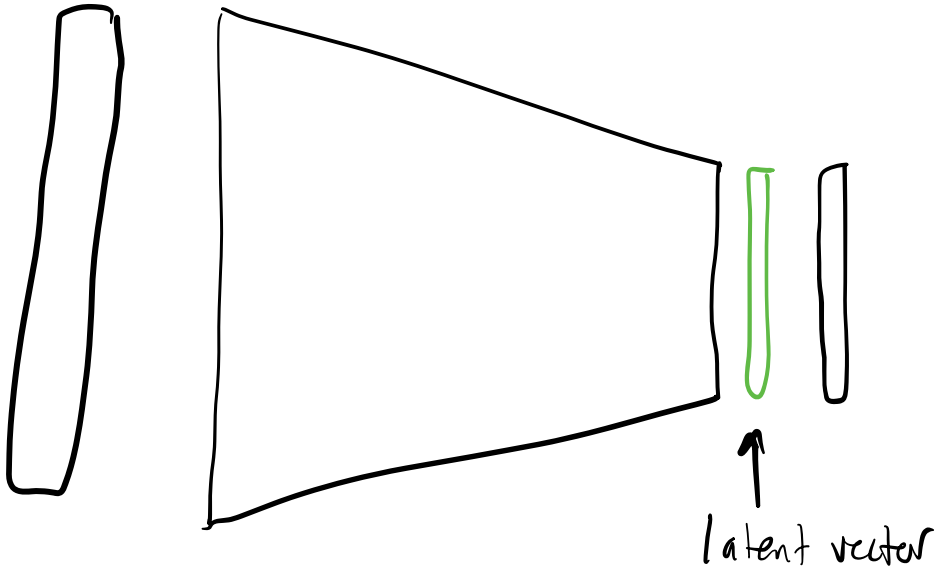
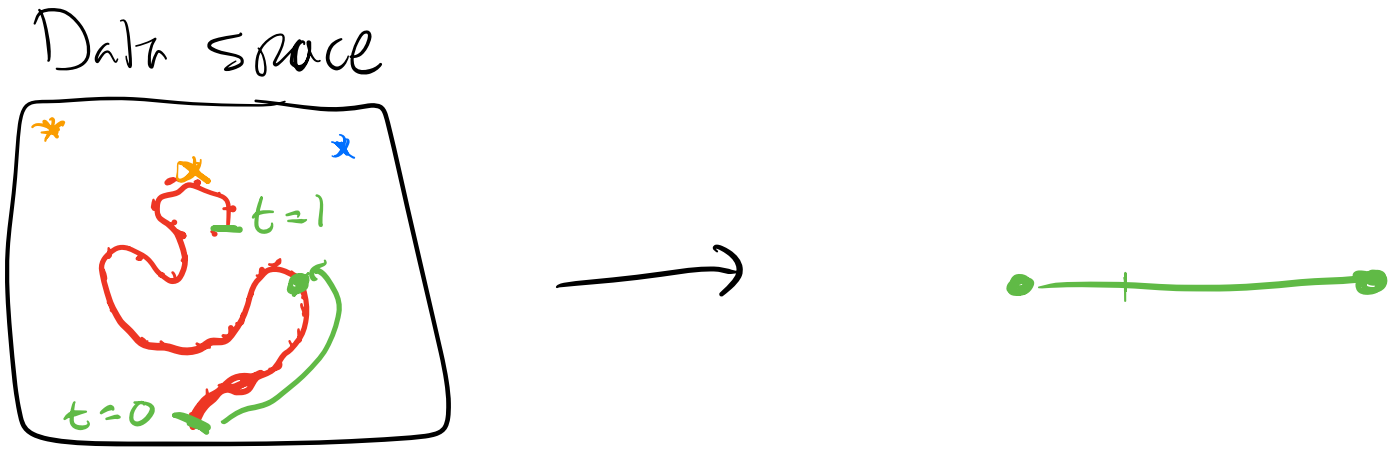
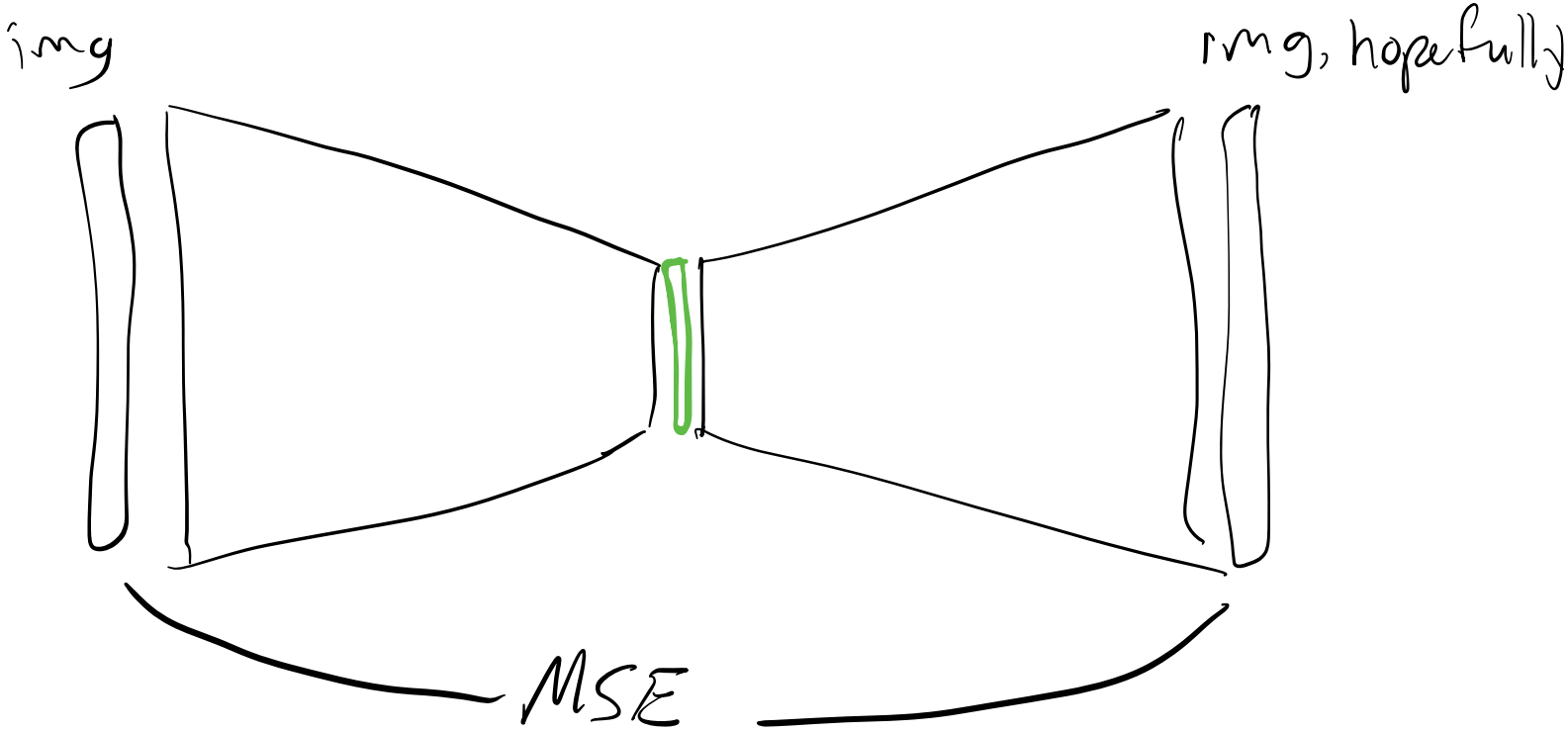
Okay but the data...

