知能工学特別講義 第７講

担当：和田山 正

名古屋工業大学

深層生成モデル

- 深層生成モデルとは
- 拡散モデル
- Stable Diffusionの内部構造

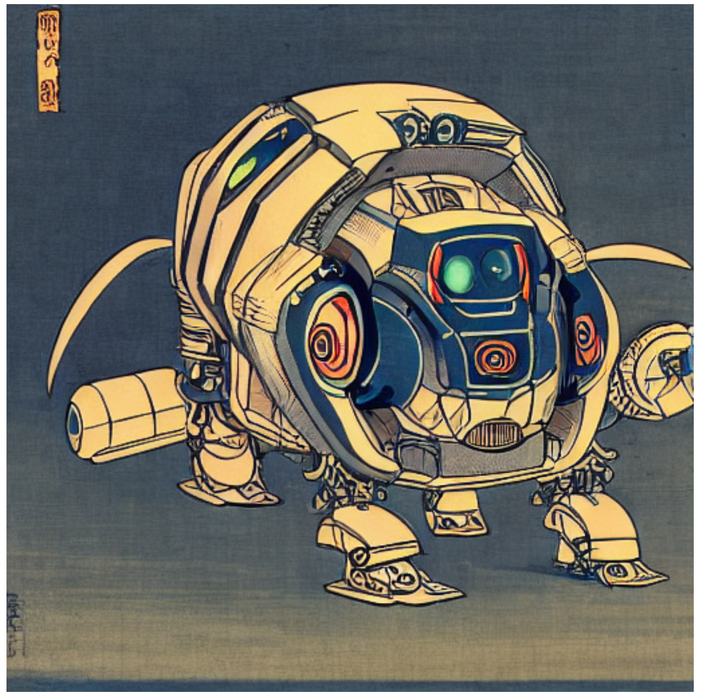
深層生成モデルによる画像生成の登場 (2022)

Stable Diffusion

https://huggingface.co/spaces/stabilityai/stable-diffusion

prompt:
mecha robot hamster
in photo style

prompt:
mecha robot hamster
in Hokusai style



prompt:
hamster
in Claude Monnet style

深層生成モデルによる画像生成の登場 (2022)

乗馬する宇宙飛行士の写真

Stable Diffusion の元ネタ論文(2022, Apr)

High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach[1] * Andreas Blattmann[1] * Dominik Lorenz[1] Patrick Esser[R] Björn Ommer[1]

