Embeddings, Manifold Learning, and Autoencoders
Generative Modeling

\[ \text{Disc: } p(y|x) \]
\[ \text{Gen: } p(x, y) \]

Sample

Embedding latent space

E

D

rand
Generative Adversarial Networks

- Real images
- Sample
- Discriminator
- Generator

Random input

Loss functions:
- Discriminator loss
- Generator loss
Diffusion Models

Some other good visuals: https://www.chenyang.co/diffusion.html
UNet - a more detailed picture

**Convolutional Encoder-Decoder**

**Skip connection**  
Transporting information at the same resolution.

**Down (sampling) side Encoder**

**Up (sampling) side Decoder**

- **conv 3x3, ReLU**
- **copy & crop**
- **max pool 2x2**
- **up-sampling 2x2**
- **conv 1x1**
Stable Diffusion
(without the conditioning)
Vision and Language
Case study: CLIP
"Attention"

The $e_i$ fit $t_i$ is $f(t_i)$.
Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or diffusion prior to produce an image embedding, and then this embedding is used to condition a diffusion decoder which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.
Stable Diffusion
(with the conditioning)