(Sharp?) left turn:
Embeddings, Manifold Learning, and Autoencoders



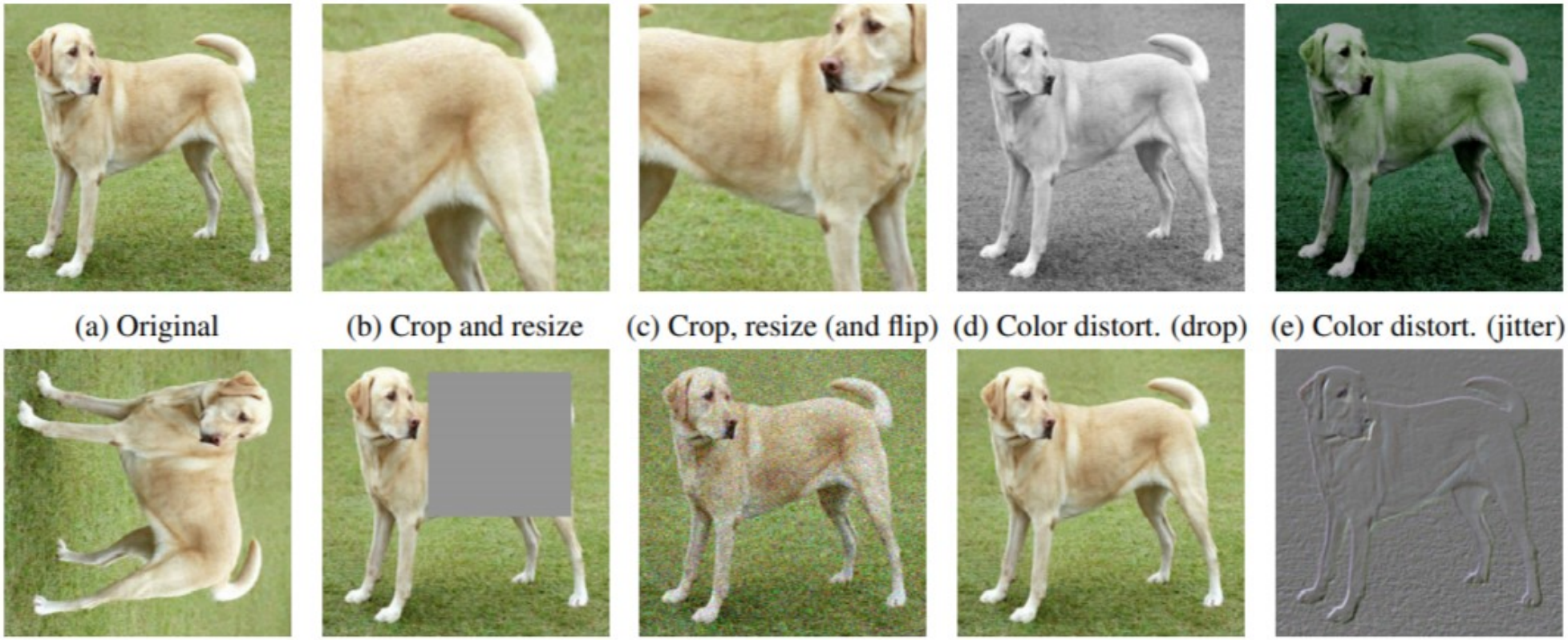


Auto encoder



Unsupervised / self-supervised learning case study: SimCLR

A Simple Framework for Contrastive Learning of Visual Representations

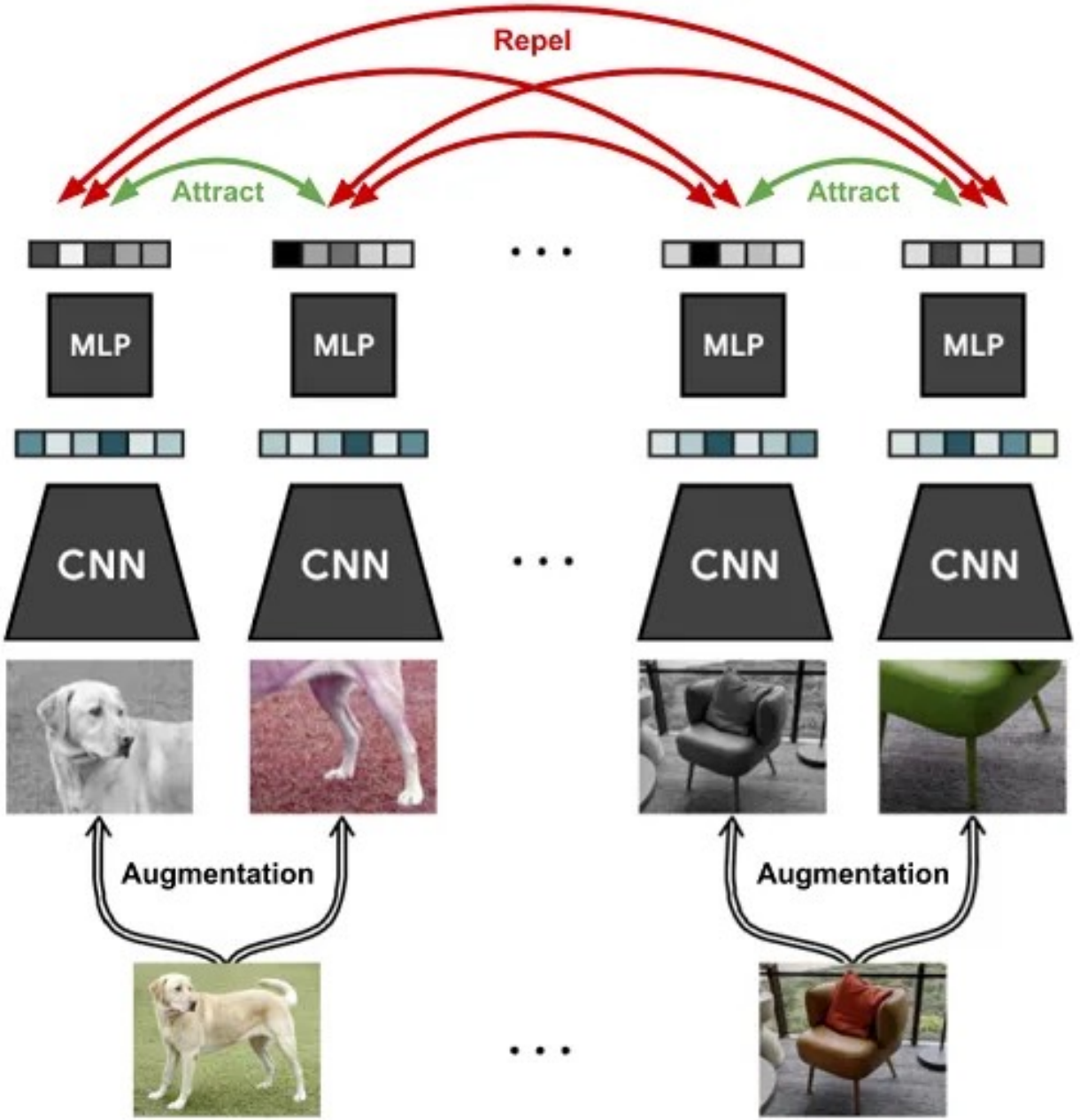


(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

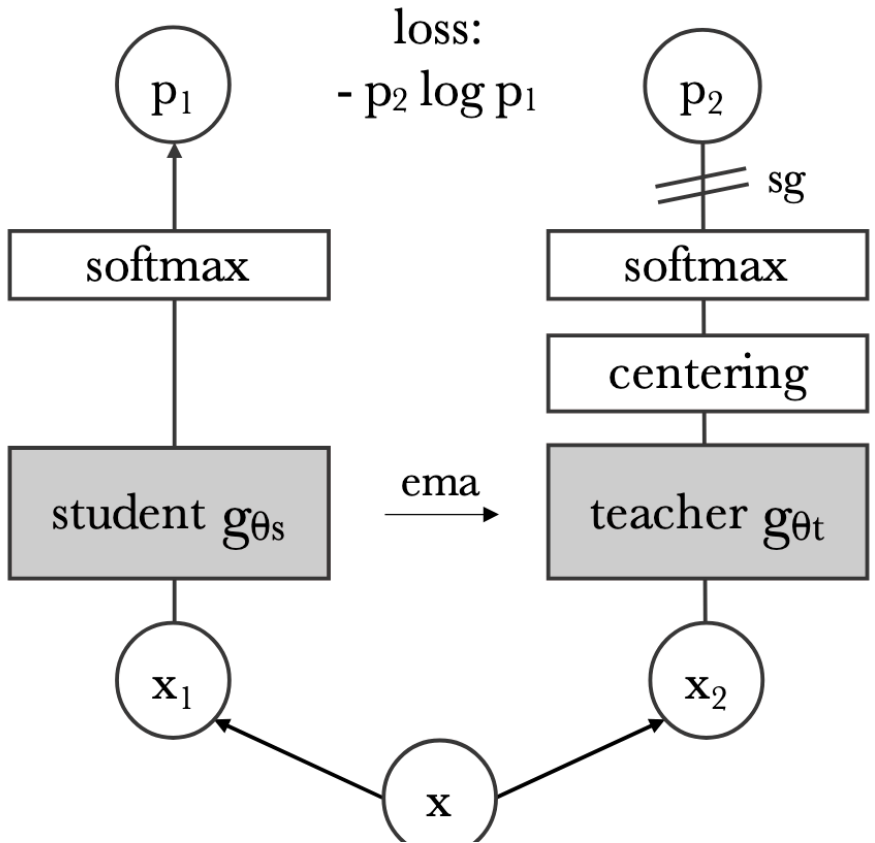
