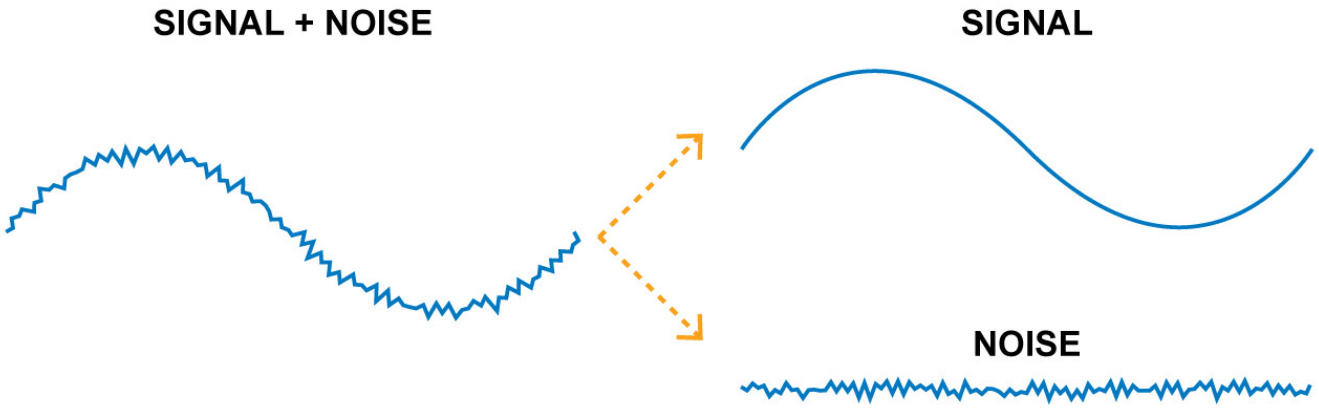
Noise, Pattern, and Image

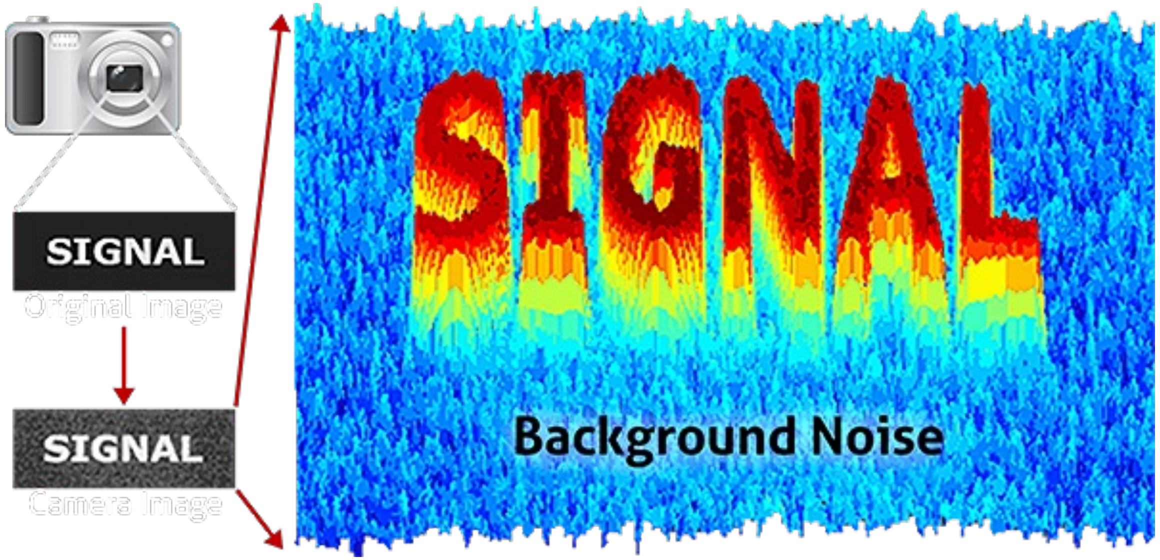
Noise, Perception, and Learning: Applications in AI Art

IAP 2023

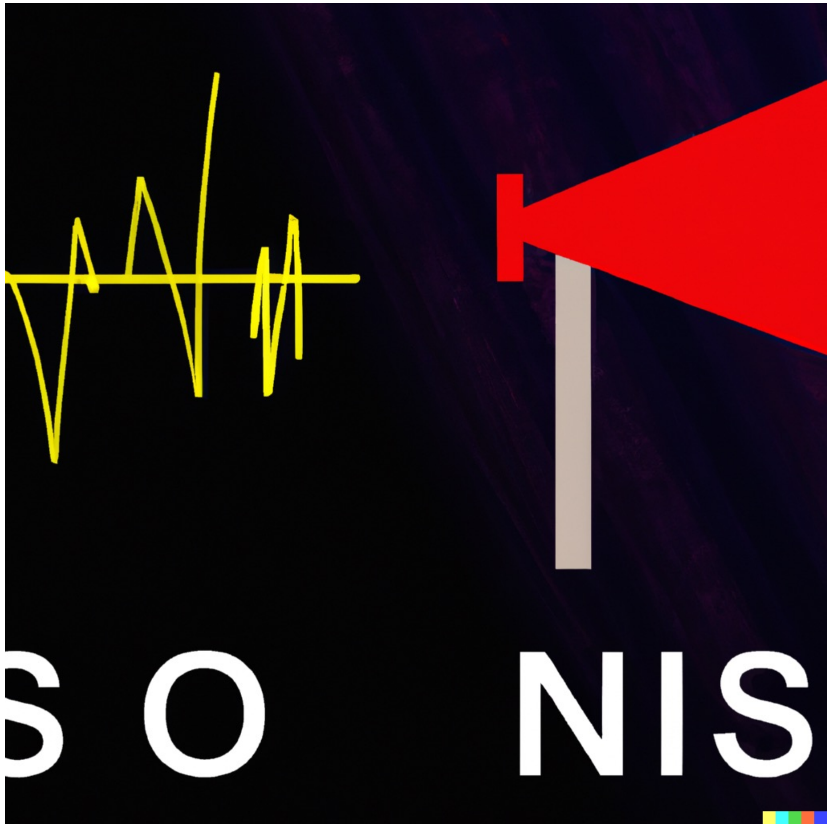
Sarah Muschinske
01/25/2023

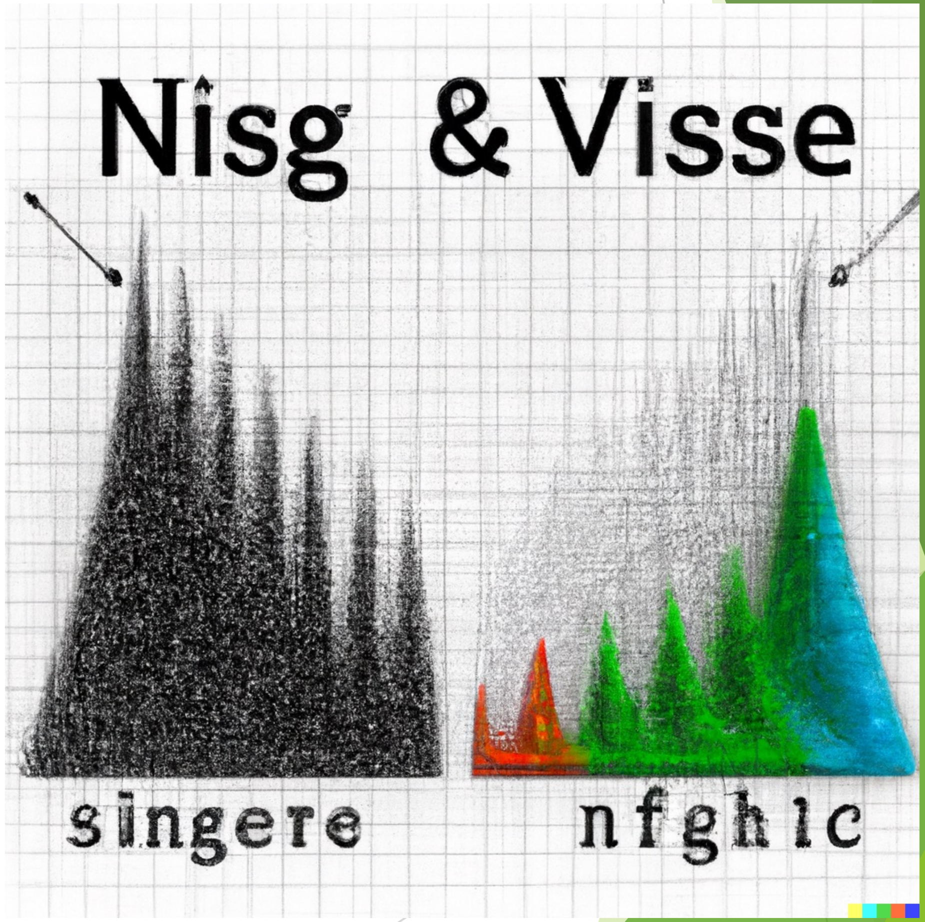
What is Noise?



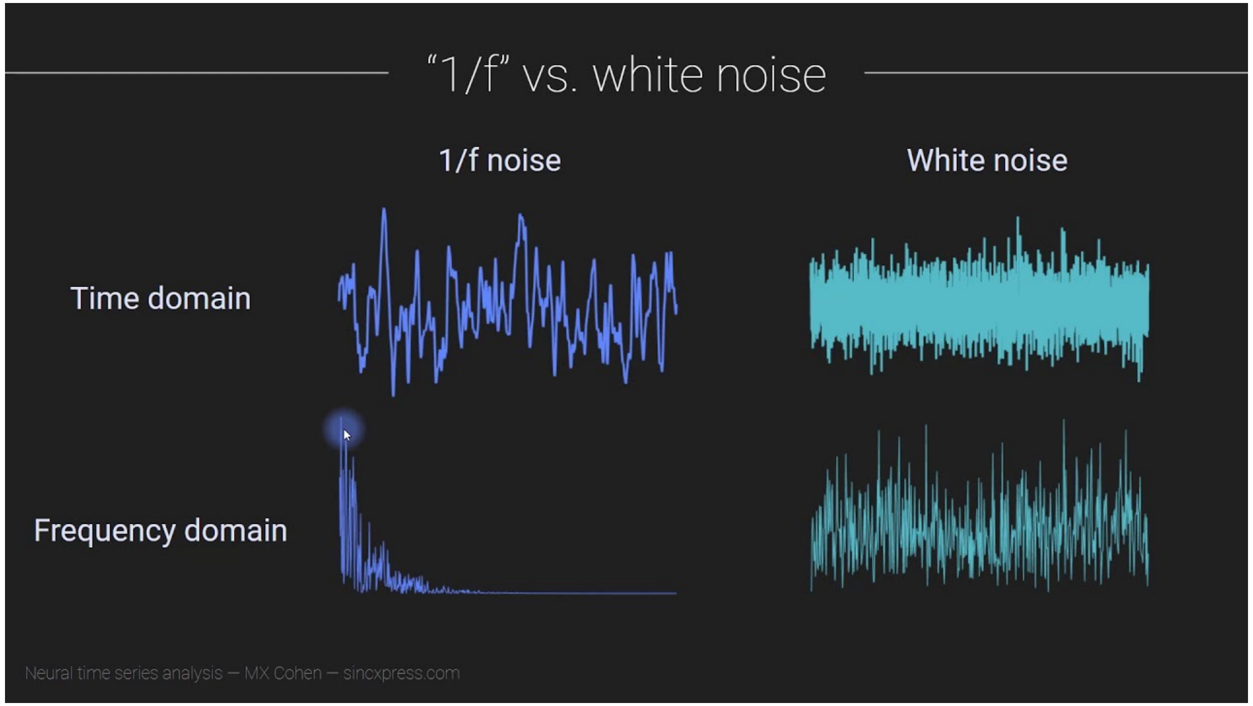


What does DALL-E 2 think signal-to-noise is?



