Dream Fusion

Text-to-3D с помощью 2D диффузии

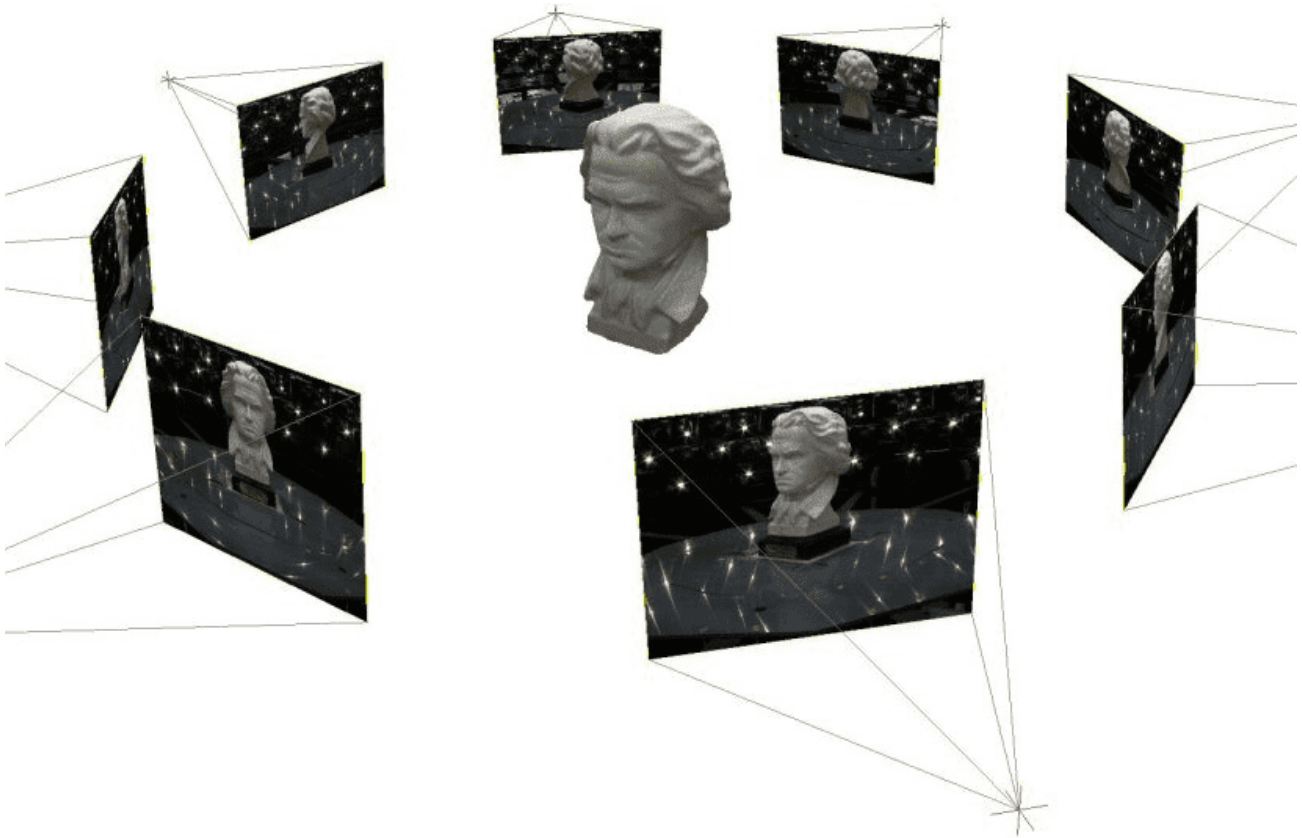
Идея

- Инициализируем 3D-модель
- Оптимизируем градиентным спуском, уменьшая ошибку 2D-рендеринга со случайных углов

Что мы хотим?

