Лаборатория компьютерной
графики и мультимедиа
ВМК МГУ имени М.В. Ломоносова

Курс «Компьютерное зрение»

«Синтез изображений, ч.2»

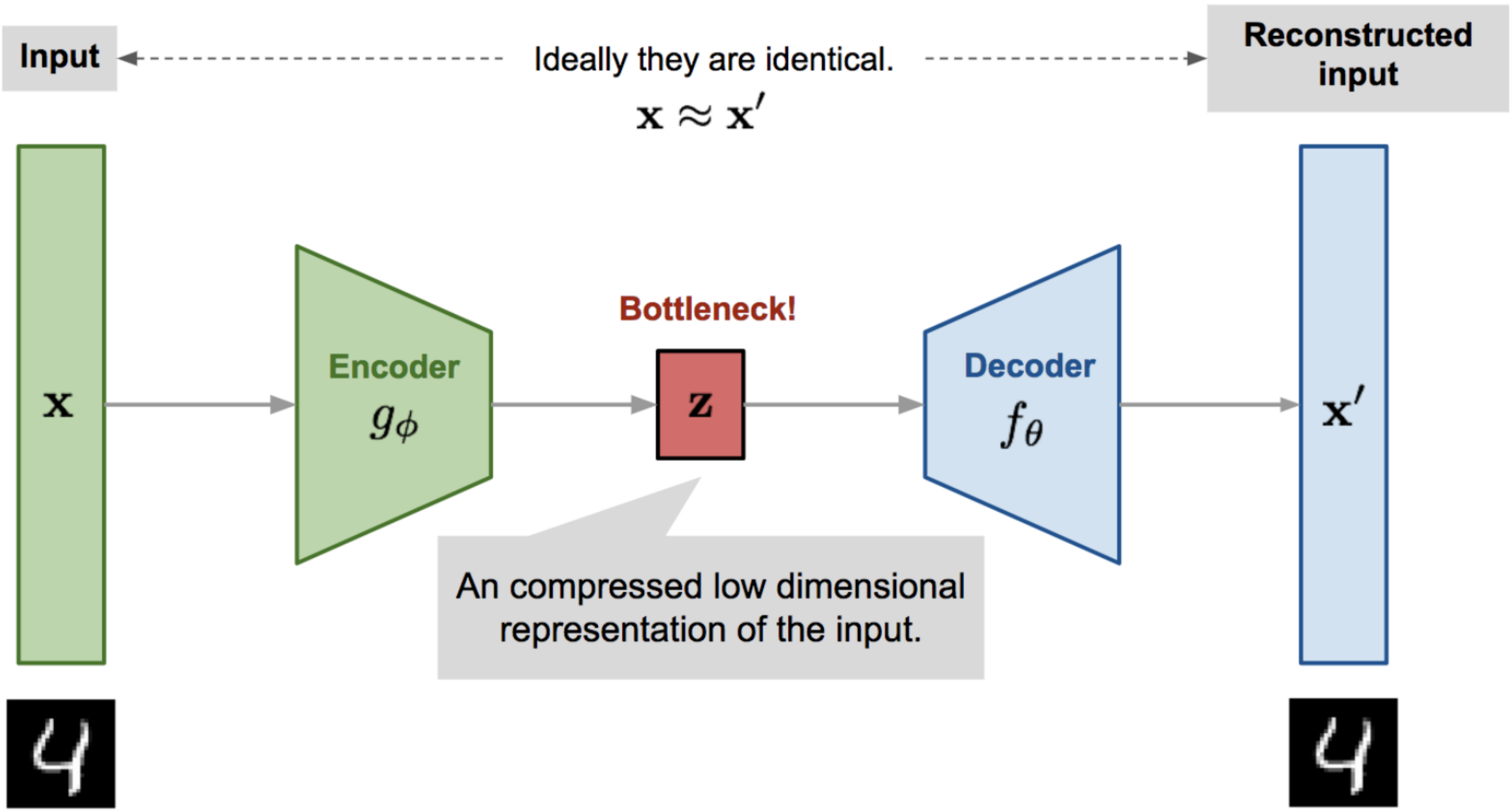
Антон Конушин и Тимур Мамедов

2025 год

План лекции

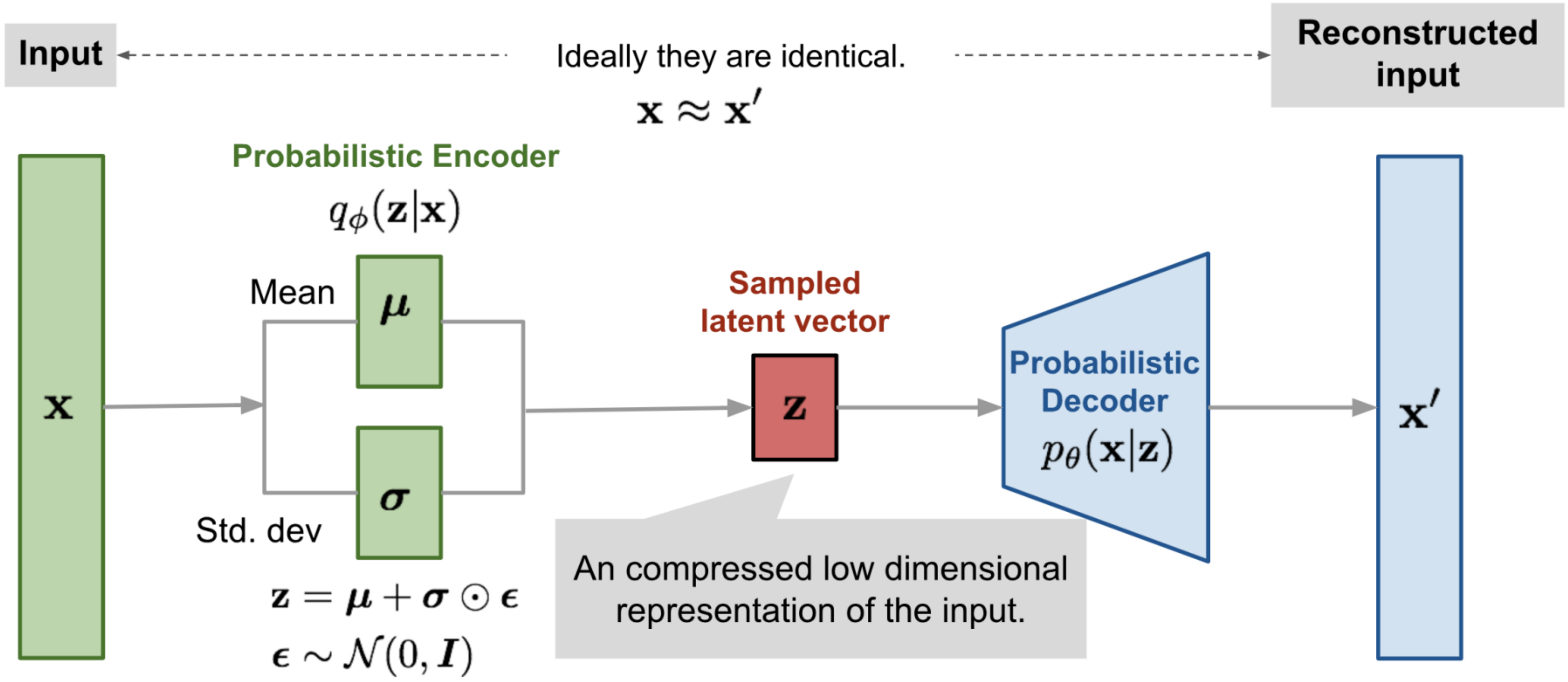
1. Вариационный автокодировщик (VAE)
2. Векторные квантованные VAE (VQ-VAE)
3. Диффузионные модели

Автокодировщик



Hinton et al. Reducing the dimensionality of data with neural networks. Science 2006