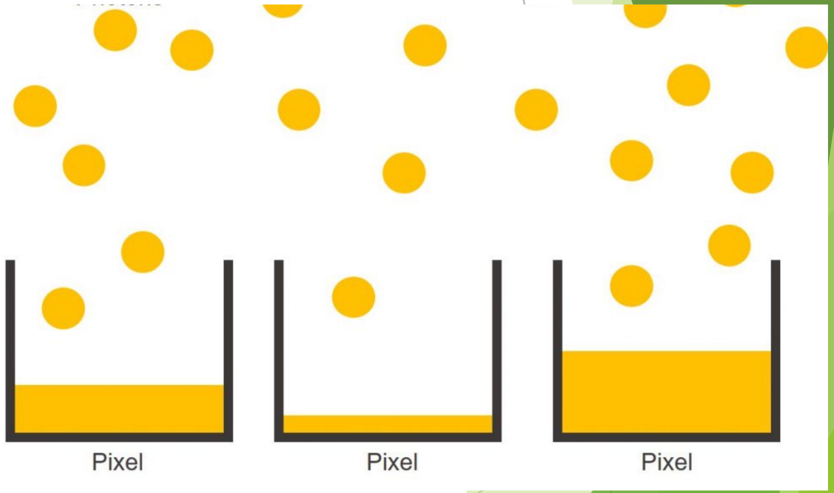


Quantum noise



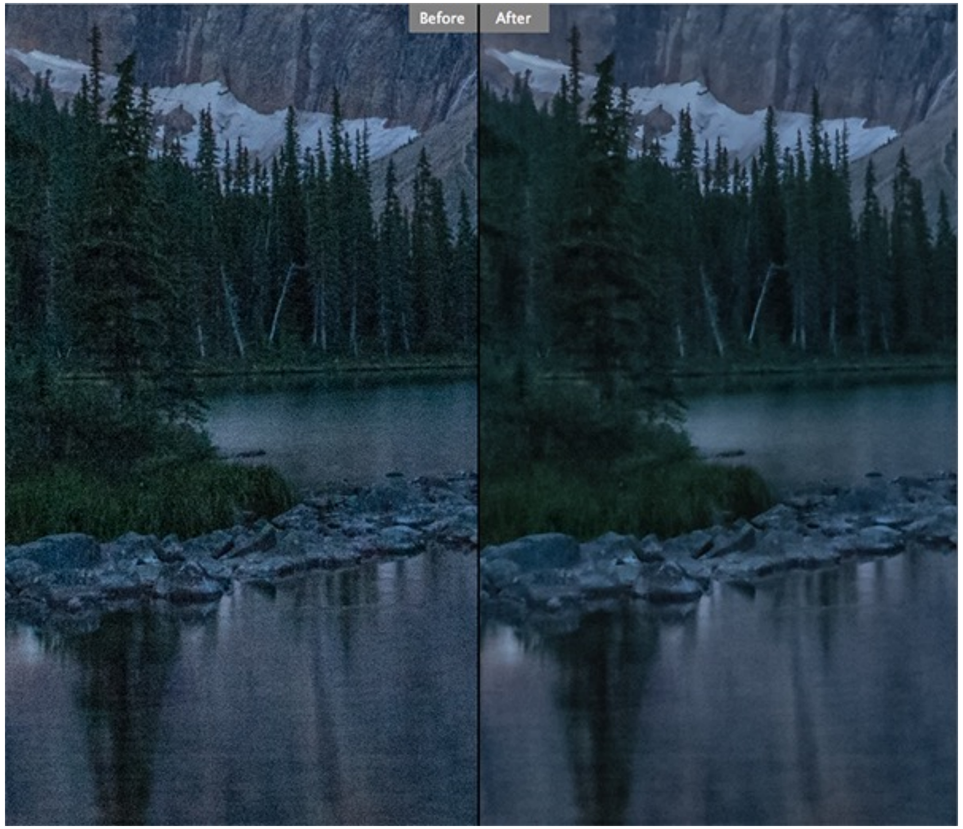
Neural time series analysis — MX Cohen — sincxpress.com