- 3D модель (которая генерируют фигуры исходя из ее параметров, которые мы будем оптимизировать)
- Предобученную text-to-image диффузионку (чтобы сравнивать проекции 3D модели с результатами диффузионной модели)
- Функцию потерь

NeRF (3D модель)



NeRF (3D модель)

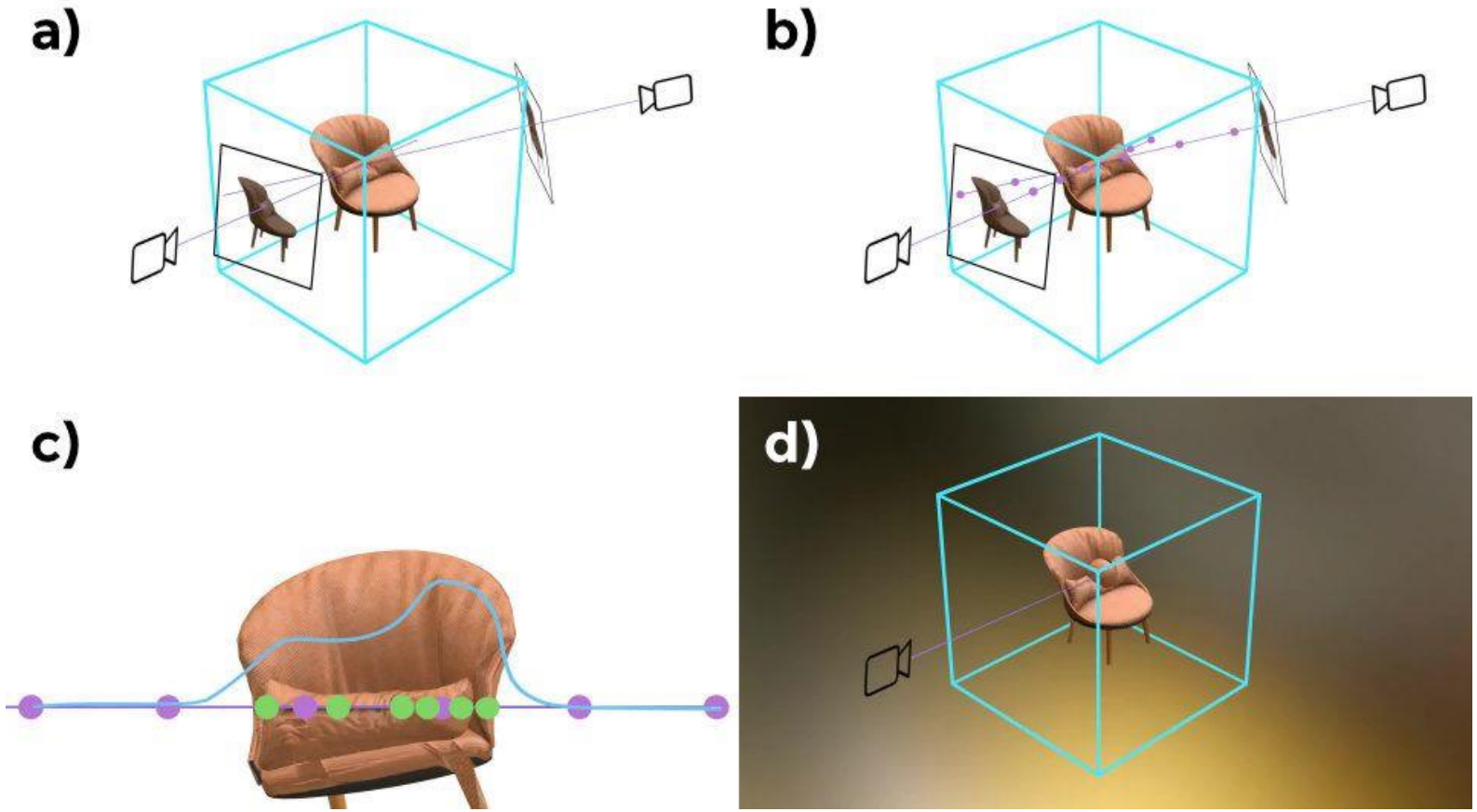


Imagen (txt-2-img diffusion model)