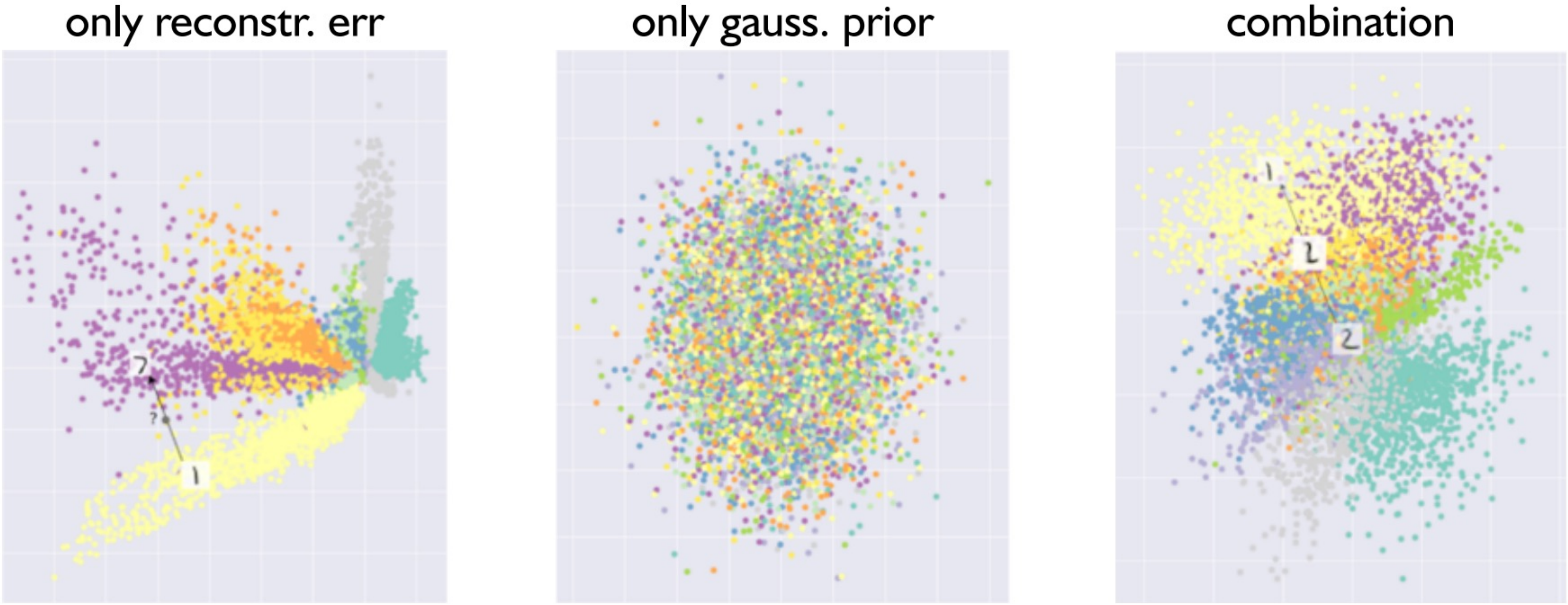
Вариационный автокодировщик



Kingma, Welling. Auto-Encoding Variational Bayes. ICLR 2014

Функция потерь VAE

$$\mathscr{L} = \underbrace{d(\boldsymbol{x}, \boldsymbol{x}')}_{\text{reconstr. err}} + \underbrace{KL(q_\phi(\boldsymbol{z}\,|\,\boldsymbol{x}) \,||\, \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}))}_{\text{gauss. prior on latent distribution}}$$