Photon shot noise

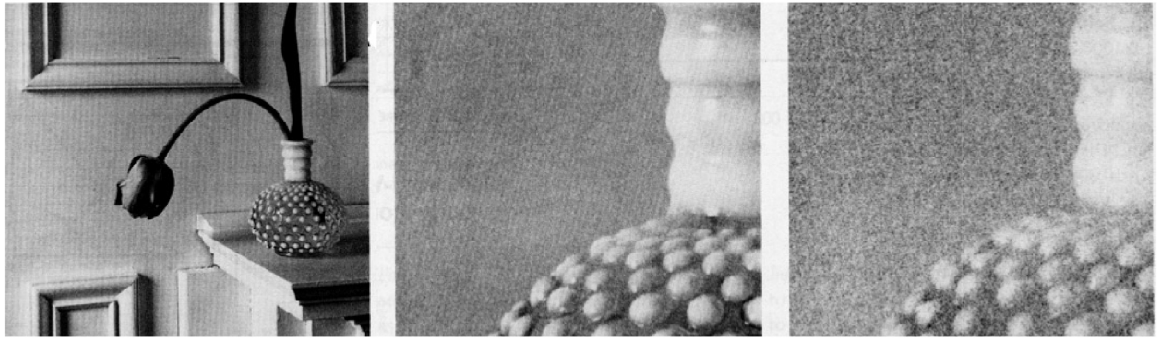


Noise in Image

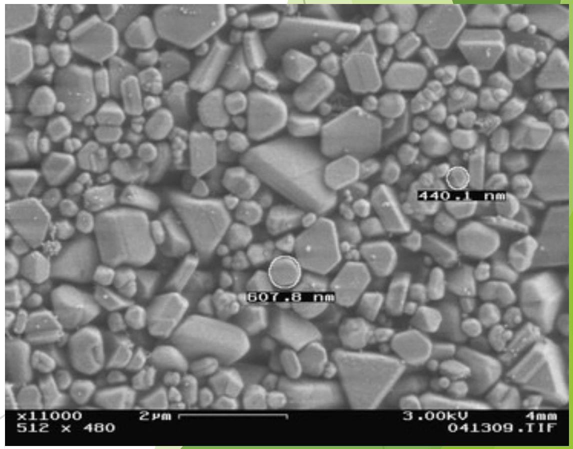
Digital



Film

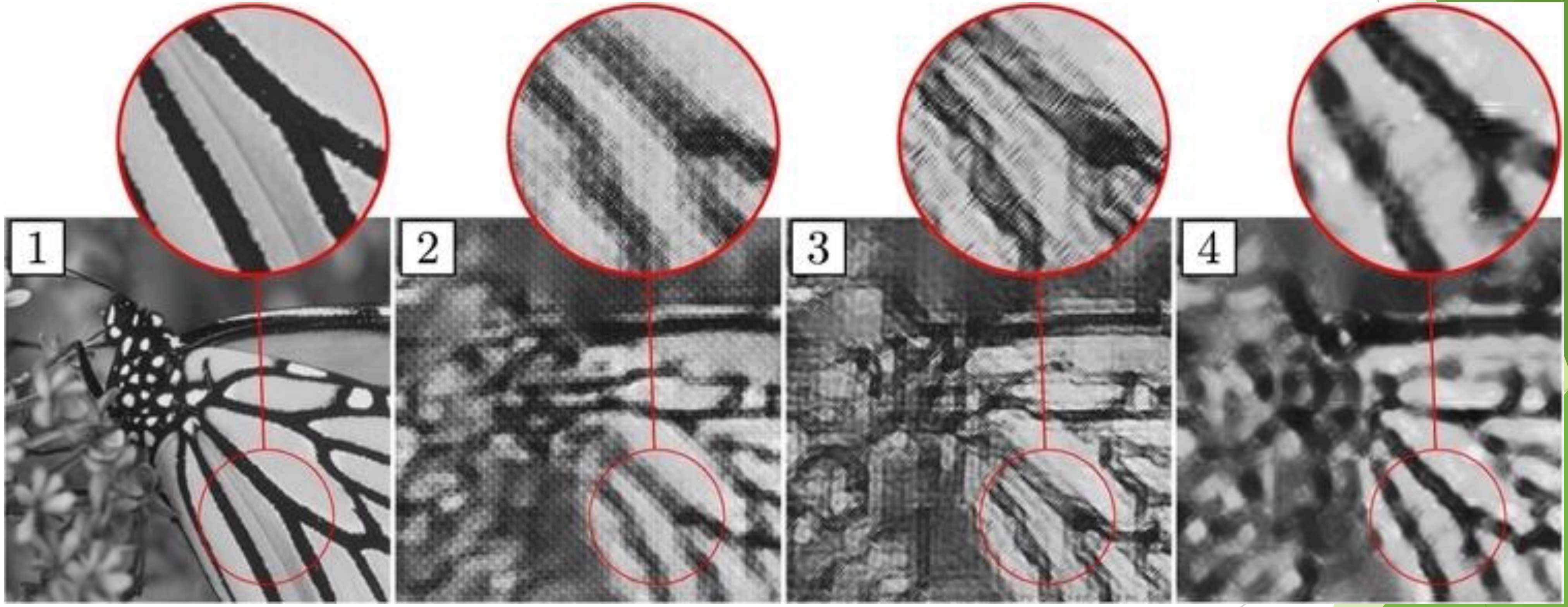


Kodak H1 film grain a.) zoomed out b.) fine grain c.) coarse grain

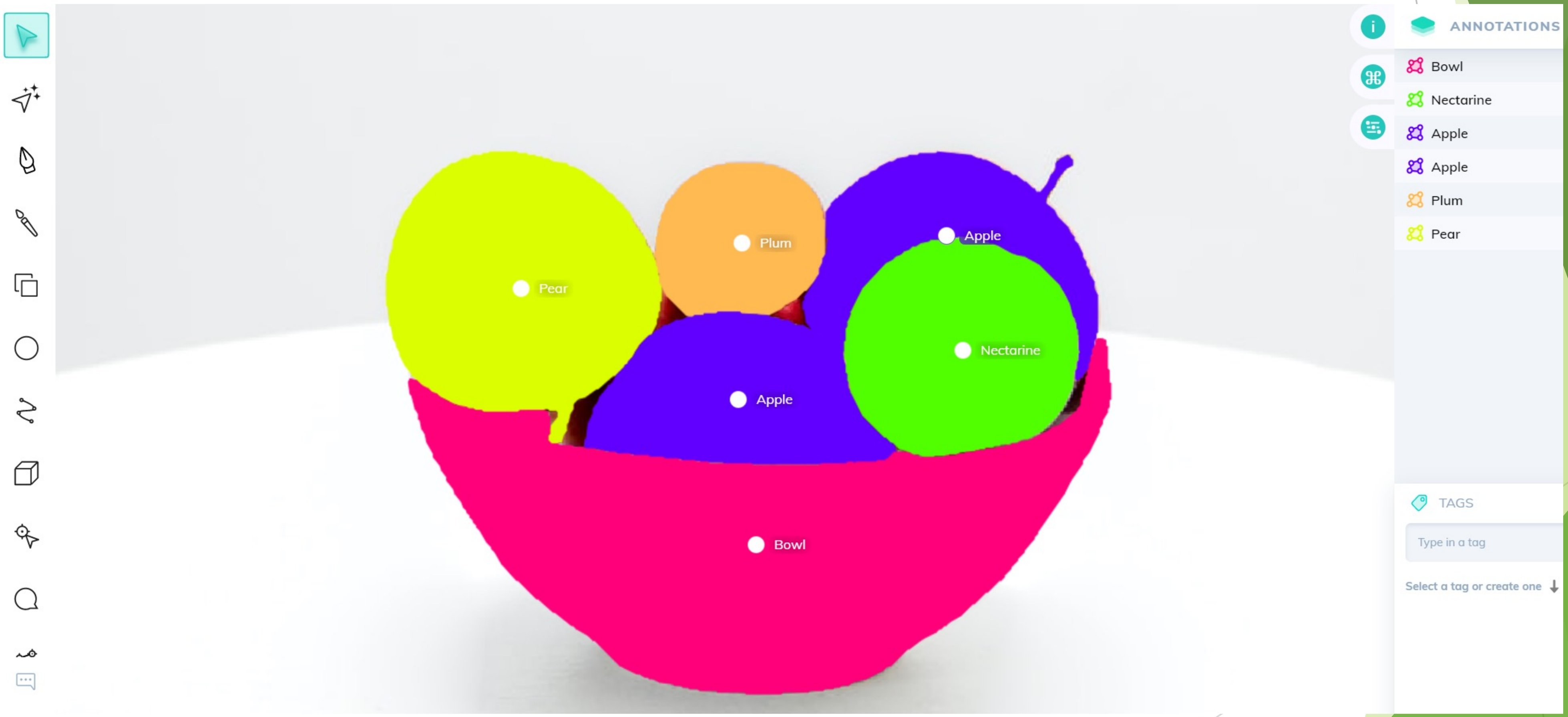


SEM of film grain

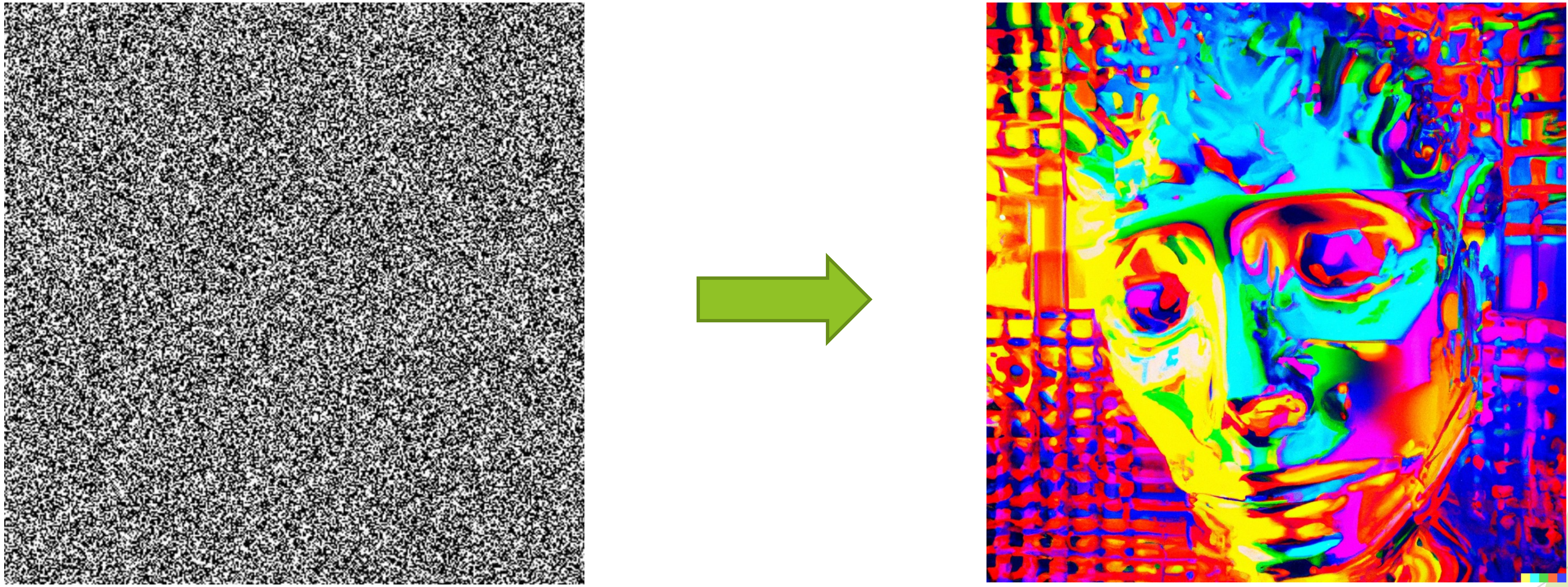
Perceptual Compression



Semantic Segmentation

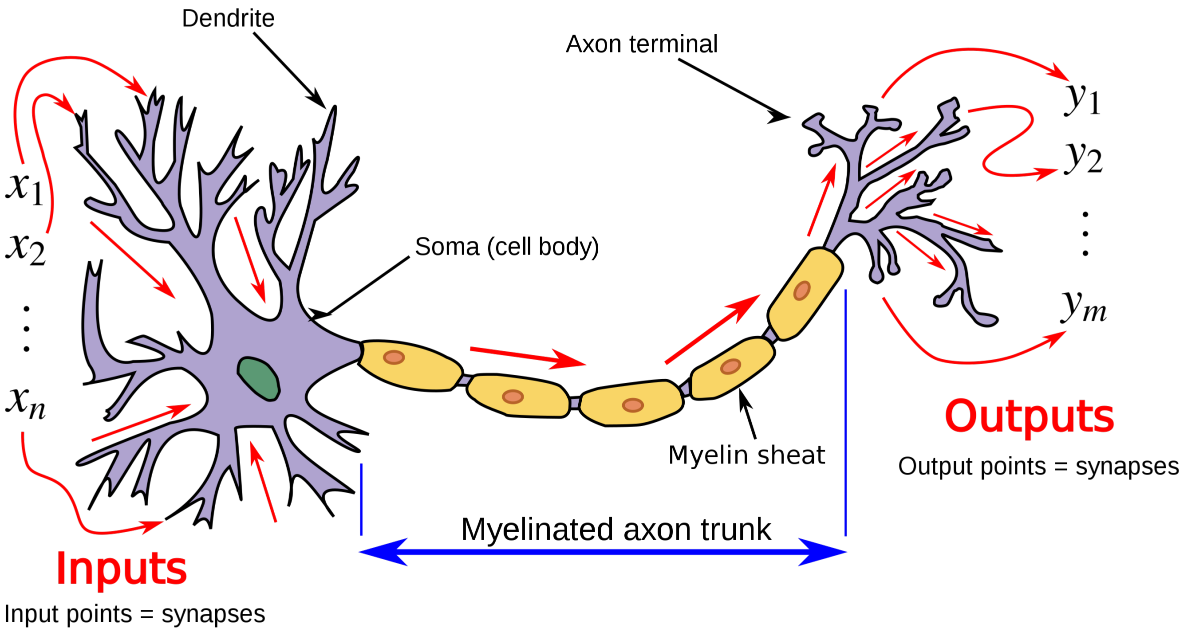


Text-to-Image Generation

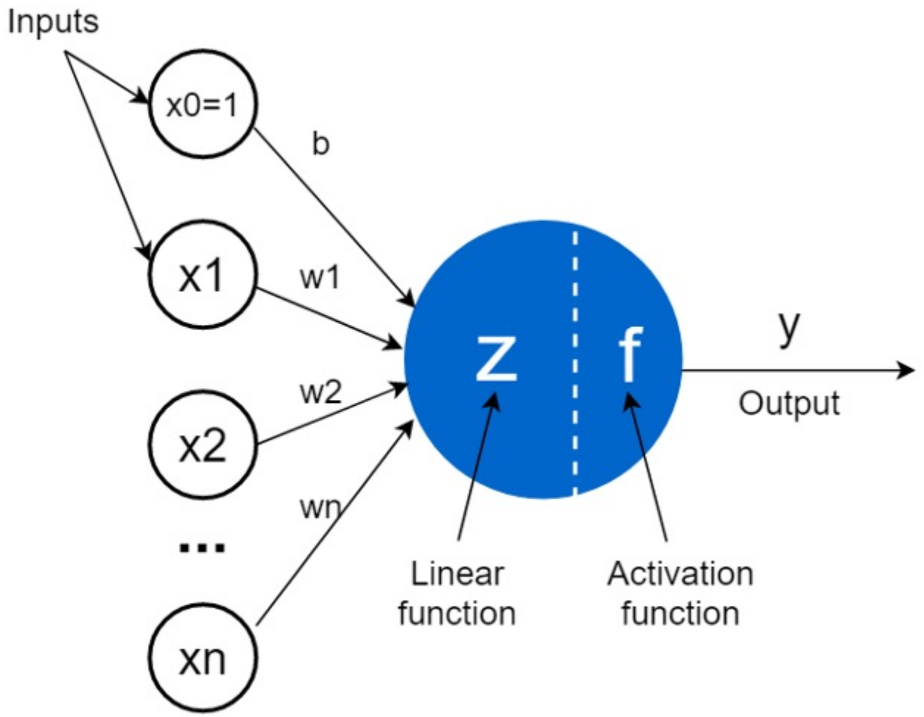


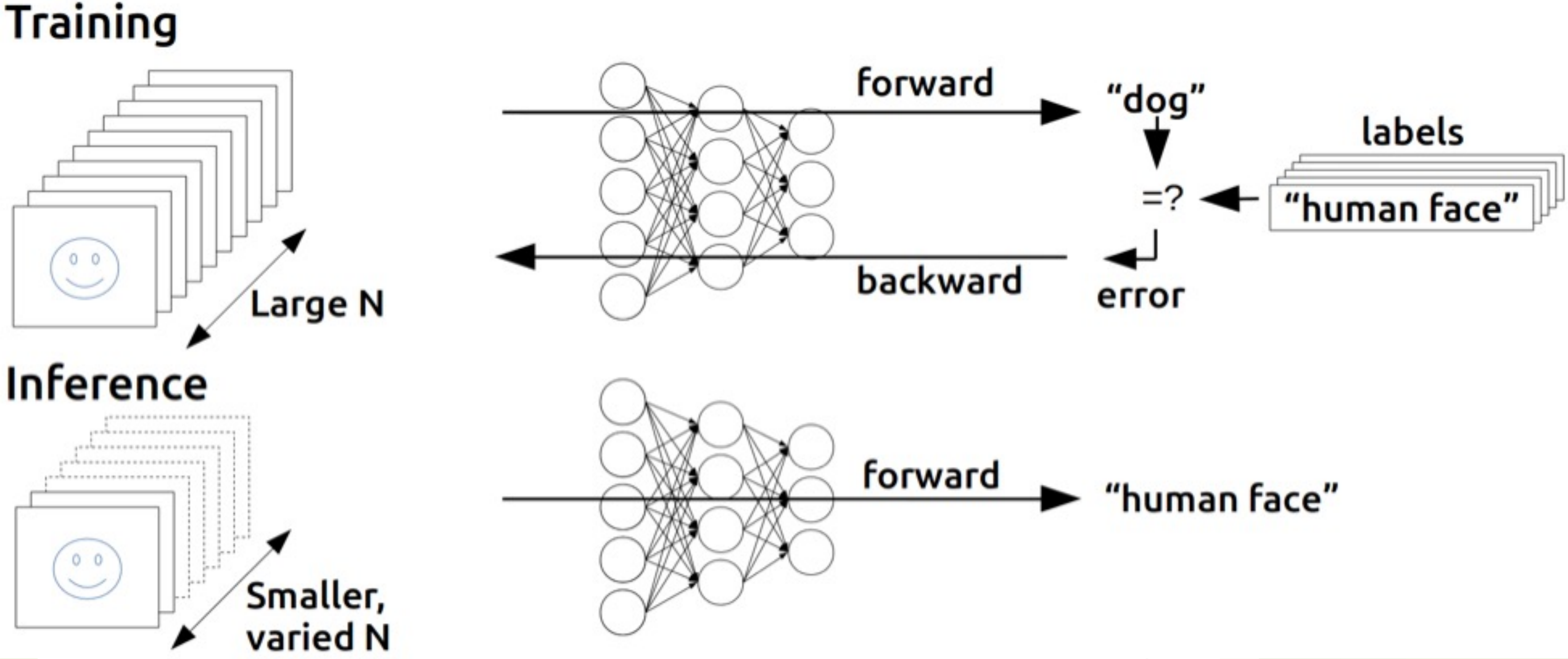
Neural Networks

Natural



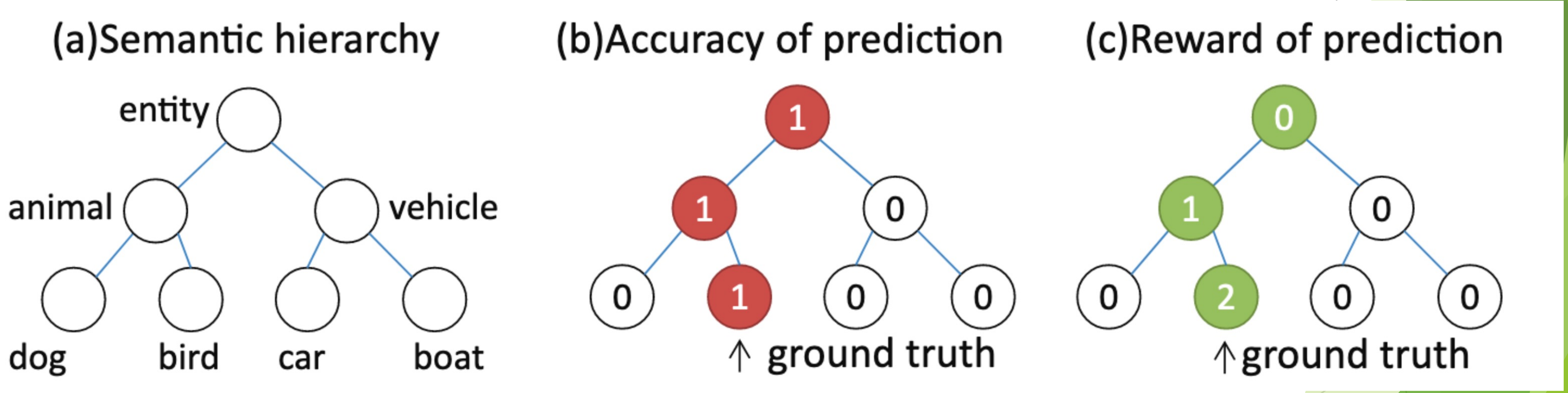
Artificial



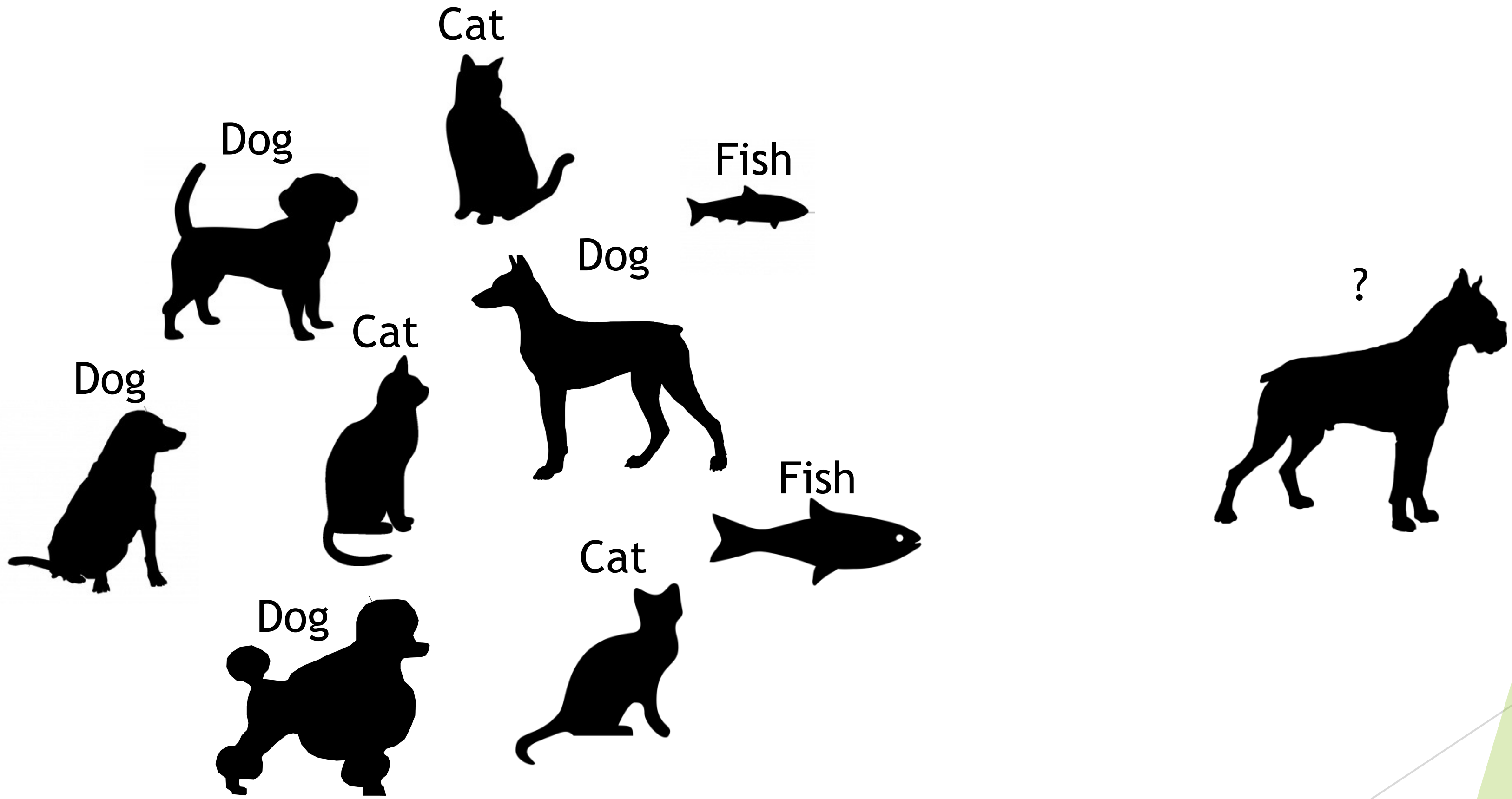

