Lightricks

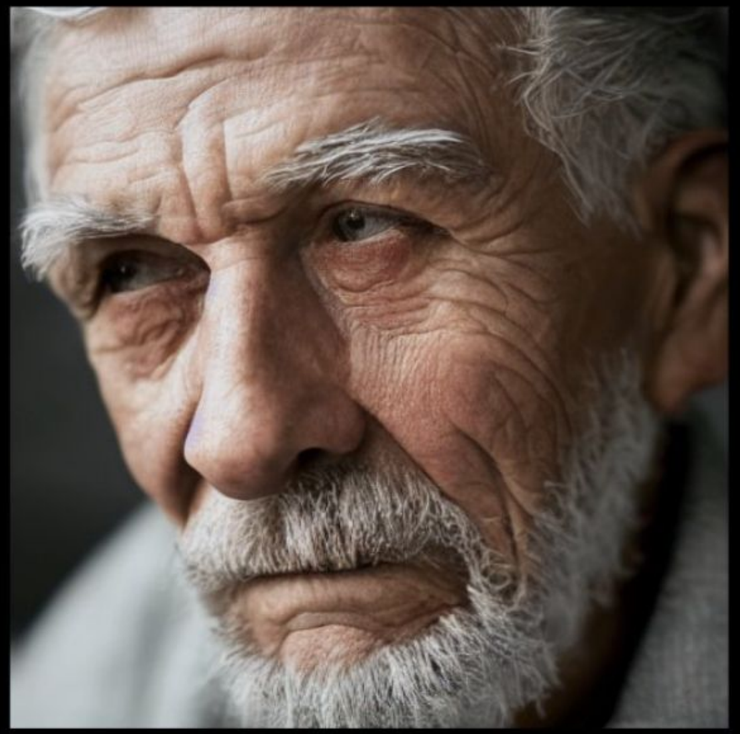
Controlling Generative AI

Naomi Ken Korem
Researcher

MIDJOURNEY

DALL-E 2

STABLEDIFFUSION



film still, portrait of an old man, wrinkles, dignified look, grey silver hair, peculiar nose, wise, eternal wisdom and beauty, incredible lighting and camera work, depth of field, bokeh, screenshot from a hollywood movie

LAION
@laion_ai

Guiding Stable Diffusion with our CLIP H:
"Professional HDR photo of a polar bear holding a lollipop on a rooftop in Hong Kong looking up at a UFO in the night sky. A UFO flies above the polar bear. The polar bear holds a lollipop on a rooftop. The background shows Hong Kong."

It is all about the control

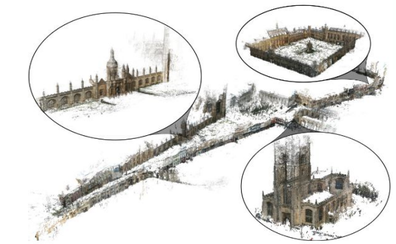
“Add fireworks to the sky”

“Replace the fruits with cake”

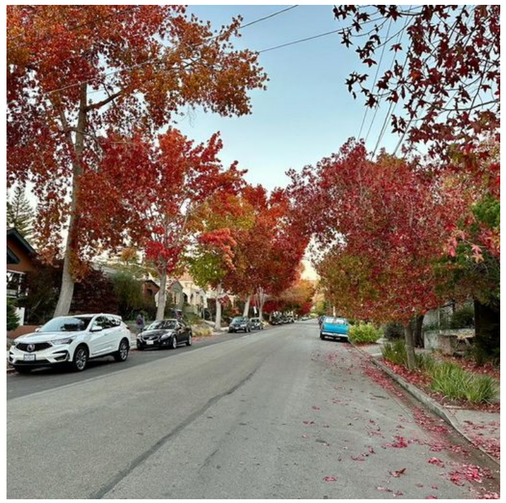
“Turn it into a still from a western”

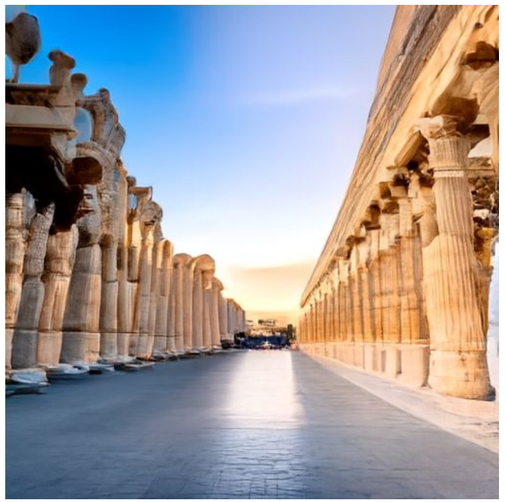
“Make his jacket out of leather”

