Score Distillation via Reparametrized DDIM

MIT, University of Oxford, IBM, Toyota Research Institute, Meta, MIT, MIT

Artem Lukoianov, Haitz S de Ocáriz Borde, Kristjan Greenewald, Vitor Guizilini, Timur Bagautdinov, Vincent Sitzmann, Justin Solomon

TL;DR

Score Distillation Sampling (SDS) [1] is a promising technique that allows to use pre-trained 2D diffusion models for 3D generation. However, the quality of the generated 3D assets is limited. In this work we:

- 🔎 Theoretically show that SDS ≈ 2D Diffusion
- 🚨 Reveal that the noise term in SDS is the reason for over-smoothing
- 🛠 Suggest a fix
- ✅ Improve the quality of 3D generation

2D Diffusion (DDIM) | Score Distillation (SDS) | Ours (SDI)

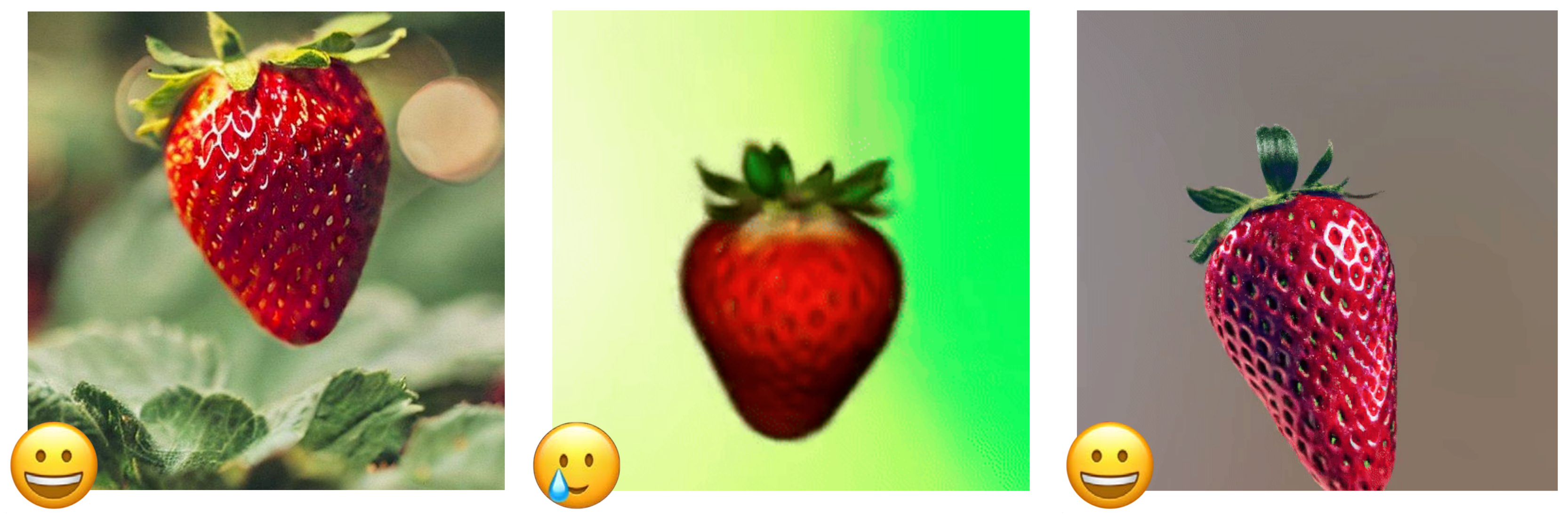


Image Diffusion generates crisp, high-quality images | Score Distillation Sampling uses Image Diffusion, but the results are blurry | We provide a theoretical analysis and suggest a fix