Training
forward
"dog"
labels
=?
"human face"
backward
error
Large N
Inference
forward
"human face"
Smaller,
varied N

ImageNet

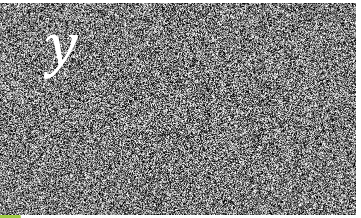
Wordnet

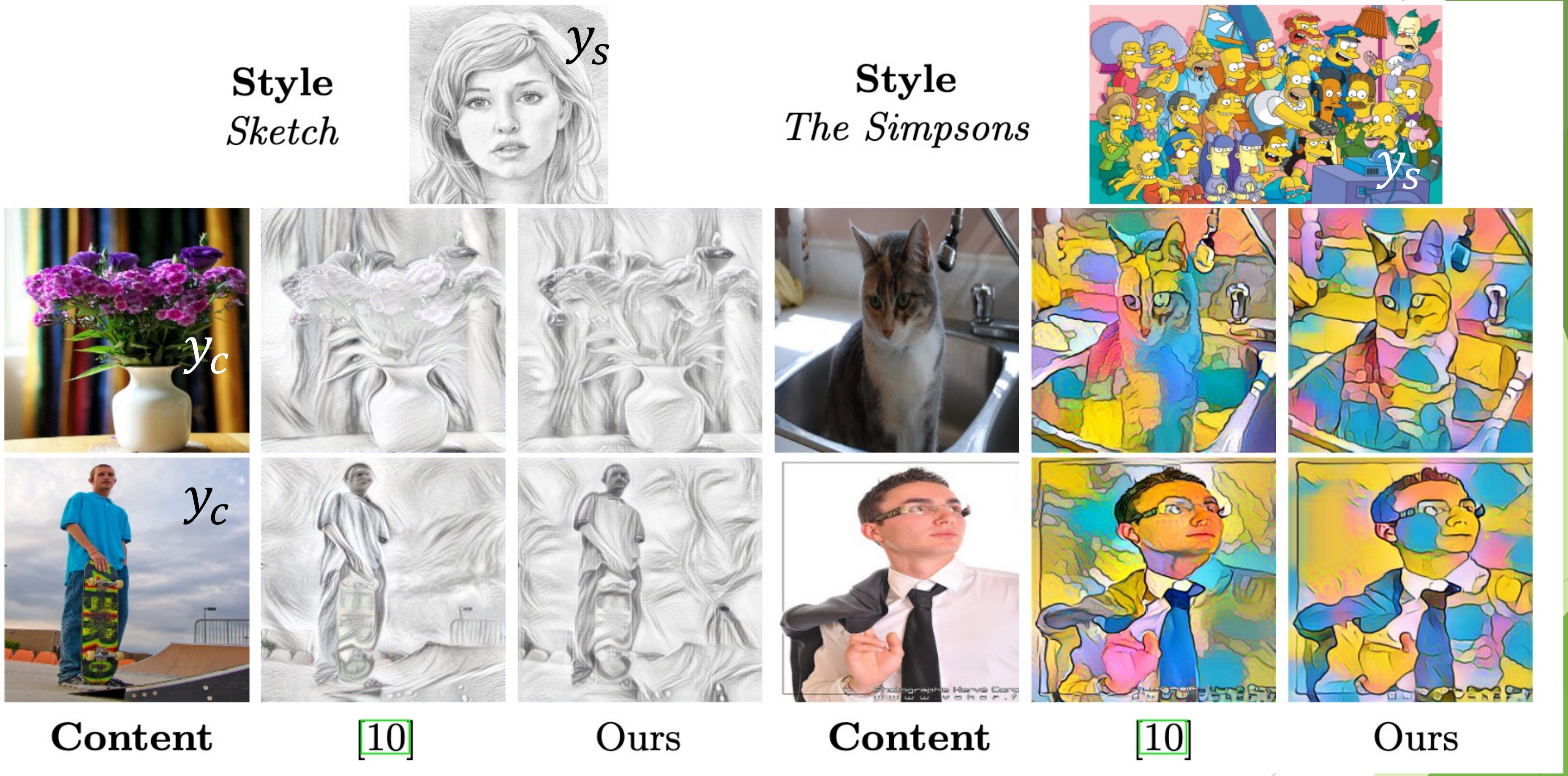

(a)Semantic hierarchy
entity
animal
vehicle
dog
bird
car
boat
(b)Accuracy of prediction
1
1
0
0
1
0
0
↑ ground truth
(c)Reward of prediction
0
1
0
0
2
0
0
↑ground truth

Classifiers



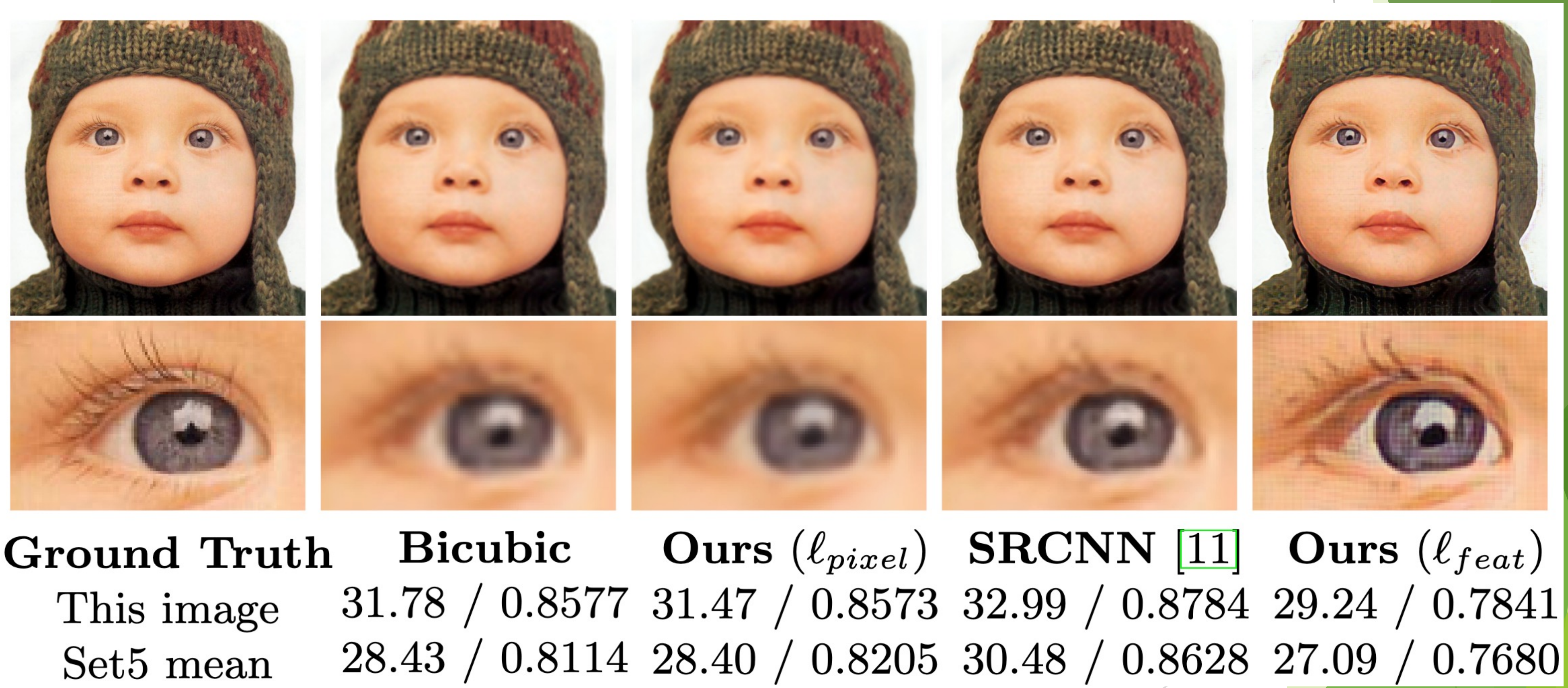
Style Transfer





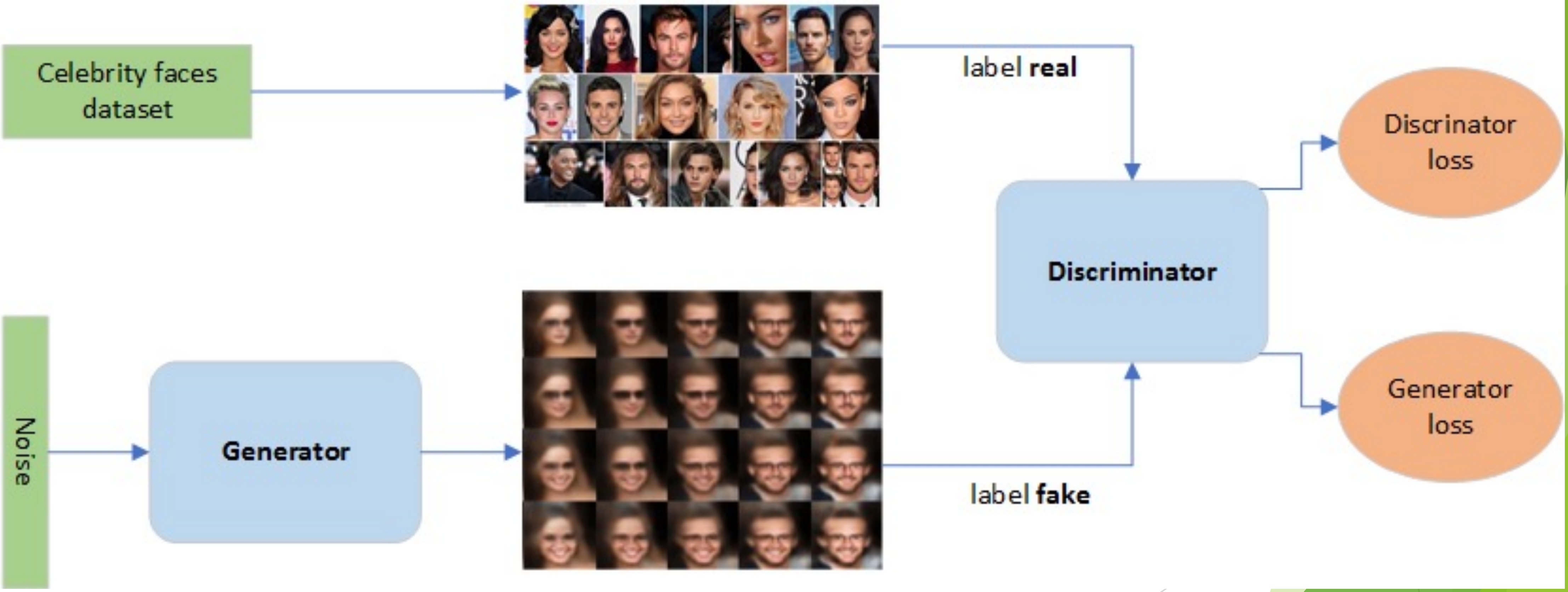
$$\hat{y} = \arg\min_{y} \lambda_c \ell_{feat}^{\phi,j}(y, y_c) + \lambda_s \ell_{style}^{\phi,J}(y, y_s) + \lambda_{TV}\ell_{TV}(y)$$

