representations such as meshes and NeRFs. An additional advantage is its capacity to efficiently modify the appearance of 3D models using text, with examples illustrated in Fig. 1. Empirical evaluations further reveal that our approach can accelerate training convergence, which in turn, reduces the number of required training steps and results in comparable overall training time.

We provide comprehensive experimental results, both qualitative and quantitative, to validate the effectiveness of our methods. Our empirical findings show that our proposed method significantly improves the baseline models in terms of texture detail, geometry, and fidelity between text prompts and the resulting 3D objects.

Related Work

In this section, we review the literature related to text-to-3D generation. Our examination centers on two principal areas:

the emergence of diffusion models for image generation and the evolution of text-to-3D generation methods.

Text-to-Image Diffusion Models

The recent surge in generative model research has been prominently characterized by the rise of diffusion models. Incorporating textual prompts within the diffusion model has proven to be a critical step toward creating high-caliber text-to-image diffusion models. Renowned models in this domain such as GLIDE (Nichol et al. 2021), DALL·E 2 (Ramesh et al. 2021), Imagen (Saharia et al. 2022), and StableDiffusion (Rombach et al. 2022), leverage textual prompts as guiding parameters during the image generation process. As a result, these models have demonstrated significant competence in creating high-quality images that mirror the descriptions provided by the accompanying text.

Text-to-3D Generation

The groundbreaking work of DreamFusion (Poole et al. 2022) led the way by introducing the Score Distillation Sampling (SDS) method. This innovative technique optimizes 3D models using diffusion prior. Subsequent models built upon these pioneering developments, seeking to further enhance and refine the 3D generation process (Lin et al. 2022; Chen et al. 2023; Wang et al. 2023; Zhu and Zhuang 2023; Wang et al. 2022; Tang et al. 2023; Seo et al. 2023). For instance, Magic3D (Lin et al. 2022) sought to boost the resolution and learning speed of DreamFusion by utilizing a two-stage optimization strategy incorporating a sparse 3D hash grid structure. Similarly, DreamBooth3D integrated DreamBooth (Ruiz et al. 2023) and DreamFusion to facilitate the personalization of text-to-3D generative models using a few subject images. Additionally, ProlificDreamer (Wang et al. 2023) presented variational score distillation (VSD), a principled particle-based variational framework, to improve text-to-3D generation.

The work most closely related to ours is DreamBooth3D (Raj et al. 2023). Similar to our approach, DreamBooth3D involves generating datasets based on renderings of coarse models and enhancing them using image-to-image pipelines. However, there are distinct differences between our methods and DreamBooth3D in two main aspects. Firstly, DreamBooth3D primarily focuses on image-to-3D and requires a few subject images to fine-tune a DreamBooth model (Ruiz et al. 2023), resulting in a low-variance dataset for a single subject. In contrast, our method is geared towards enhancing text-to-3D generation. We generate high-variance datasets without the need for any tuning on large diffusion models. Secondly, our methods make more efficient use of the generated datasets by leveraging a 3D GAN. This approach proves superior to using an L2 loss with view and content-inconsistent datasets, as illustrated in Fig. 3.

Preliminary

Dreamfusion

The text-to-3D generation achieved by DreamFusion (Poole et al. 2022) is composed of two key components: a neural scene representation, referred to as the 3D model, and a pre-trained text-to-image diffusion-based generative model that provides diffusion prior.

The 3D model, denoted as a parametric function $x = g(\theta)$, can produce images x at specified camera poses. In this context, g refers to the chosen volumetric renderer, while θ stands for a coordinate-based MLP representing a 3D volume. The diffusion model ϕ is accompanied by a learned denoising function $\epsilon_\phi(x_t; y, t)$, which predicts sampled noise ϵ based on the noisy image x_t, noise level t, and text embedding y. This denoising function supplies the gradient direction to update θ, aiming to guide all rendered images towards high-probability density regions conditioned on the text embedding under the diffusion prior.

DreamFusion introduces Score Distillation Sampling (SDS) to compute the gradient as follows:

$$\nabla_\theta \mathcal{L}_{SDS}(\phi, g(\theta)) = \mathbf{E}_{t,\epsilon}\left[w(t)(\epsilon_\phi(x_t; y, t) - \epsilon)\frac{\partial x}{\partial \theta}\right]. \quad (1)$$

Here, $w(t)$ denotes a weighting function. The scene model g and diffusion model ϕ are modular elements within the framework, allowing for flexible choices.