[1]Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany [R]Runway ML

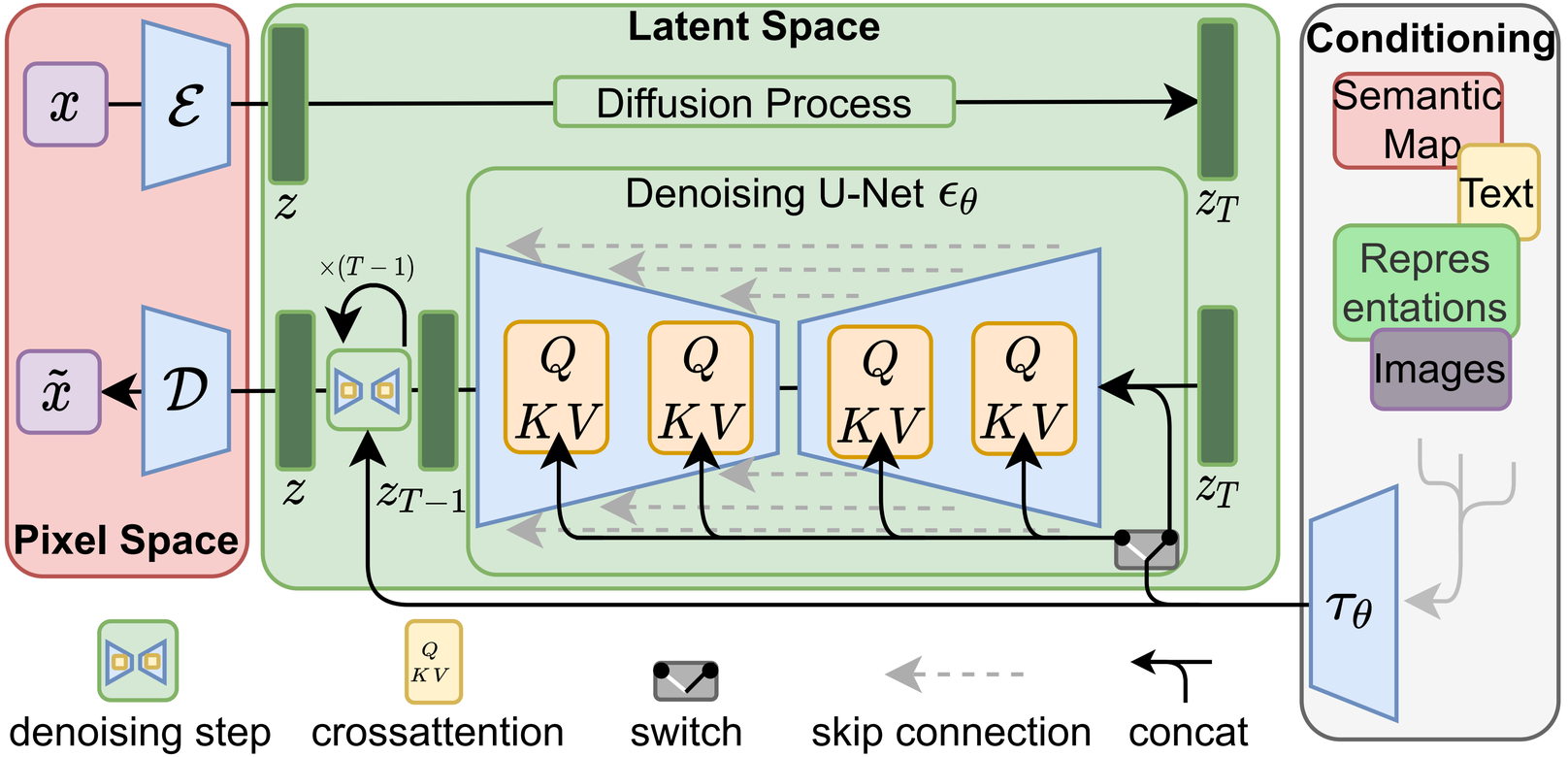
https://github.com/CompVis/latent-diffusion

Abstract

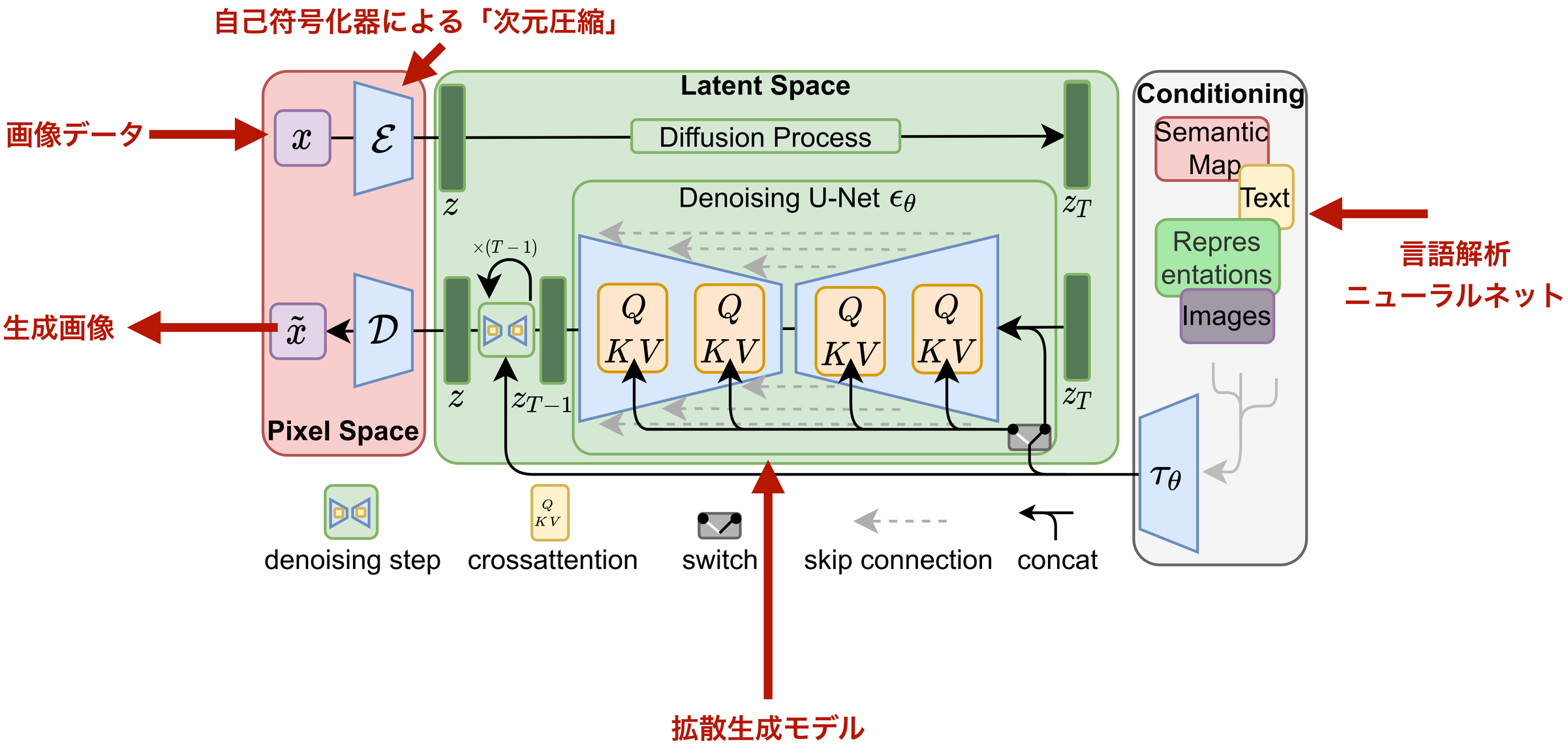
By decomposing the image formation process into a sequential application of denoising autoencoders, diffusion models (DMs) achieve state-of-the-art synthesis results on image data and beyond. Additionally, their formulation allows for a guiding mechanism to control the image generation process without retraining. However, since these models typically operate directly in pixel space, optimization of powerful DMs often consumes hundreds of GPU days and inference is expensive due to sequential evaluations. To enable DM training on limited computational resources while retaining their quality and flexibility, we apply them in the latent space of powerful pretrained autoencoders. In contrast to previous work, training diffusion models on such a representation allows for the first time to reach a near-optimal point between complexity reduction and detail preservation, greatly boosting visual fidelity. By introducing cross-attention layers into the model architecture, we turn diffusion models into powerful and flexi-