Lightricks



Naomi Ken Korem





Lightricks

Bridge the gap between imagination and creation

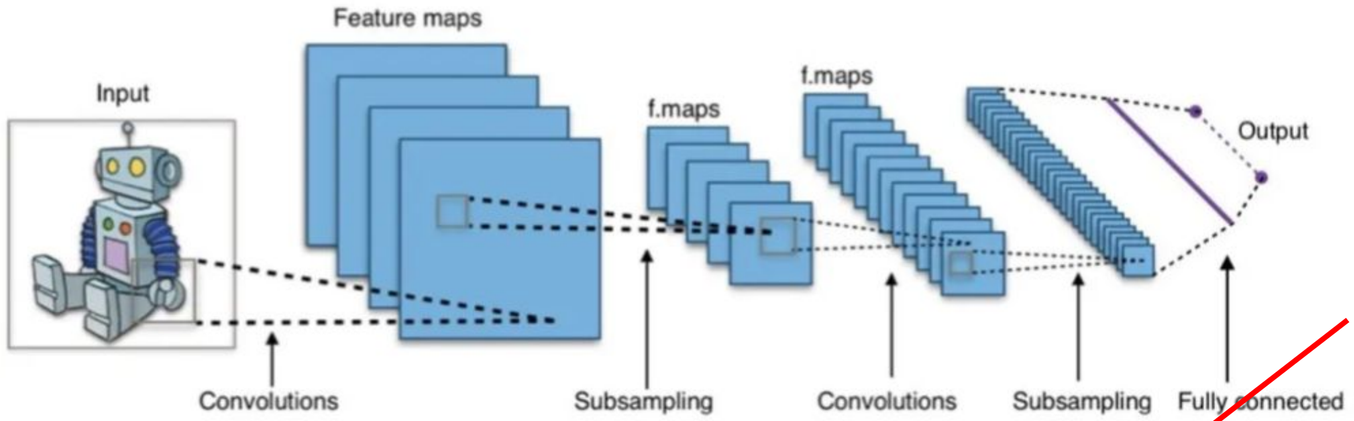
Lightricks

AGENDA

- Unet, Gans, StyleGAN and Gan Control
- Diffusion models, Text to Image models
- Controlling text 2 Image models