(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

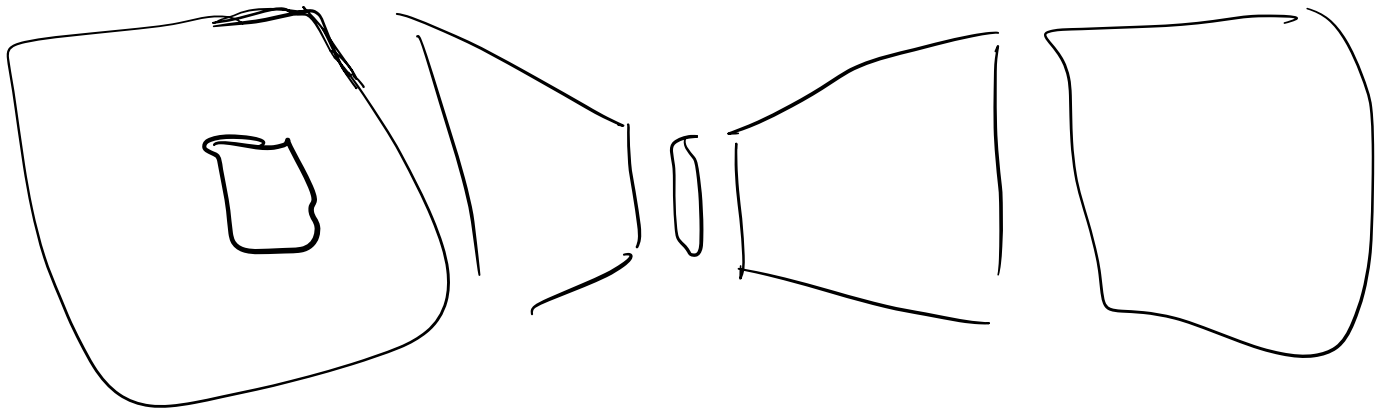
Figure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we *only* test these operators in ablation, the *augmentation policy used to train our models* only includes *random crop (with flip and resize)*, *color distortion*, and *Gaussian blur*. (Original image cc-by: Von.grzanka)