Figure 1. Boosting the upper bound on achievable quality with less agressive downsampling. Since diffusion models offer excellent inductive biases for spatial data, we do not need the heavy spatial downsampling of related generative models in latent space, but can still greatly reduce the dimensionality of the data via suitable autoencoding models, see Sec. 3. Images are from the DIV2K [1] validation set, evaluated at 512^2 px. We denote the spatial downsampling factor by f. Reconstruction FIDs [29] and PSNR are calculated on ImageNet-val. [12]; see also Tab. 8.

Stable Diffusion のブロックダイアグラム



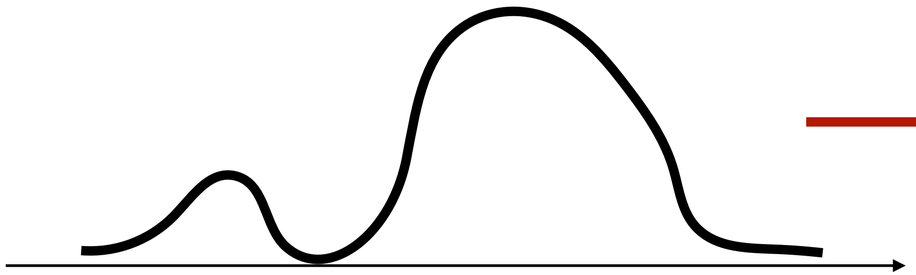
Stable Diffusion のブロックダイアグラム



生成モデルとは？

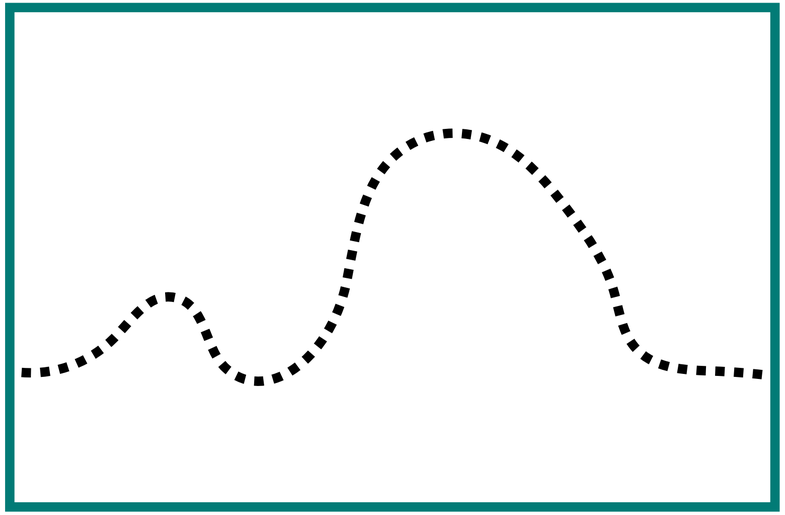
学習プロセス

データの生成分布(未知)



