# High-Resolution Image Synthesis with Latent Diffusion Models

Author: Rombach et al.

created by paper2slides

2021-12-20

https://arxiv.org/abs/2112.10752

# Executive Summary

- The paper introduces Latent Diffusion Models (LDMs) to reduce the computational costs of Diffusion Models (DMs) for image synthesis.
- LDMs operate in a lower-dimensional latent space learned by a pretrained autoencoder, maintaining visual fidelity while being computationally efficient.
- LDMs achieve state-of-the-art results in various tasks, including image inpainting, super-resolution, and text-to-image synthesis, significantly lowering training and sampling costs compared to pixel-based DMs.

# Introduction

- High-resolution image synthesis requires significant computational resources, especially with likelihood-based generative models such as DMs.
- Existing approaches involve GANs, VAEs, and ARMs, each with limitations in quality, training stability, and computational efficiency.
- The paper proposes using an autoencoder to map images to a perceptually equivalent, lower-dimensional latent space for efficient training and sampling of DMs.

# Motivation

- Pixel-based DMs consume massive computational resources for training and inference.
- The reweighted variational objective aims to address this by undersampling initial denoising steps.
- Training DMs in latent space can reduce computational demands, focusing on perceptual and semantic compression while preserving high-fidelity reconstructions.

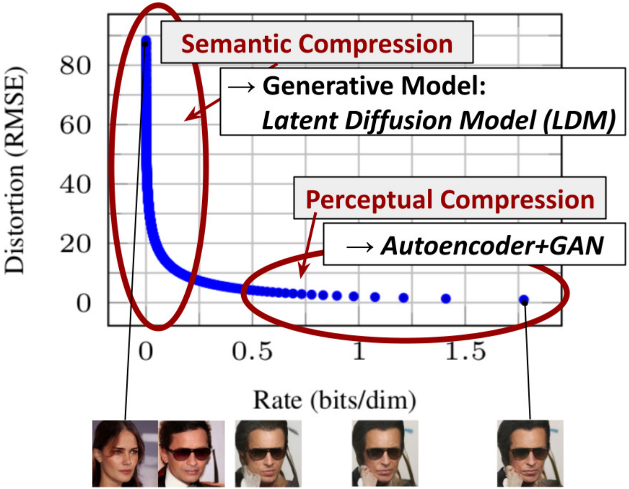

Figure: Illustrating perceptual and semantic compression in diffusion models.

# Perceptual Image Compression

- An autoencoder is trained with a combination of perceptual loss and a patch-based adversarial objective for realistic reconstructions.
- The encoder downsamples the image by a factor of $f$, producing a latent representation, while the decoder reconstructs the image from this latent space.
- Two types of regularizations are explored: KL-regularization (similar to VAEs) and vector quantization regularization (similar to VQGANs).

Figure: Comparison of reconstruction quality with different first stage models and regularizations.

# Latent Diffusion Models (LDMs)

- LDMs apply DMs in the learned latent space from the autoencoder, which reduces computational complexity and maintains image quality.
- The neural backbone of LDMs is a time-conditional UNet that operates in the latent space.
- The training objective for LDMs in latent space is similar to pixel-based DMs but focuses on the latent representations.

# Equation: Latent Diffusion Model Loss

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right] \tag{1}$$

- $\mathcal{E}(x)$ is the latent representation of the image.
- $\epsilon$ is sampled from a standard normal distribution.
- $t$ is the time step in the diffusion process.
- $\epsilon_\theta(z_t, t)$ is the denoising model at time $t$.

# Conditioning Mechanisms

- LDMs can model conditional distributions $p(z|y)$ by augmenting the underlying UNet with cross-attention mechanisms.
- This allows for various types of conditioning inputs such as class labels, text, semantic maps, and more.
- Conditioning mechanisms include cross-attention layers that map intermediate representations of the UNet to the conditioning input.

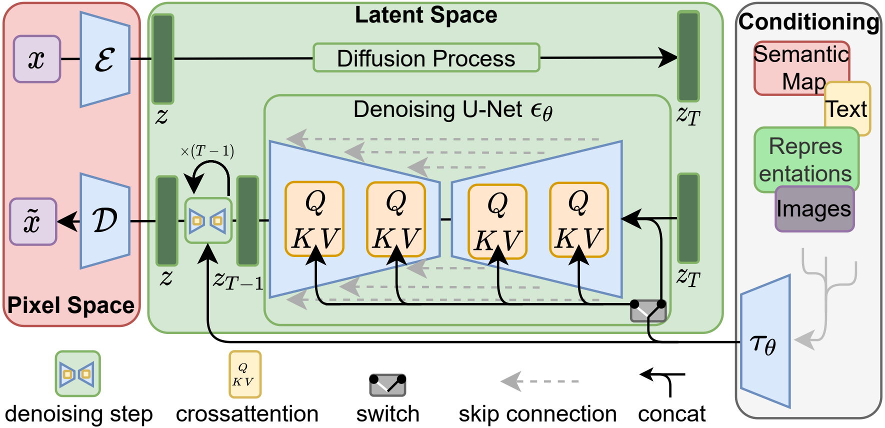

Figure: Cross-attention based conditioning mechanism in LDMs.

# Experimental Results

- Evaluated LDMs on various datasets and tasks: CelebA-HQ, FFHQ, LSUN-Churches, LSUN-Bedrooms, ImageNet, and COCO.
- Achieved state-of-the-art FID scores for image synthesis, inpainting, and super-resolution.
- Demonstrated efficient training and sampling with reduced computational requirements compared to pixel-based DMs.

Figure: Visual comparison of samples from LDM-trained models on various datasets.

# Conclusion

- LDMs significantly reduce the computational costs of training and sampling DMs while maintaining high sample quality.
- They achieve state-of-the-art results on several benchmarks including image inpainting, super-resolution, and text-to-image synthesis.
- The cross-attention conditioning mechanism allows for flexible control over the image generation process with various types of input.

# Future Directions

- Explore further reducing computational requirements for training and inference.
- Investigate applying LDMs to other generative tasks and domains beyond image synthesis.
- Study the impact of different types of latent space regularizations and their effects on model performance.