Rethinking Score Distillation as a Bridge Between Image Distributions

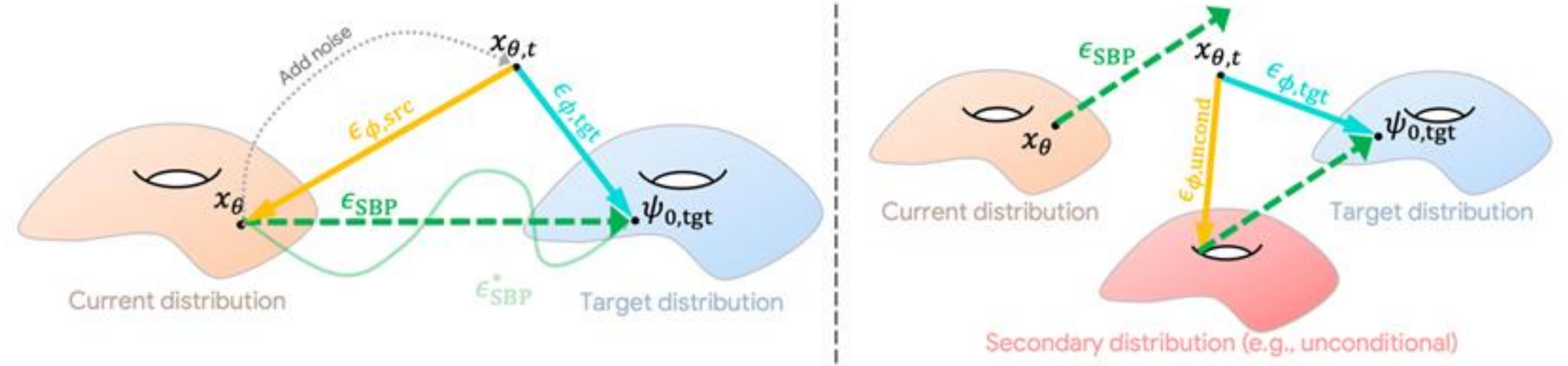
David McAllister[1*] Songwei Ge[2*] Jia-Bin Huang[2] David W. Jacobs[2]
Alexei A Efros[1] Aleksander Holynski[1] Angjoo Kanazawa[1]
[1] UC Berkeley [2] University of Maryland

2024.08.21
Jaihoon Kim

KAIST Visual AI Group

Motivation

SDS and its variants suffer from characteristic artifacts.
→ Analysis of SDS and its variants through the lens of Schrodinger Bridge problem.



(a) Score distillation as a bridge between two distributions.

(b) Distribution mismatch error.

$$\epsilon^*_{\text{SBP}} = \psi_{0,\text{tgt}} - \psi_{0,\text{src}}$$

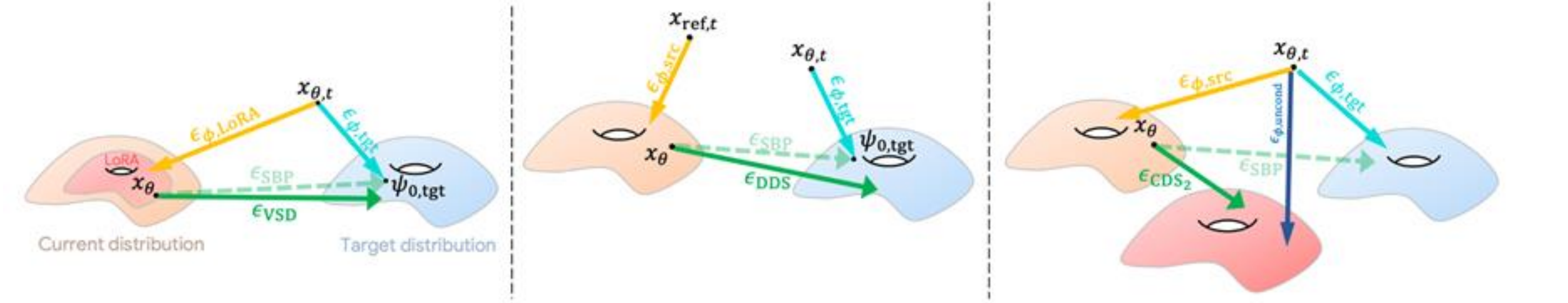
Full PF ODE

$$\epsilon_{\text{SBP}} = \epsilon_{\phi,\text{tgt}} - \epsilon_{\phi,\text{src}}$$