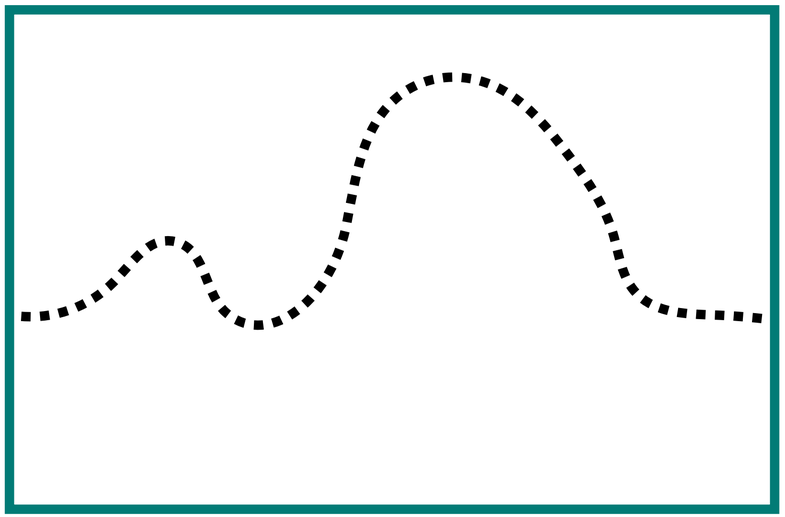
データサンプル

生成モデル(確率分布を学習)



生成プロセス

ランダムサンプル



生成モデルとは？

学習プロセス

データの生成分布(未知)

Imagenetの画像
に対応する
確率分布

→

データサンプル

→

生成モデル(確率分布を学習)

