Progressive Rendering Distillation

Adapting Stable Diffusion for Instant Text-to-Mesh Generation without 3D Data
(Accepted by CVPR 2025)

Zhiyuan Ma
The Hong Kong Polytechnic University
March 31, 2025

About the Author

Zhiyuan Ma
PhD Candidate
Department of Computing
The Hong Kong Polytechnic University

Advised by:
Prof. Lei Zhang (PolyU)
Prof. Zhen Lei, Xiangyu Zhu (CASIA)

Contact & Links:
GitHub
Google Scholar
LinkedIn
Personal Website

Contents

What is Text-to-3D Generation?

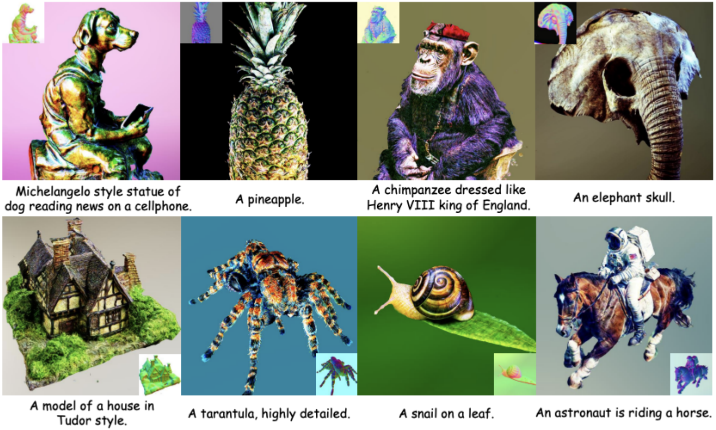
- **Task**: Generate 3D content from natural language descriptions
  - Input: Text prompt (e.g., "a red cartoon car")
  - Output: 3D model (ideally textured mesh)
- **Goal**: High-quality 3D meshes that accurately reflect text descriptions
- **Applications**: Gaming, AR/VR, content creation, product design

Instant Text-to-3D Generation from Natural Language

Current Methods: Optimization-based Methods

- **Optimization-based methods** - High quality but SLOW
  - DreamFusion: Uses Score Distillation Sampling (SDS) to optimize NeRF
  - MVDream: Multi-view diffusion models for 3D consistency
  - Takes minutes to hours (thousands of optimization iterations)
  - Intensive computation: rendering + backpropagation at each step

Optimization-Based Methods: High Quality but SLOW



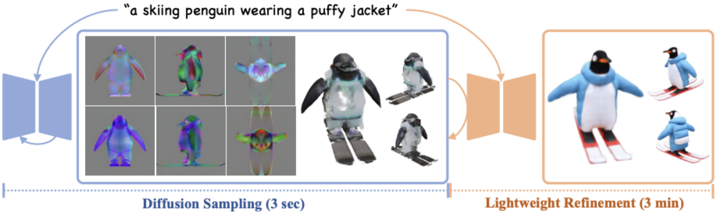
(a) ProlificDreamer can generate meticulously detailed and photo-realistic 3D textured meshes.

Example of optimization-based approach (ProlificDreamer)

Current Methods: Direct Generation Methods

- **Direct generation methods** - FAST but low quality
  - Train large models to directly output 3D representation
  - Generation within one minute
  - Limited by insufficient 3D training data
  - Struggle with complex prompts and geometric details

Direct Generation Methods: Fast but Lower Quality

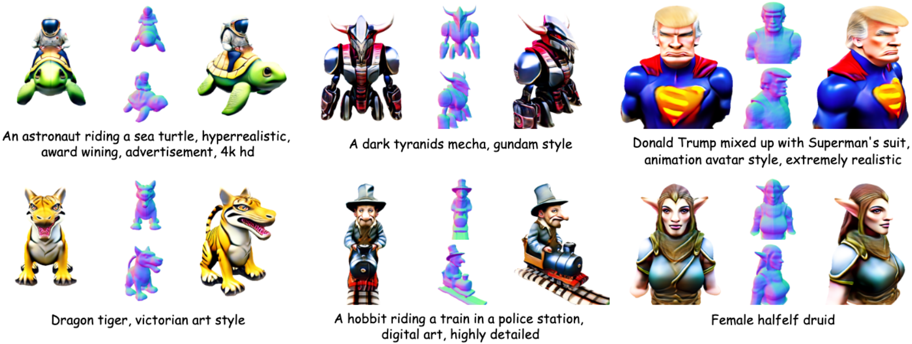


Example of direct generation approach (PI3D)