Unsupervised / self-supervised learning case study: SimCLR



DINO - Self-Supervised Learning Image Features

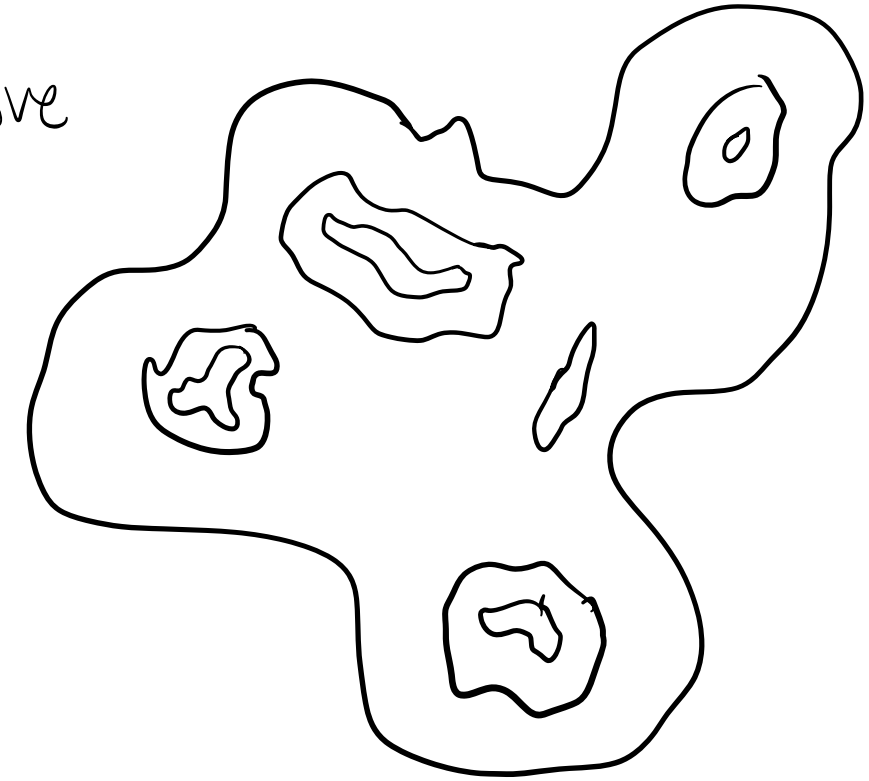




Generative Modeling

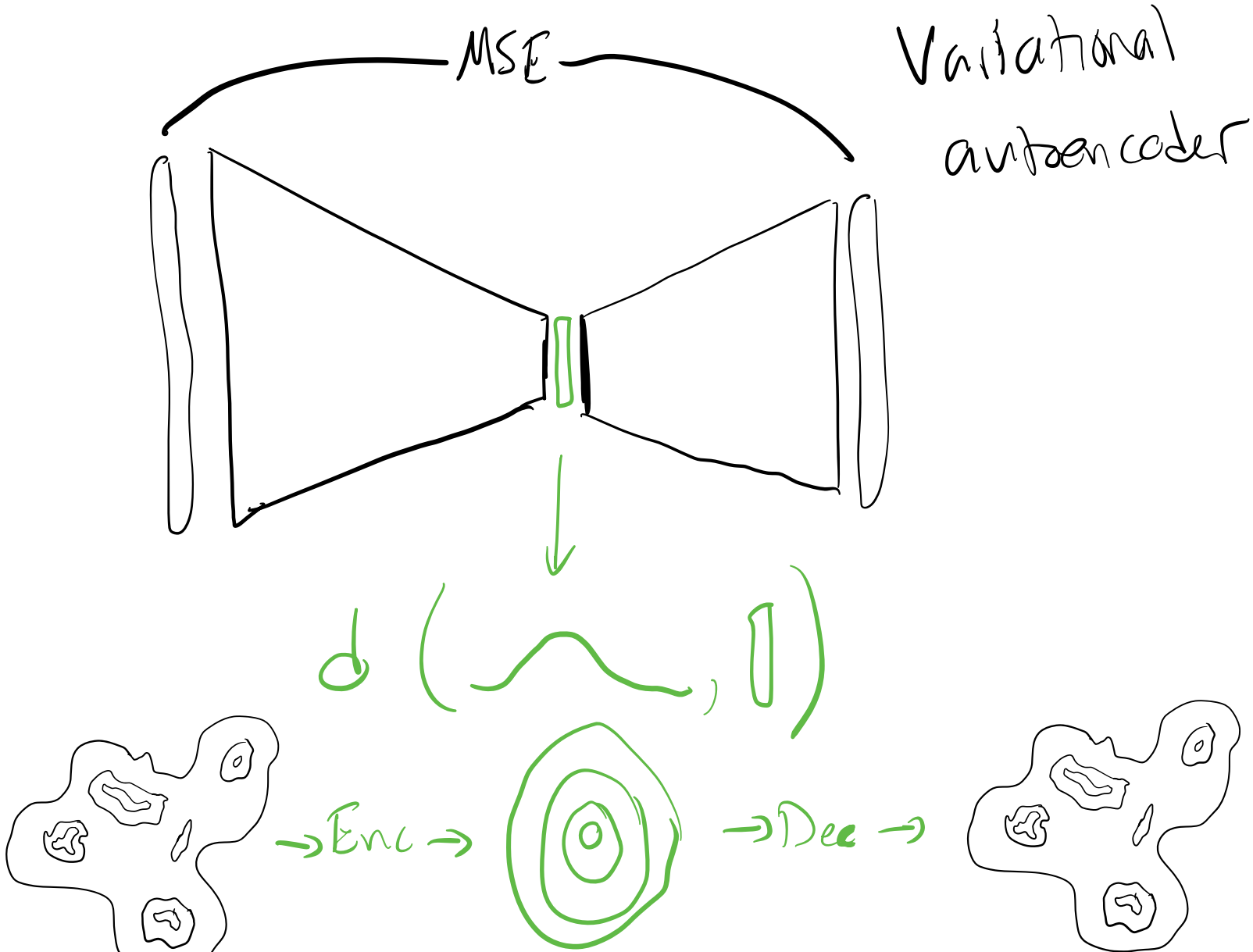
$P(y \mid x)$ discriminative