Imagenetの画像
に対応する
確率分布の近似

生成プロセス

ランダムサンプル

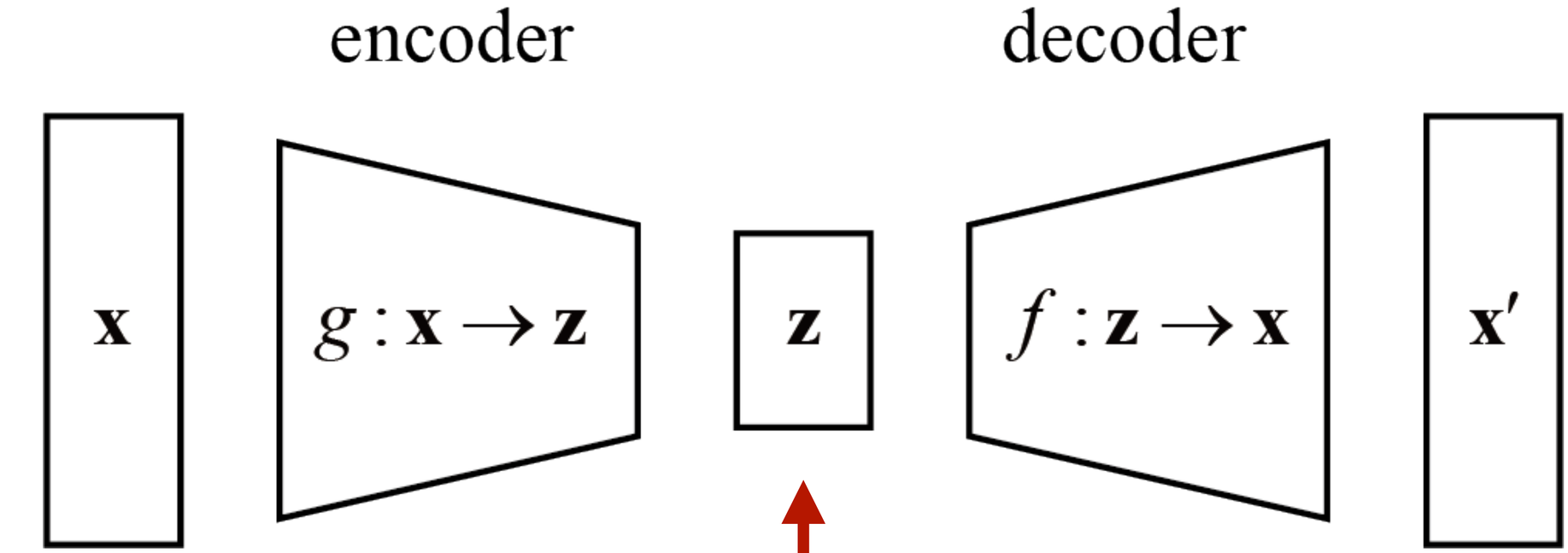
←

Imagenetの画像
に対応する
確率分布の近似

代表的な深層生成モデル

- 変分自己符号化器
- GAN(Generative Adversarial Network, 敵対的生成ネットワーク)
- フローモデル (確率密度関数を少しづつ可逆変換していく）
- 拡散モデル

変分自己符号化器(Variational Autoencoder)