Super-resolution

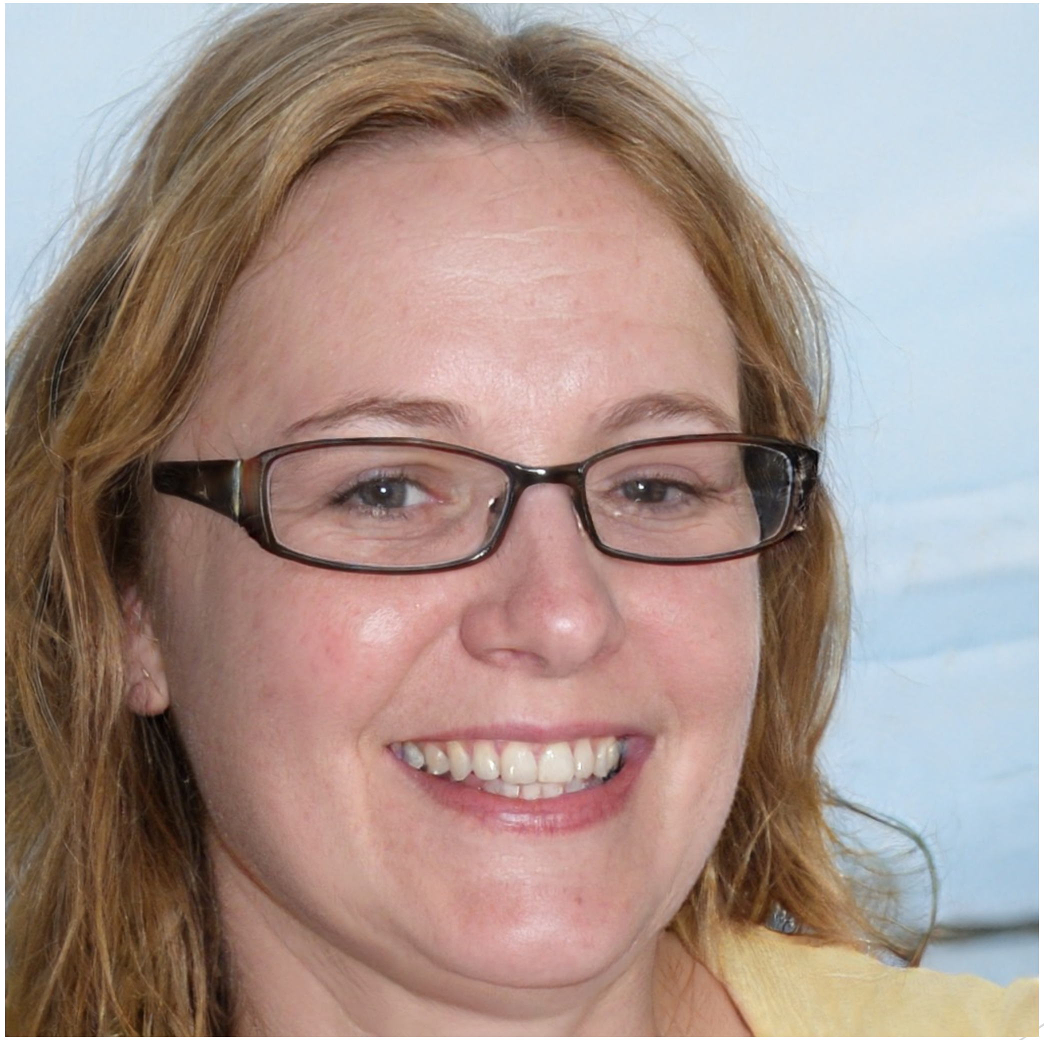


Ground Truth	Bicubic	Ours (ℓ_{pixel})	SRCNN [11]	Ours (ℓ_{feat})
This image	31.78 / 0.8577	31.47 / 0.8573	32.99 / 0.8784	29.24 / 0.7841
Set5 mean	28.43 / 0.8114	28.40 / 0.8205	30.48 / 0.8628	27.09 / 0.7680

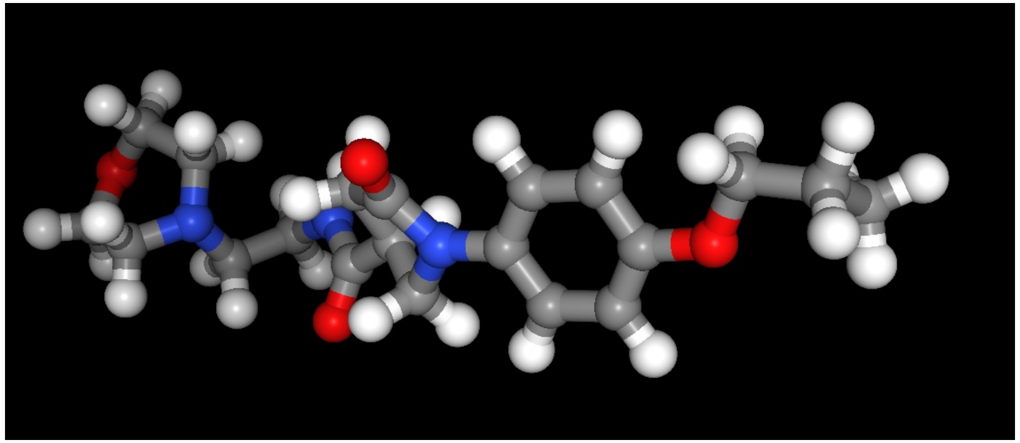
GANs


Celebrity faces dataset
label real
Discrinator loss
Discriminator
Noise
Generator
Generator loss
label fake

thispersondoesnotexist.com

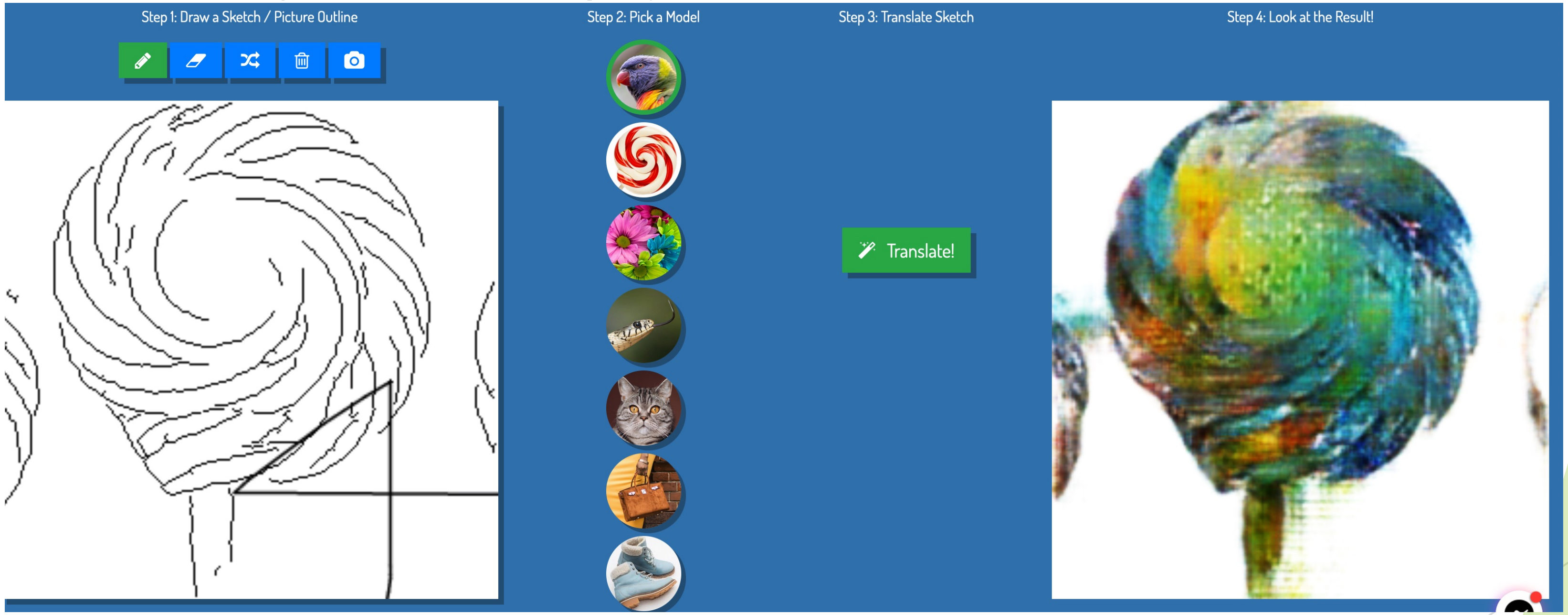


This x does not exist



GAN Play

- https://mitmedialab.github.io/GAN-play/



Step 1: Draw a Sketch / Picture Outline

Step 2: Pick a Model

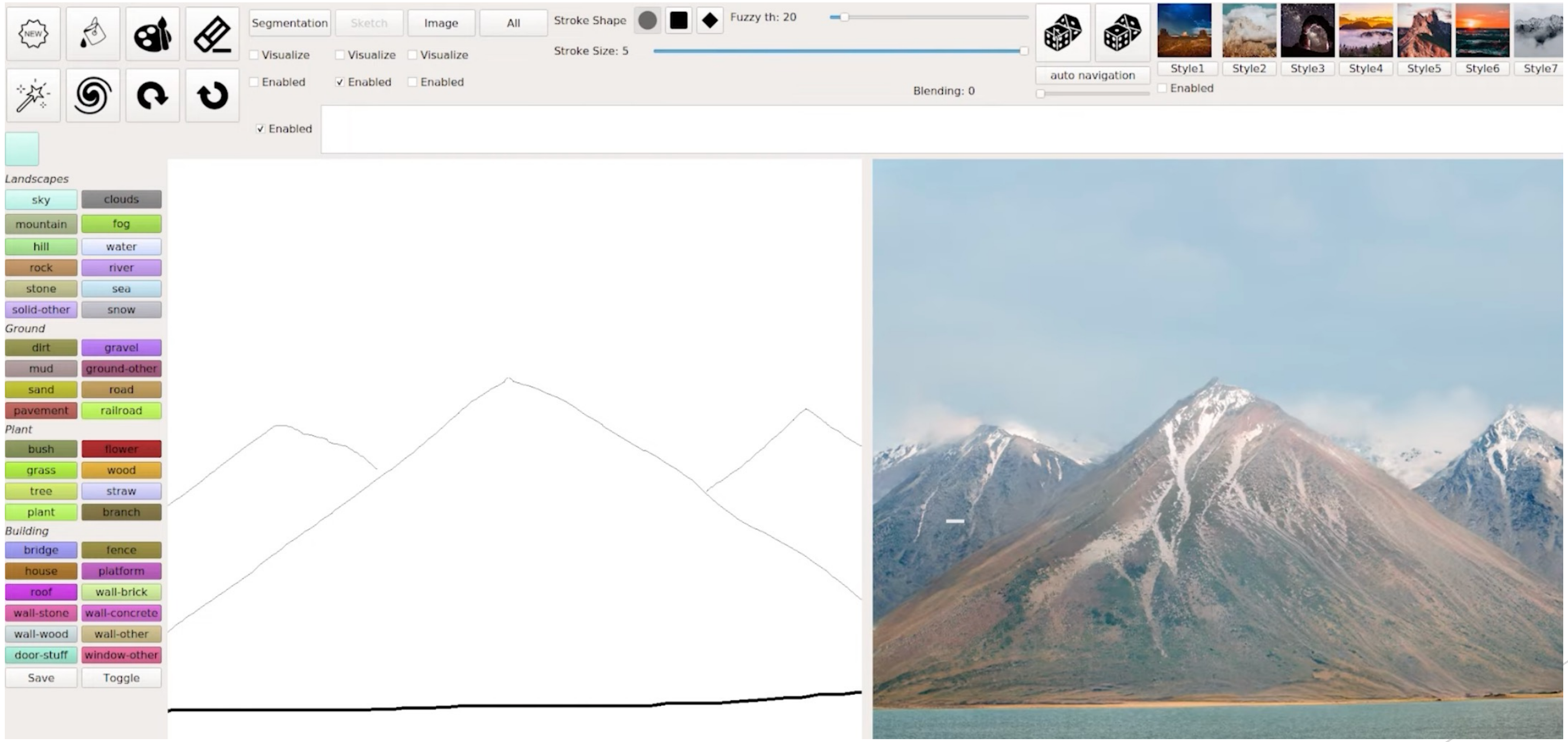
Step 3: Translate Sketch

Step 4: Look at the Result!

