Poster # 74

ControlDreamer: Blending Geometry and Style in Text-to-3D

Yeongtak Oh[1], Jooyoung Choi[1], Yongsung Kim[2], Minjun Park[2], Chaehun Shin[1], Sungroh Yoon[1,2]

Department of Electrical and Computer Engineering, Seoul National University

Interdisciplinary Program in Artificial Intelligence, Seoul National University

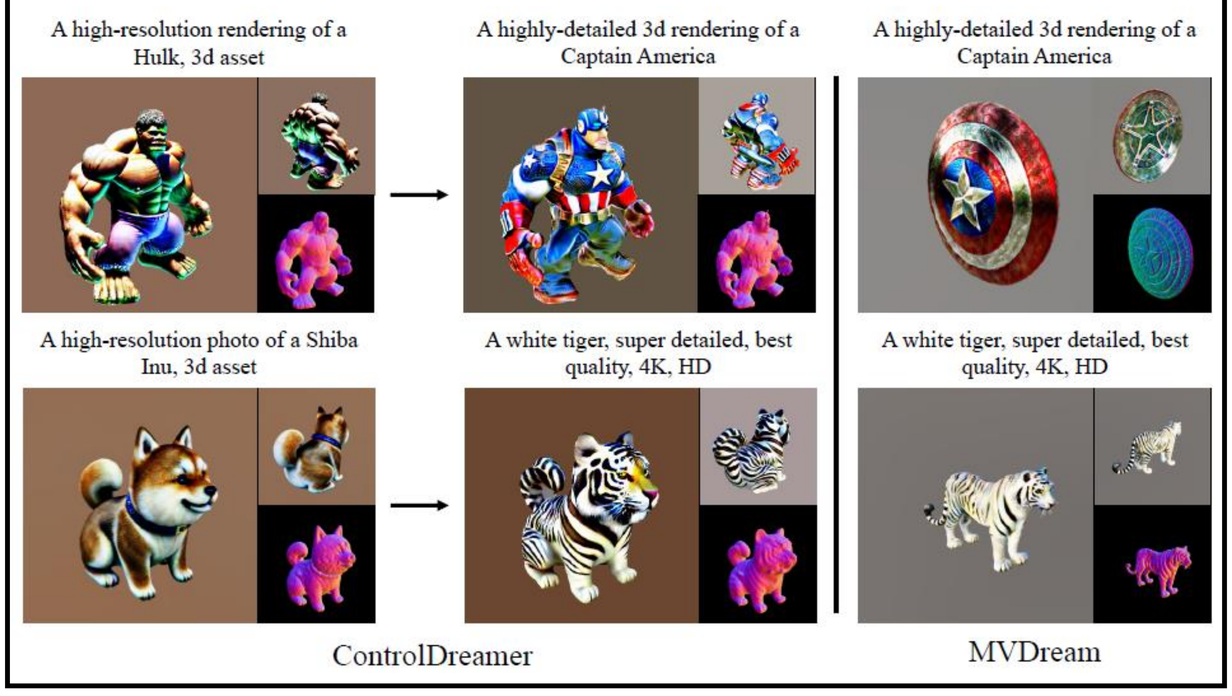
1 Background: Text-to-3D Generation

- **Score Distillation Sampling (SDS)**
  - $\nabla_\theta L_{SDS} = E_{t,\epsilon}[w(t)(\hat{\epsilon}_\phi(z_t; y, c, t) - \epsilon)]\frac{\partial z}{\partial x}\frac{\partial x}{\partial \theta}$
  - $\hat{\epsilon}_\phi(z_t; y) = \epsilon_\phi(z_t; y) - \epsilon_\phi(z_t)$ (guidance)
  - Mode-seeking nature
- **3D Representations**
  - NeRF: Neural network learns volume density and color
  - DMTet: Tetrahedral mesh for 3D surface reconstruction
- **Multi-view diffusion model**
  - MVDream: Generates multi-views from text input

[1] Poole et al., DreamFusion: Text-to-3D using 2D Diffusion, ICLR, 2023.
[2] Mildenhall et al., NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV, 2020.
[3] Shen et al., Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis, NeurIPS, 2019.
[4] Shi et al., MVDream: Multi-view Diffusion for 3D Generation, ICLR, 2024.
[5] Deitke et al., Objaverse-XL: A Universe of 10M+ 3D Objects, NeurIPS, 2023.