Lightricks

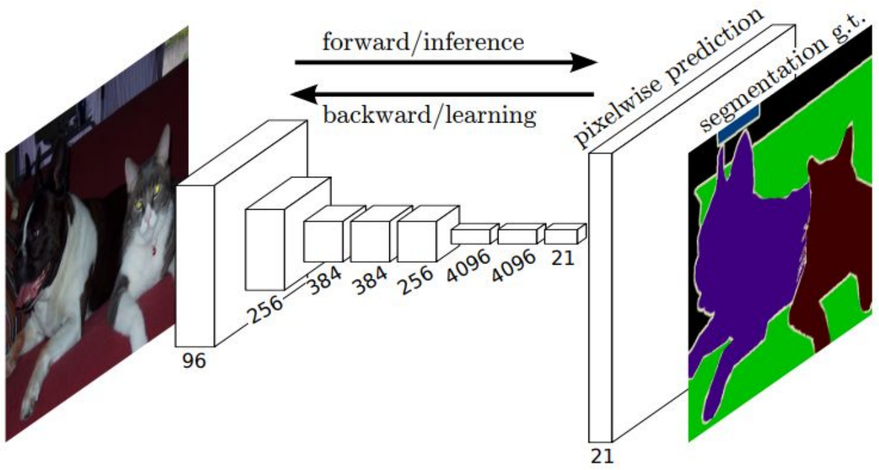
Back to 2015

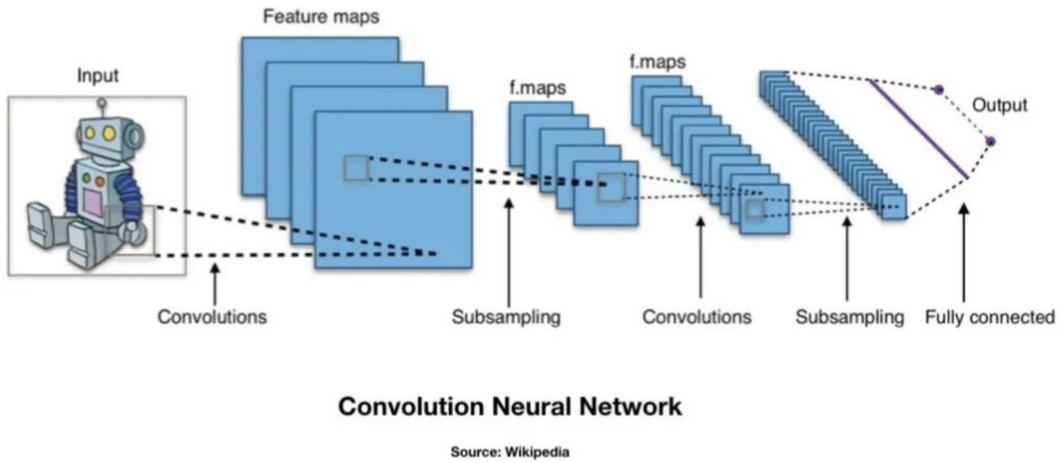


Convolution Neural Network

Source: Wikipedia

Image 2 Image?
Fully Convolutional Models



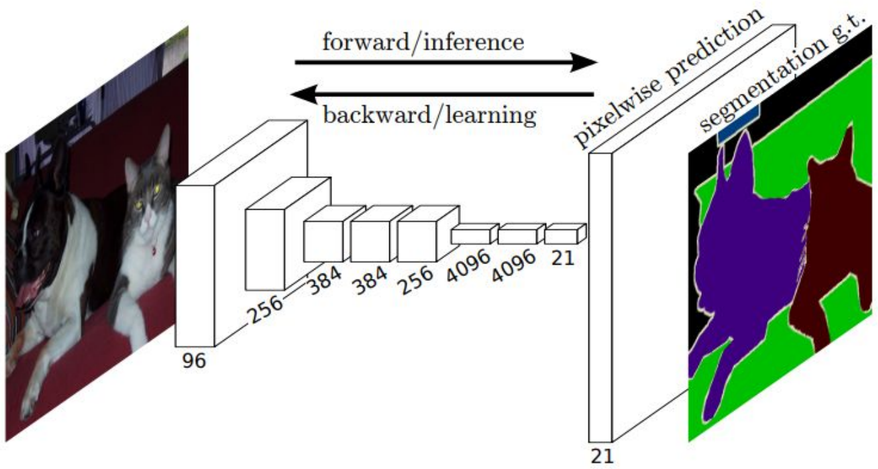


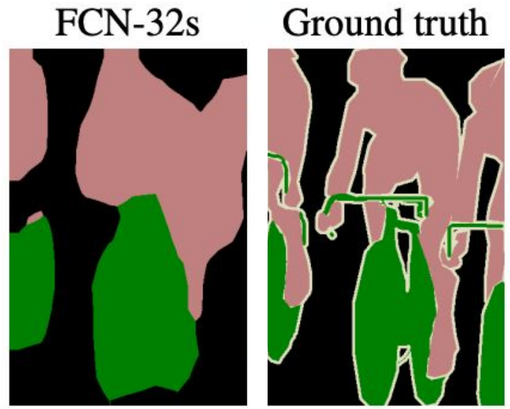
Convolution Neural Network

Source: Wikipedia

[Fully Convolutional Networks for Semantic Segmentation]

Back to 2015 – Fully Convolutional Models

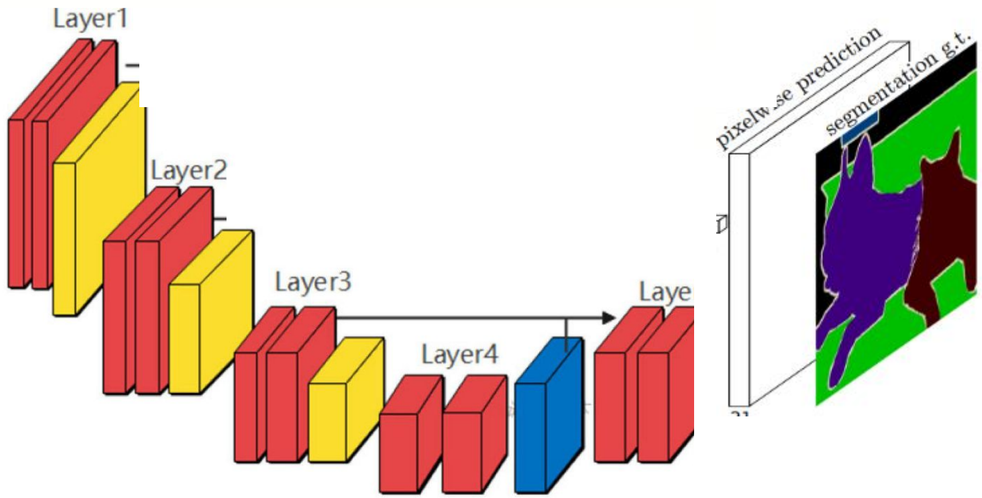


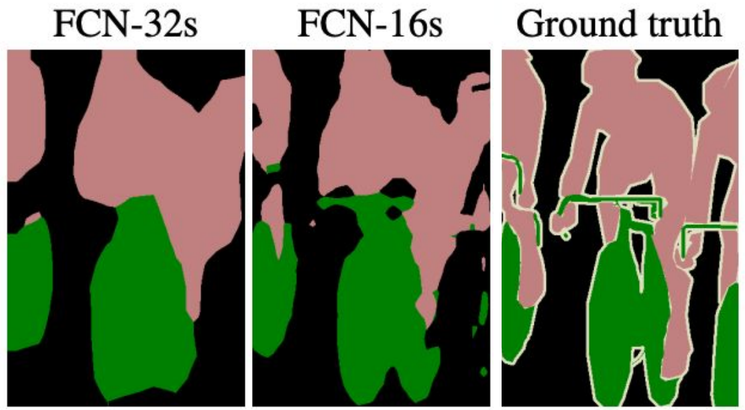


[Fully Convolutional Networks for Semantic Segmentation]