$P(x, y)$ generative

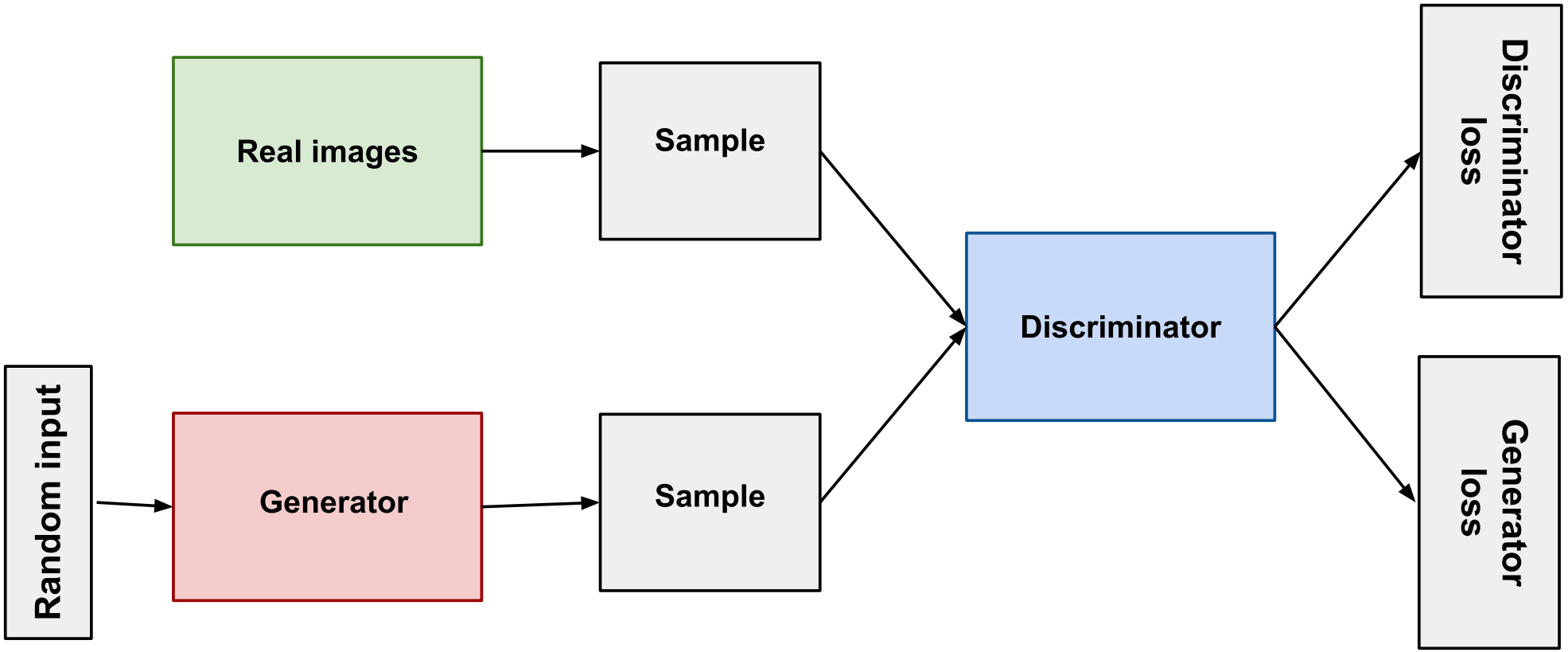


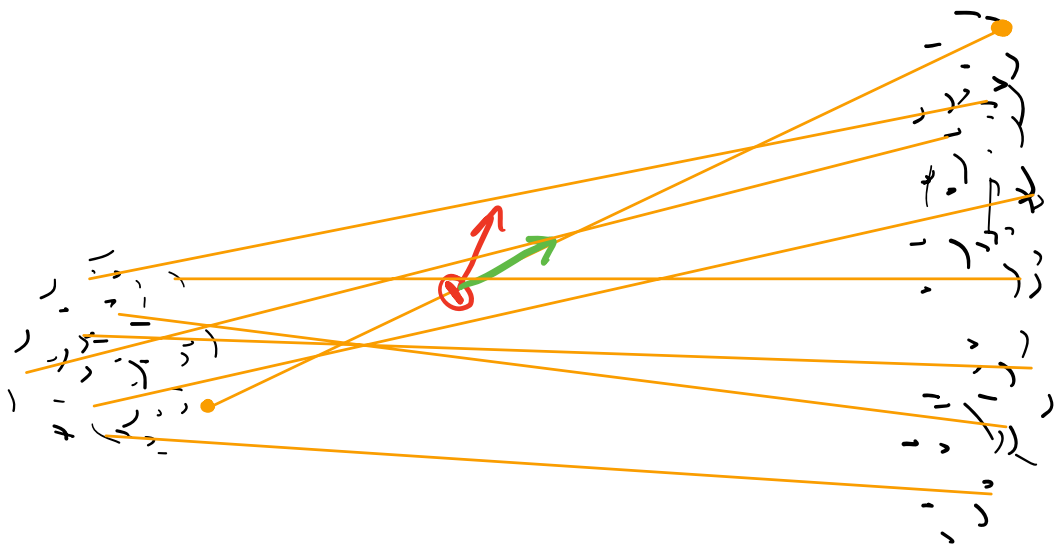
$P(\mathbf{x}, y)$

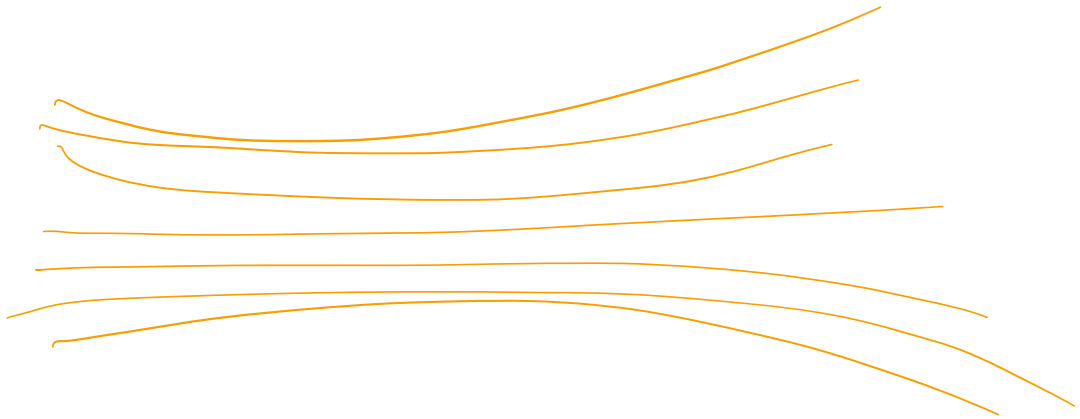
...with autoencoders



Generative Adversarial Networks

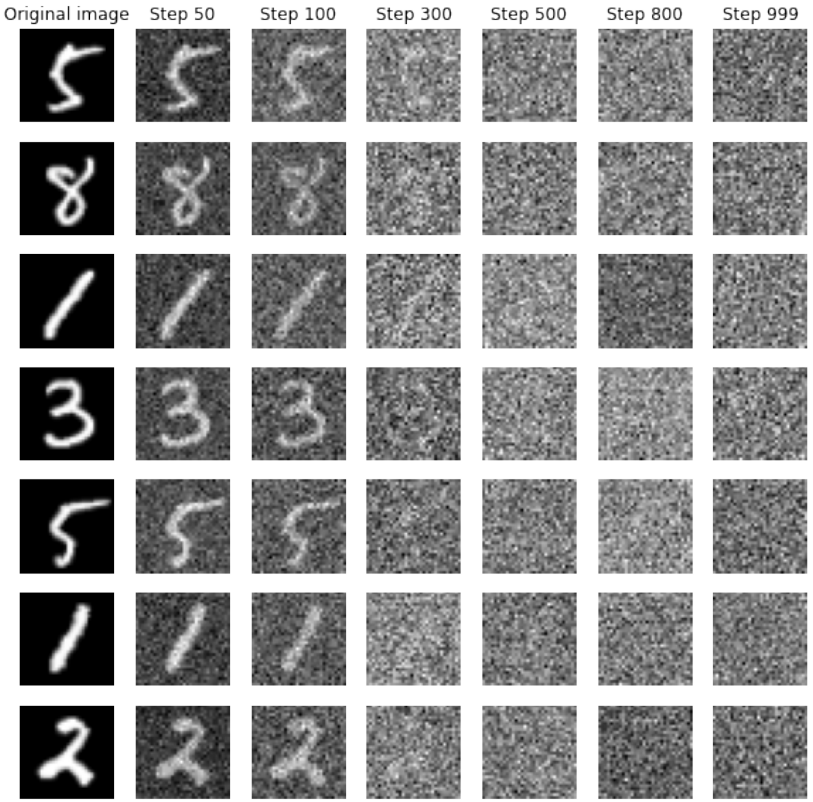