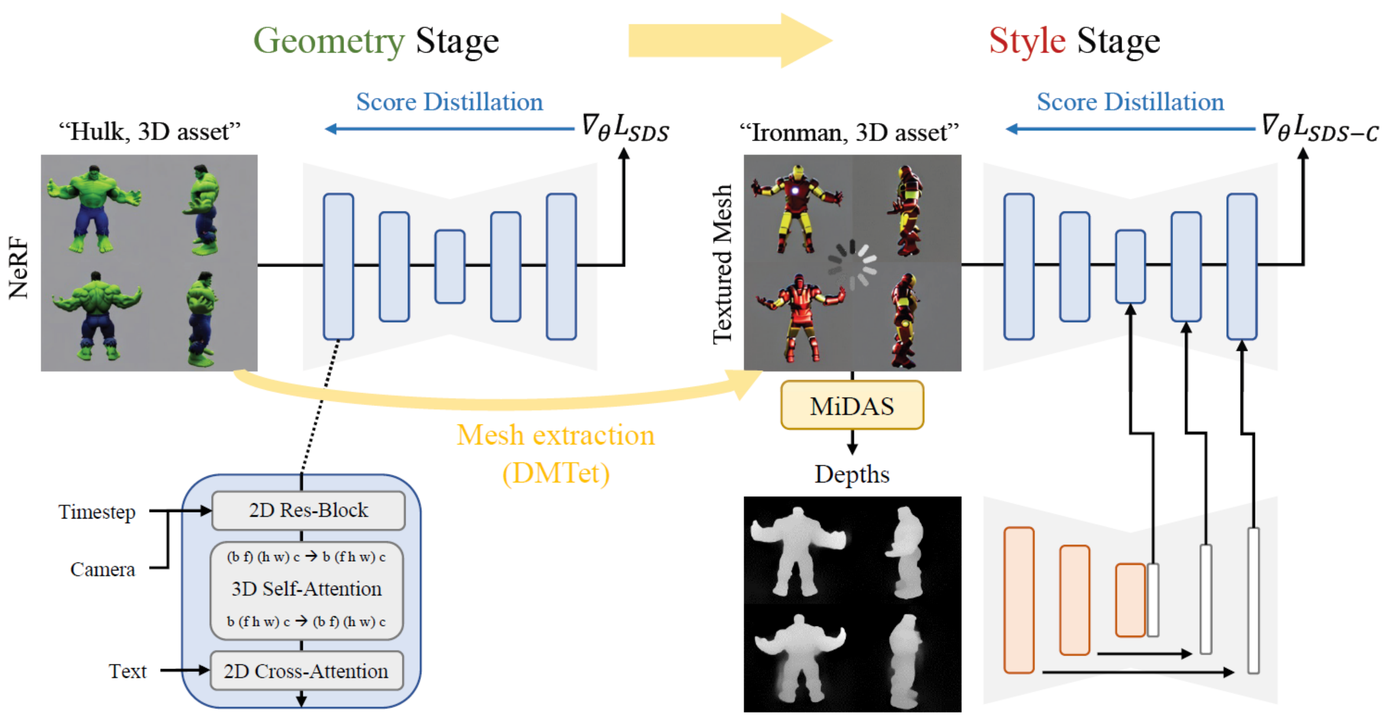
2 Blending Geometry and Style

- **Limitation: geometry bias**
  - MVDream trained on the Objaverse dataset



3 ControlDreamer

- **Two-Stage Pipeline**
  - Geometry: Optimize NeRF with multi-view diffusion
  - Style: Optimize DMTet with **MV-ControlNet**
  - $\hat{\epsilon}_\phi \propto \nabla_{z_t}(logp(y_{style}|z_t, D_{geo}) + logp(z_t|D_{geo}))$
    - $logp(y_{style}|z_t, D_{geo})$: Style blending
    - $logp(z_t|D_{geo})$: Geometry