Kingma, Welling. Auto-Encoding Variational Bayes. ICLR 2014

Reparametrization trick

To compute gradients ∇_ϕ through sampling, substitute sampling

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}_\phi, \boldsymbol{\sigma}^2_\phi)$$

with deterministic function that depends on separate noise variable

$$\boldsymbol{z} = \boldsymbol{\mu}_\phi + \boldsymbol{\sigma}_\phi \odot \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

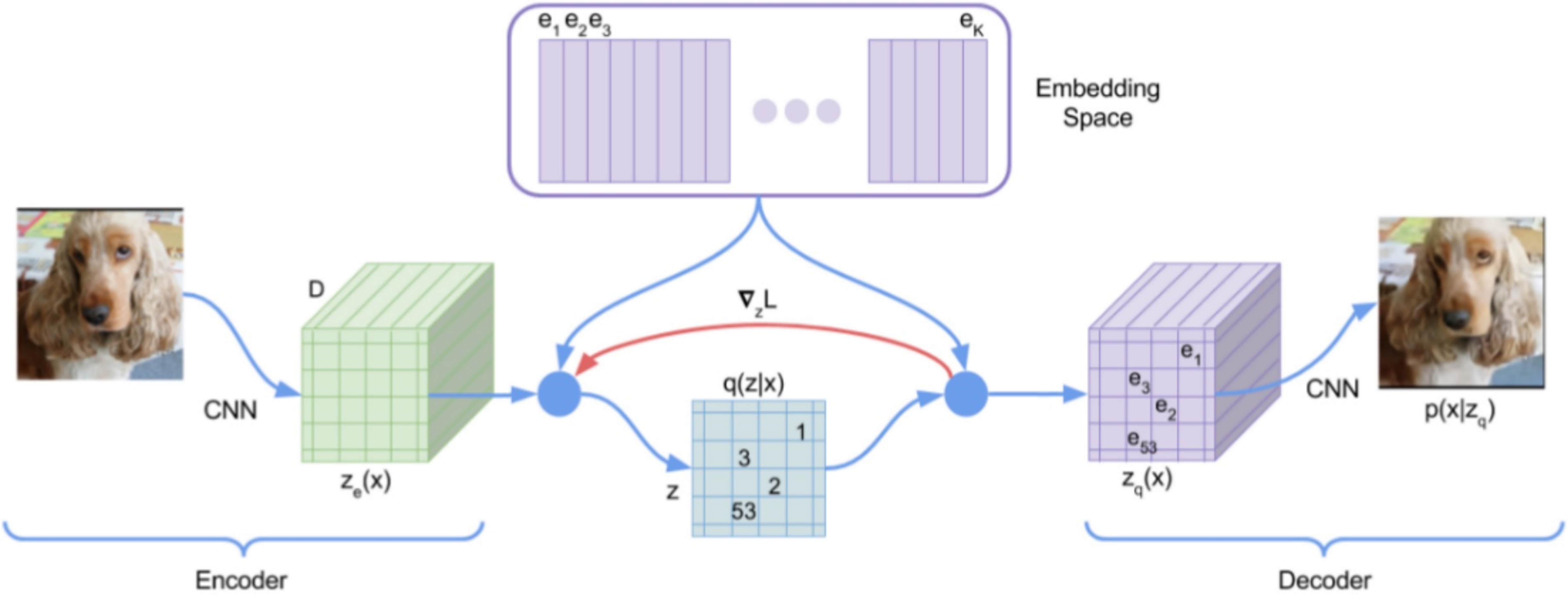
This way sampled values $\boldsymbol{\varepsilon}$ during backpropagation may be seen as constant that don't depend on encoder parameters ϕ

Kingma, Welling. Auto-Encoding Variational Bayes. ICLR 2014

План лекции

1. Вариационный автокодировщик (VAE)
2. Векторные квантованные VAE (VQ-VAE)
3. Диффузионные модели

VQ-VAE



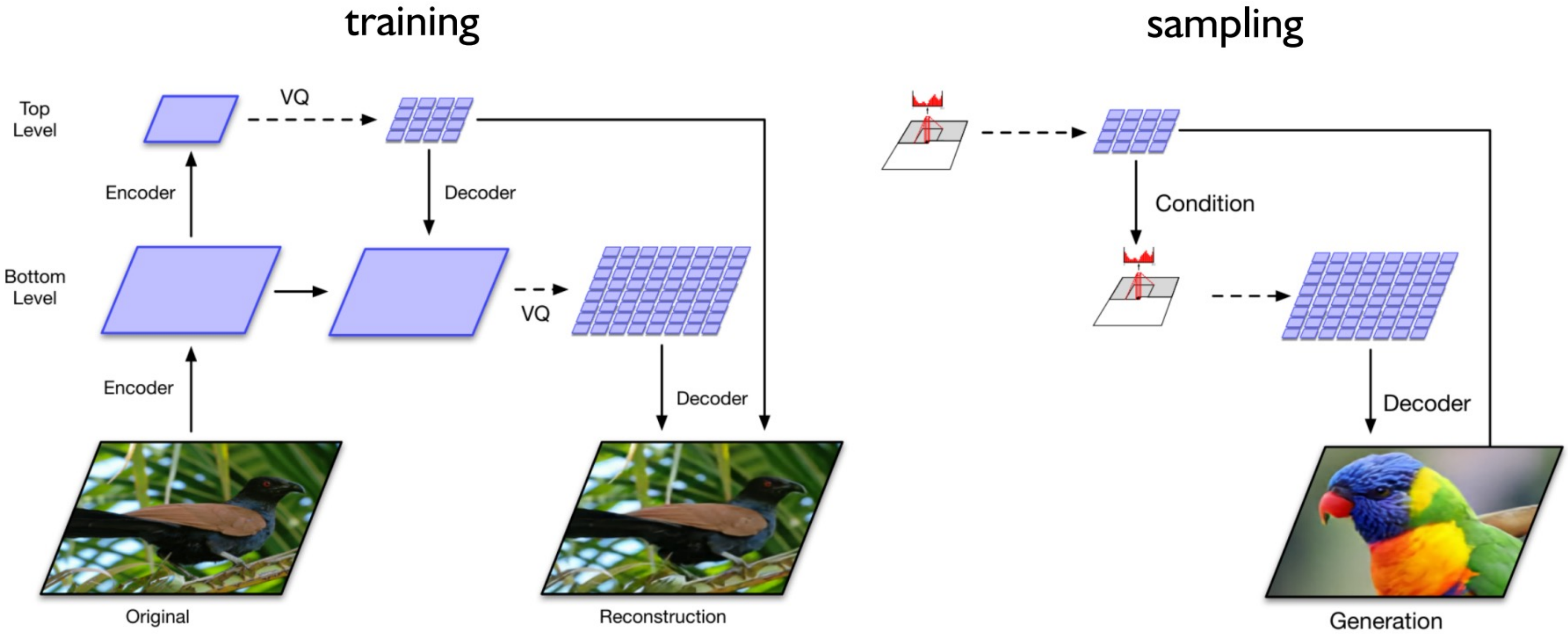
Oord et al. Neural Discrete Representation Learning. NeurIPS 2017