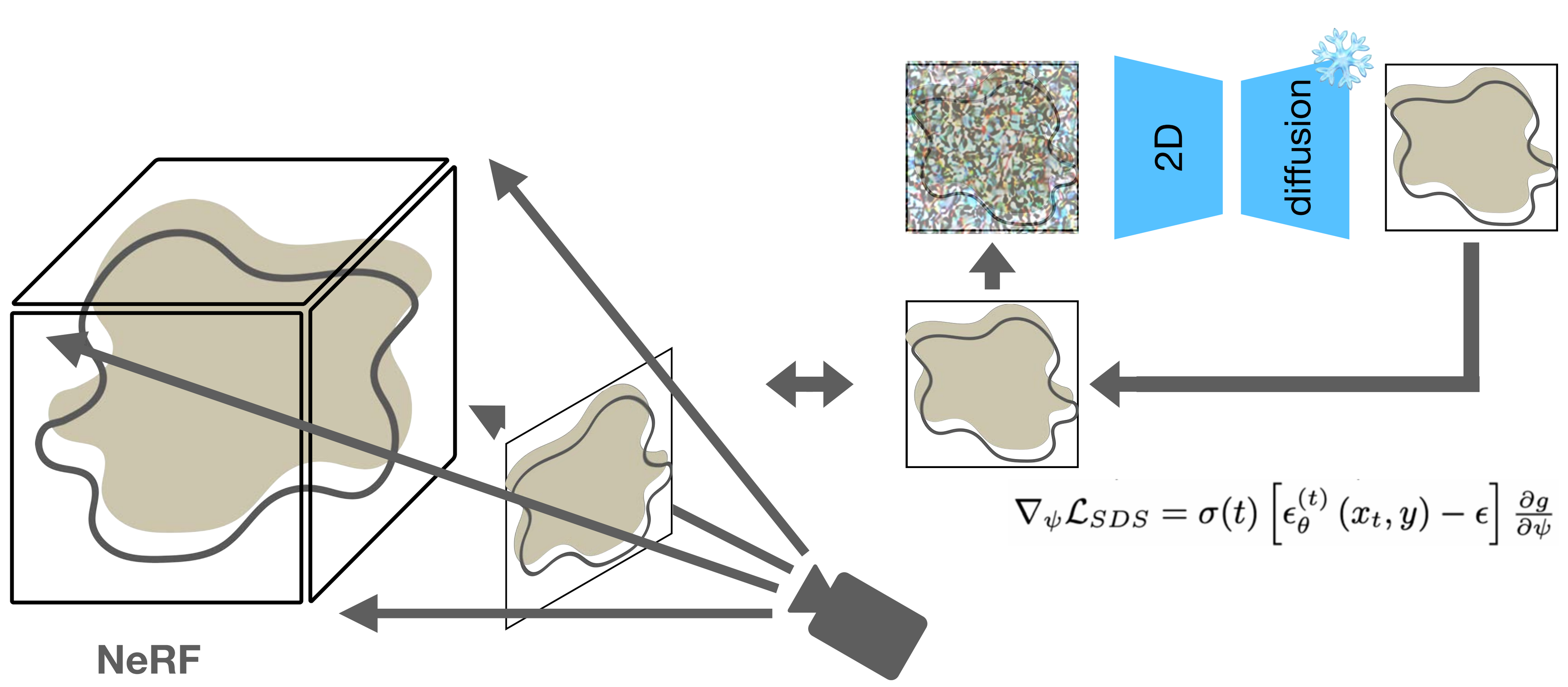
How does Score Distillation work?

Proposed in DreamFusion [1] and Score Jacobian Chaining [2], Score Distillation is a method for generating 3D shapes using a pre-trained and frozen 2D diffusion model.

I. Initialize a differentiable 3D representation
  1. Sample a random camera pose 📷
  2. Render a view of the object 🖼
  3. Add noise to the rendering
  4. Denoise the image with the 2D diffusion model
  5. Optimize the parameters of the 3D representation to match the denoised image 🎯

II. Repeat steps 1-5 until convergence 🔄

Often the generated shapes are over-smoothed and over-saturated.



$$\nabla_\psi \mathcal{L}_{SDS} = \sigma(t)\left[\epsilon_\theta^{(t)}(x_t, y) - \epsilon\right]\frac{\partial g}{\partial \psi}$$

SDS is a dual process of DDIM

A single upate step of Denoising Diffusion Implicit Models (DDIM) [3] sampling algorithm removes a portioin of the predicted noise to match the noise level of the next step:

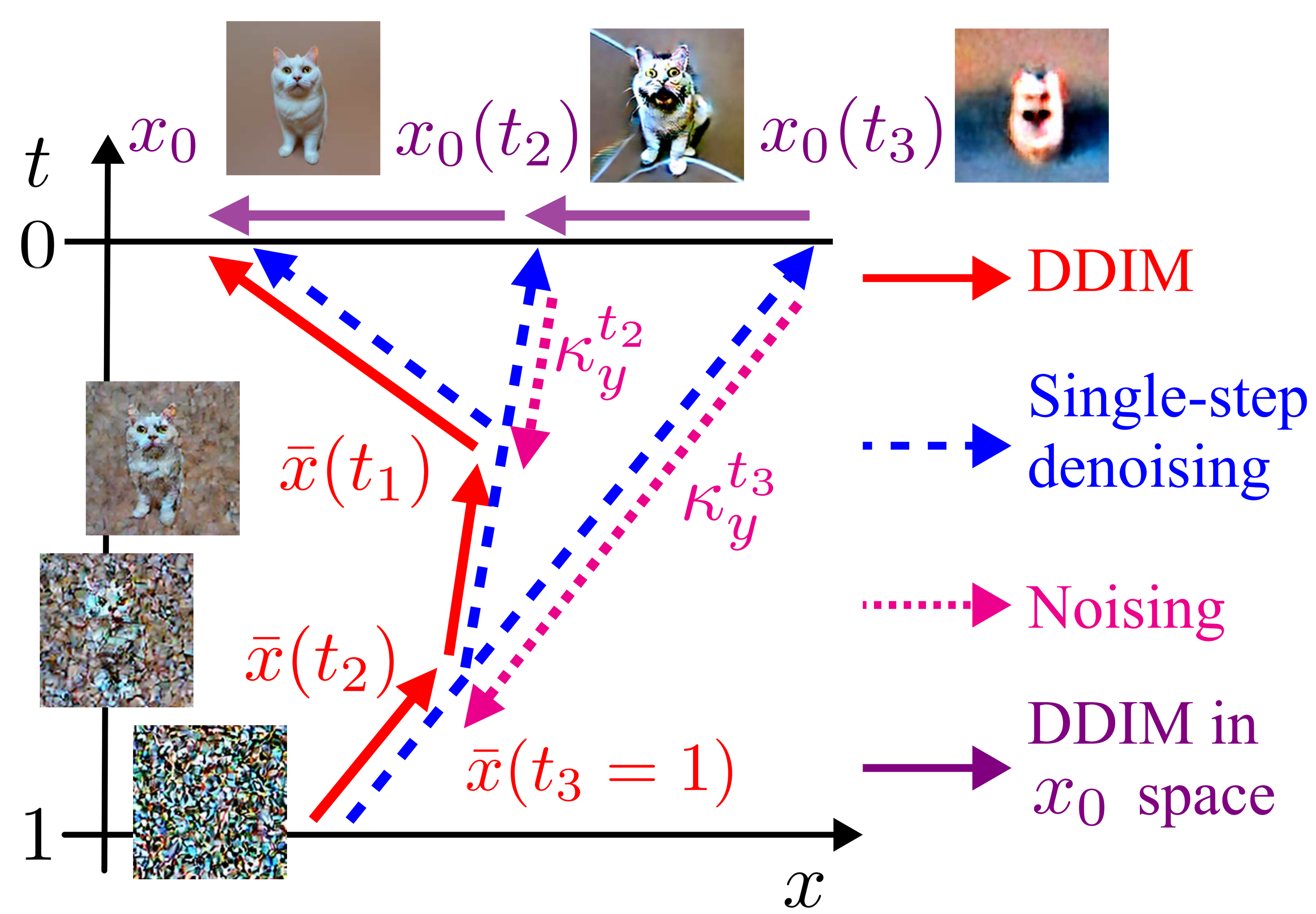
$$\bar{x}(t-\tau) = \bar{x}(t) + \epsilon_\theta^t\left(\sqrt{\alpha(t)}\bar{x}(t), y\right)\left[\sigma(t-\tau) - \sigma(t)\right].$$

Let's consider a new variable that is formed by denoising the images in a single step:

$$x_0(t) \triangleq \bar{x}(t) - \sigma(t)\epsilon_\theta^t(x(t), y).$$

With a simple change of variable in the DDIM update rule, we obtain a dual process of DDIM defined for the space of noise-free images:

$$x_0(t-\tau) = x_0(t) - \sigma(t-\tau)\Big[\underbrace{\epsilon_\theta^{t-\tau}\big(\overbrace{\sqrt{\alpha(t-\tau)}x_0(t) + \sqrt{1-\alpha(t-\tau)}\kappa_y^t(x_0(t))}^{x_0 \text{ noised with } \kappa_y^t \text{ to time } t-\tau}, y\big)}_{\text{predicted noise}} - \underbrace{\kappa_y^t(x_0(t))}_{\text{noise sample } \kappa_y^t}\Big].$$