Single-step approximation
Source distribution mismatch

Key Ideas

Seeing SDS and its variants through the lens of SB.



(a) Variational Score Distillation (b) Delta Distillation Sampling (c) Classifier Score Distillation (2nd term)

$$\epsilon_{\text{VSD}} = \epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right) + s \cdot \left(\epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\text{tgt}}, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right)\right) - \epsilon_{LoRA}\left(\mathbf{x}_{\theta,t}; y_{\text{tgt}}, t\right)$$

$$\epsilon_{\text{DDS}} = \epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\text{tgt}}, t\right) - \epsilon_\phi\left(\mathbf{x}_{\text{ref},t}; y_{\text{src}}, t\right)$$

$$\epsilon_{\text{CSD}} = w_1 \cdot \left(\epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\text{tgt}}, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right)\right) + w_2 \cdot \left(\epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\text{src}}, t\right)\right)$$

Method

"...simply describing image corruptions with a text prompt, we can improve our estimate of the source distribution…"

1. Use SDS loss to produce an output with the method's characteristic artifacts.

$$\epsilon_{\text{SDS}} = \epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right) + s \cdot \left(\epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\text{tgt}}, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; \varnothing, t\right)\right) - \epsilon$$

2. Append the descriptors to better describe the source distribution and generate using the proposed loss function.

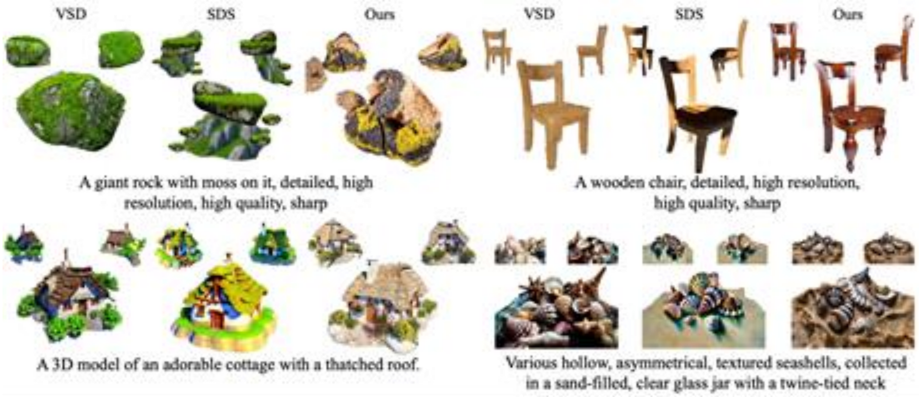
$$\epsilon_{\text{ours}} = w \cdot \left(\epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\text{tgt}}, t\right) - \epsilon_\phi\left(\mathbf{x}_{\theta,t}; y_{\text{src}}, t\right)\right).$$

Negative prompt: *", oversaturated, smooth, pixelated, cartoon, foggy, hazy, blurry, bad structure, noisy, malformed"*

Experiments

Text-to-Image

Text-to-3D



	DDIM (lower bound)	SDS [46]	NFSD [28]	CSD [74]	VSD [68]	Ours
Zero-Shot FID (↓)	49.12	86.02	91.70	89.96	**59.22**	67.89
Zero-Shot CLIP FID (↓)	16.56	28.39	29.25	27.07	**18.86**	20.31
Time per Sample (mins)	0.05	**4.48**	7.20	6.21	16.02	**4.48**

Experiments

Sketch-to-3D

Ambiguous Image