VQ-VAE learning

$$\mathcal{L} = \underbrace{d(x, x')}_{\text{reconstr. err}} + \underbrace{\left\| \mathrm{sg}\left[z_e(x)\right] - e \right\|_2^2}_{\text{dict learning}} + \underbrace{\beta \left\| z_e(x) - \mathrm{sg}\left[e\right] \right\|_2^2}_{\text{regularization of image features}}$$

Here sg is a stop gradient function, which makes operand non-updated constant. k-means step with EMA may be used to update dictionary instead of the second term

VQ-VAE-2



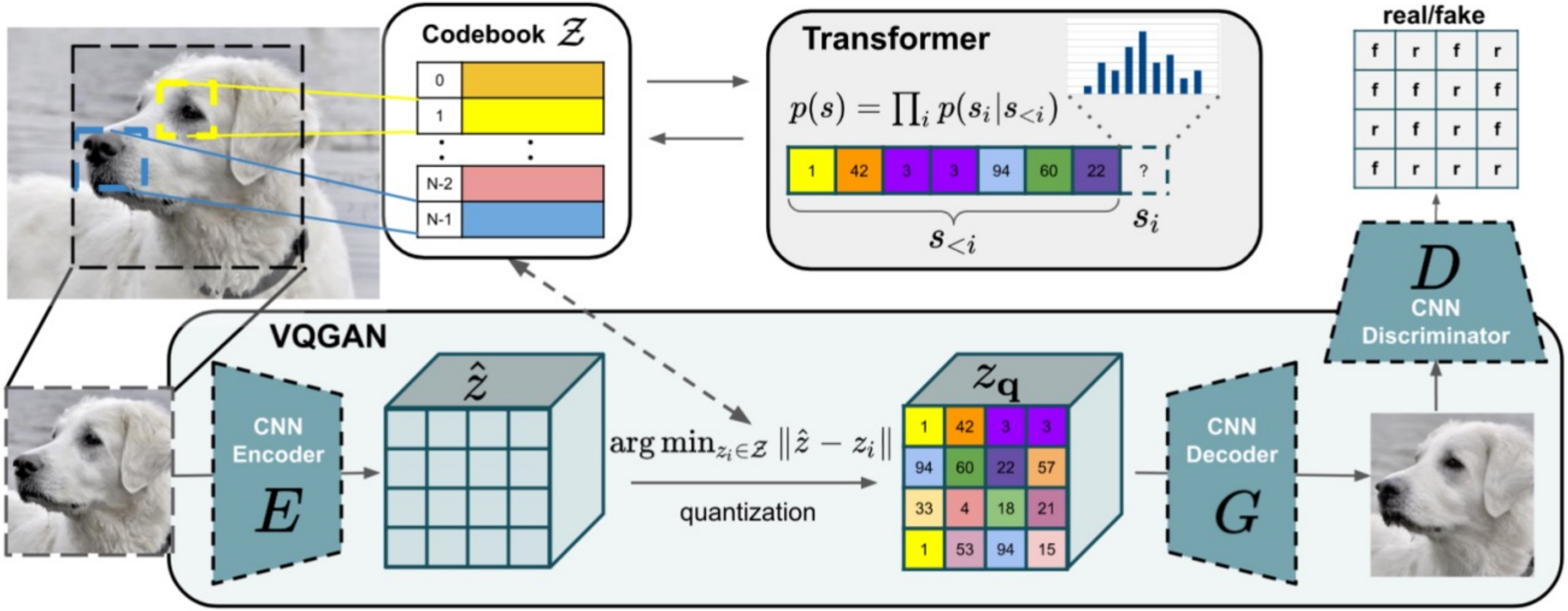
Razavi et al. Generating Diverse High-Fidelity Images with VQ-VAE-2. NeurIPS 2019

Пример реконструкции VQ-VAE-2

h_{top} | $h_{\text{top}}, h_{\text{middle}}$ | $h_{\text{top}}, h_{\text{middle}}, h_{\text{bottom}}$ | Original