https://phillipi.github.io/pix2pix/ -- the relevant paper 2017.
https://affinelayer.com/pixsrv/
Pix to pix

GAUGAN

- http://gaugan.org/gaugan2/

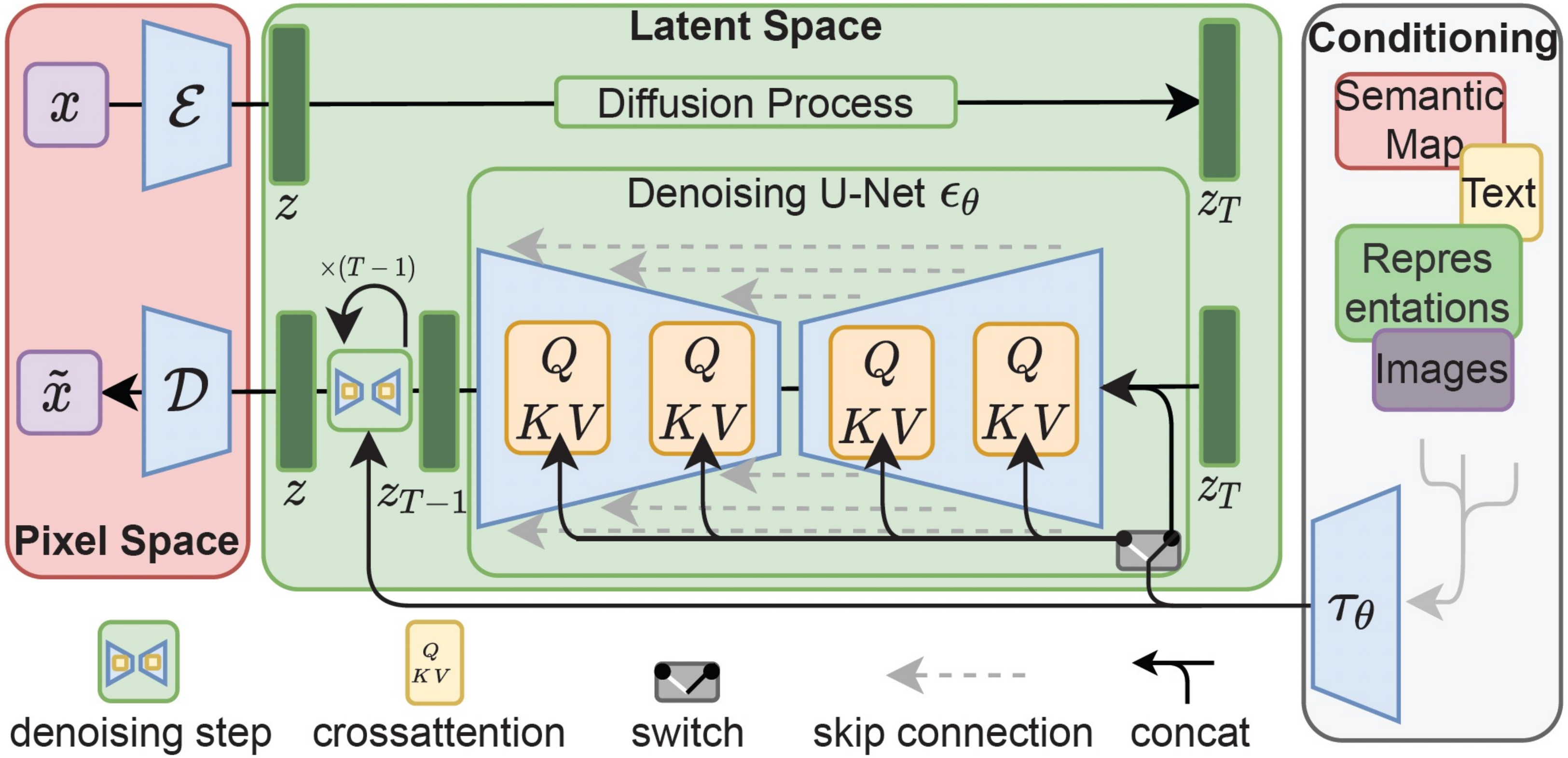


Stroke Size: 5

Fuzzy th: 20

Blending: 0

Stable Diffusion



Diffusion

- Predicts the score function $\nabla_x \log p(x)$ for an unconditional model
- Adding conditioning:

$$\nabla_x \log p(x|y) = \nabla_x \log p(y|x) + \nabla_x \log p(x)$$

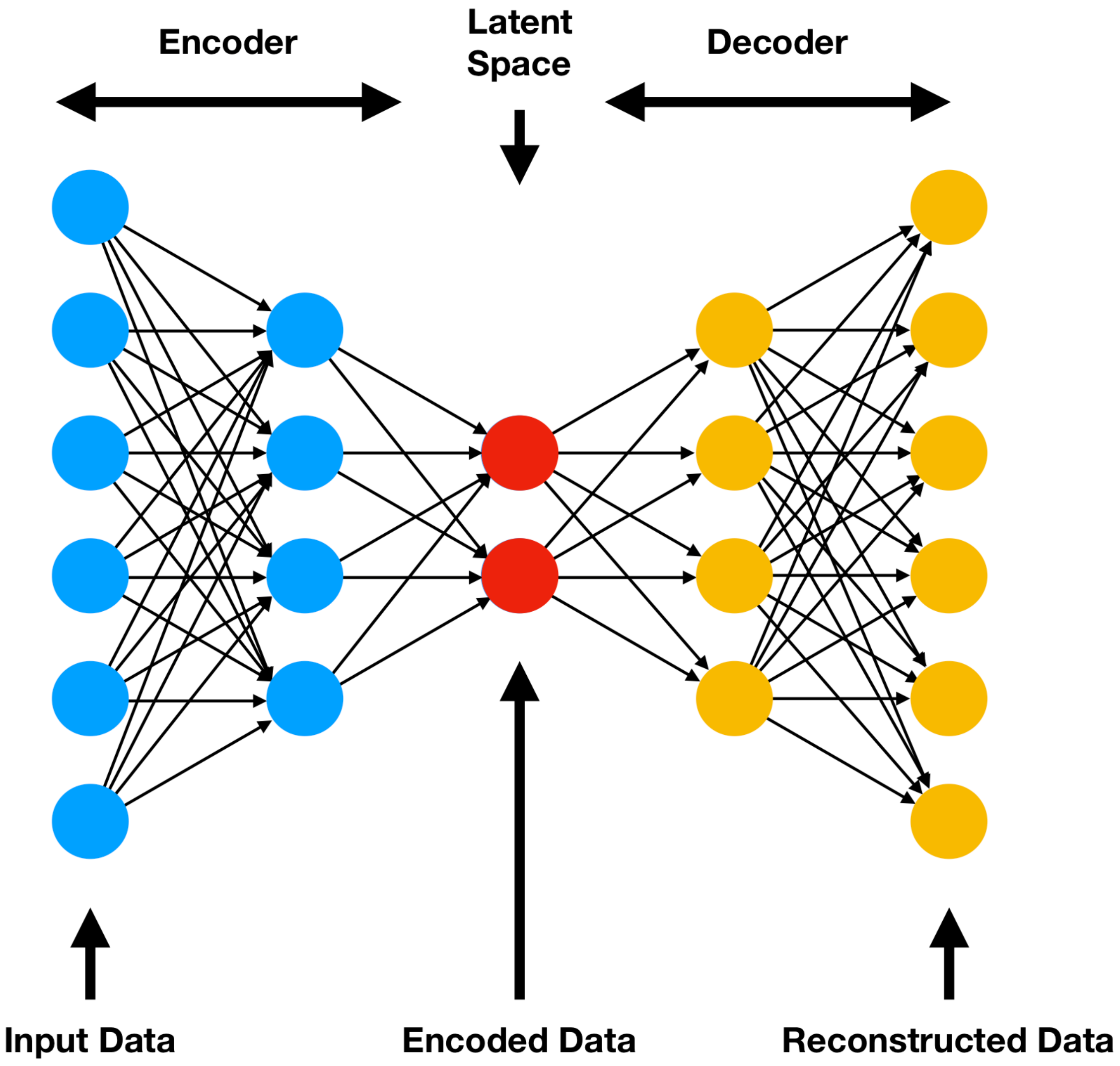
Where y is your conditioning i.e your text input

$$\nabla_x \log p_\gamma(x|y) = \gamma \nabla_x \log p(y|x) + \nabla_x \log p(x)$$

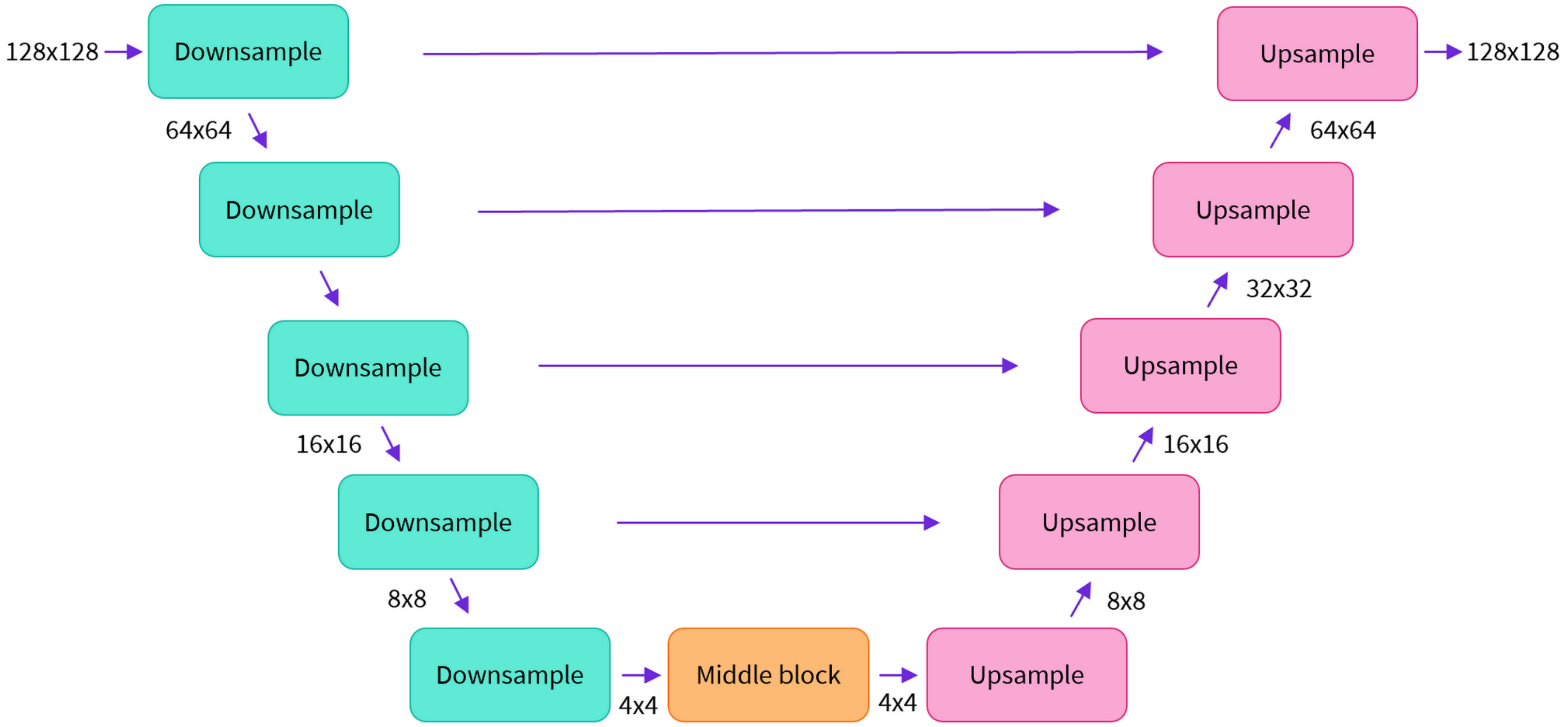
Where γ is the guidance scale

$$\nabla_x \log p_\gamma(x|y) = (1-\gamma)\nabla_x \log p(x) + \gamma \nabla_x \log p(x|y)$$