Despite the success of DreamFusion, over saturation, inadequate detailing, and unrealistic outputs can often be observed in text-to-3D method based on diffusion prior.

Method

This study aims to address prevalent issues in prior text-to-3D methods, namely over-saturation, lack of details, and unrealistic outputs (Poole et al. 2022; Lin et al. 2022; Wang et al. 2023). While these challenges persist within the field of text-to-3D, substantial progress has been achieved in mitigating them by employing 2D image generation techniques driven by diffusion models. Building on this insight, we seek to tackle these concerns through the incorporation of 2D image Wgeneration techniques (Fig. 2). We will show that high-quality posed datasets that are free from the aforementioned issues can be generated at a low cost (Section 4.1).

However, the challenges of view inconsistency and content inconsistency persist due to the inherent generative nature of large diffusion models (Zhang and Agrawala 2023; Rombach et al. 2022). This inherent nature leads to considerable difficulties in effectively leveraging these generated images. To harness the potential of the generated dataset more effectively, we propose an innovative approach involving the integration of a discriminator (Chan et al. 2022). This discriminator serves to distill knowledge from the generated datasets, thereby guiding the training process of the 3D model (Section 4.2).

Despite the proficiency of 3D GANs in accommodating high variance datasets, the computational expense associated with training a 3D GAN to simulate realistic textures and intricate details renders it impractical for text-to-3D applications (Chan et al. 2022). To expedite the training process, we confine the size of the generated dataset to approximately two hundred samples. However, this diminutive dataset proves inadequate for training a discriminator that can solely fine-tune the 3D model, as the discriminator quickly overfits to such limited data (as depicted in

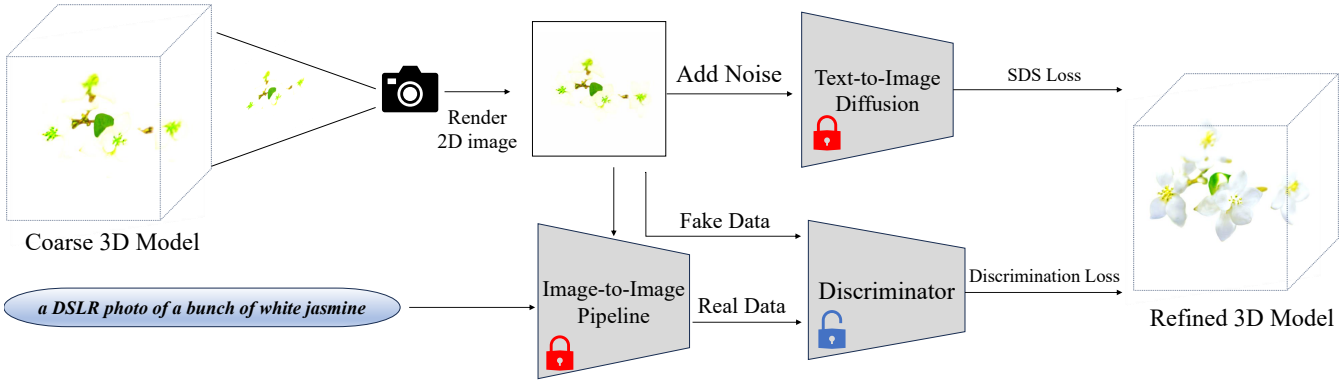


Figure 2: Pipeline of IT3D. Begin from a coarse 3D model, IT3D first generate a tiny posed dataset leveraging image-to-image pipeline conditioning on rendering of the coarse 3D model. Then incorporate a randomly initialized discriminator to distill knowledge form the generated dataset and update the 3D model with discrimination loss and SDS loss