Lightricks

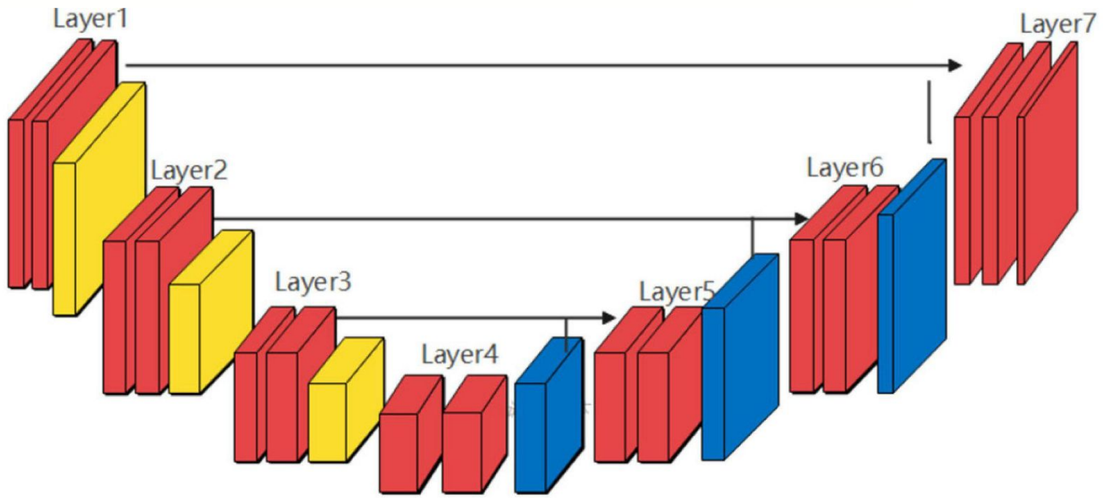
Back to 2015 – Fully Convolutional Models

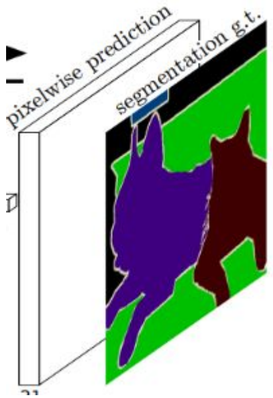


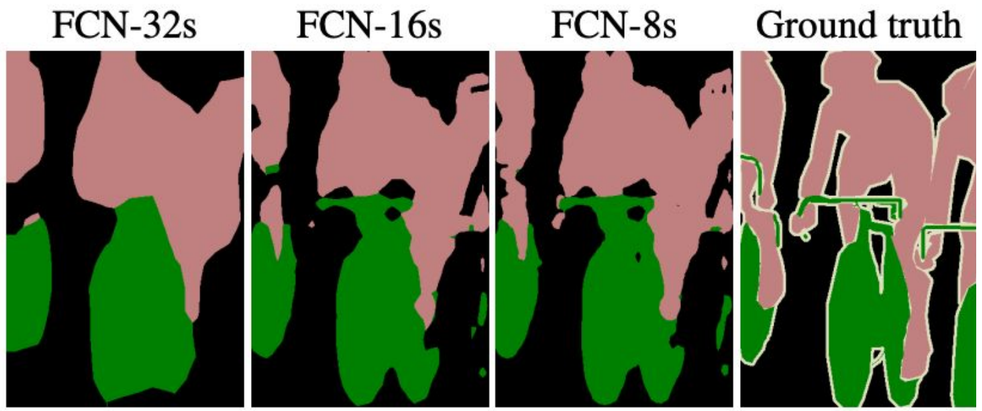


[Fully Convolutional Networks for Semantic Segmentation]

Back to 2015 – Fully Convolutional Models







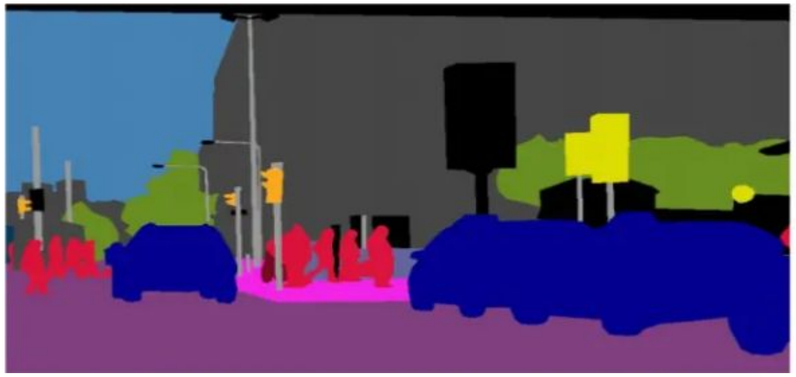
[Fully Convolutional Networks for Semantic Segmentation]