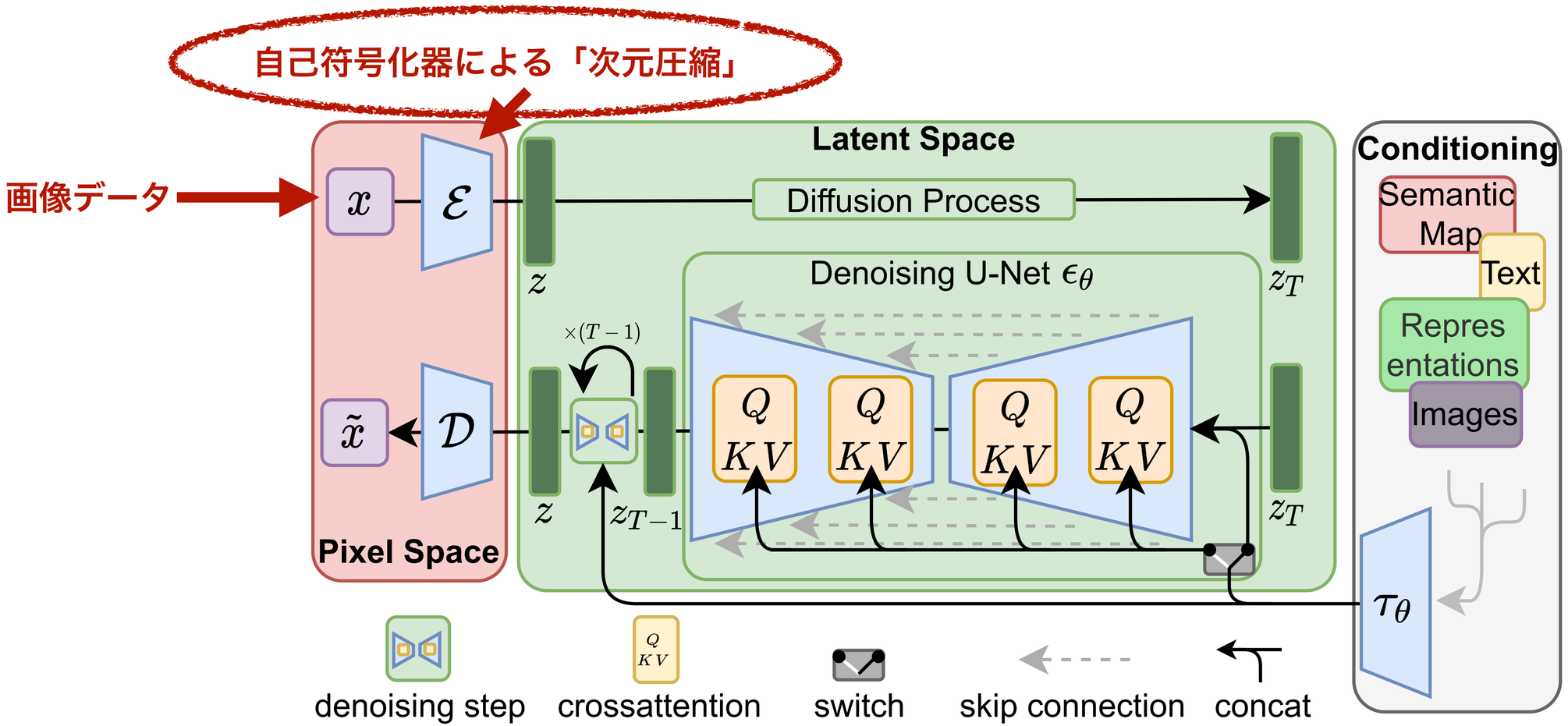


入力ベクトル

ガウス分布
N(0,I)となるように
学習される

出力ベクトル
入力ベクトル ≒ 出力ベクトル
となるように学習される

Stable Diffusion のブロックダイアグラム



拡散モデルの構造

Denoising Diffusion Probabilistic Models

J.Ho, A. Jain, and P. Abbeel, 2020

逆拡散プロセス

マルコフ連鎖

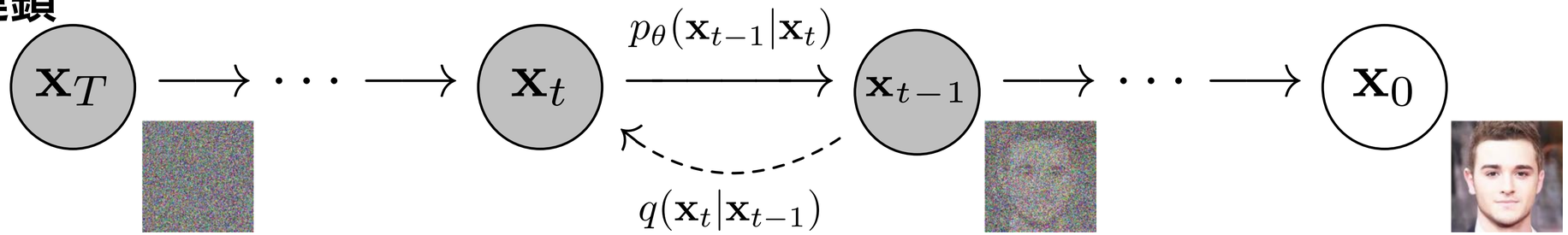


Figure 2: The directed graphical model considered in this work.