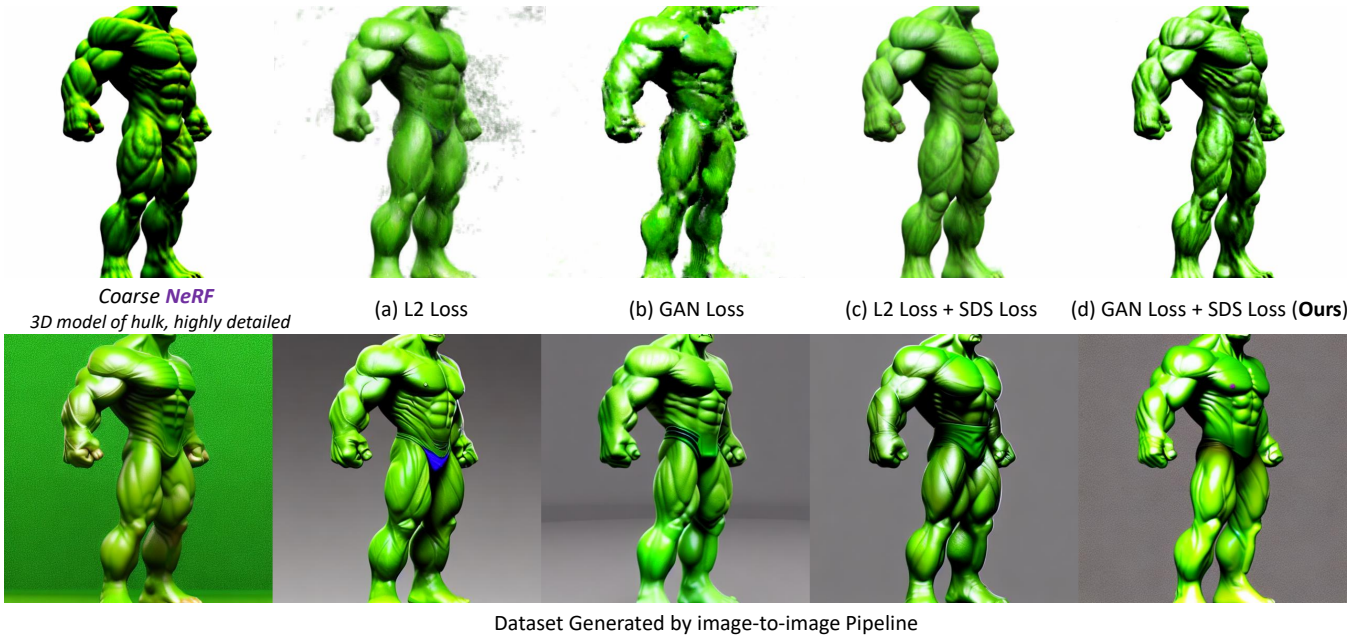


Figure 3: Top: Results for different loss setting. Bottom: Data samples generated through image-to-image pipeline. Prompt: a 3D model of hulk, highly detailed, full body. The generated dataset incorporate 60 camera views and 5 samples for each view. Given such a tiny but high variance dataset, (a) and (b) tend to result in noised refinement without the help of strong denoising ability of SDS loss. Comparing (c) and (d), (c) compromises to the average of generated dataset while (d) successfully bridge the distribution gap between coarse NeRF and generated dataset.

Fig. 3). To surmount this challenge, we introduce an innovative training strategy that employs a combination of diffusion prior and GAN loss (Section 4.3). This strategy results in a notable reduction in the training burden of the GAN component.

Dataset Generation

In this section, we describe our data generation strategy in detail. IT3D starts with a coarse 3D model parameterized by θ, derived using a text-to-3D baseline method that conditions on the prompt T. Firstly, we render the coarse 3D model across a range of camera poses represented by C, ultimately resulting in the creation of a dataset comprising posed images denoted as D. Subsequent to this, we employ image-to-image pipelines to generate a dataset of enhanced quality, termed D', which is conditioned on both the prompt T' and the original coarse dataset D. Users may either maintain T' as equivalent to T to refine the 3D model θ, or select a distinct T' to edit the 3D model θ. It is noteworthy that IT3D does not necessitate a well-trained θ. A preliminary 3D model exhibiting the foundational shape, albeit lacking extensive texture, is sufficient for providing depth maps or

normal maps as conditioning images for the image-to-image pipeline.

The collection of camera poses C envelops the 3D model in a uniform manner, with the same number of images being generated from the image-to-image pipeline for each viewpoint. This uniform distribution of poses is crucial for achieving an unbiased pose distribution within the dataset, a factor of high importance for effective GAN training. In our experimentation, the size of C ranges from 24 to 60, and 3 to 7 images are sampled for each viewpoint, contingent upon the complexity of the 3D model.