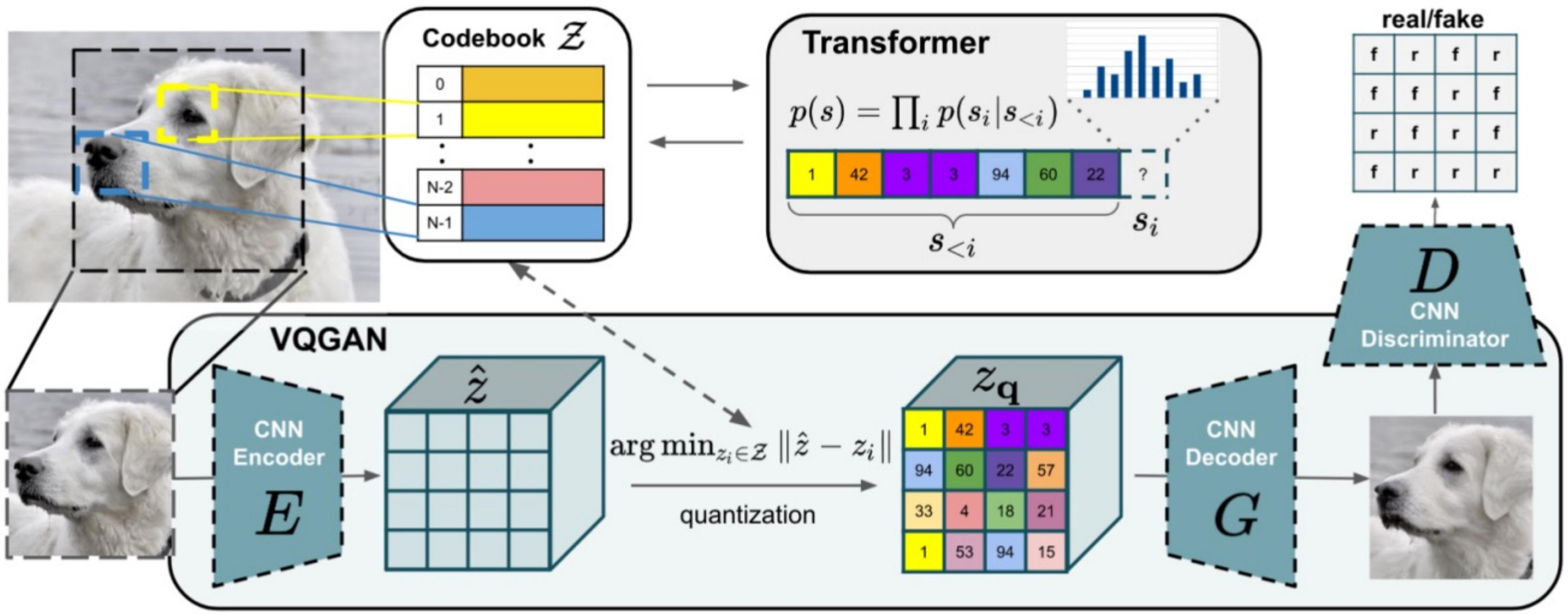
Razavi et al. Generating Diverse High-Fidelity Images with VQ-VAE-2. NeurIPS 2019