,where κ_y^t is a noise term added to the current image. For the new process to match DDIM exactly, it should satisfy the following stationary-point equation:

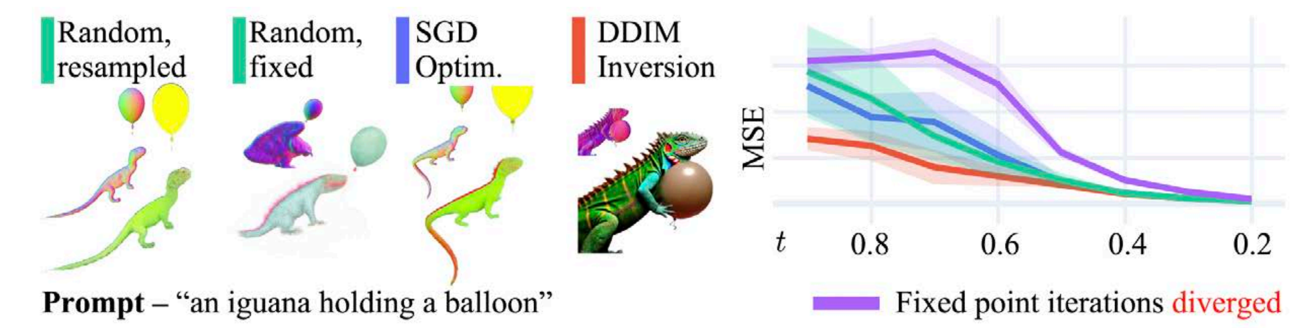
$$\kappa_y^t = \epsilon_\theta^t\left(\sqrt{\alpha(t)}x_0(t) + \sqrt{1-\alpha(t)}\kappa_y^t, y\right)$$

Solving this equation precisely is very hard as it involves inverting a higly non-linear and high-dimentional Neural Network. A naive solution that we can adopt is to sample the noise term randomly from the Normal Gaussian distribution.

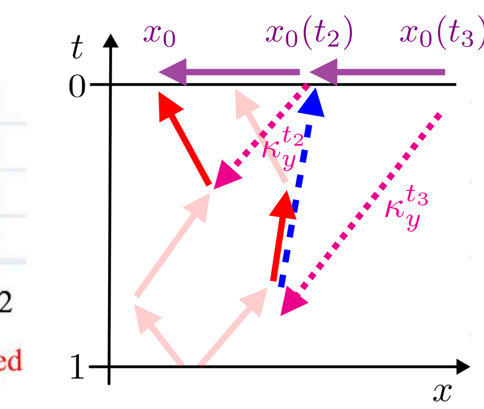
$$\kappa_y^t \sim \mathcal{N}(0, I)$$

In this case the re-parametrization of DDIM precisely matches SDS update rule!

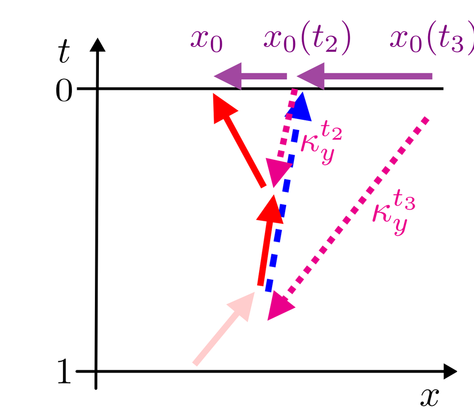
Random noise is the reason of blurriness



Prompt – “an iguana holding a balloon”



Due to the randomness in choise of κ_y^t SDS's optimization process is being averaged acrross multiple trajectories of DDIM, what causes bluriness. In this work we explore multiple approaches to solving the stationary point equation and find that **running DDIM in the inverse direction yields the most accurate solution.** In this case Score Distillation optimisation process becomes time- and spatial-consistent.



3D generation quality gets closer to 2D models

See the paper for detailed comparisons with ProlificDreamer, Noise-Free SD, HiFA, Lucid Dreamer, and other amazing works in Score Distillation

References

[1] Poole, Ben, et al. "DreamFusion: Text-to-3D using 2D Diffusion." *The Eleventh International Conference on Learning Representations.*

[2] Wang, Haochen, et al. "Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2023.

[3] Song, Jiaming, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models." *arXiv preprint arXiv:2010.02502* (2020).

MORE ABOUT THE WORK

lukoianov.com/sdi