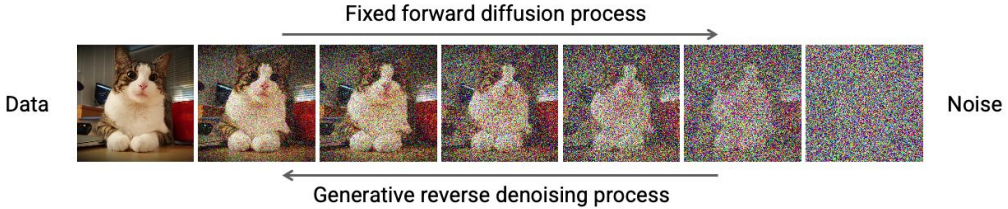
DreamFusion:
Text-to-3D using 2D Diffusion

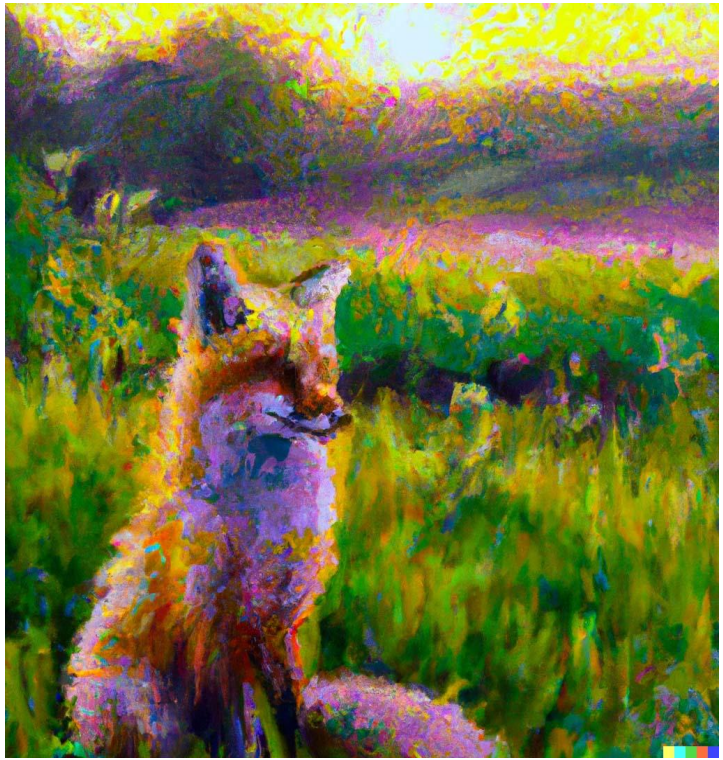
Обзор статьи: Алексей Шишков, БПМИ192

Что уже умеем?

Diffusion models: по тексту генерировать картинку

- Обучаются на парах (описание, картинка)
- При обучении сначала зашумляем, потом пытаемся предсказывать добавленный шум
- Для генерации из шума пытаемся понять, какая могла быть картинка





A painting of a fox sitting in a field at sunrise in the style of Claude Monet

Тогда давайте учить 3d-диффузию! Или нет?..

Проблемы обучения 3d-диффузионной модели:

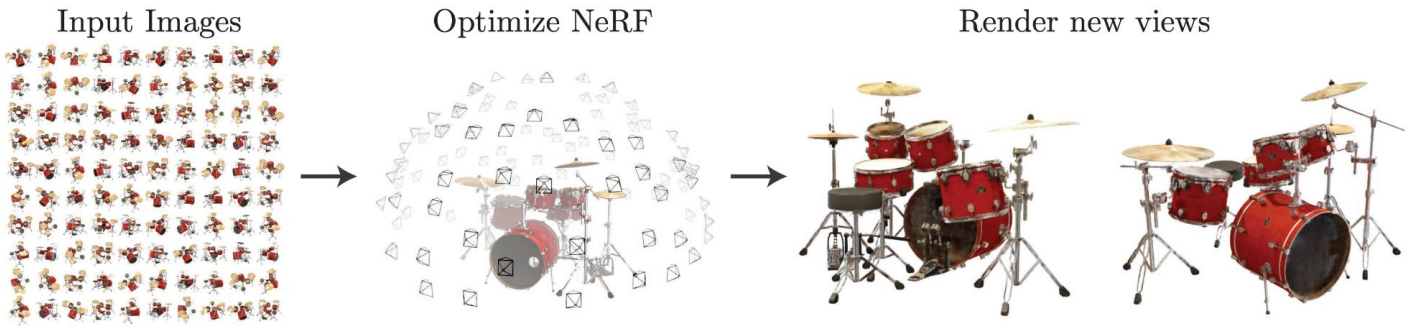
- Нужно много размеченных пар (описание, 3d-модель)
- Существующие архитектуры на 3d-данных работают плохо