For the image-to-image pipeline, we opt for ControlNet (Zhang and Agrawala 2023) and Stable Diffusion (Rombach et al. 2022). We extract feature maps from images in the coarse datasets, encompassing depth maps, normal maps, and soft-edge maps. These extracted features serve as conditioning images for ControlNet, ensuring the generation of data that retains viewpoint consistency. In the context of the Stable Diffusion image-to-image pipeline (Rombach et al. 2022), random Gaussian noise is initially introduced to the rendered images. Subsequently, these noise-affected images are denoised with vanilla 2D diffusion denoising steps, ultimately yielding images of elevated quality.

Handling Inconsistencies with GAN

Although the image-to-image pipeline is conditioned on feature maps extracted from the original renderings during the generation process, it is inevitable that the generated dataset D' exhibits inconsistencies in terms of content and camera views. Naively applying a plain L2 loss would yield suboptimal results due to its inclination to align the 3D model θ with the average content of the highly variable dataset (refer to Fig. 3).

However, the capabilities of Generative Adversarial Networks (GANs) shine in scenarios involving datasets characterized by high variance (Chan et al. 2022). GANs have the ability to learn both geometry and texture-related knowledge from such datasets, subsequently guiding the 3D model to converge towards the same high-quality distribution exhibited by the generated dataset.

In our approach, we designate the 3D model as a 3D generator. We then integrate a discriminator initialized with random values. In this setup, the generated dataset D' is treated as real data, while the renderings of the 3D model θ represent fake data. The role of the discriminator involves learning the distribution discrepancy between the renderings and D', subsequently contributing to the discrimination loss which in turn updates the 3D model θ.

Diffusion-GAN Dual Training Strategy

Instead of relying solely on the discrimination loss, we propose an innovative Diffusion-GAN dual training strategy that combines the strengths of the diffusion prior and discrimination loss. This strategy utilizes the discrimination loss to guide the updating direction and leverages the diffusion prior to provide intricate geometry and texture details.

During the refinement phase, we fine-tune the coarse 3D model θ using both the discriminator loss and the diffusion prior. As the discriminator is randomly initialized, the discriminator loss introduces a considerable amount of noise along with meaningful content to the 3D model. However, due to the robust denoising capabilities inherent in large diffusion models, the diffusion prior effectively eliminates the added noise without erasing the valuable information acquired from the discriminator.

We found that this training strategy significantly reduce the training burden of the 3D GAN. Training a 3D GAN to achieve full convergence is highly time-consuming (Chan et al. 2022). However, the proposed training strategy accomplishes full convergence within a mere 0.5 to 1.0 GPU hours. This abbreviated refinement period, however, is insufficient for the discriminator to thoroughly comprehend the nuances of fine textures and geometry, as seen in result (b) in Fig. 3, which was trained for over 10.0 GPU hours. Nonetheless, the effectiveness of our training strategy is empirically validated.

We attribute this efficiency to the robust denoising capabilities and content dreaming capacity of the diffusion prior. Leveraging the diffusion prior empowers the generator to closely track the discriminator's updates at an accelerated pace. As a result, the discriminator can focus on learning higher-level distribution discrepancies without spending time on mastering minute details. Consequently, both the generator and discriminator update at significantly higher speeds compared to traditional vanilla 3D GANs.

It is important to note that the entire refinement process is iterative in nature. After enhancing the 3D model's quality, the same process of dataset generation and refinement can be iteratively applied to the improved 3D model, resulting in further quality improvements. Additionally, the incorporation of a Dataset Update (DU) strategy (Shao et al. 2023; Haque et al. 2023) can complement our training approach.

Experiments

Implementation Details

As our baseline method, we opt for the default configuration of text-to-3D in the Stable DreamFusion repository (Tang 2022). This baseline approach embraces classical SDS loss and the regularization losses outlined in DreamFusion (Poole et al. 2022). It is noteworthy that our IT3D method serves as a plug-and-play refinement technique and can also be seamlessly applied to any text-to-3D method based on 2D diffusion priors.