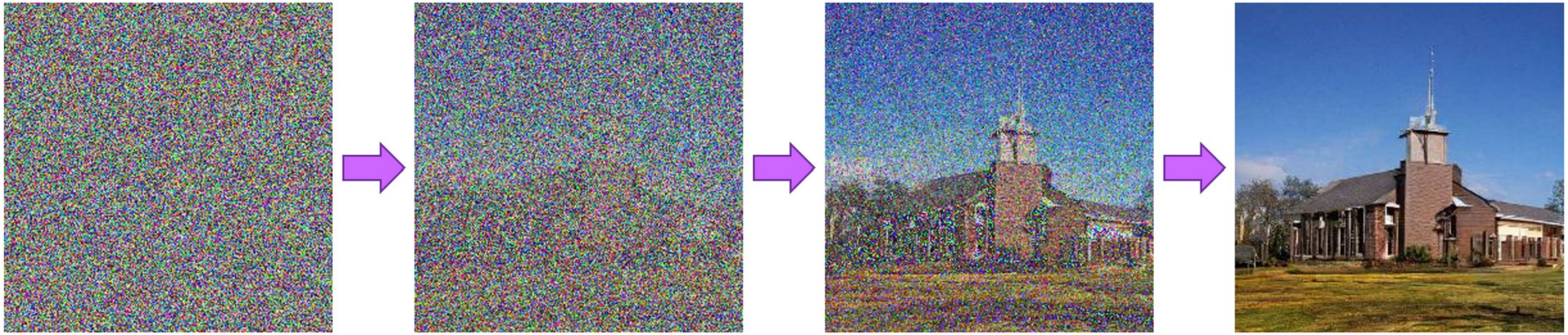
VAE (Variational Autoencoder)



U-Net


128x128
Downsample
64x64
Downsample
Downsample
16x16
Downsample
8x8
Downsample
4x4
Middle block
4x4
Upsample
8x8
Upsample
16x16
Upsample
32x32
Upsample
64x64
Upsample
128x128

Denoising



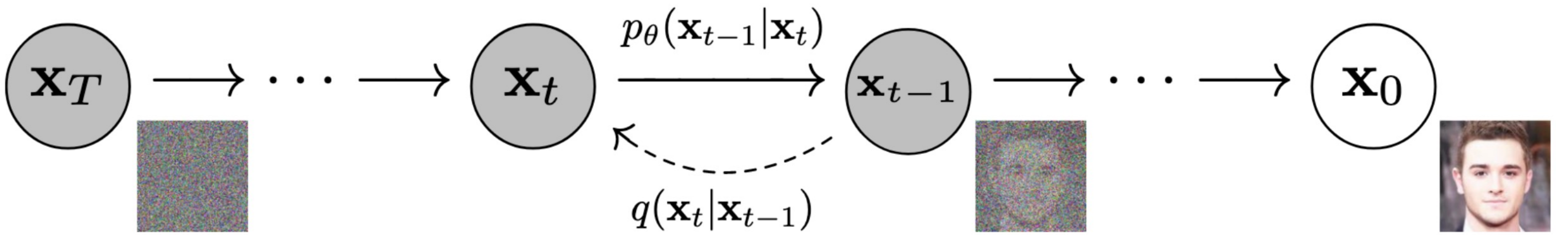
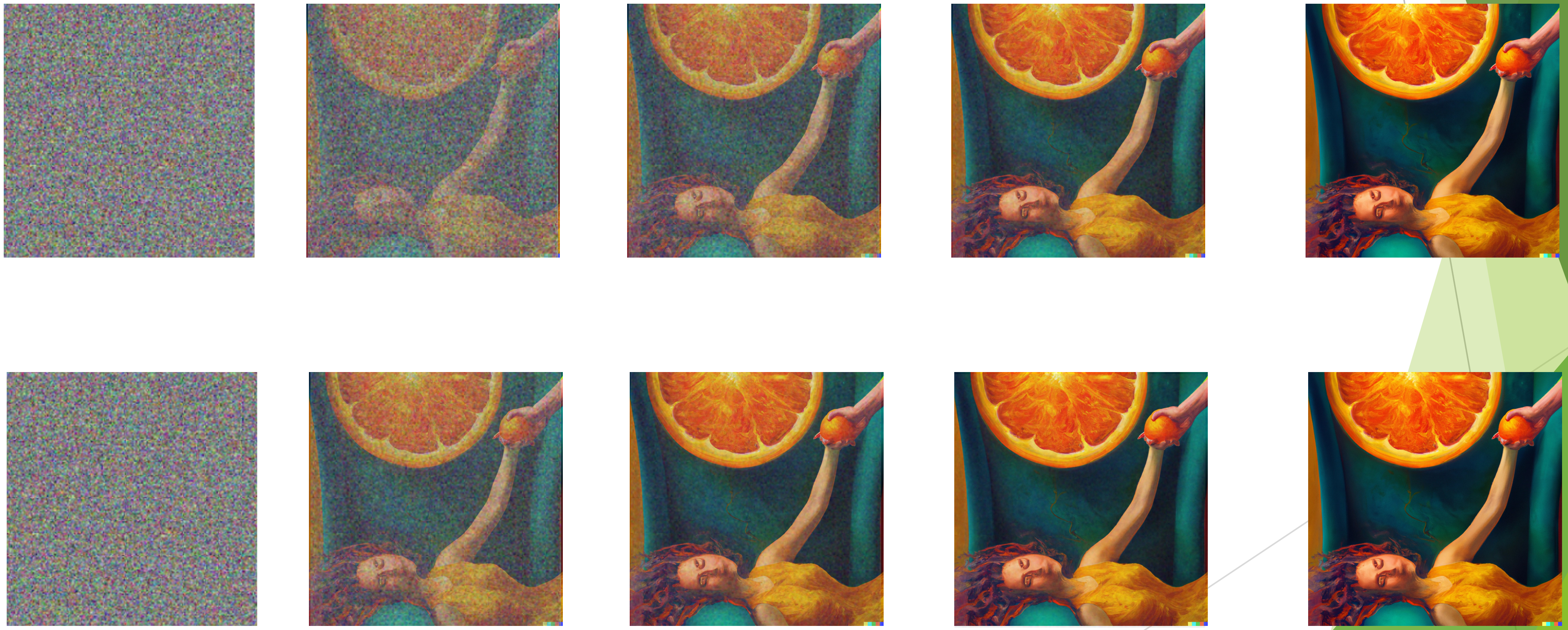

$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_{t-1}$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
$\mathbf{x}_0$

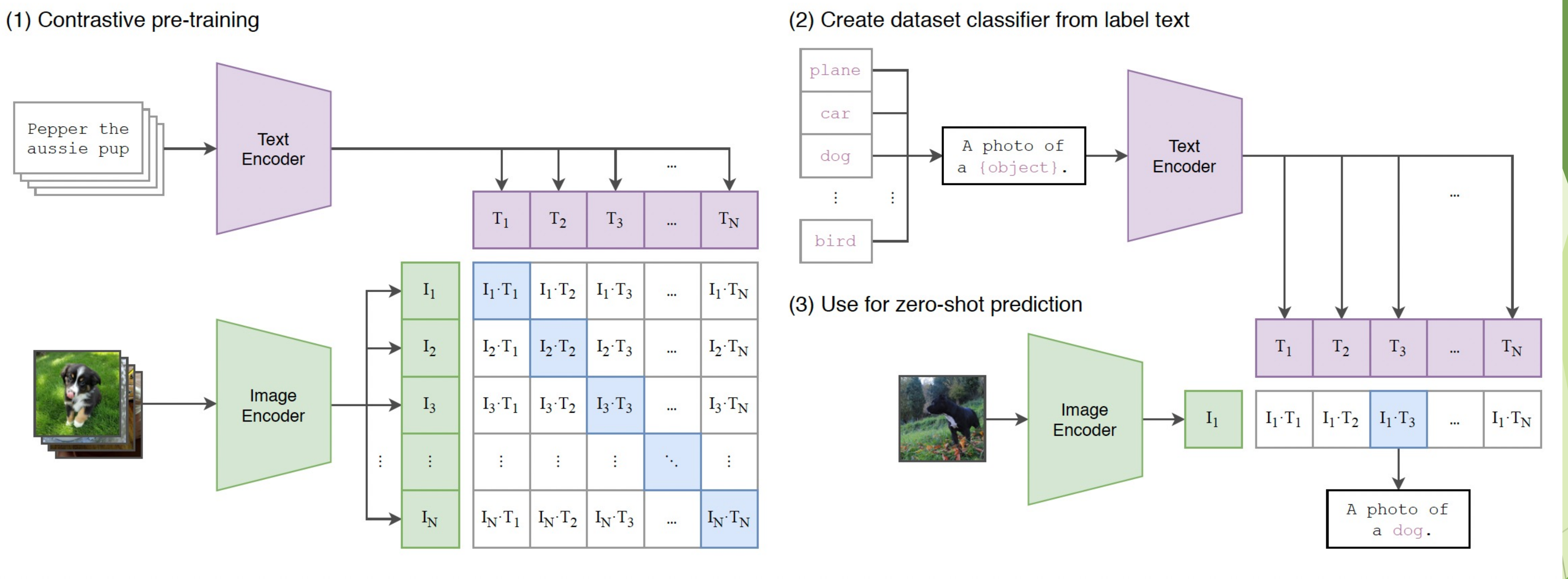
Figure 2: The directed graphical model considered in this work.

Schedulers

- Sets how much noise the decoder tries to remove with each step



CLiP



(1) Contrastive pre-training

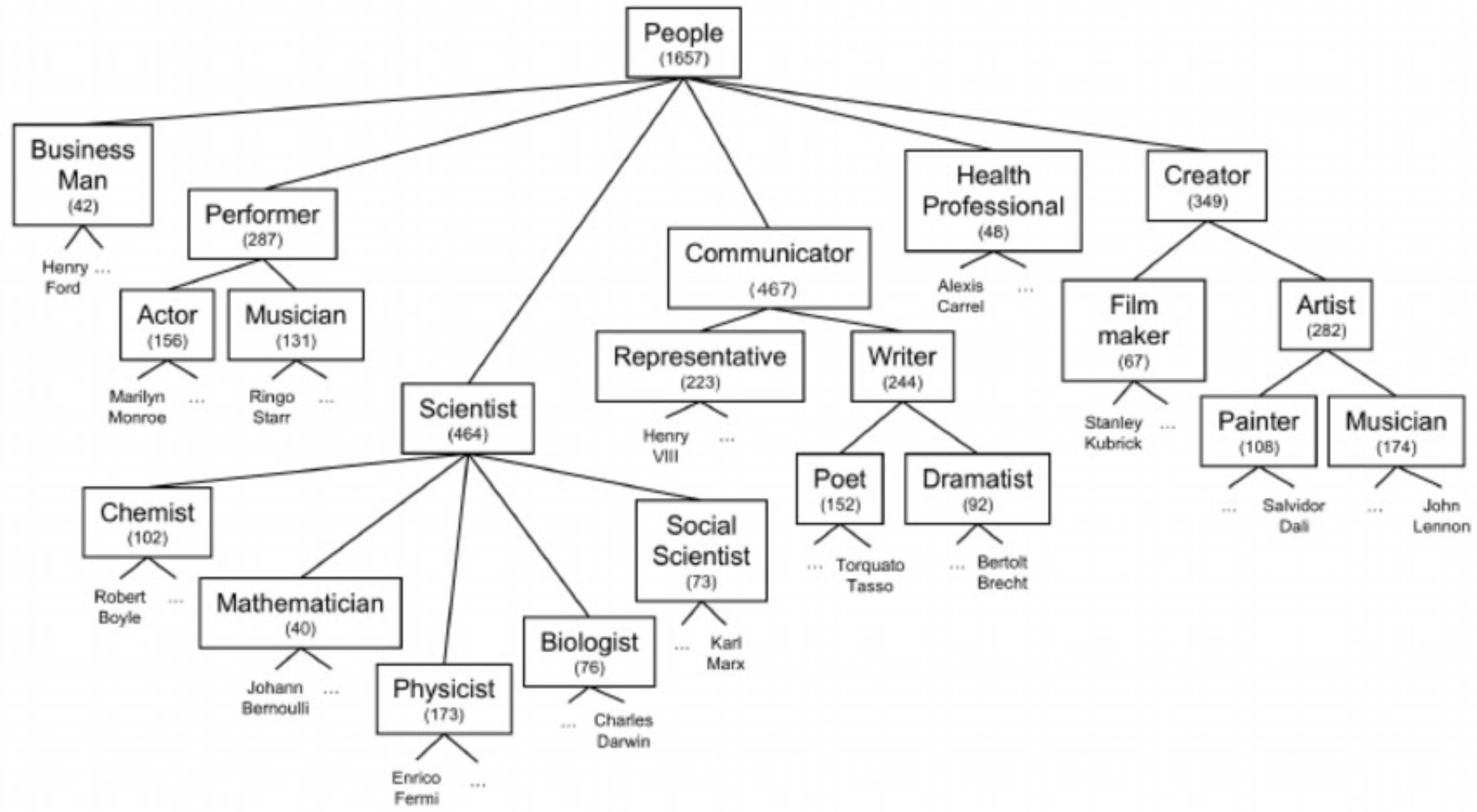
	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