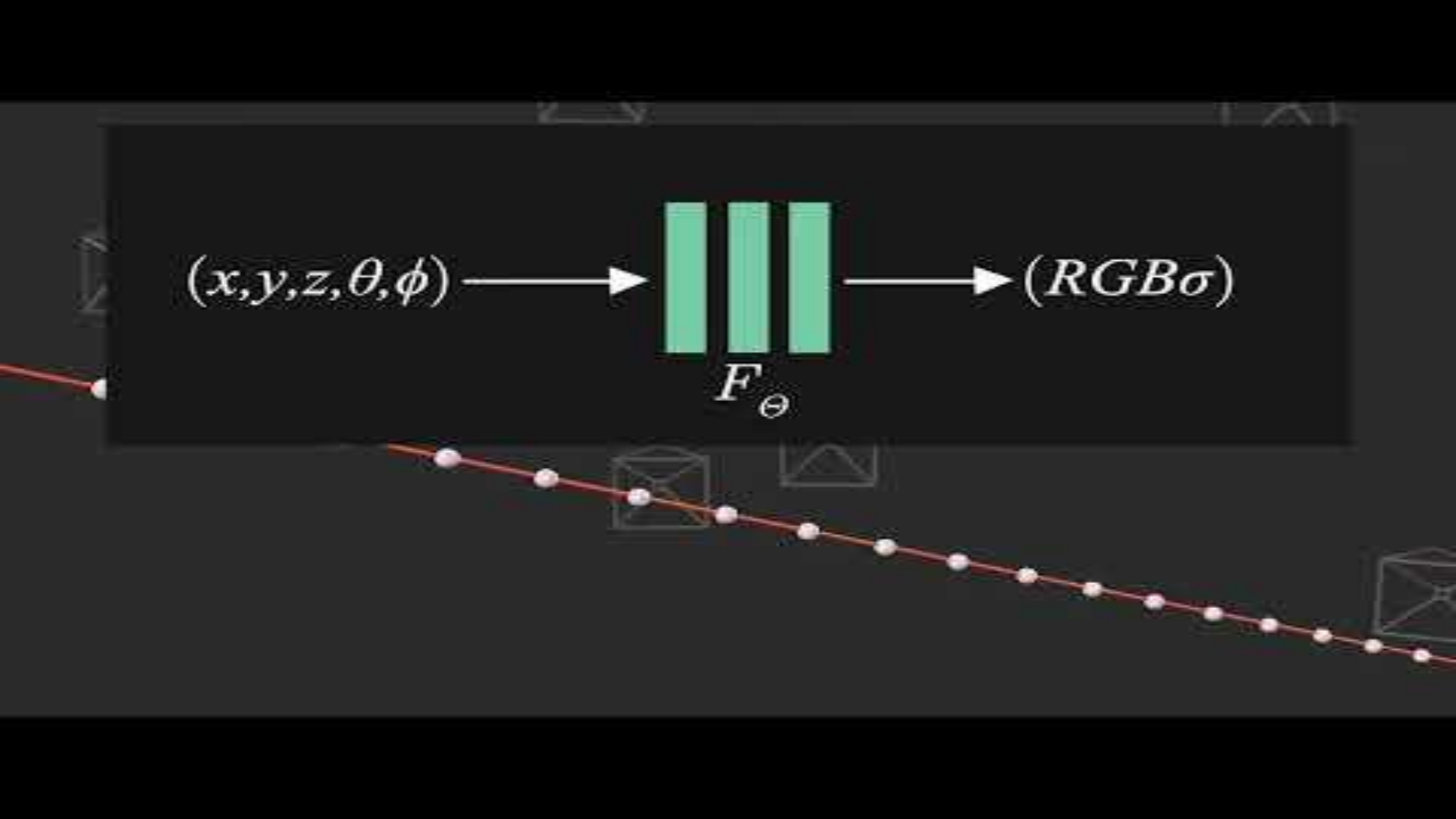
Figure 3. Examples of aligned models in the chair, laptop, bench, and airplane synsets.