For our baseline method, complex prompts such as avatars typically require approximately 20k steps to achieve complete convergence, while simpler prompts like flowers converge around 15k steps. To ensure a fair comparison, we train the baseline method for 25K steps (1.5 to 2.5 GPU hours) to guarantee the full convergence of the 3D model. As for our method, we resume training from 5k to 15k steps of the baseline method, varying based on the complexity of the prompt. Following the resumption and dataset generation steps, we proceed to initialize a discriminator using the same architecture as EG3D, commencing our Diffusion-GAN dual training strategy. The learning rates for both the



Figure 4: More Results of IT3D. IT3D can refine the texture and geometry (body posture of Deadpool) of 3D models generated by text-to-3D method. IT3D can also edit 3D models given a prompt.

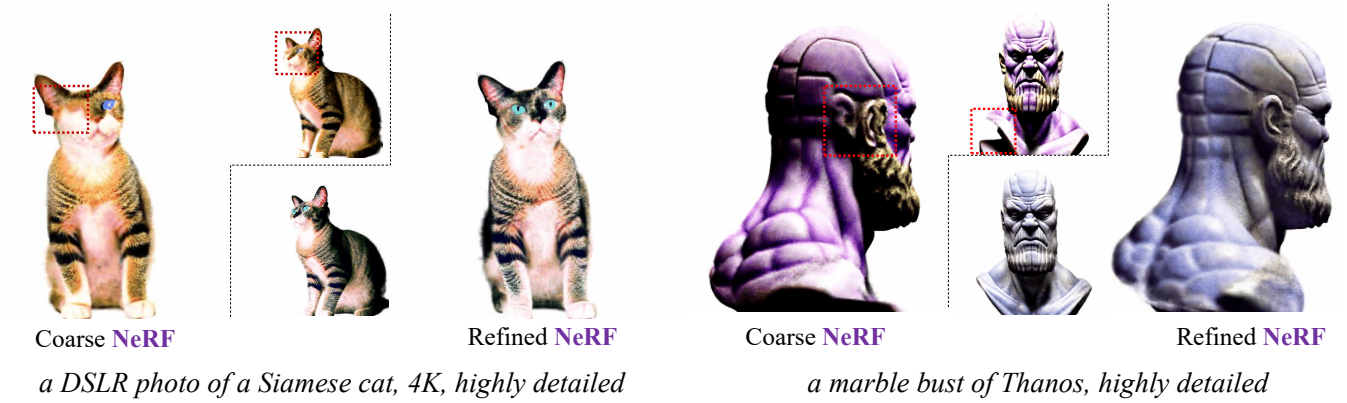


Figure 5: IT3D can recover NeRF with mild geometry or Janus problem. Current text-to-3D methods are potentially to generate failure cases like below examples, our method are capable of recovering failure cases with mild geometry or Janus problem.