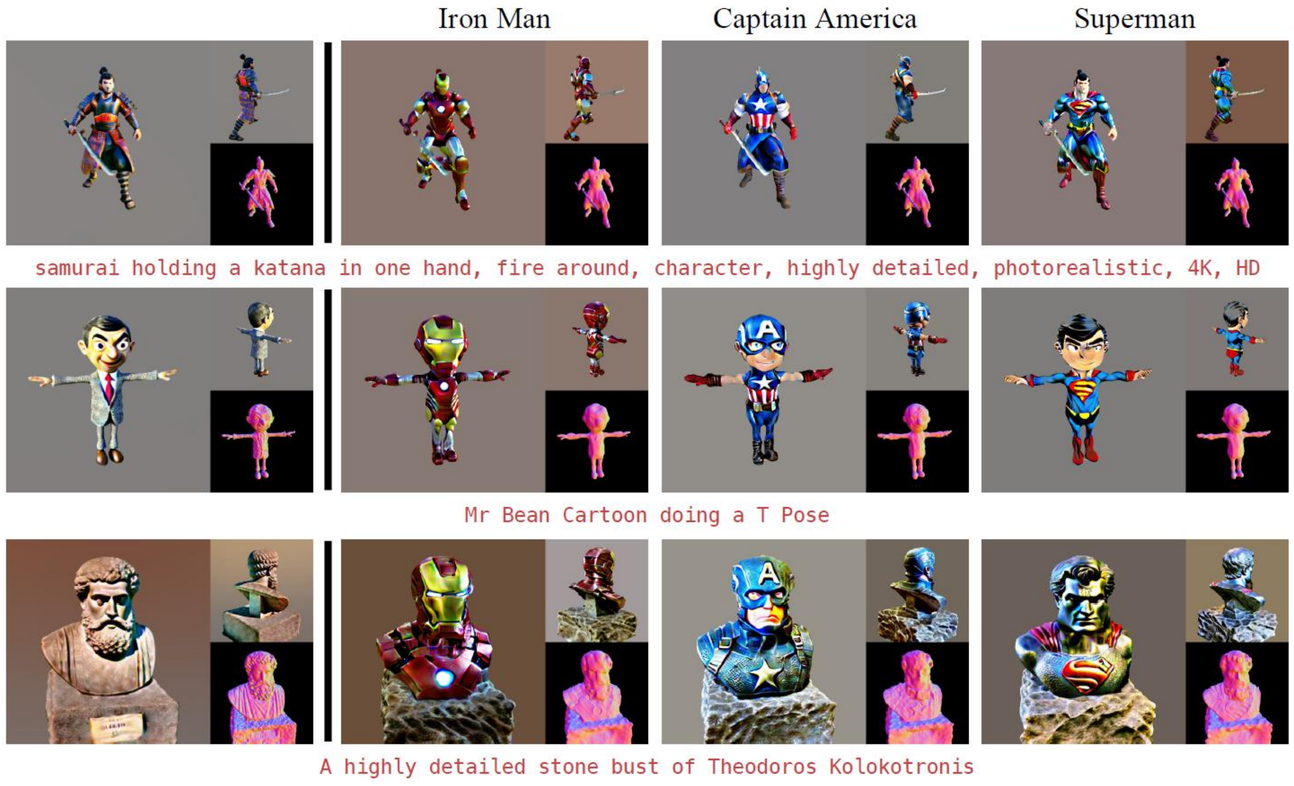


- **MV-ControlNet**
  - Depth encoder initialized from multi-view diffusion
    - Leverages the 3D knowledge of multi-view diffusion
  - Generated dataset of (text, depth, image) triplets
    - Distills knowledge from diffusion & depth estimator
    - Rich text captions refined using GPT-4

[6] Zhang et al., Adding Conditional Control to Text-to-Image Diffusion Models, ICCV, 2023.

4 Results

- **Main results**
  - Generates a blend of geometry and style that was previously unattainable with multi-view diffusion alone.



- **Ablation studies**
  - Outperforms various two-stage pipelines
  - Depth surpasses normal map and edge variants