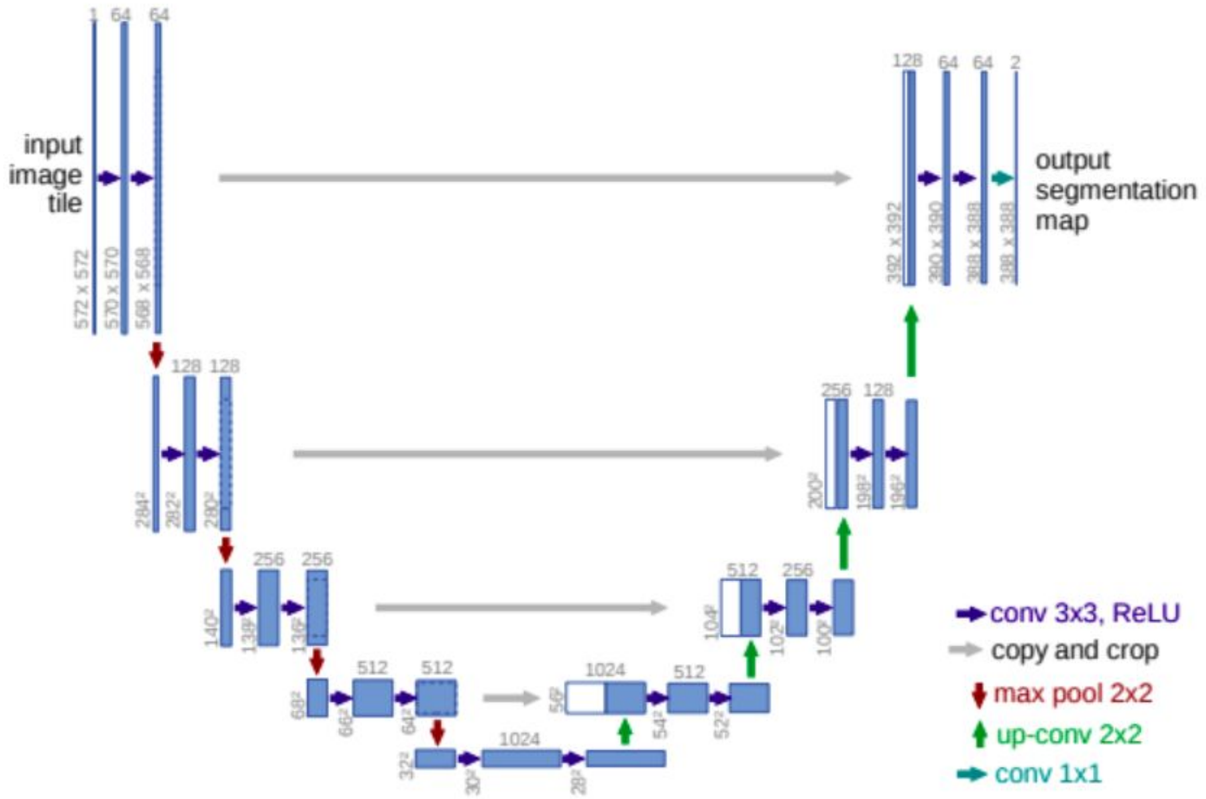
Lightricks

Unet Architecture

Input:

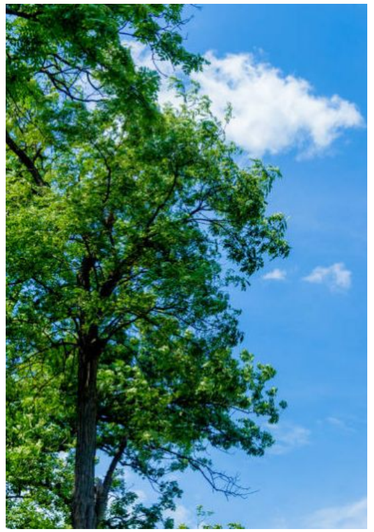
Output:

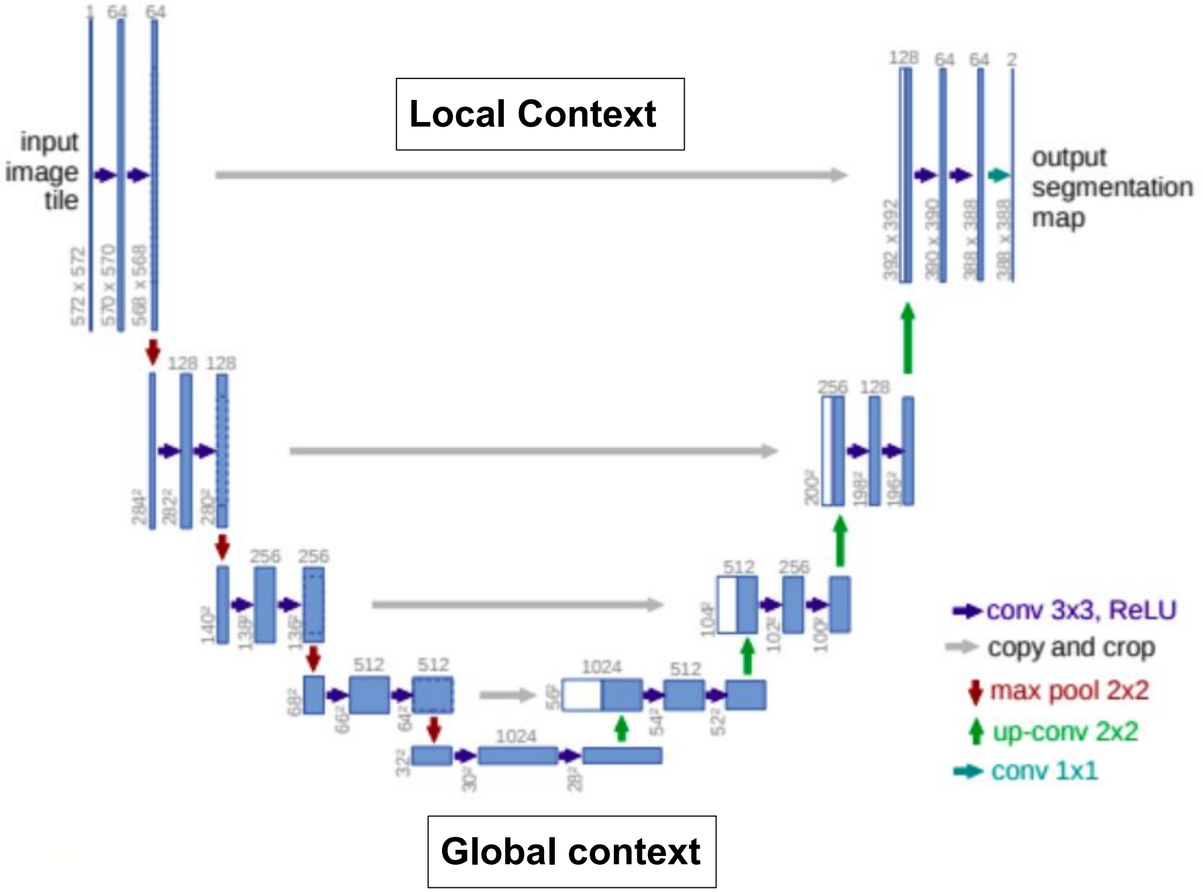




[Unet 2015]