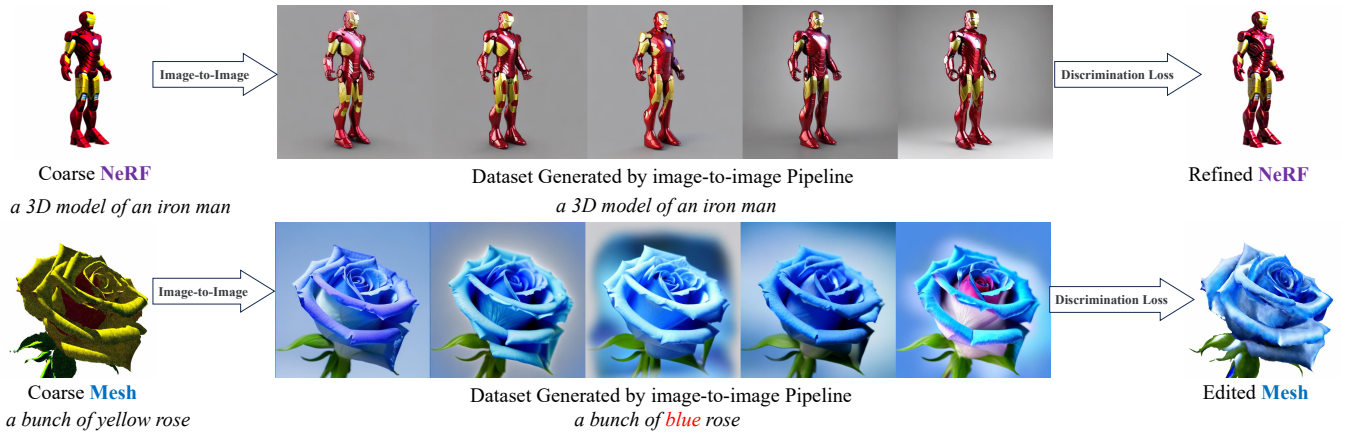


Figure 6: Examples for generated dataset. Our method can work on high-variance dataset and can tolerate failure cases of image-to-image pipeline (see texture of iron man dataset).

NeRF and the discriminator are set to 1e-3 and 2e-3, respectively.

3D Model Refinement and Editing

In this section, we present our results for 3D model refinement and editing, as depicted in Fig. 1 and Fig. 4.

We demonstrate that even with a low-quality coarse NeRF (Fig. 5) and a generated dataset that includes failure cases (Fig. 6), IT3D is capable of achieving a significant enhancement in quality.

In Fig. 5, IT3D demonstrates success in rectifying slight Janus problems and effectively corrects erroneous geometry, such as character body poses.

User Study

We conduct a user study with the baseline method over 15 prompts. The renderings utilized for this study align with the settings employed in the earlier experiments. The study involved 45 participants, resulting in 675 pairwise comparisons. On average, our methods were preferred by 89.92% in comparison to the baseline method.

Conclusion

In this paper, we propose a new plug-and-play refinement method for text-to-3D generation. Given a coarse 3D model, IT3D first generates posed image datasets by leveraging an image-to-image diffusion pipeline. It then refines the coarse 3D model with a novel Diffusion-GAN training strategy. To the best of our knowledge, our work is the first to combine GAN and diffusion prior to improve the text-to-3D task.

Limitations. The performance of our proposed method is limited by the performance of image-to-image pipelines. Current image-to-image pipelines fail to generate high-quality results when the input prompt is too complicated.

Acknowledgements

This study is supported under the RIE2020 Industry Alignment Fund–Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). This research is also supported by the MoE AcRF Tier 2 grant (MOE-T2EP20220-0007).

References

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16123–16133.

Chen, R.; Chen, Y.; Jiao, N.; and Jia, K. 2023. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. *arXiv preprint arXiv:2303.13873*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Haque, A.; Tancik, M.; Efros, A. A.; Holynski, A.; and Kanazawa, A. 2023. Instruct-nerf2nerf: Editing 3d scenes with instructions. *arXiv preprint arXiv:2303.12789*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2022. Magic3D: High-Resolution Text-to-3D Content Creation. *arXiv preprint arXiv:2211.10440*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Raj, A.; Kaza, S.; Poole, B.; Niemeyer, M.; Ruiz, N.; Mildenhall, B.; Zada, S.; Aberman, K.; Rubinstein, M.; Barron, J.; et al. 2023. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Seo, J.; Jang, W.; Kwak, M.-S.; Ko, J.; Kim, H.; Kim, J.; Kim, J.-H.; Lee, J.; and Kim, S. 2023. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*.

Shao, R.; Sun, J.; Peng, C.; Zheng, Z.; Zhou, B.; Zhang, H.; and Liu, Y. 2023. Control4D: Dynamic Portrait Editing by Learning 4D GAN from 2D Diffusion-based Editor. *arXiv preprint arXiv:2305.20082*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2256–2265. PMLR.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. arXiv:2011.13456.

Tang, J. 2022. Stable-dreamfusion: Text-to-3D with Stable-diffusion. Https://github.com/ashawkey/stable-dreamfusion.

Tang, J.; Wang, T.; Zhang, B.; Zhang, T.; Yi, R.; Ma, L.; and Chen, D. 2023. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*.

Wang, H.; Du, X.; Li, J.; Yeh, R. A.; and Shakhnarovich, G. 2022. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. *arXiv preprint arXiv:2212.00774*.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *arXiv preprint arXiv:2305.16213*.

Zhang, L.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Zhu, J.; and Zhuang, P. 2023. HiFA: High-fidelity Text-to-3D with Advanced Diffusion Guidance. *arXiv preprint arXiv:2305.18766*.