А что мы умеем делать в 3d?

NeRF — Neural Radiance Fields

- На вход: 3 координаты в пространстве и 2 координаты камеры — (x, y, z, θ, φ)
- На выход: цвет и прозрачность пикселя с координатами (x, y, z), если на него смотреть из (θ, φ)

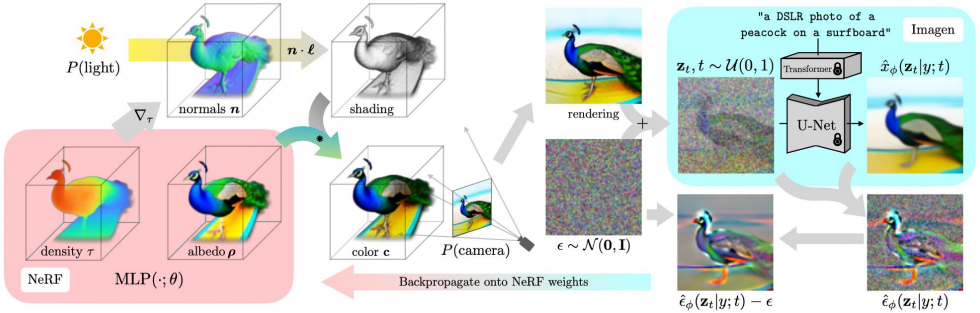



(x,y,z,θ,ϕ)
$(RGB\sigma)$
F_Θ

Давайте обучим NeRF! Но где взять изображения?

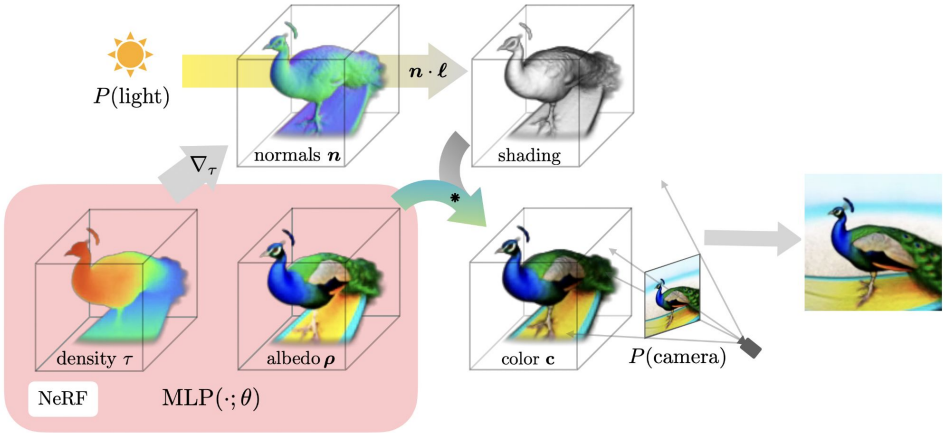
Этап обучения:

- Инициализирует NeRF случайными весами
- Генерируем изображение с виртуальной камеры
- Спрашиваем у DM, насколько изображение подходит под текст
- Делаем градиентный спуск для параметров NeRF-а (DM не обновляем)