Our Approach

- **Goal**: Combine speed of direct methods with quality of optimization methods
- **Key insight**: Adapt existing 2D generative models (Stable Diffusion) for 3D
- **Our solution**: Progressive Rendering Distillation
  - Adapt Stable Diffusion into a native 3D generator
  - No 3D training data required
  - Fast inference: generate high-quality 3D in seconds

Our Results: Instant High-Quality Text-to-3D Generation



All generated in just 1-2 seconds

Try our demo: huggingface.co/spaces/ZhiyuanthePony/TriplaneTurbo

Key Contributions

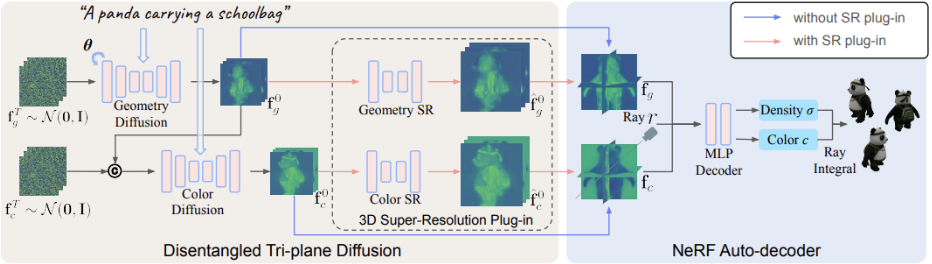
- **Progressive Rendering Distillation (PRD)**
  - First method to adapt pretrained SD into a native 3D generator without 3D data
  - Distills knowledge from multi-view diffusion models
- **Parameter-Efficient Triplane Adapter (PETA)**
  - Adds only 2.5% trainable parameters to frozen SD
  - First parameter-efficient training for direct 3D content generation
- **State-of-the-art performance**
  - Generates high-quality textured meshes in just 1.2 seconds
  - Better quality and generalization to complex prompts

Contents

Related Work: Key Methods in Text-to-3D Generation

- **DIRECT-3D**: Learning on Massive Noisy 3D Data **(CVPR 2024)**
  - Trains on large-scale noisy 3D datasets with iterative cleaning
  - Uses tri-plane diffusion model for efficient 3D generation
- **PI3D**: Pseudo-Image Diffusion for Text-to-3D **(CVPR 2024)**
  - Adapts Stable Diffusion to generate pseudo-images for 3D
  - Leverages 2D diffusion models for 3D generation
- **ATT3D**: Amortized Text-to-3D Object Synthesis **(ICCV 2023)**
  - Introduces amortized optimization across text prompts
  - Shifts from per-prompt optimization to a universal generator

DIRECT-3D: Method Overview

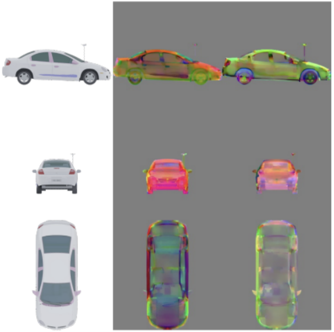


Tri-plane diffusion model architecture from DIRECT-3D

DIRECT-3D: Learning on Massive Noisy 3D Data

- **Noisy Data Training**: Addresses data scarcity challenge **(CVPR 2024)**
  - Training on large-scale noisy and unaligned 3D datasets
  - Iterative optimization to automatically clean and align data
- **Tri-Plane Diffusion Model**
  - Disentangles object geometry and color features
  - Enhances efficiency and provides important geometry priors
- **3D Super-Resolution**: Enhances resolution from 128^3 to 512^3
- **Geometry Consistency**: Reduces issues like the Janus problem

