High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach   Andreas Blattmann   Dominik Lorenz   Patrick Esser   Bjorn Ommer

小组成员：卢宇航 齐林 蒋卫 刘梦茵 李企峥

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.
- 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.
- By introducing cross-attention layers into the model architecture, we turn diffusion models into powerful and flexible generators for general conditioning inputs such as text or bounding boxes and high-resolution synthesis becomes possible in a convolutional manner.
- Our latent diffusion models (LDMs) achieve new state of the art scores for image inpainting and class-conditional image synthesis and highly competitive performance on various tasks, including unconditional image generation, text-to-image synthesis, and super-resolution, while significantly reducing computational requirements compared to pixel-based DMs.

Introduction

Image synthesis is one of the computer vision fields with the most spectacular recent development, but also among those with the greatest computational demands. Recently, diffusion models, which are built from a hierarchy of denoising autoencoders, have shown to achieve impressive results in image synthesis and beyond, and define the state-of-the-art in class-conditional image synthesis and super-resolution.Moreover, even unconditional DMs can readily be applied to tasks such as inpainting and colorization or stroke-based synthesis , in contrast to other types of generative models.

1.Democratizing High-Resolution Image Synthesis

DMs belong to the class of likelihood-based models, whose mode-covering behavior makes them prone to spend excessive amounts of capacity (and thus compute resources)on modeling imperceptible details of the data.Reducing the computational demands of DMs without impairing their performance is, therefore, key to enhance their accessibility.

2.Departure to Latent Space

We aim to first find a perceptually equivalent, but computationally more suitable space, in which we will train diffusion models for high-resolution image synthesis. we separate training into two distinct phases: First, we train an autoencoder which provides a lower-dimensional (and thereby efficient) representational space which is perceptually equivalent to the data space. The reduced complexity also provides efficient image generation from the latent space with a single network pass. We dub the resulting model class Latent Diffusion Models (LDMs).

(i) Our method scales more graceful to higher dimensional data and provides more faithful and detailed reconstructions than previous work and can be efficiently applied to high-resolution synthesis of megapixel images.

(ii) We achieve competitive performance on multiple tasks and datasets while significantly lowering computational costs and inference costs.

(iii) We ensure extremely faithful reconstructions and requires very little regularization of the latent space.

(iv) Our model can be applied in a convolutional fashion and render large, consistent images of ∼ 1024^2 px.

(v) Moreover, we design a general-purpose conditioning mechanism based on cross-attention, enabling multi-modal training. We use it to train class-conditional, text-to-image and layout-to-image models.

Related Work

1.Generative Models for Image Synthesis

The high dimensional nature of images presents distinct challenges to generative modeling.To scale to higher resolutions, several two-stage approaches use ARMs to model a compressed latent image space instead of raw pixels.

2.Diffusion Probabilistic Models

DM have achieved state-of-the-art results in density estimation as well as in sample quality .We proposed LDMs, which work on a compressed latent space of lower dimensionality. This renders training cheaper and speeds up inference with almost no reduction in synthesis quality.

3.Two-Stage Image Synthesis

To mitigate the shortcomings of individual generative approaches, a lot of research has gone into combining the strengths of different methods into more efficient and performant models via a two stage approach.

Method

To lower the computational demands of training diffusion models towards high-resolution image synthesis, we observe that although diffusion models allow to ignore perceptually irrelevant details by undersampling the corresponding loss terms , they still require costly function evaluations in pixel space, which causes huge demands in computation time and energy resources.

We propose to circumvent this drawback by introducing an explicit separation of the compressive from the generative learning phase To achieve this, we utilize an autoencoding model which learns a space that is perceptually equivalent to the image space, but offers significantly reduced computational complexity.

1. **Perceptual Image Compression**

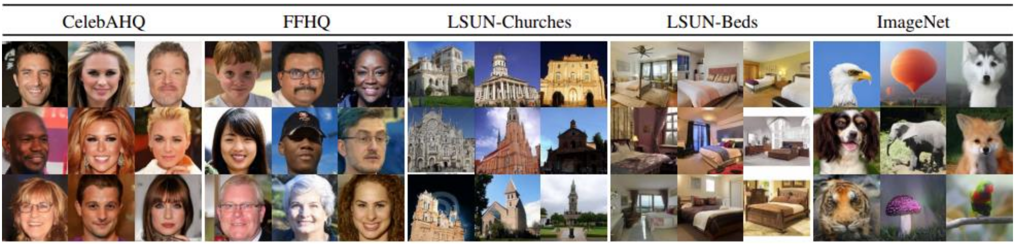
   Our perceptual compression model is based on previous work and consists of an autoencoder trained by combination of a perceptual loss and a patch-based adversarial objective. This ensures that the reconstructions are confined to the image manifold by enforcing local realism and avoids bluriness introduced by relying solely on pixel-space losses such as L2 or L1 objectives.

2. **Latent Diffusion Models**

   Diffusion Models [79] are probabilistic models designed to learn a data distribution p(x) by gradually denoising a normally distributed variable, which corresponds to learning the reverse process of a fixed Markov Chain of length T.

3. **Conditioning Mechanisms**

   Similar to other types of generative models ,diffusion models are in principle capable of modeling conditional distributions of the form p(z|y). This can be implemented with a conditional denoising autoencoder ϱθ(zt, t, y) and paves the way to controlling the synthesis process through inputs y such as text semantic maps or other image-to-image translation tasks .



Samples from LDMs trained on CelebAHQ ,FFHQ , LSUN-Churches , LSUN-Bedrooms and class-conditional ImageNet], each with a resolution of 256 × 256. Best viewed when zoomed in. For more samples cf . the supplement.

Conclusion

We have presented latent diffusion models, a simple and efficient way to significantly improve both the training and sampling efficiency of denoising diffusion models without degrading their quality. Based on this and our crossattention conditioning mechanism, our experiments could demonstrate favorable results compared to state-of-the-art methods across a wide range of conditional image synthesis tasks without task-specific architectures.

杭州電子科技大學
HANGZHOU DIANZI UNIVERSITY