Embeddings, Manifold Learning, and Autoencoders
Generative Modeling

\[ D_{\text{3}} : p(y \mid x) \]

\[ \text{Gen} : p(x, y) \]
Diffusion Models

Some other good visuals: https://www.chenyang.co/diffusion.html
UNet - a more detailed picture
Stable Diffusion
(without the conditioning)
Vision and Language
Case study: CLIP
"Attention"

This dog is cute.
unCLIP aka DALL-E 2

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)