Triplane Representation for 3D Content



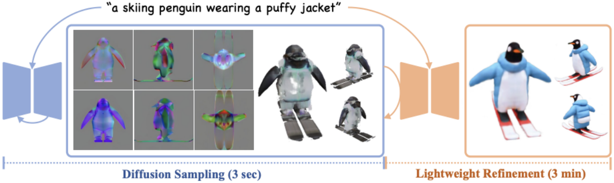
Orthogonal Views Pseudo-Images

A 3D representation can be decomposed into three orthogonal planes: left-right (xy), front-back (xz), and up-down (yz)

PI3D: Based on Stable Diffusion

- **Stable Diffusion**: State-of-the-art text-to-image model
  - Latent diffusion model (LDM) architecture
  - Works by denoising random noise guided by text conditioning
  - Operates in compressed latent space for efficiency
  - Trained on billions of image-text pairs
- **PI3D adapts Stable Diffusion for 3D generation**
  - Leverages powerful 2D priors from Stable Diffusion
  - Fine-tunes SD to output tri-plane representation instead of images
  - Reuses SD's text understanding capabilities
  - Maintains generation speed advantages of diffusion models

PI3D: Limitations in Output Quality



Examples of PI3D generation results

Key Limitations:

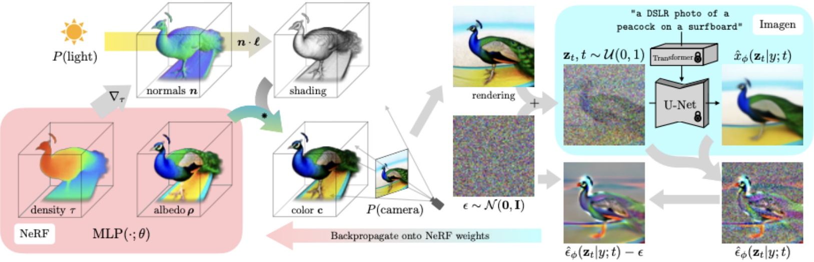
- Insufficient geometric details
- Limited texture quality
- Poor representation of complex concepts
- Multi-view consistency issues
- Still dependent on 3D training data

Root Causes:

- Insufficient and low-quality 3D training data

Note: PI3D employs Score Distillation Sampling (SDS) as a lightweight refinement step to improve results.

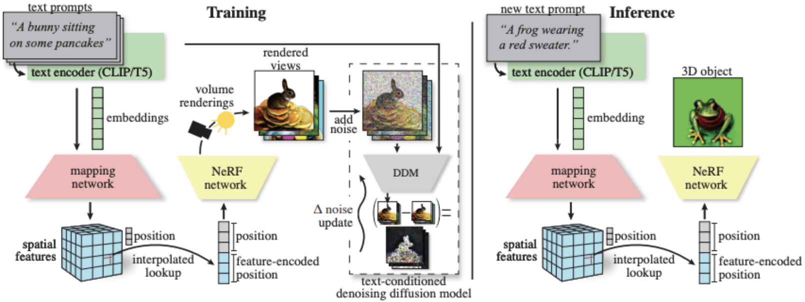
Score Distillation Sampling (SDS) in DreamFusion



Score Distillation Sampling optimizes 3D representation by matching rendered views with diffusion model predictions **(CVPR 2023)**

- **Key insight**: Uses Stable Diffusion as a guiding model for 3D generation
  - Measures consistency between 3D renderings and text description
  - Provides gradient signals to update 3D representation parameters
  - Leverages knowledge from 2D diffusion models trained on billions of images

ATT3D: Paradigm Shift from Optimization to Generation



ATT3D's core innovation: Shifting from per-prompt optimization to training a universal text-to-3D generator

ATT3D: Amortized Text-to-3D Object Synthesis

- **Key Innovation**: Amortized optimization over text prompts **(ICCV 2023)**
  - Trains a single model for multiple prompts simultaneously
  - Shares computation across prompts, reducing training time
  - Generalizes to unseen prompts without additional optimization
- **Uses Score Distillation Sampling (SDS)**
  - Adopts DreamFusion's score distillation technique
  - Transfers knowledge from 2D diffusion models to 3D
  - But applies it across multiple prompts simultaneously
- **Prompt Interpolation** enables smooth transitions between text prompts
  - Generates novel assets and simple animations
  - Achieved by interpolating text embeddings during inference