拡散プロセス

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

拡散モデルのアイデア

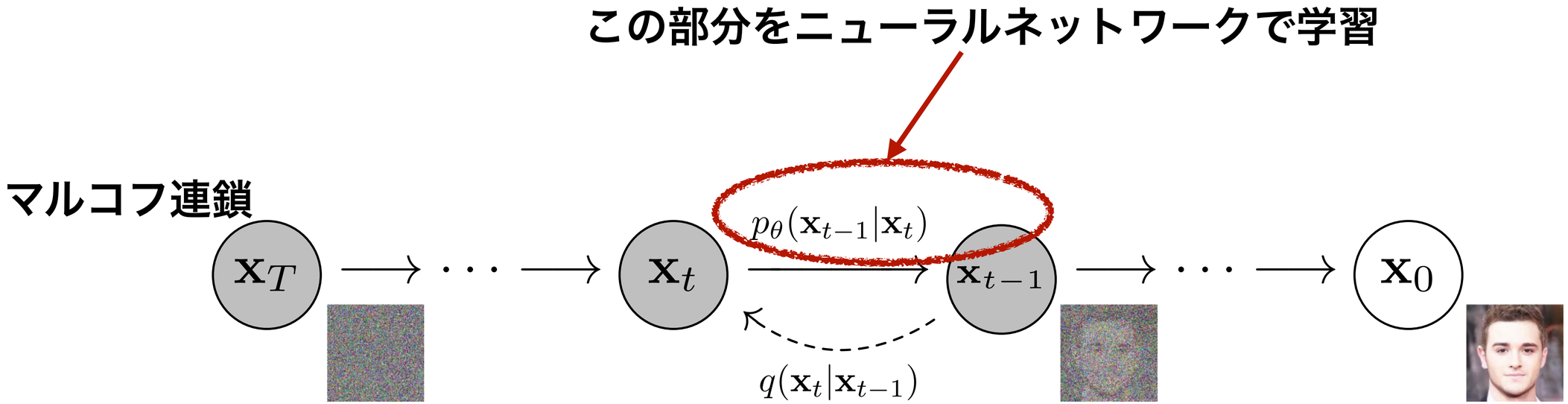
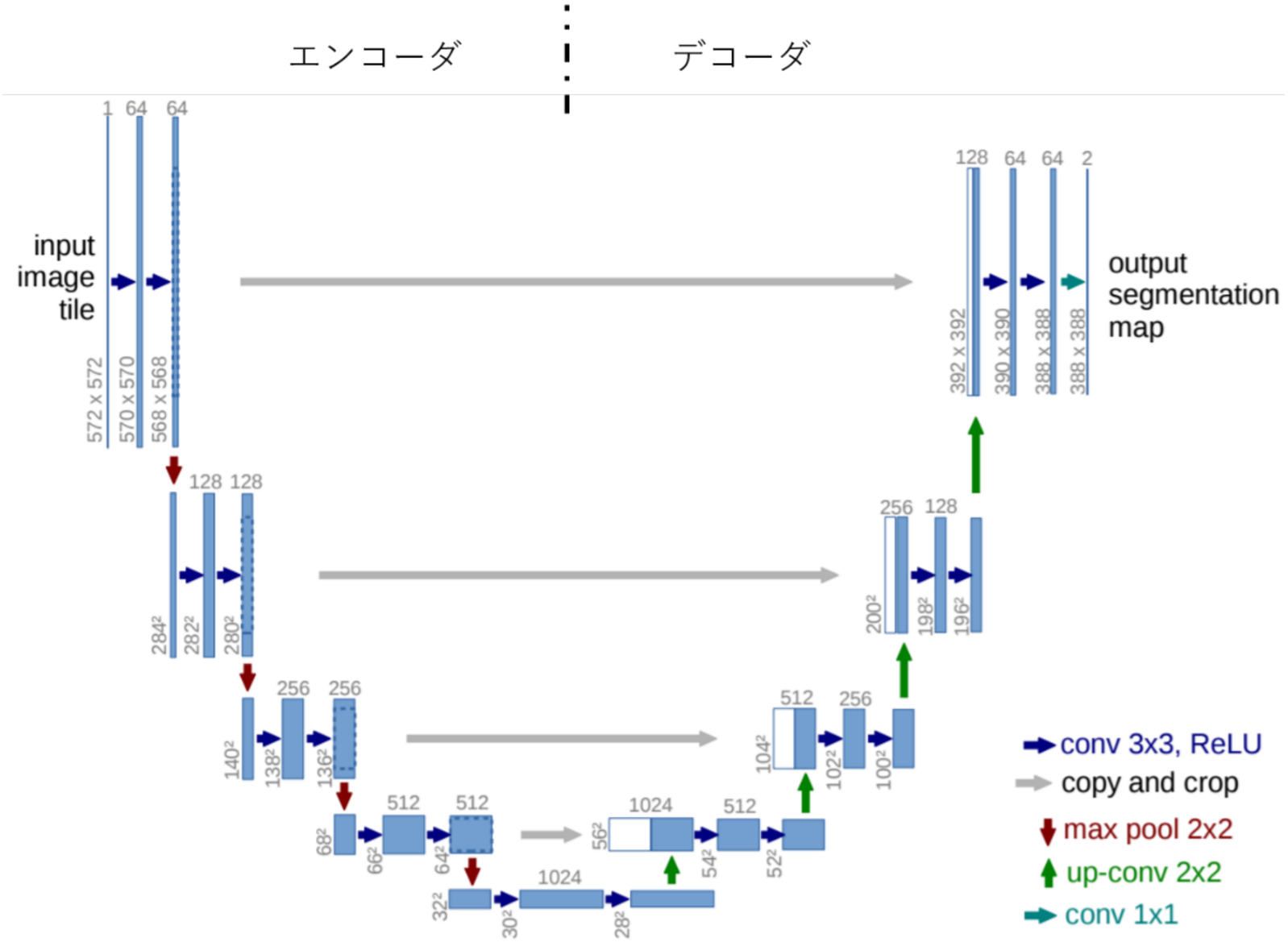


Figure 2: The directed graphical model considered in this work.