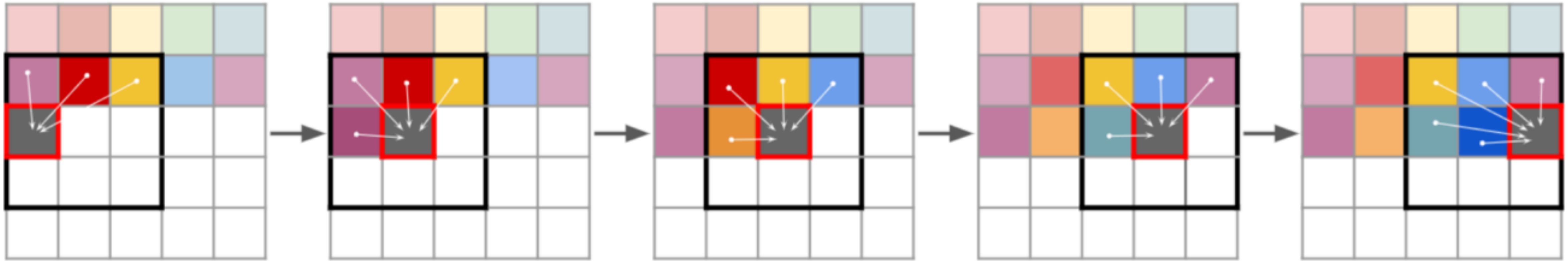
VQ-GAN



Esser et al. Taming Transformers for High-Resolution Image Synthesis. CVPR 2021

VQ-GAN

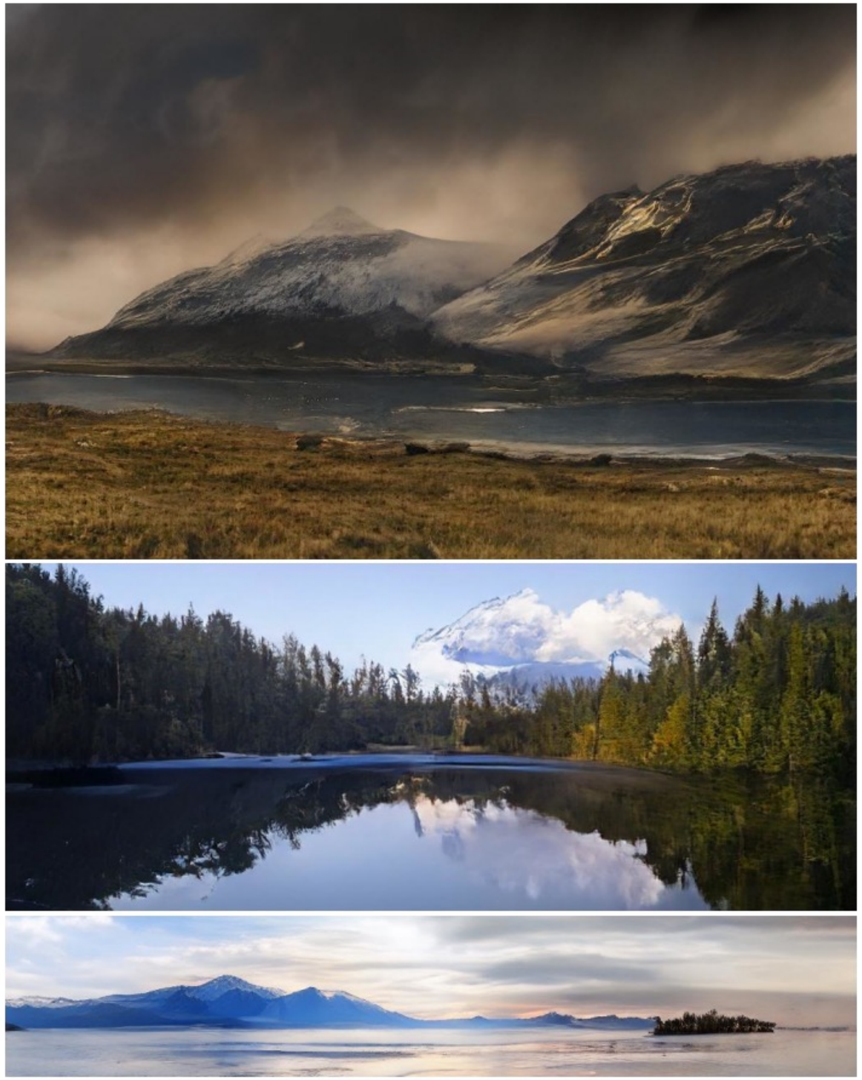




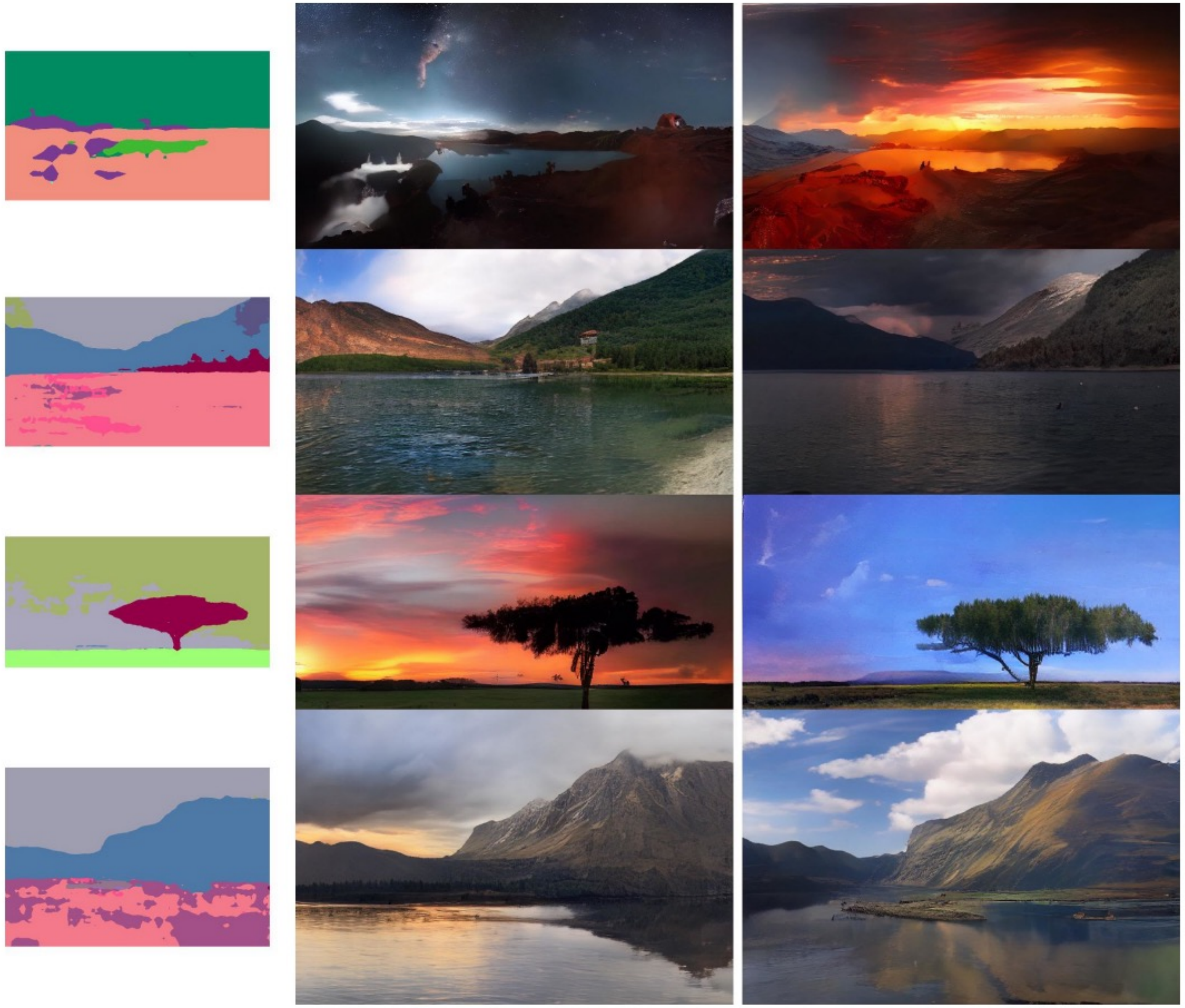
Esser et al. Taming Transformers for High-Resolution Image Synthesis. CVPR 2021

Примеры результатов VQ-GAN

unconditional



conditional



DALL·E

Large-scale dVAE (similar to VQ-VAE) + transformer prior:
• 250M (image, text) pairs
• image VAE was trained on 64×V100 GPUs
• ResNets as image encoder and decoder
• image dict size K = 8192, 256×256 → 32×32
• transformer (12B params) was trained on 1024×V100 GPUs
• efficient implementation of distributed mixed precision training
(params, activations and Adam moments are stored in 16 bits)

Ramesh et al. Zero-Shot Text-to-Image Generation. arXiv:2102.12092

DALL·E

(a) a tapir made of accordion. a tapir with the texture of an accordion.