From SDS to ATT3D: The Paradigm Shift

SDS in DreamFusion:

- Optimizes a specific 3D representation for each text prompt
- Per-prompt optimization process
- Hours of computation for each new prompt
- Formula: $\text{Text} \rightarrow \text{Optimize}_{\text{hours}}(\theta) \rightarrow \text{3D}$
- Not reusable across different prompts

ATT3D's Approach:

- Trains a generator that maps text to 3D
- One-time training process for many prompts
- Fast inference for new prompts (seconds)
- Formula: $\text{Train}_{\text{once}}(G_\phi) \rightarrow [\text{Text} \rightarrow G_\phi \rightarrow \text{3D}]$
- Generator knowledge shared across prompts

Limitations of Current Methods

- **Key limitations of existing approaches:**

Comparison of Text-to-3D Methods		
	Training from Scratch	**Adapted from SD**
Data-driven	Direct3D (Limited by data)	PI3D (Still needs 3D data)
Score Distillation	ATT3D (Low quality)	**Our Method (Best of both worlds)**

- **Rows**: Training approach (Data-driven vs. Score Distillation)
- **Columns**: Model initialization (From Scratch vs. Adapted from SD)
- **Our approach**: Combine score distillation training with SD adaptation
  - No need for 3D data + Leverages powerful SD priors

Contents

Preliminaries: Diffusion Models

Stable Diffusion Training:

- Forward diffusion: gradually add noise to images
- Train model to reverse this process (denoising)
- Predict noise ϵ at each step $\epsilon_\theta(z_t, t, y)$
  - z_t: Noisy latent at timestep t
  - z_0: Clean ground-truth latent
  - ϵ: Noise added to latent
  - y: Text prompt embedding
- Loss: $\mathbb{E}_{z,y,t,\epsilon}[||\epsilon - \epsilon_\theta(z_t, t, y)||^2]$

Image Generation:

- Start with random noise $z_T \sim \mathcal{N}(0, I)$
- Iteratively denoise to generate image
- Conditioned on text embedding y

Score Distillation Sampling (SDS):

- Core technique in DreamFusion
- Transfers knowledge from 2D diffusion to 3D
- Gradient: $\nabla_\phi \mathcal{L}_{SDS} = \mathbb{E}_{t,\epsilon}[w(t)(\epsilon_\theta(z_t, t, y) - \epsilon)\frac{\partial z_t}{\partial \phi}]$
  - ϕ: 3D representation parameters
  - θ: Diffusion model parameters
  - $w(t)$: Time-dependent weight
- **3D Representations**:
  - NeRF: Neural Radiance Fields (density + color)
  - Mesh: Vertices and faces with textures
- Slow process: requires

Motivation: Why Progressive Rendering Distillation?

- **Challenge 1: 3D Data Scarcity**
  - Existing 3D datasets are much smaller than image datasets
  - ◦ 5B text/image pairs vs. 50K text/3D pairs
  - Poor texture quality and inconsistent object poses
  - Cannot generalize well to diverse text prompts
- **Challenge 2: Adapting SD for 3D Generation**
  - Traditional SD adaptation requires 3D ground-truth data
  - This conflicts with our goal to eliminate 3D training data dependency
  - No previous attempt to adapt SD without 3D data
- **Our Solution: Progressive Rendering Distillation**
  - Enables 3D-data-free distillation
  - Accelerates generation through few-step inference
  - Uses multiple teachers for high-quality supervision

Progressive Rendering Distillation (PRD)

Key innovation:

- Eliminates need for 3D ground-truth data
- Denoises latent from random noise
- Uses multi-view teachers for supervision
- Progressive steps allow few-shot generation