Unet Architecture





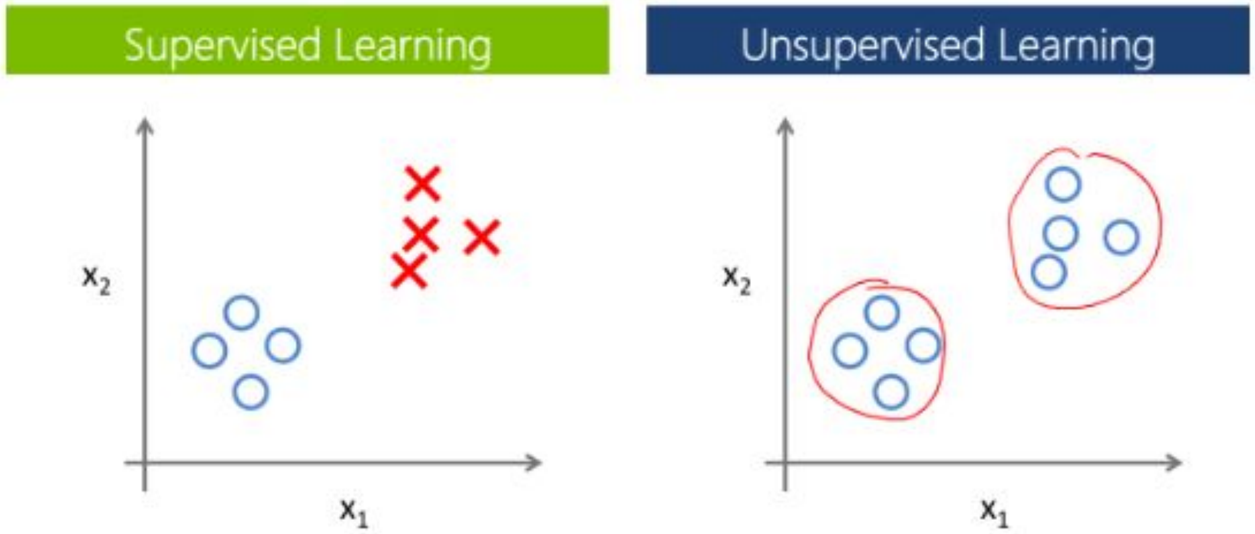
Unet 2015

Lightricks

Generative Adversarial Network

Supervised/ Unsupervised Learning

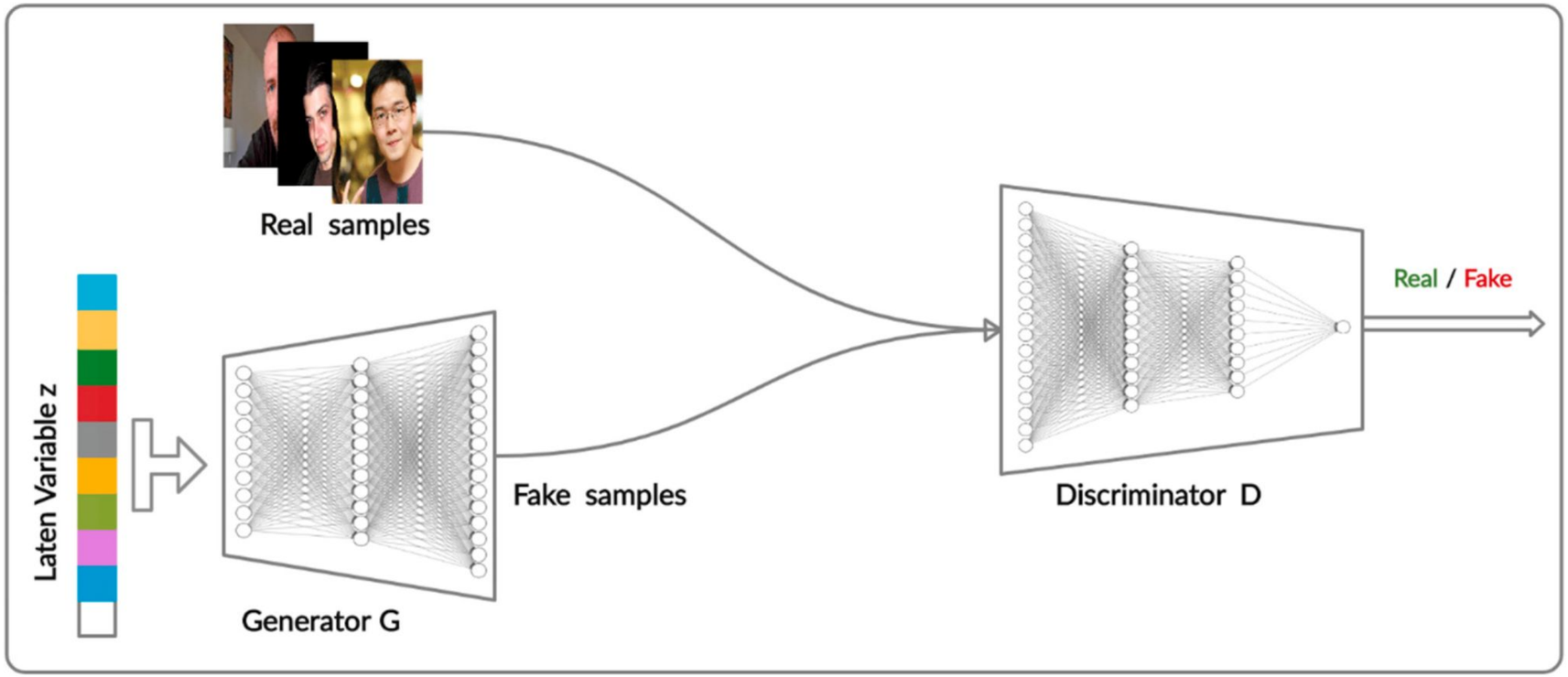
Discriminative Model



Generative Model

credit

Generative Adversarial Network (GAN)



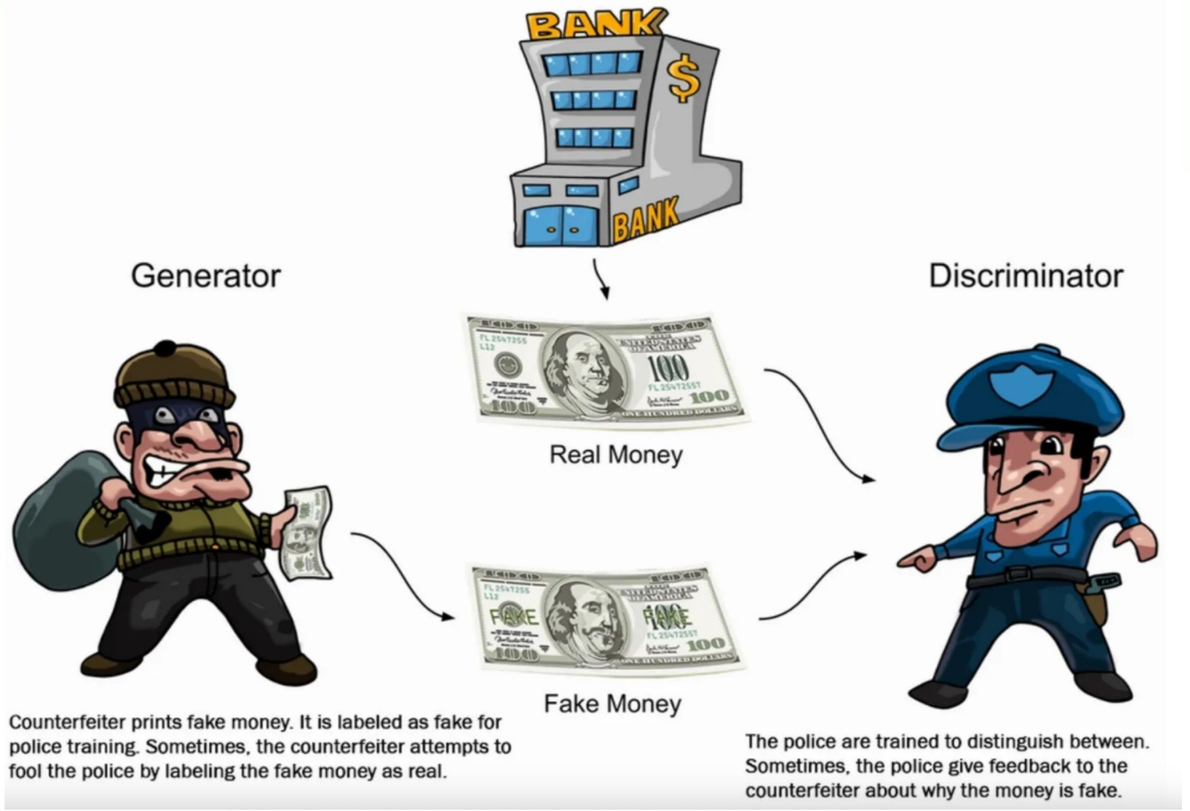
Generative Adversarial Networks, Ian Goodfellow, Yoshua Bengio