Функция потерь

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x} = g(\theta)) = \nabla_\theta \mathbb{E}_t \left[\sigma_t / \alpha_t w(t) \mathbf{KL}(q(\mathbf{z}_t | g(\theta); y, t) \| p_\phi(\mathbf{z}_t; y, t))\right]$$

ϕ - параметры диффузионной модели

$x = g(\theta)$ - изображение сгенерированное NeRF с параметрами θ

t - временной шаг

σ, α из диффузионной модели

$w(t)$ - вес

p, q - вероятнотсные распределения из диффузионной модели

Эксперименты

Table 1: Evaluating the coherence of DreamFusion generations with their caption using different CLIP retrieval models. We compare to the ground-truth MS-COCO images in the object-centric subset of Jain et al. (2022) as well as Khalid et al. (2022). †Evaluated with only 1 seed per prompt. Metrics shown in parentheses may be overfit, as the same CLIP model is used during training and eval.

Method	R-Precision ↑					
	CLIP B/32		CLIP B/16		CLIP L/14	
	Color	Geo	Color	Geo	Color	Geo
GT Images	77.1	–	79.1	–	–	–
Dream Fields (reimpl.)	68.3 **78.6**	– 1.3	74.2 (99.9)	– (0.8)	– **82.9**	– 1.4
CLIP-Mesh	67.8	–	75.8	–	74.5†	–
DreamFusion	75.1	**42.5**	**77.5**	46.6	79.7	**58.5**

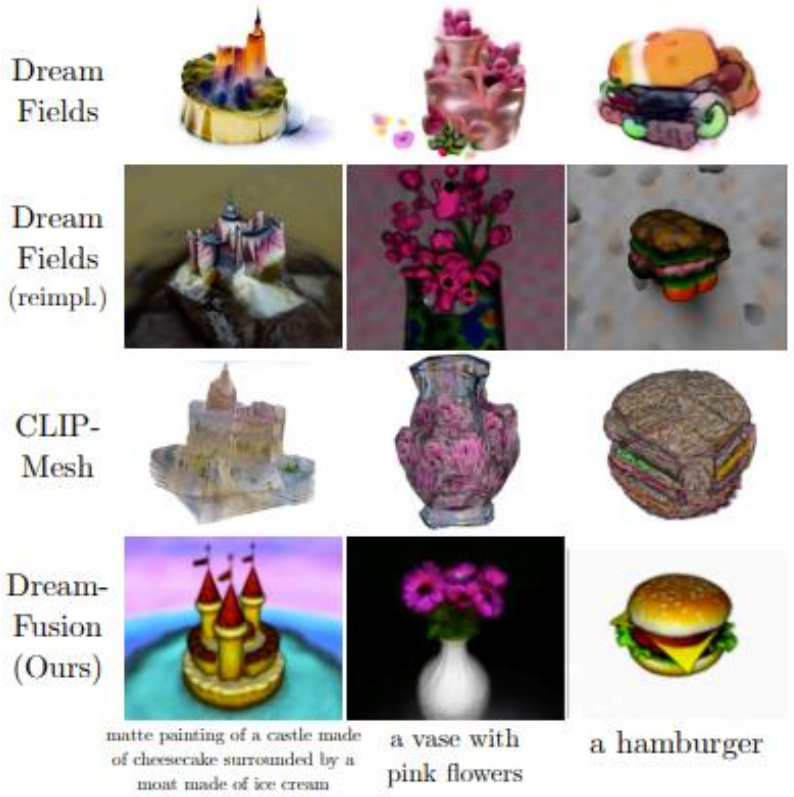


Figure 5: Qualitative comparison with baselines.

