High-Resolution Image Synthesis with Latent Diffusion Models

DISCLAIMER: Summarized by AI

Problem they are trying to solve / Purpose of method

Latent Diffusion Models (LDMs) address the challenges of high-resolution image synthesis with deep generative models, particularly:

- **High computational cost**: Traditional diffusion models operate in pixel space, which is computationally expensive for high-resolution images due to the large dimensionality.
- **Limited scalability**: Prior models struggle to scale to high resolutions (e.g., 512×512 or higher) without incurring significant resource usage or requiring architectural compromises.
- **Semantic limitations**: Models directly working in pixel space often fail to capture high-level semantic structure and context.

The purpose of introducing LDMs is to enable high-resolution and semantically-aware image synthesis while dramatically reducing computational complexity.

How does it differ from other methods?

Latent Diffusion Models differ in the following key ways:

- **Latent space diffusion**: Unlike traditional methods that apply the diffusion process in pixel space, LDMs perform it in a *learned lower-dimensional latent space* obtained from a pretrained autoencoder. This reduces computational load significantly.
- **Modularity**: The architecture decouples the image compression and generation processes by using a separately trained autoencoder and a U-Net-based diffusion model in latent space.
- **Conditioning flexibility**: LDMs support various conditioning methods (e.g., text, semantic maps), enabling controlled generation.

These features make LDMs both *efficient* and *flexible* for high-quality image generation and editing tasks.

How the method works

Simple Overview:

1. Compress the image into a lower-dimensional latent representation using a pretrained autoencoder.
2. Apply the diffusion model (a denoising process) in this latent space.
3. Decode the generated latent representation back to the image space using the decoder.

More Details:

- The autoencoder is trained to reconstruct images while preserving perceptual quality.
- The latent space is used as the domain for diffusion, dramatically reducing dimensionality and training costs.
- The diffusion model is based on a U-Net with cross-attention mechanisms, supporting conditional generation.
- Various conditioning methods (e.g., CLIP text embeddings) can be used to steer generation in a semantically meaningful way.

This approach balances quality, efficiency, and flexibility, achieving state-of-the-art results for high-resolution image synthesis.