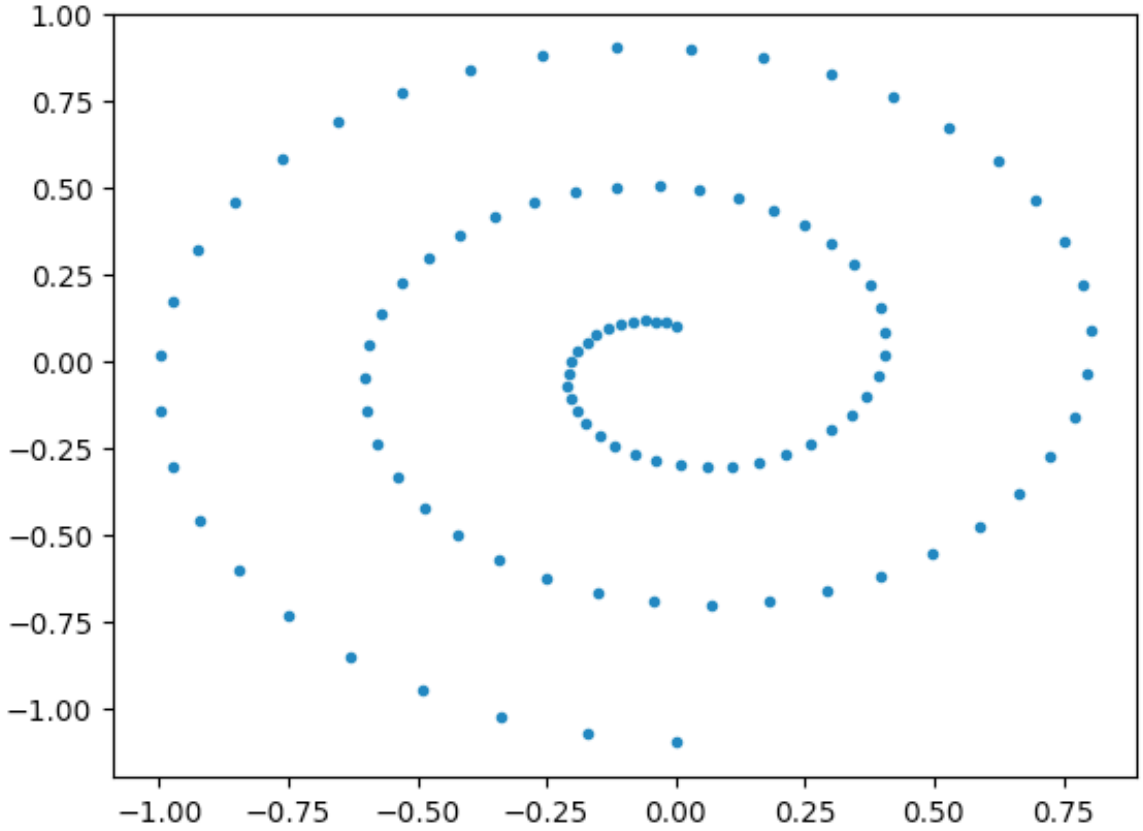


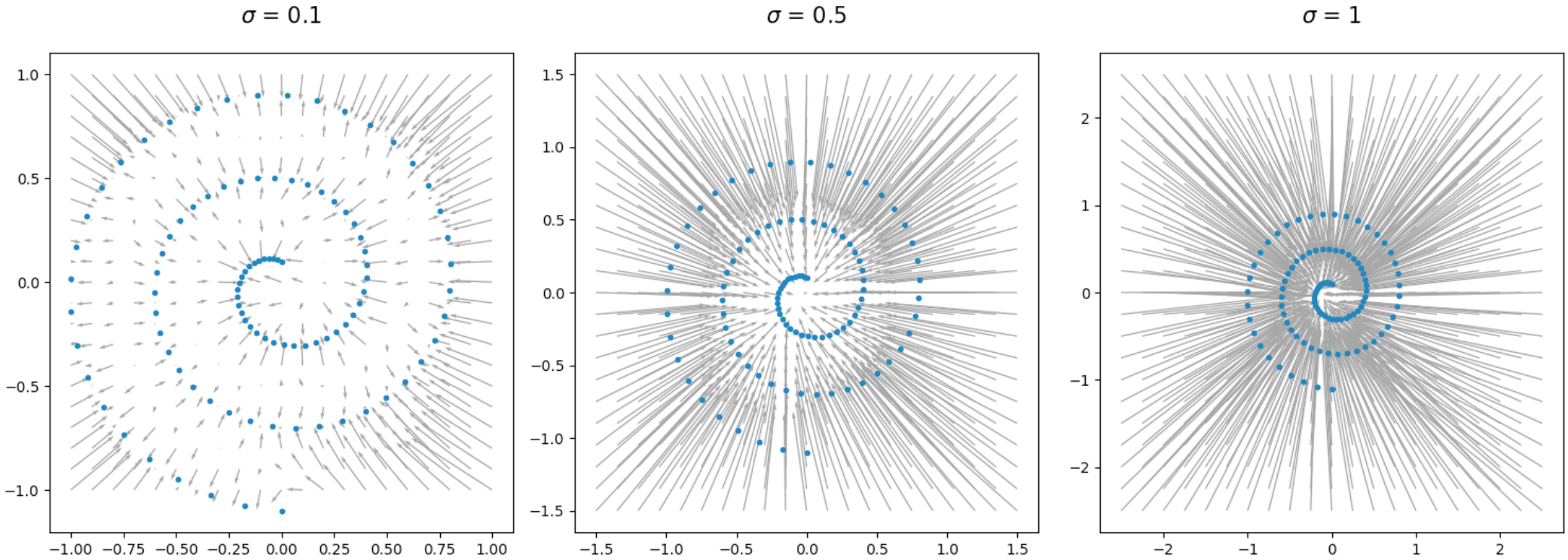

pixers

Diffusion Models



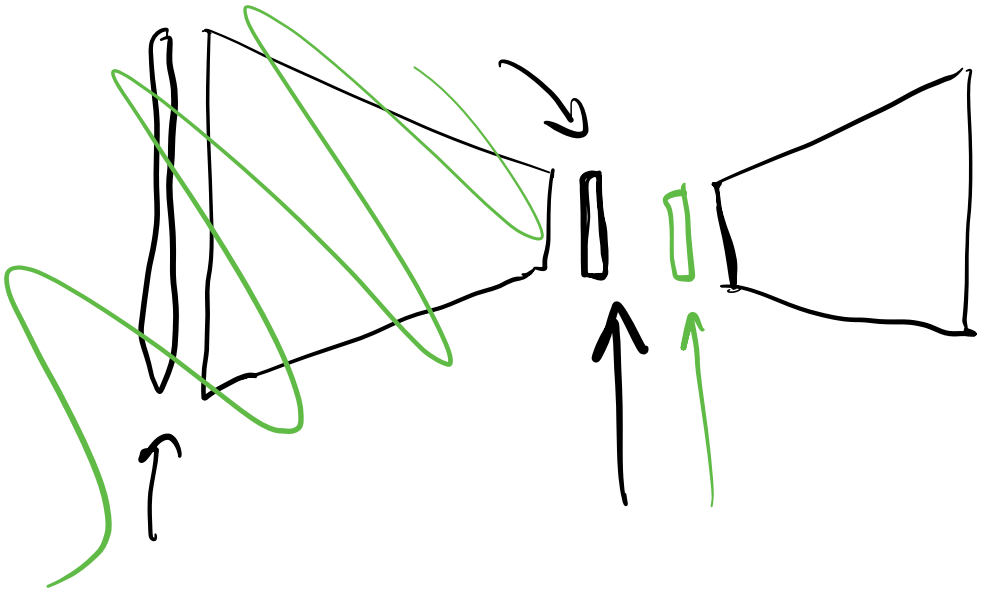
Some other good visuals: https://www.chenyang.co/diffusion.html