(2) Create dataset classifier from label text

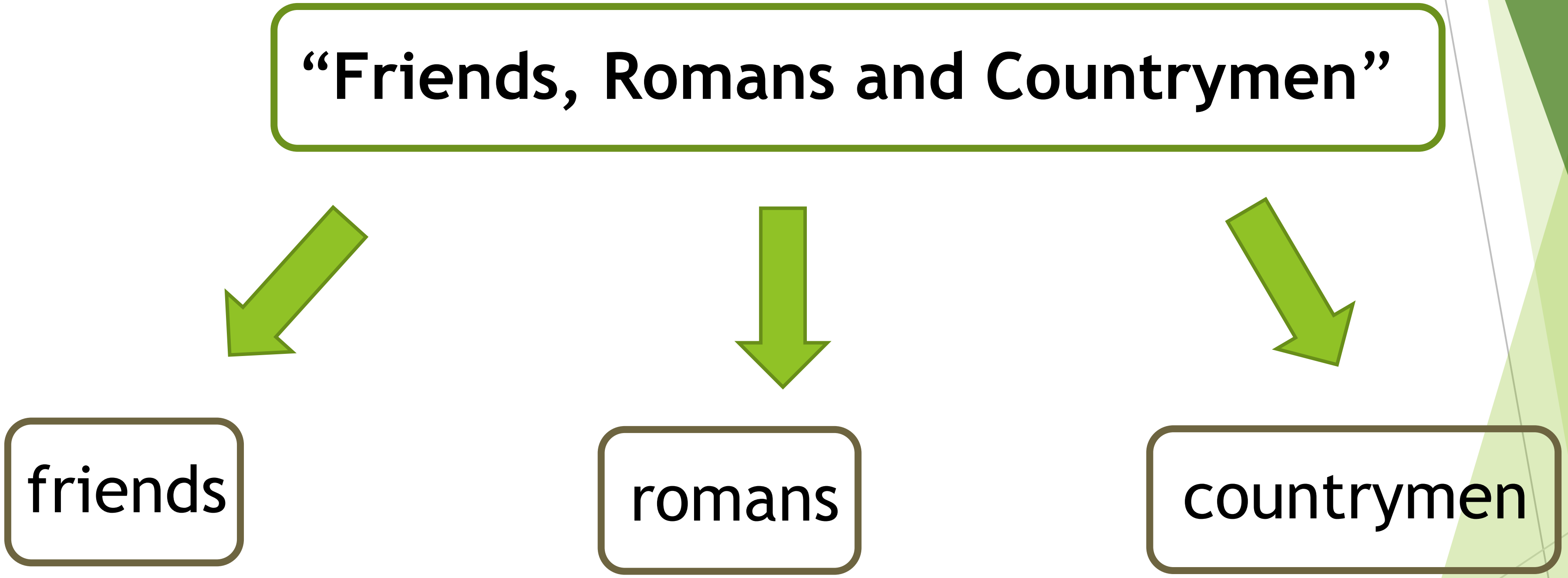
(3) Use for zero-shot prediction

$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$

Ontological Model



Tokenizer



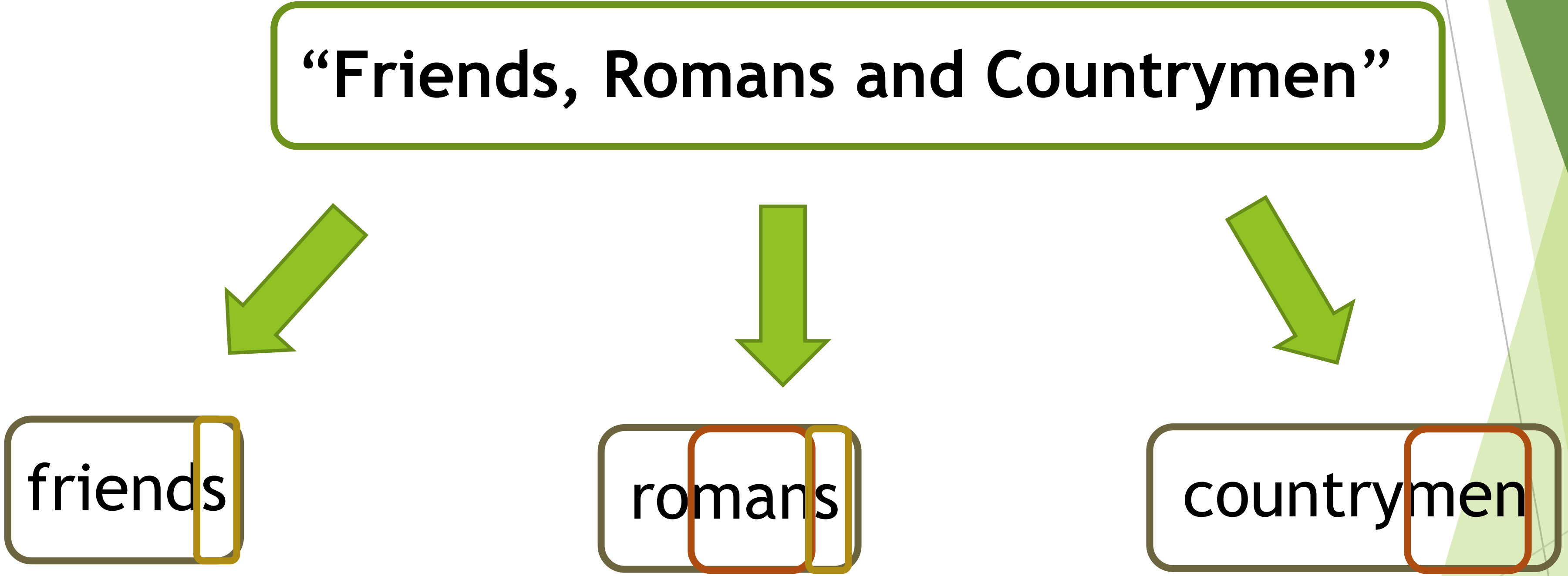
“Friends, Romans and Countrymen”

friends

romans

countrymen

Tokenizer



“Friends, Romans and Countrymen”

friends

romans

countrymen

Text Embedding

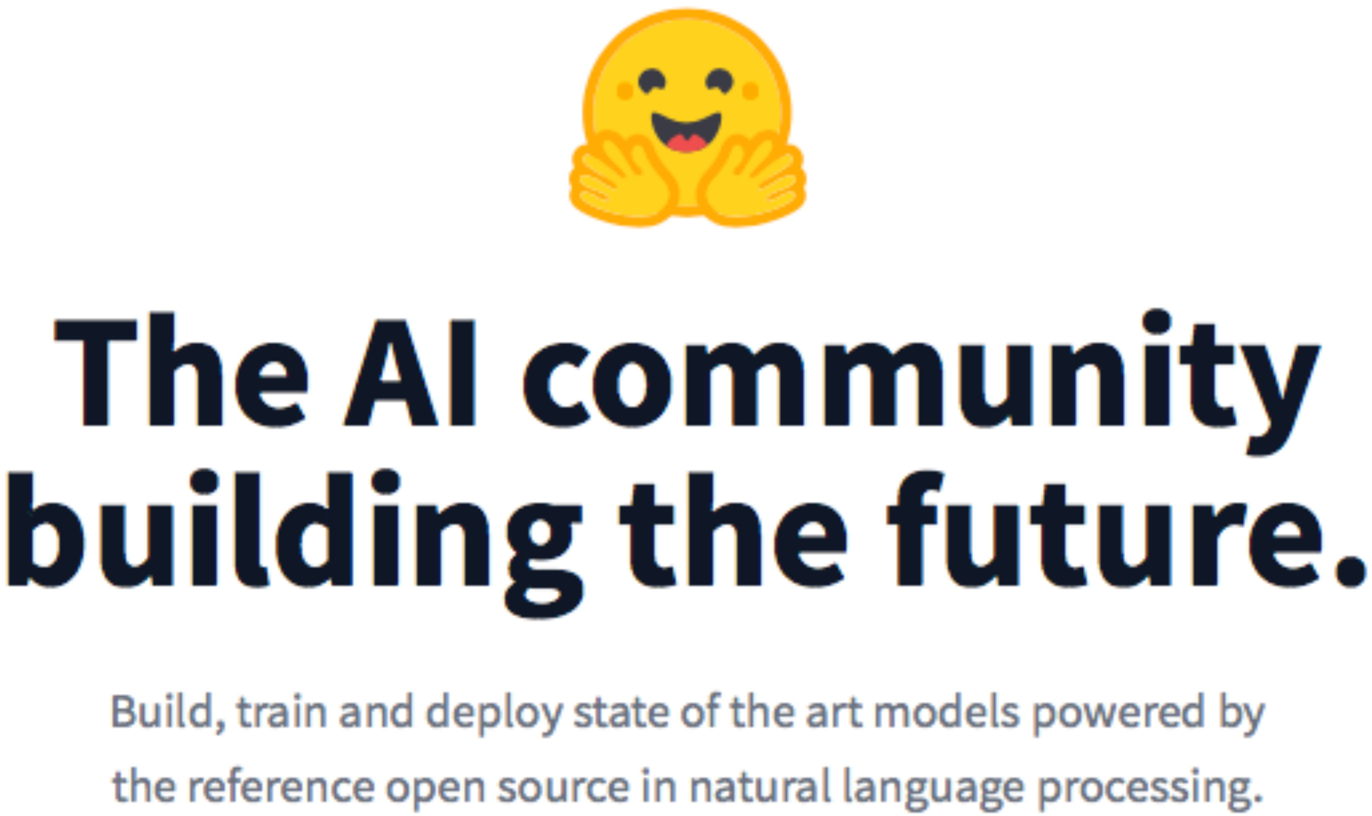
family

friends

romans

Italy

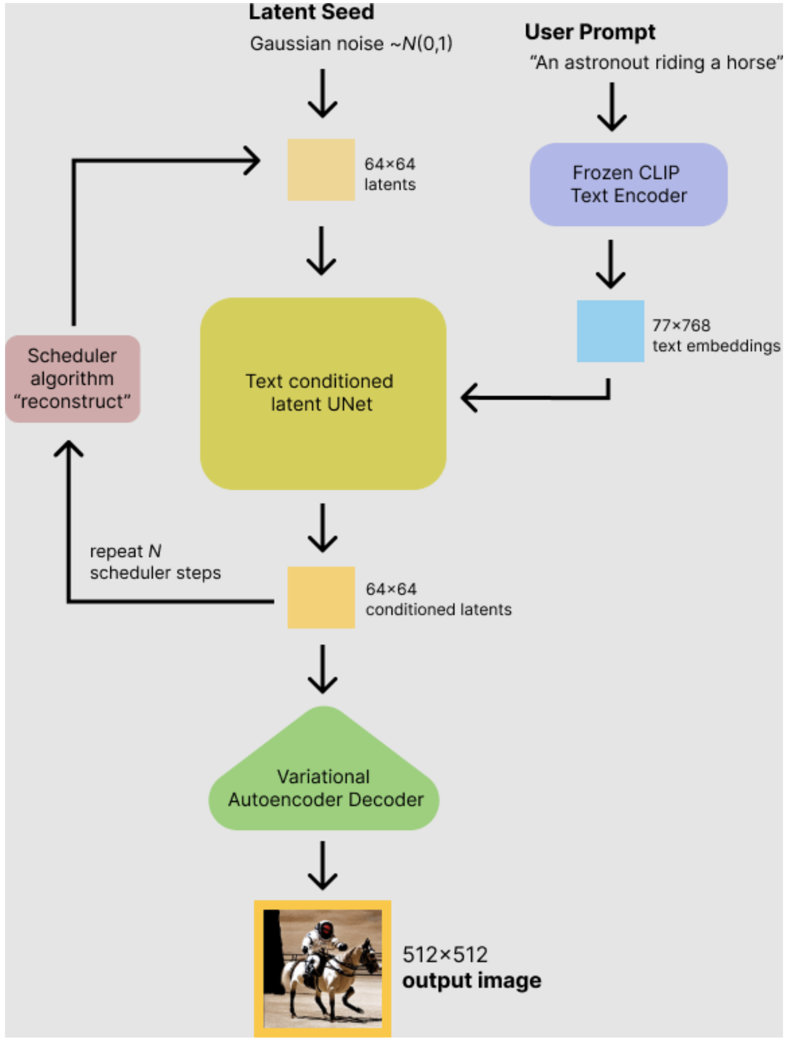
Hugging Face



The AI community building the future.

Build, train and deploy state of the art models powered by the reference open source in natural language processing.

Stable Diffusion API with HuggingFace



Let’s try it out

Imagen

- Discovered language model trained only on text data are good text encoders for text-to-image
- Increasing the size of a text-only language model improves output quality more efficiently than increasing the size of an image diffusion model
- Dynamic thresholding

Latent Space (mathematically)

- items resembling each other are positioned closer to one another in the latent space
- Embedding $f: X \rightarrow Y$
  - Def: An instance of a mathematical structure that is contained within another instance such as the rational numbers within integers
  - Must be injective (i.e. 1:1)

$$f(x_1) = f(x_2)\ implies\ x_1 = x_2$$

Latent Space (intuitively)