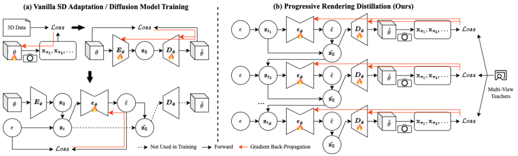


Figure: PRD Scheme

PRD Scheme Detail

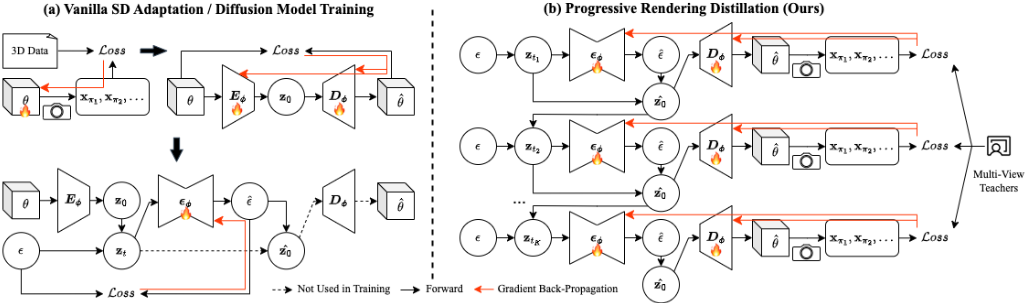


Figure: Progressive Rendering Distillation (PRD) Scheme

Progressive Rendering Distillation Algorithm

Require: Text prompt y, number of progressive steps K
Ensure: Generated triplane representation T

```
z_T ~ N(0, I)                                        ▷ Initialize with random noise
for k = 1 → K do
    Sample random camera parameters c
    ẑ_{t_{k-1}} ← DenoisingUNet(z_{t_k}, t_k, y, c)  ▷ Teacher denoising
    Render multi-view images from triplane T
    Compute distillation loss and update parameters
end for
return Triplane T for mesh extraction
```

Parameter-Efficient Triplane Adaptation (PETA)

Design principles:

- Triplane representation (geometry + texture)
- LoRA adaptation for convolution and cross-attention layers
- Plane-specific LoRA for self-attention
- Only 2.5% additional parameters

LoRA (Low-Rank Adaptation): An efficient fine-tuning technique that updates weights via $W = W_0 + AB$, where W_0 is frozen pre-trained weights and A, B are small low-rank matrices.

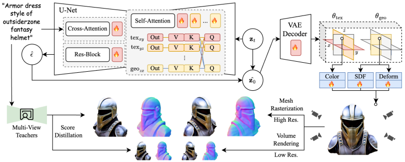


Figure: PETA architecture

PETA Architecture in Detail

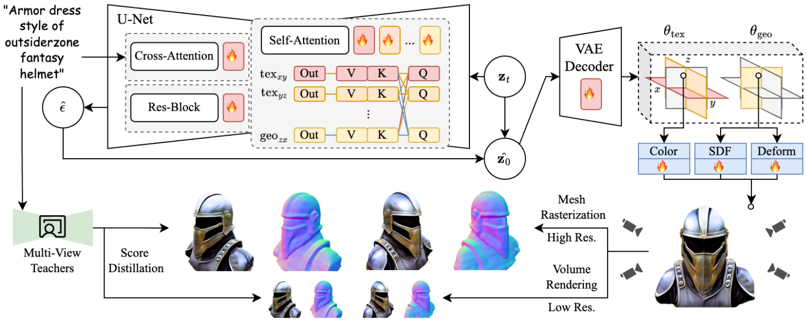


Figure: Detailed view of Parameter-Efficient Triplane Adapter (PETA) architecture

MVDream: Multi-View Diffusion for 3D Generation

Core Innovations:

- Generates multi-view consistent images for 3D supervision
- Based on Stable Diffusion architecture
- Dilated 3D self-attention mechanism connecting all views
- Combines 3D rendering datasets and 2D image-text pairs for training

Technical Details:

- Uses 2-layer MLP for camera parameter embedding
- Camera embedding added as residual to time embedding
- Combines multi-view diffusion loss and image diffusion loss