(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

(c) a neon sign that reads “backprop”. a neon sign that reads “backprop”. backprop neon sign

(d) the exact same cat on the top as a sketch on the bottom

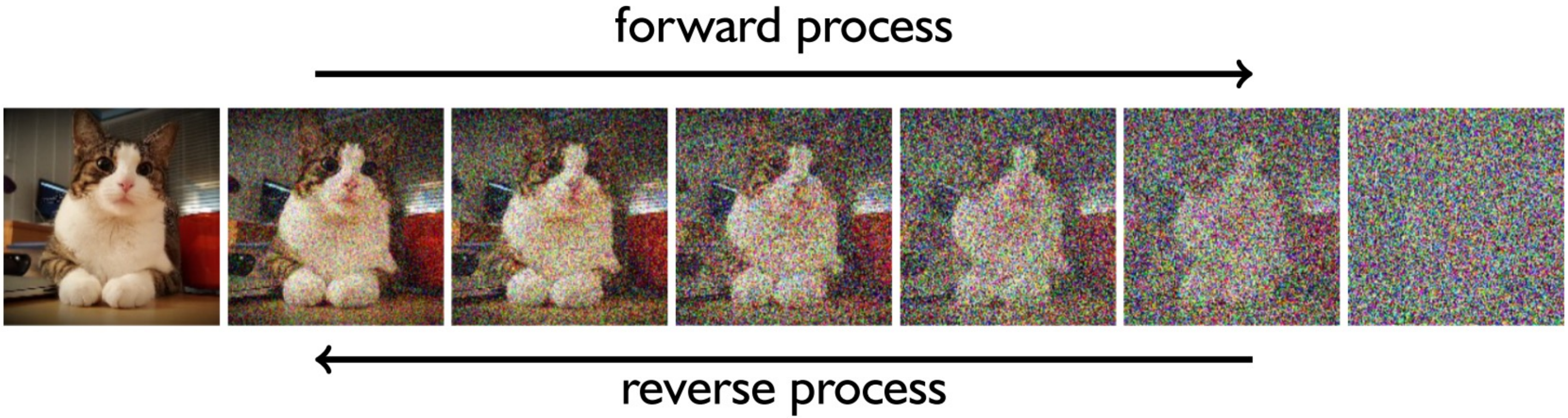
Quality of samples may be improved using CLIP scoring

Ramesh et al. Zero-Shot Text-to-Image Generation. arXiv:2102.12092

План лекции

1. Вариационный автокодировщик (VAE)
2. Векторные квантованные VAE (VQ-VAE)
3. Диффузионные модели

Denoising diffusion probabilistic models

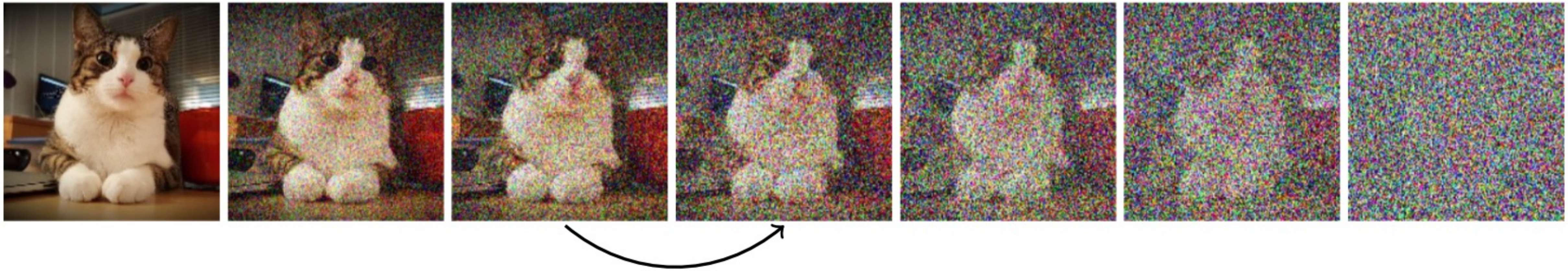


DDPM consists of two processes:

- forward diffusion process gradually adds gaussian noise to input
- reverse denoising process that learns to generate data by denoising

Ho et al. Denoising Diffusion Probabilistic Models. NeurIPS 2020

Diffusion (forward process)



