<https://arxiv.org/pdf/2112.10752.pdf>

Stable diffusion based on this! https://github.com/CompVis/stable-diffusion

**Background**

Decomposing image formation process into sequential application of denoising autoencoders (diffusion), image synthesis

* Guiding mechanism to control image generation process without retraining

Currently

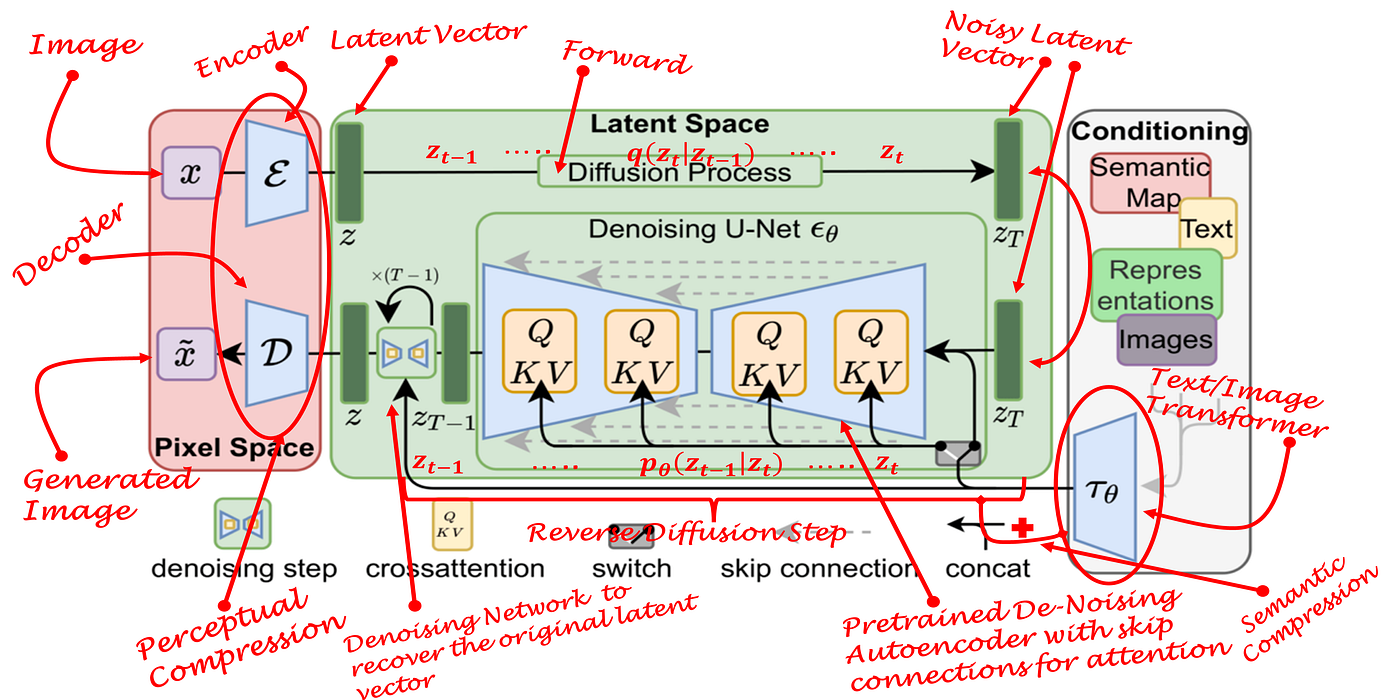
* Large likelihood based models (ARM) in autoregressive transformers, billions of parameters -> low resolution
  + mode covering behavior makes them spend excessive amounts of capacity modeling imperceptible details -> High computational cost
* Also likelihood: VAE and flow based sample quality is worse than GANs but can render multi modal
* GAN confined to data with limited variability due to adversarial learning not easily scaling to model complex, multi modal distributions
* Democratizing High Resolution Image Synthesis for DMs
  + Diffusion probabilistic models are also likelihood based models
  + No mode collapse and instability in GANs
  + Parameter sharing so doesn’t need billions of parameters
  + Training in pixel space is hard because high computational cost
* Two stage image synthesis which combines the strengths of different methods into more efficient and performant models
  + VQ- VAE: autoregressive to learn prior in latent space
  + VQGANs: first stage with adversarial and perceptual objective to scale autoregressive transformer
    - Computationally expensive scaling

**Idea (Latent Diffusion Models)**

Training in pixel space is expensive and requires sequential evaluation

* Perceptual compression: removes high frequency details and a bit of semantic variation
  + Theirs is lower dimensional than data space
  + Do not need to rely on excessive spatial compressions
  + Efficient image generation with single network pass as a result
  + “Universal autoencoding stage”: reuse for multiple DM trainings
  + Perceptual loss + patch based adversarial objective
    - Confined to image manifold by enforcing local realism and avoid blurriness introduced by relying solely on pixel space losses
  + KL regularization to avoid high variance latent spaces
    - Mild compression with 2D latent space (previous used 1D, ignored spatial structure)
* Semantic compression: Generative model learns semantic and conceptual composition of data
  + Connects transformers to DM’s Unet backbone
  + Denoising Unet with transformers (combine conditioning info with noisy latent space)
* Find a perceptually equivalent but computationally more suitable space (latent)
* Latent space instead with pre trained autoencoders
  + Complexity reduction and detail preservation -> better visual fidelity (focus on semantics)
  + Does not require delicate weighting of reconstruction (autoencoding) and generative (diffusion) abilities
    - Prior work needs to learn encoder/ decoder and score based prior

Cross attention layers for general conditioning inputs (text, bounding boxes) + synthesis can be possible in a convolutional manner



Explicit separation of compressive from generative learning phase

* Autoencoding model which learns a space perceptually equivalent to image space but reduced complexity
* Exploit inductive bias of DMs (UNet architecture): effective for data with spatial structure and alleviate need for compression

Isn’t VQ regularized latent space the 1D space?

* LDMs in VQ latent space achieve better sample quality even though reconstruction quality is worse

**Tasks**

* Image inpainting
* Class conditioned image synthesis
* Text to image, super resolution, etc.

**Limitations**

* Still require sequential sampling but does reduce computational requirements
* Use of LDMs questionable when high precision is required
  + Table 5 (section 4.4)