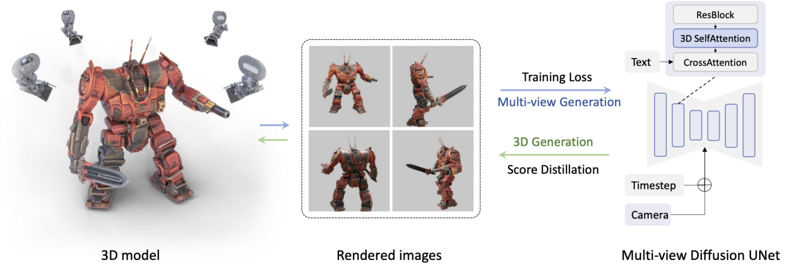
Key Problems Solved:

- Effectively resolves the "Janus problem" (multi-faced objects)
- Eliminates content drift between different views
- Improves stability and consistency of 3D generation
- Maintains correspondence between generated content and text prompts

Application in Our Method:

- Serves as teacher model for multi-view consistency supervision
- Guides triplane representation to generate consistent visual content
- Combines with RichDreamer and SD

MVDream: Method Overview



Overview of MVDream's multi-view diffusion architecture

RichDreamer: Normal-Depth Diffusion Model

Core Innovations:

- Addresses detail richness in text-to-3D generation
- Diffusion model based on normal and depth maps
- Pre-trained on LAION large-scale dataset
- Fine-tuned on Objaverse for enhanced object-level 3D generation

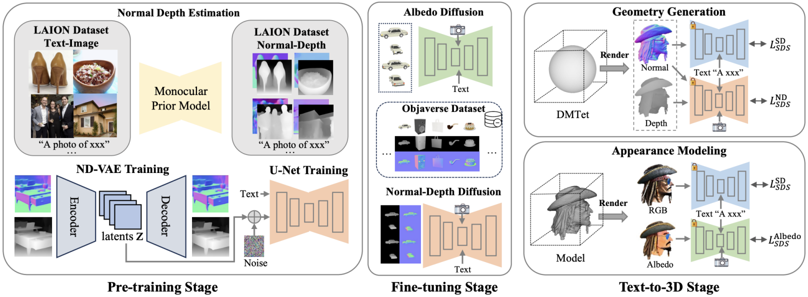
Advantages:

- Provides stronger geometric priors and guidance
- Resolves material-lighting entanglement in traditional methods
- Supports DMTet and NeRF

Application in Our Method:

- Provides geometric supervision signals
- Guides 3D geometry generation via normal and depth maps
- Improves surface details and topological correctness
- Combines with multi-view consistency to enhance generation quality

RichDreamer: Method Overview



Overview of RichDreamer's normal-depth diffusion architecture

RichDreamer: Demo Results

Demonstration of RichDreamer's generation capabilities

Contents

Comparison with Existing Methods

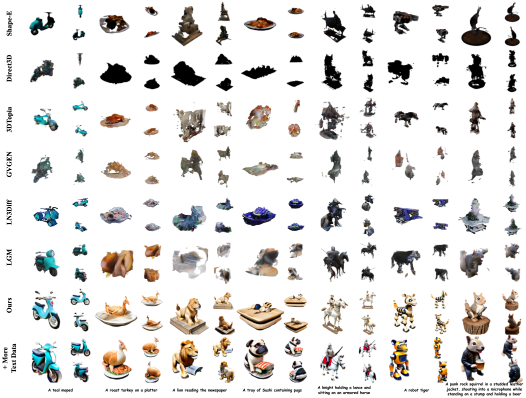


Figure: Qualitative comparison with competing methods

Quantitative Results

	C.S. ↑	R@1 ↑	Latency (s)
Shape-E	55.1	27.1	13.0
Direct3D	60.8	4.33	16.0
3DTopia	59.7	11.2	23.7
PI3D	65.9	25.2	3.00
GVGEN	51.1	2.44	49.2
LN3Diff	55.9	5.09	8.16
LGM	67.4	28.3	56.1
Ours	68.2	32.3	**1.23**
+More Text Data	**75.1**	**46.0**	**1.23**

Key advantages:

- Better quality (CLIP score)
- Higher accuracy (R@1)
- 2-40x faster inference
- Scales well with more data

Scaling with More Text Data

- Training without 3D data allows scaling to 1.7M text prompts
- First method to train on more than 1M creative text prompts
- Better handling of challenging concepts
- Improved generation quality

Figure: Sample results

Sample Results: Scaling with Text Data

More examples of our method's generation results with 1.7M text prompts

More Results: Scaling with Text Data

Ablation Studies

The effect of progressive steps (K):

- K=1: Poor 3D structure (equivalent to vanilla generator)
- K=2: Suboptimal but acceptable results
- K=4: Best trade-off between quality and efficiency

Figure: Effect of progressive steps (K): K=1 fails to generate proper 3D structures, K=2 produces acceptable results, while K=4 achieves the best balance between quality and efficiency.

