HIGH-RESOLUTION IMAGE SYNTHESIS WITH LATENT DIFFUSION MODELS

Robin Rombach[1]   Andreas Blattmann[1]   Dominik Lorenz[1], Patrick Esser[2] *and* Bjorn Ommer[1]

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

[2] Runway ML

What ?

We introduce a model to synthesis image by multi-modal conditional, in which we have:

- Proposed a latent space for diffusion model to image synthesis.

- Built the general model trained on some dataset: CelebAHQ, FFHQ, LSUN-Churches, LSUN-Bedrooms.

- Evaluated synthesis process with multi condition input: text, semantic maps, image-to-image translation task.

Why ?

- Recent advancements in computer vision have led to impressive progress in image synthesis. However, GANs, popular generative models, struggle with modeling complex, multi-modal distributions using their adversarial learning approach. In contrast, diffusion models (DMs) built from denoising autoencoders show promising results in various image synthesis tasks but are *computationally demanding*, limiting accessibility and raising environmental concerns.

Overview

Latent Diffusion Model   Layout to Image   Text to Image
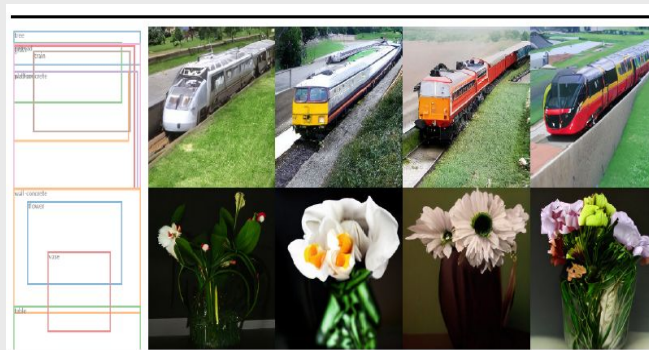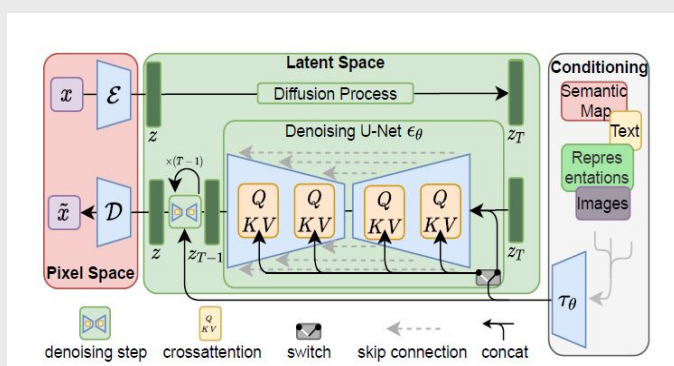
Description

1. Image Reconstruction

- Diffusion models offer inductive biases for spatial data, reduce the dimensionality of the data via autoencoding models, without spatial downsampling of related generative models in latent space.
- Boosting the upper bound on achievable quality with less aggressive downsampling.

Figure 2. Samples from LDMs trained on CelebAHQ, FFHQ, LSUN-Churches, LSUN-Bedrooms and class-conditional ImageNet

3. Conditional Latent Diffusion

Figure 1. Reconstruction FIDs and PSNR are calculated on ImageNet-val.

2. Image Generation with LDM

- We train unconditional models of 2562 images and evaluate the sample quality and their coverage of the data manifold using FID and Precision-Recall.
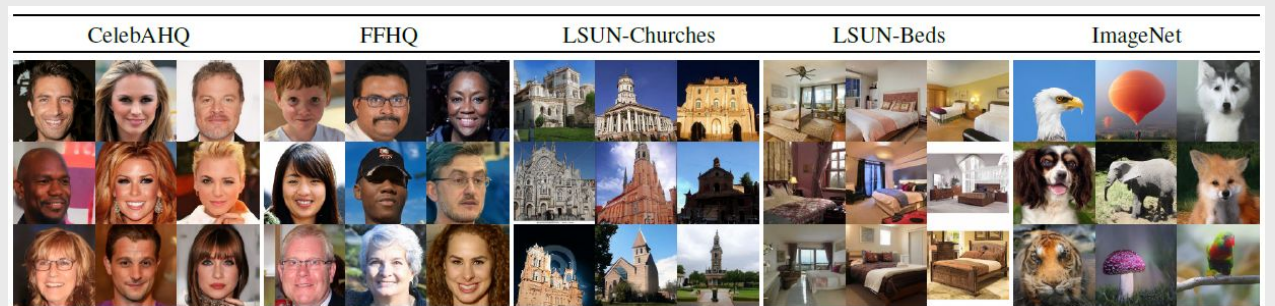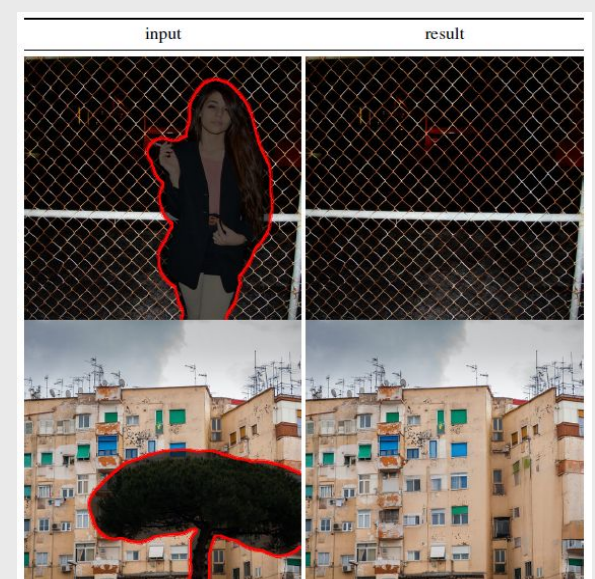
Figure 3. Super-Resolution with Latent Diffusion.

Figure 4. Inpainting with Latent Diffusion.