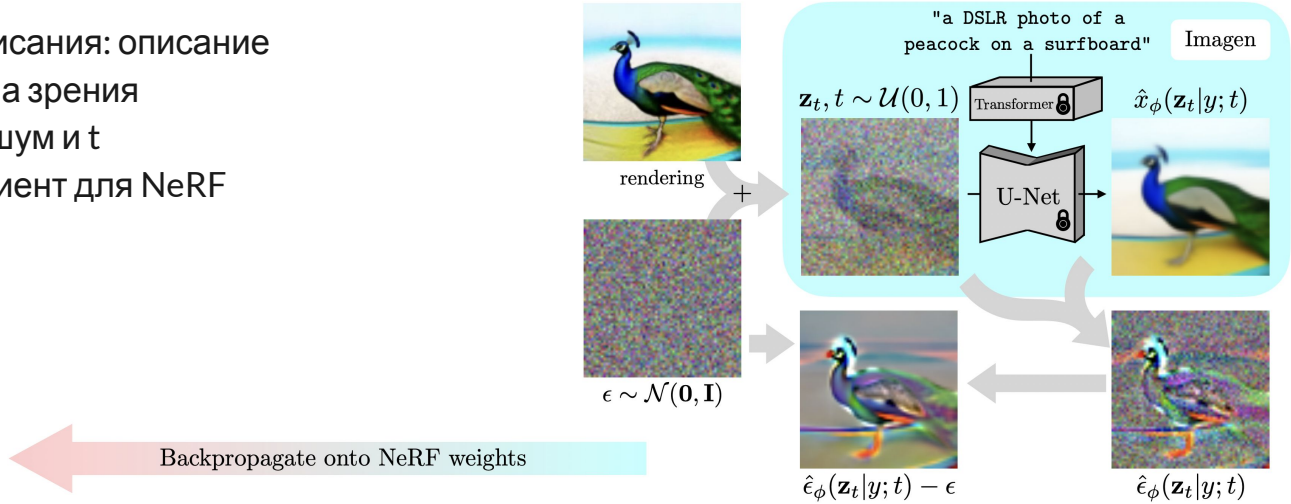
Генерация изображения с виртуальной камеры

- Shading: предсказываем альбедо и плотность
- Случайная замена альбедо на белый цвет: генерируем только тени
- Второй MLP для фона: позиционные кодировки направления луча



Обратная связь от DM

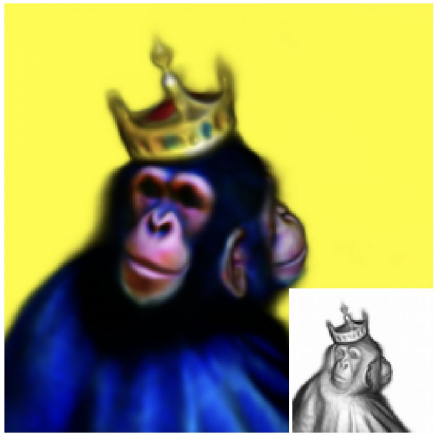
- Текстовые описания: описание картинки + угла зрения
- Сэмплируем шум и t
- Считаем градиент для NeRF



Обратная связь от DM: текстовые описания

Как заставлять модельку поощрять за правильно повёрнутую картинку?

- В текстовое описание добавляем эмбеддинг поворота изображения
  - overhead view, front view, back view
  - Взвешиваем в зависимости от угла обзора
- Работает, но неидеально



Обратная связь от LDM: функция потерь

Стандартная функция потерь:

$$\nabla_\theta \mathcal{L}_{\text{Diff}}(\phi, \mathbf{x} = g(\theta)) = \mathbb{E}_{t,\epsilon}\left[w(t) \underbrace{(\hat{\epsilon}_\phi(\mathbf{z}_t; y, t) - \epsilon)}_{\text{Noise Residual}} \underbrace{\frac{\partial \hat{\epsilon}_\phi(\mathbf{z}_t; y, t)}{\mathbf{z}_t}}_{\text{U-Net Jacobian}} \underbrace{\frac{\partial \mathbf{x}}{\partial \theta}}_{\text{Generator Jacobian}} \right]$$

По заявлению авторов, якобиан U-Net плохо обусловлен и считать его дорого, поэтому используетс

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x} = g(\theta)) \triangleq \mathbb{E}_{t,\epsilon}\left[w(t) (\hat{\epsilon}_\phi(\mathbf{z}_t; y, t) - \epsilon) \frac{\partial \mathbf{x}}{\partial \theta} \right]$$