More Ablation Studies

Effect of Multiple Teachers:

- **Stable Diffusion (SD)**: Ensures high-fidelity textures and text consistency
- **MVDream (MV)**: Provides multi-view consistency, reduces Janus problem
- **RichDreamer (RD)**: Improves geometry supervision through normal/depth maps
- **Combined**: Maximizing strengths of all teachers yields optimal results

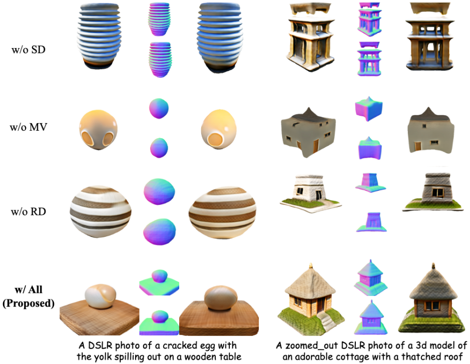


Figure: Effect of multiple teachers: Combining SD, MV and RD models yields optimal results.

Dual Rendering for Multi-View Distillation

Why Dual Rendering?

- **Volumetric Rendering**:
  - Complete 3D space supervision
  - Ensures geometric consistency
  - Handles complex topology
- **Mesh Rasterization**:
  - High-resolution texture details
  - Faster rendering speed
  - Better surface quality
- **Combined Benefits:**
  - Ensures training stability
  - Improves output quality
  - Balances efficiency and quality

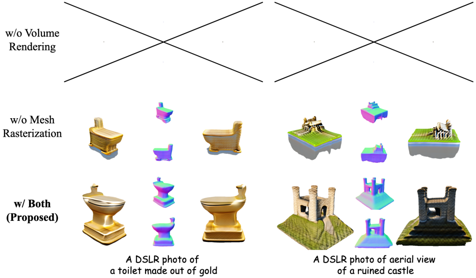


Figure: Dual rendering approach: Volumetric rendering provides complete 3D supervision while mesh rasterization enables high-resolution texture details

Effect of LoRA Rank

Impact of LoRA Rank:

- **Rank Selection Trade-off**:
  - Lower rank: More parameter efficient but limited capacity
  - Higher rank: Better quality but increased parameters
- **Our Findings**:
  - Rank=8: Insufficient for complex geometry
  - Rank=16: Optimal balance of quality and efficiency
  - Rank=32: Marginal improvements, excessive parameters

Figure: Ablation study on LoRA rank: Rank=16 achieves the best trade-off between generation quality and parameter efficiency

Contents

Conclusion and Future Work

Summary:

- First method to adapt SD for 3D generation without 3D data
- Parameter-efficient approach (only 2.5% additional parameters)
- State-of-the-art performance in both quality and speed
- Scales well with more text data

Limitations and Future Work:

- Challenges with generating precise numbers of multiple objects
- Limited facial and hand details for full-body humans
- Potential extension to 3D scene generation and image-to-3D tasks
- Apply to other pre-trained models (e.g., DiT)

Thank you!

Questions?

Paper	arXiv	Demo	Code
Link	arxiv.org/abs/2403.15319	HuggingFace	GitHub

```
@article{ma2025progressive,
title={Progressive Rendering Distillation: Adapting Stable Diffusion for Instant
Text-to-Mesh Generation without 3D Data},
author={Ma, Zhiyuan and Liang, Xinyue and Wu, Rongyuan and Zhu, Xiangyu and Lei, Zhen
and Zhang, Lei},
booktitle={Proceedings of the IEEE/CVF conference on computer vision and pattern
recognition},
```

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Opening Minds • Shaping the Future • 啟迪思維 • 成就未來