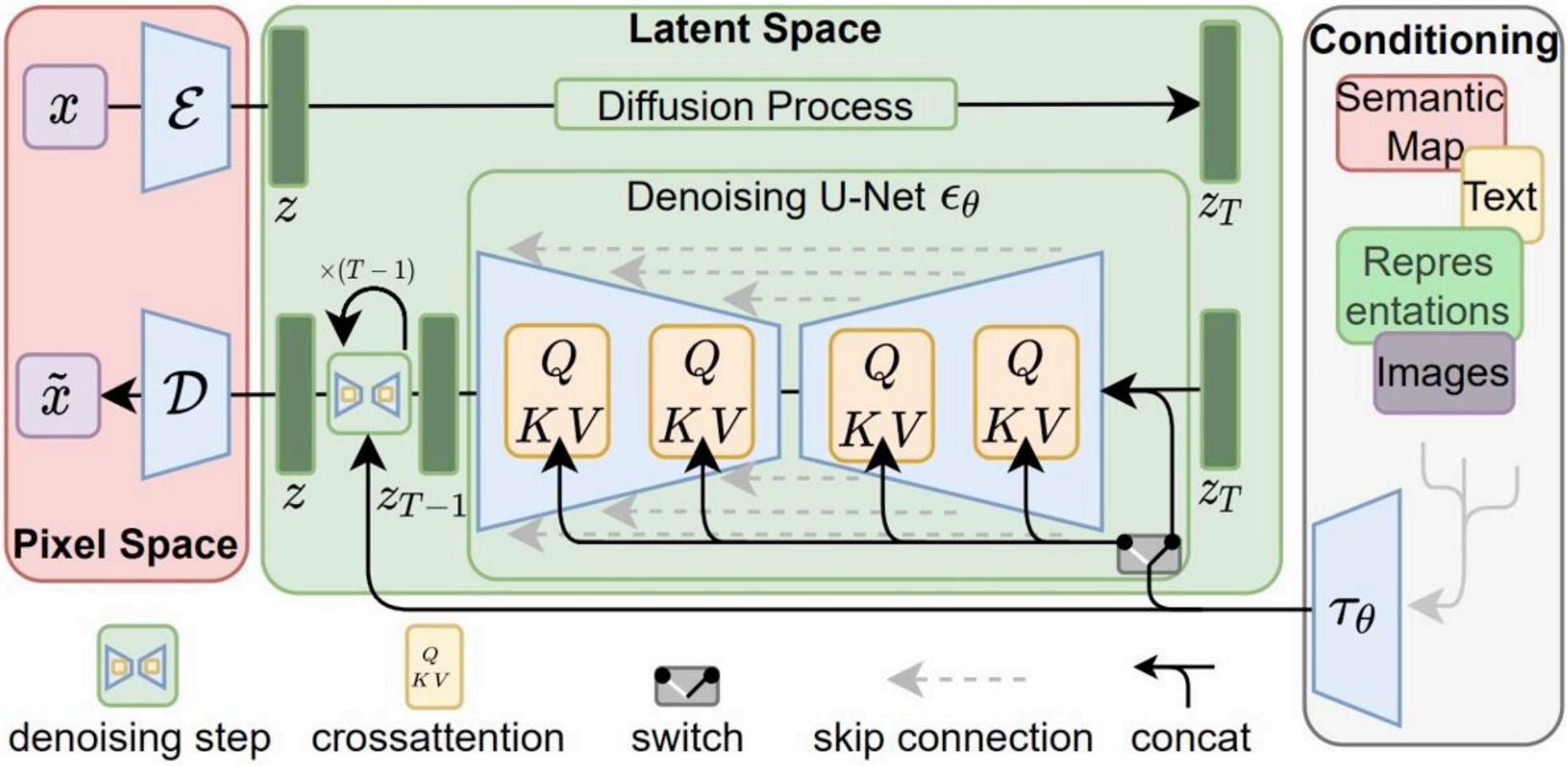

1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75


$\sigma = 0.1$
$\sigma = 0.5$
$\sigma = 1$
1.0
0.5
0.0
−0.5
−1.0
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5 −1.0 −0.5 0.0 0.5 1.0 1.5
2
1
0
−1
−2
−2 −1 0 1 2