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

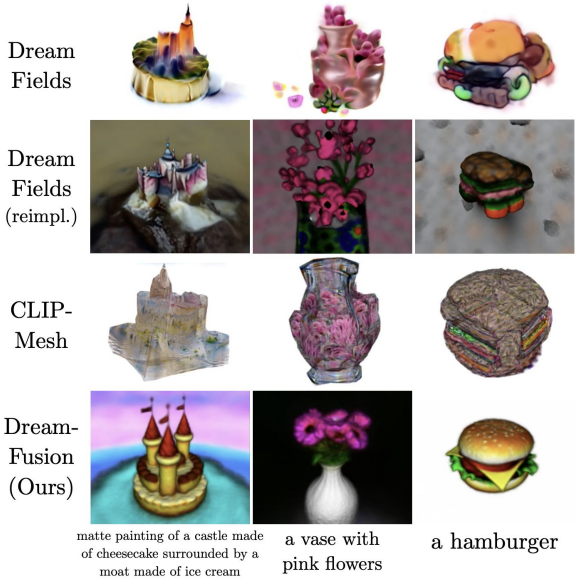
Как использовать обученную модель?

Генерация семпла – обучение NeRF-а с нуля!

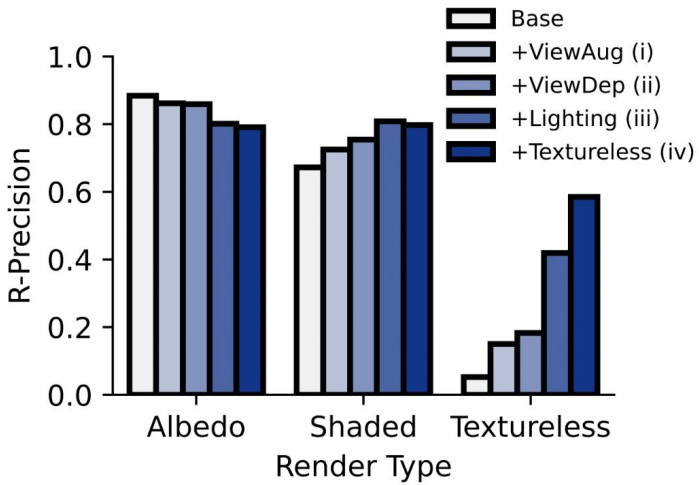
- Выкидываем DM
- Семплим картинки
- При желании можно построить 3d модель

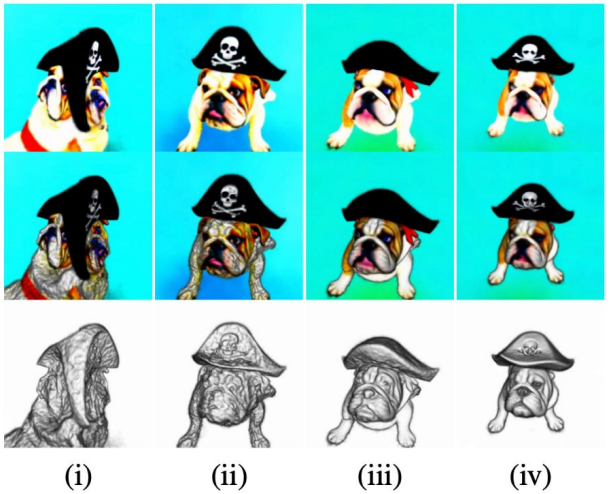
Эксперименты: сравнение с другими моделями