- ランダムベクトルを初期値として左から右に計算を進める
- 一種のノイズ除去ネットワークの学習とみることもできる
- ベクトルと時刻を入力とするネットワークが必要

U-Net


エンコーダ
デコーダ
input image tile
output segmentation map
conv 3x3, ReLU
copy and crop
max pool 2x2
up-conv 2x2
conv 1x1

学習過程とサンプル過程

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

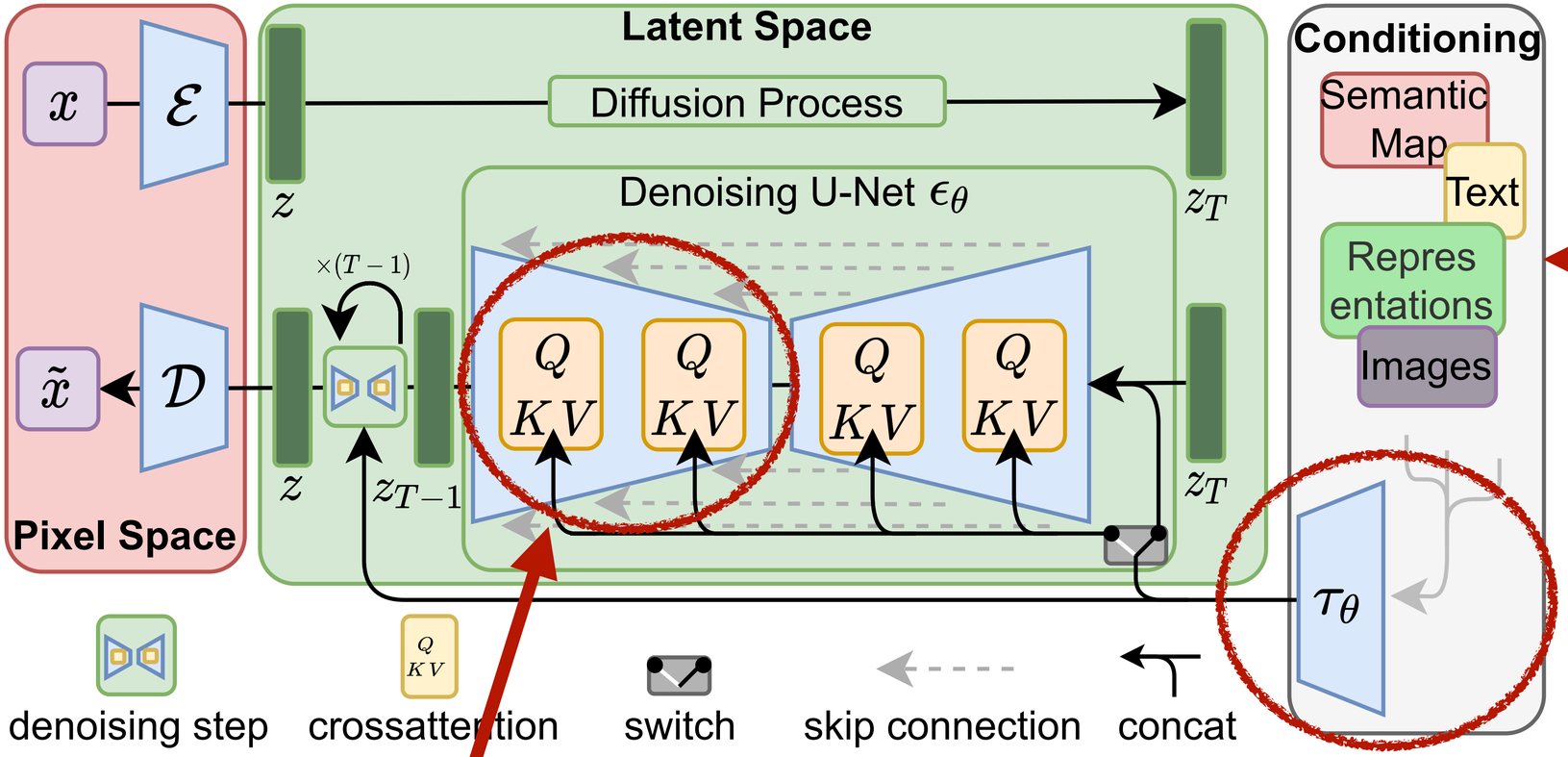
Text-to-Imageはどうやって？

$$P(X|Y)$$

条件部

$$P(\text{画像}|\text{テキスト})$$

Stable Diffusion のブロックダイアグラム



BERT （大規模言語モデル）に基づく言語モデル

言語解析
ニューラルネット

アテンション(注意)
メカニズム

P(画像|テキスト)

まとめ

- Stable diffusionならびに拡散モデルの紹介をした
- 最新の研究成果について、理論(ArXiv)、コード(github)、解説(各種ブログ記事・Youtubeビデオ)などが公開されている
- その気になれば最新の事例を学ぶことも可能
- (英語が少し大変かもしれないが)最新情報のキャッチアップも結構面白いので試してみてください