Flow Models

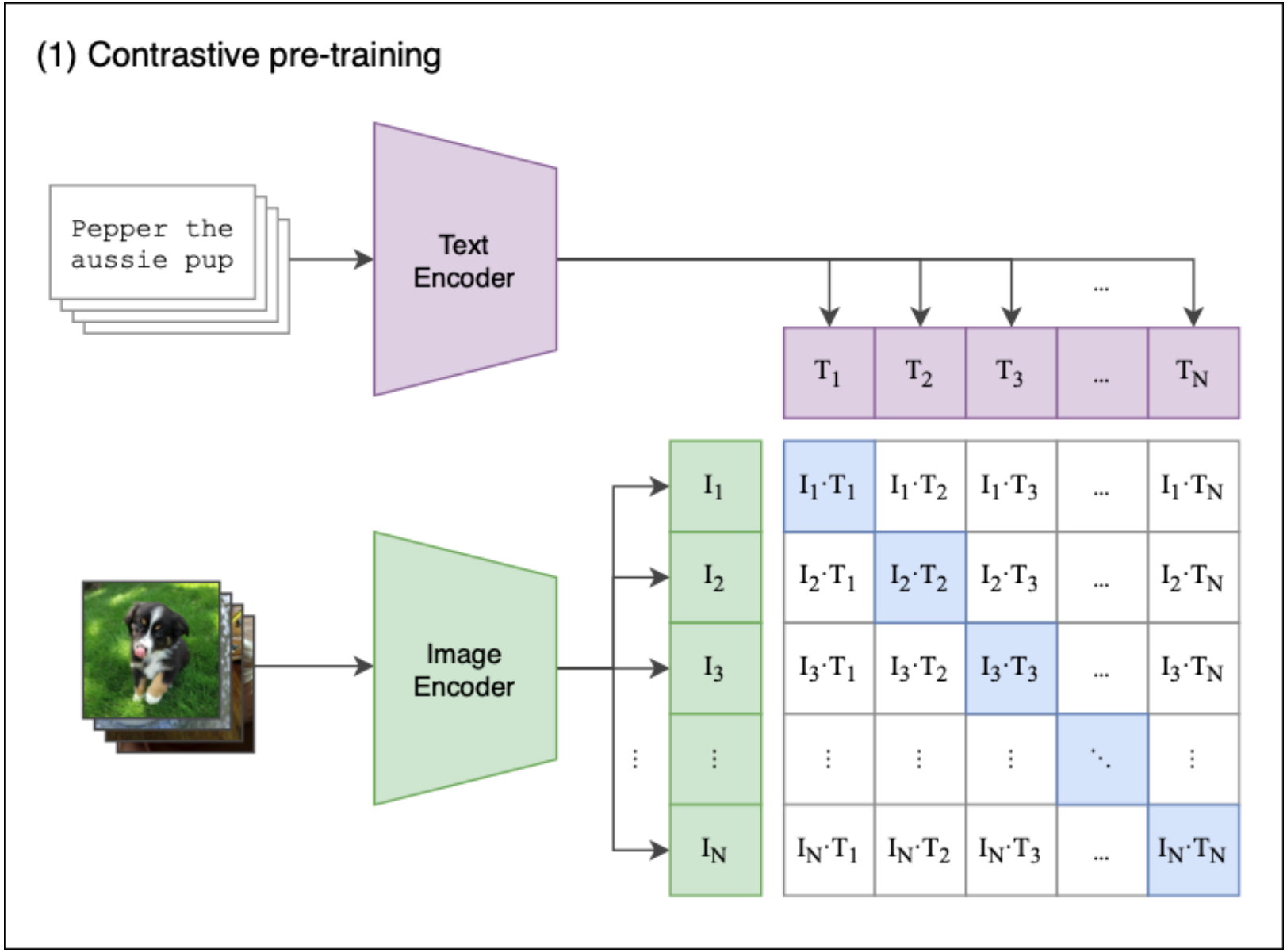
Stable Diffusion
(without the conditioning)




Latent Space
Diffusion Process
Denoising U-Net ϵ_θ
Conditioning
Semantic Map
Text
Repres entations
Images
Pixel Space
x
$\mathcal{E}$
z
z_T
$\times(T-1)$
$\tilde{x}$
$\mathcal{D}$
z
z_{T-1}
Q K V
z_T
τ_θ
denoising step
crossattention
switch
skip connection
concat

Vision and Language

Case study: CLIP



unCLIP aka DALL-E 2

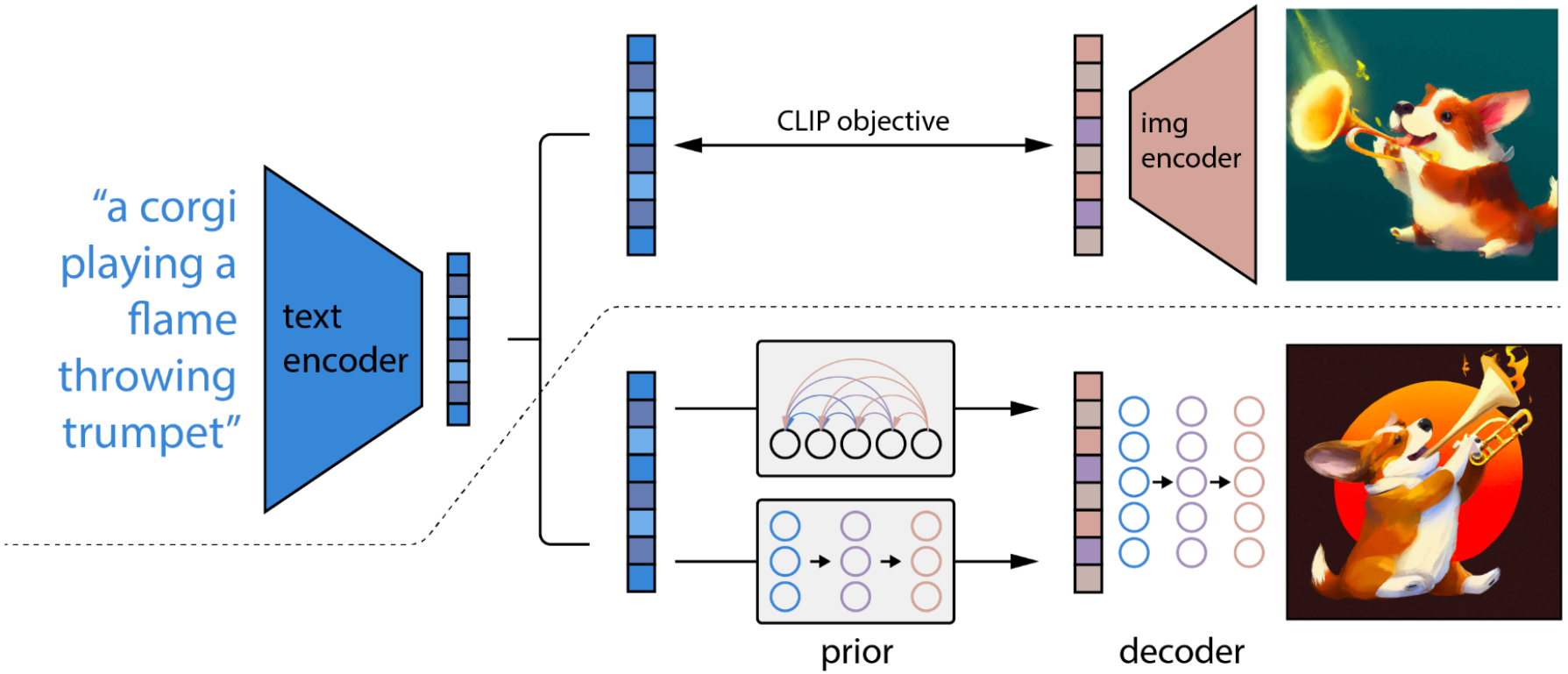


Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive 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)