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**

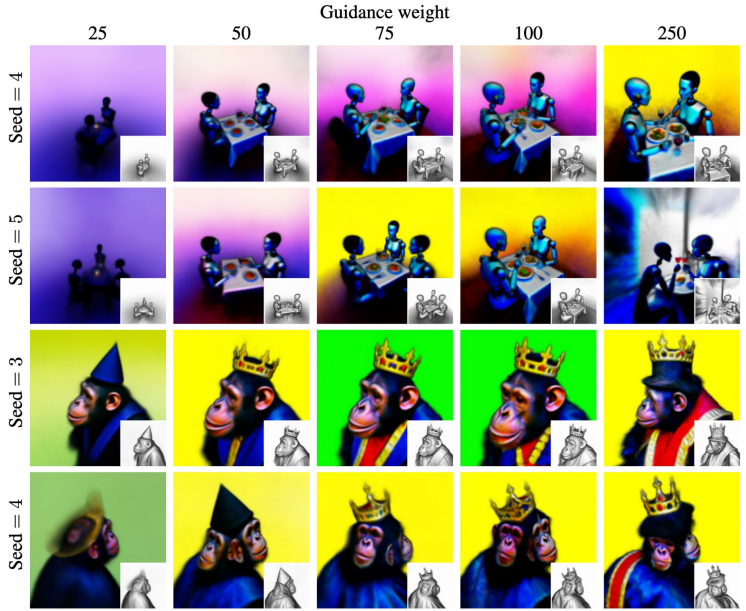


Эксперименты: изучение прироста качества от улучшений

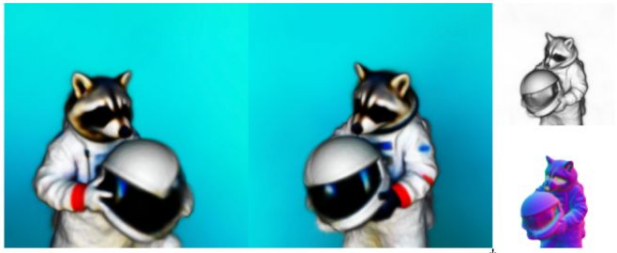




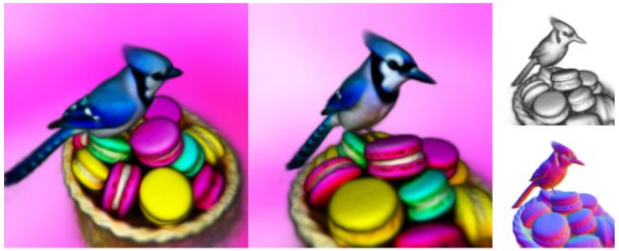
Эксперименты: изучение параметра guidance



an orangutan making a clay bowl on a throwing wheel*

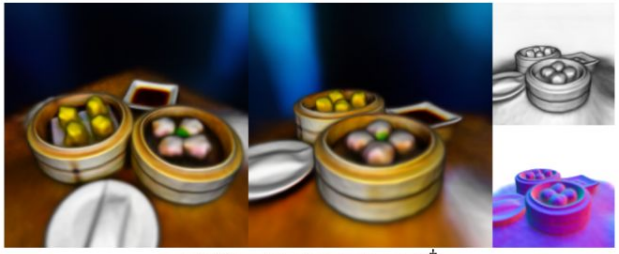


a raccoon astronaut holding his helmet†



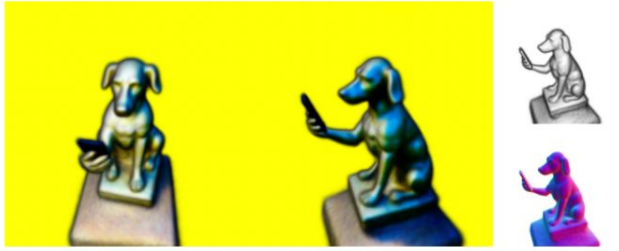
a blue jay standing on a large basket of rainbow macarons*

a corgi taking a selfie*



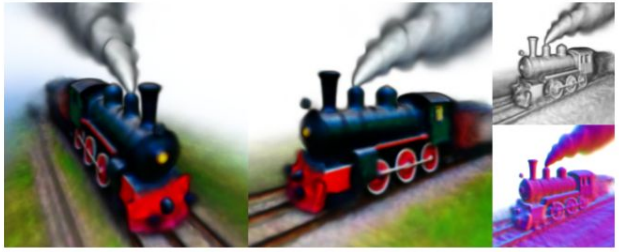
a table with dim sum on it†

a lion reading the newspaper*



Michelangelo style statue of dog reading news on a cellphone

a tiger dressed as a doctor*



a steam engine train, high resolution*

a frog wearing a sweater*

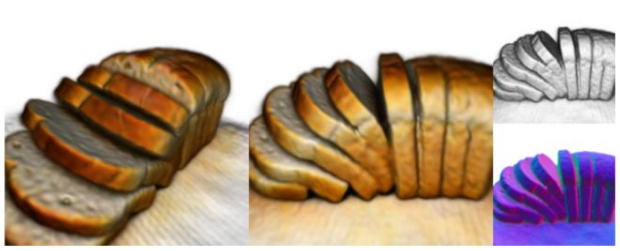
a humanoid robot playing the cello*

Sydney opera house, aerial view†

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*

Спасибо!

- Статья: https://arxiv.org/pdf/2209.14988.pdf
- Демонстрация: https://dreamfusion3d.github.io/