Эксперименты

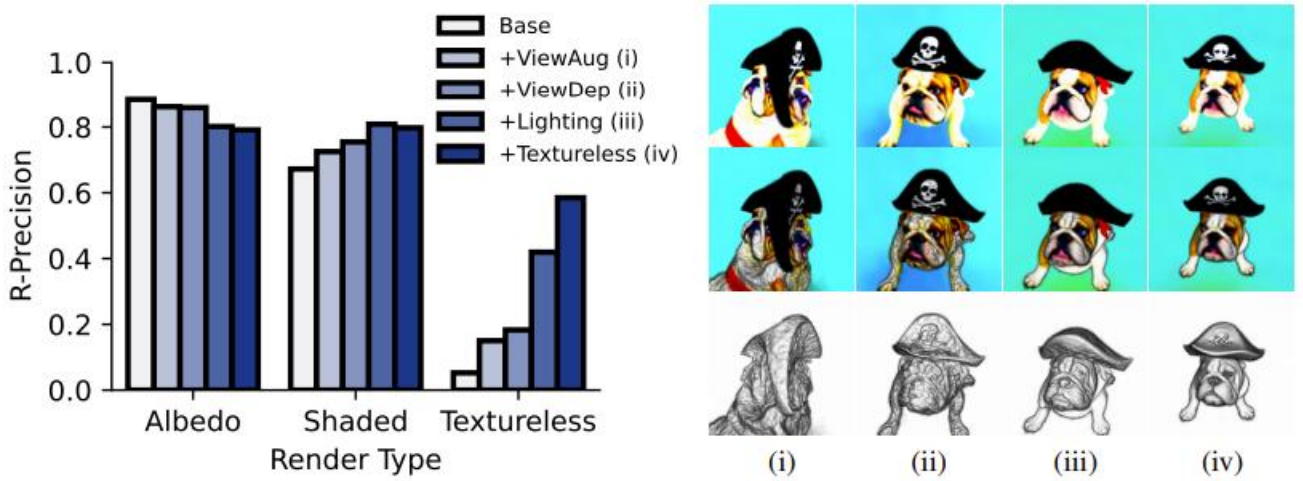


Figure 6: An ablation study of DreamFusion. **Left**: We evaluate components of our unlit renderings on albedo, full shaded and illuminated renderings and textureless illuminated geometry using CLIP L/14 on object-centric COCO. **Right**: visualizations of the impact of each ablation for *"A bulldog is wearing a black pirate hat."* on albedo (top), shaded (middle), and textureless renderings (bottom). The base method (i) without view-dependent prompts results in a multi-faced dog with flat geometry. Adding in view-dependent prompts (ii) improves geometry, but the surfaces are highly non-smooth and result in poor shaded renders. Introducing lighting (iii) improves geometry but darker areas (e.g. the hat) remain non-smooth. Rendering without color (iv) helps to smooth the geometry, but also causes some color details like the skull and crossbones to be "carved" into the geometry.

an all-utility vehicle driving across a stream†

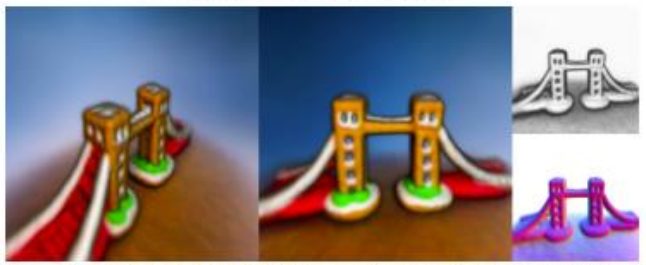
a chimpanzee dressed like Henry VIII king of England*

a baby bunny sitting on top of a stack of pancakes†

a sliced loaf of fresh bread

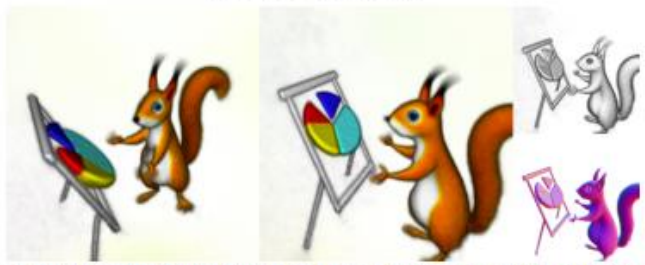
a bulldozer clearing away a pile of snow*

a classic Packard car*



zoomed out view of Tower Bridge made out of gingerbread and candy‡

a robot and dinosaur playing chess, high resolution*



a squirrel gesturing in front of an easel showing colorful pie charts