$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I})$$

Forward sampling using reparametrization trick:

$$\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

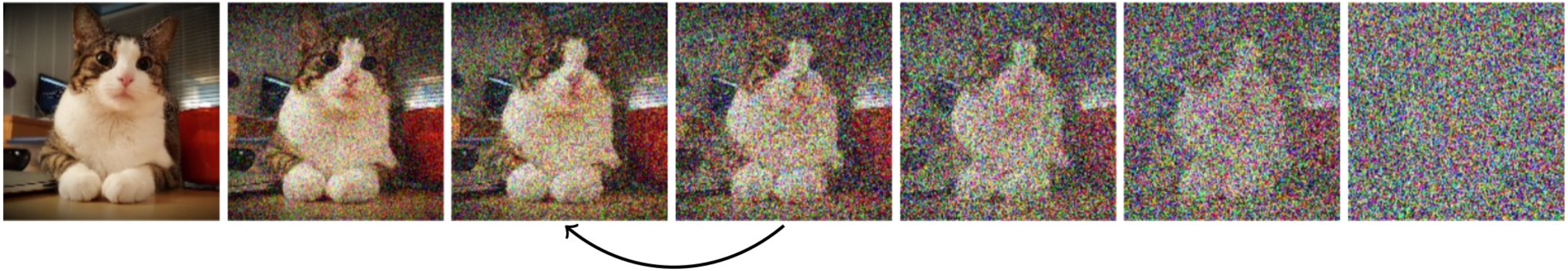
$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}, \quad \bar{\alpha}_t = \Pi_{i=1}^{t}(1-\beta_i)$$

β_t is a noise schedule designed s.t.

$$q(\boldsymbol{x}_T|\boldsymbol{x}_0) \approx \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

Ho et al. Denoising Diffusion Probabilistic Models. NeurIPS 2020

Diffusion (reverse process)



$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \sigma_t^2 \boldsymbol{I})$$

Substitute $\boldsymbol{\mu}_\theta$ with $\boldsymbol{\varepsilon}_\theta$ due to reparametrization trick. Sampling:

$\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$

for $t = T, \dots, 1$:

$\quad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$

$\quad \boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)\right) + \sigma_t \boldsymbol{z}$

return $\boldsymbol{x}_0$

$\boldsymbol{\varepsilon}_\theta$ — trained UNet

α_t, σ_t — diff. parameters

Ho et al. Denoising Diffusion Probabilistic Models. NeurIPS 2020

Learning to denoise at any step

denoise $\boldsymbol{x}_t$

repeat until converged:

$\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$

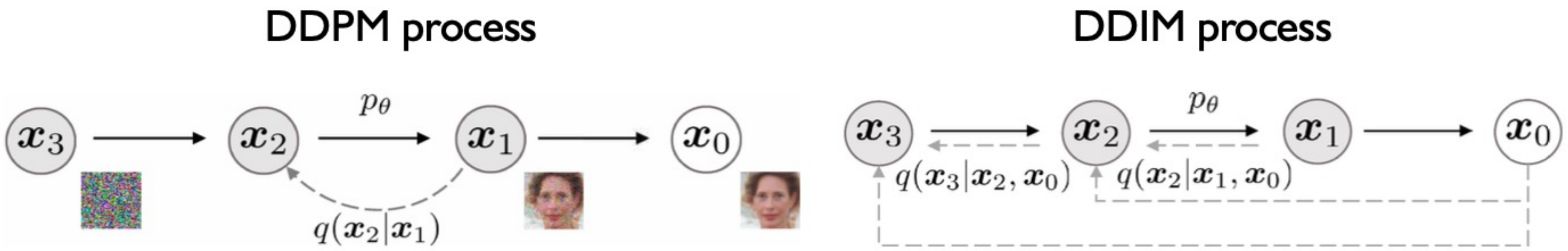
$t \sim \mathcal{U}\{1, T\}$

$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$

Take gradient step on

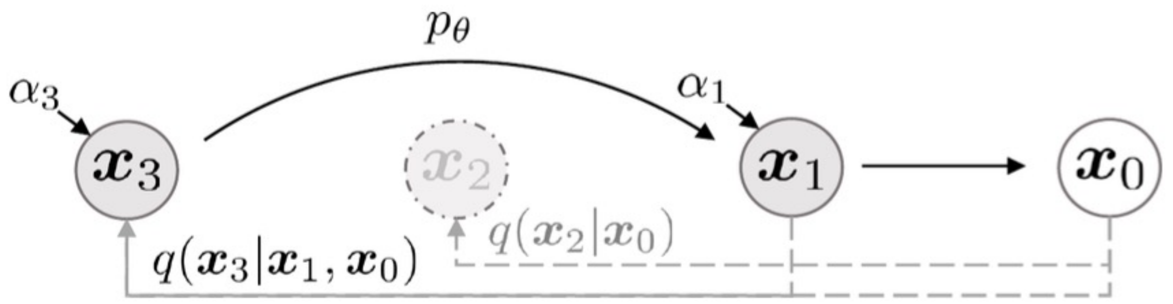
$$\nabla_\theta \left\| \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\theta \left(\sqrt{\bar{\alpha}_t} \boldsymbol{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\varepsilon}, t \right) \right\|^2$$

DDIM



Song et al. Denoising Diffusion Implicit Models. ICLR 2021

DDIM



$$\boldsymbol{x}_j = \sqrt{\alpha_j}\underbrace{\left(\frac{\boldsymbol{x}_i - \sqrt{1-\alpha_i}\cdot\boldsymbol{\varepsilon}(\boldsymbol{x}_i)}{\sqrt{\alpha_i}}\right)}_{\text{predicted } \boldsymbol{x}_0} + \underbrace{\sqrt{1-\alpha_j-\sigma_i{}^2}\cdot\boldsymbol{\varepsilon}(\boldsymbol{x}_i)}_{\text{move back towards } \boldsymbol{x}_t} + \underbrace{\sigma_i\boldsymbol{z}}_{\text{random noise}}$$

Резюме лекции

- Вариационные автокодировщики – важный этап в области генерации изображений, оказавшие сильное влияние на исследования в этом направлении
- Векторные квантованные VAE позволили генерировать изображения в высоком разрешении, а также отказаться от условия «нормальности» распределения
- Диффузионные модели – квинтэссенция идей GAN и VAE
- Диффузионные модели являются современными подходами к генерации контента