Input: random noise vector (dimensions - 1xd)

Output: Image (dimensions wxhx3)

Lightricks

Generative Adversarial Network (GAN)

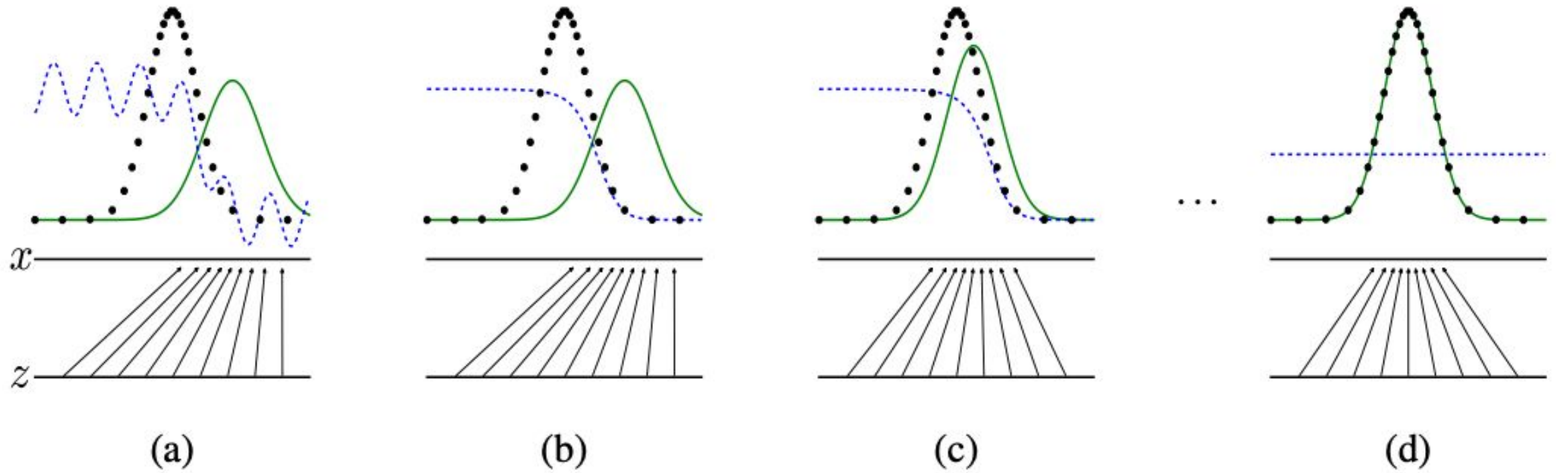


[Generative Adversarial Networks](), Ian Goodfellow, Yoshua Bengio

Lightricks

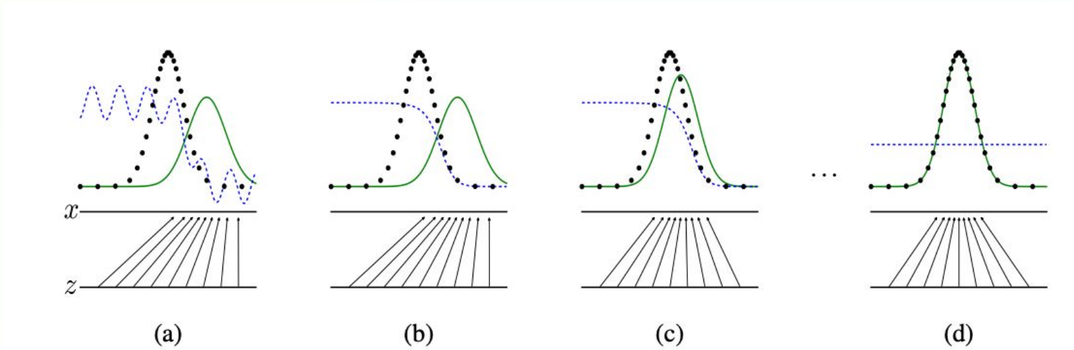
Generative Adversarial Network (GAN)

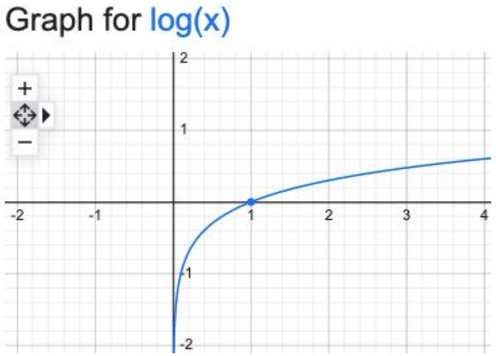
Black - Original distribution
Green - Generated distribution
Blue - Discriminator



Generative Adversarial Networks, Ian Goodfellow, Yoshua Bengio

The Objective:





Discriminator Maximizes:

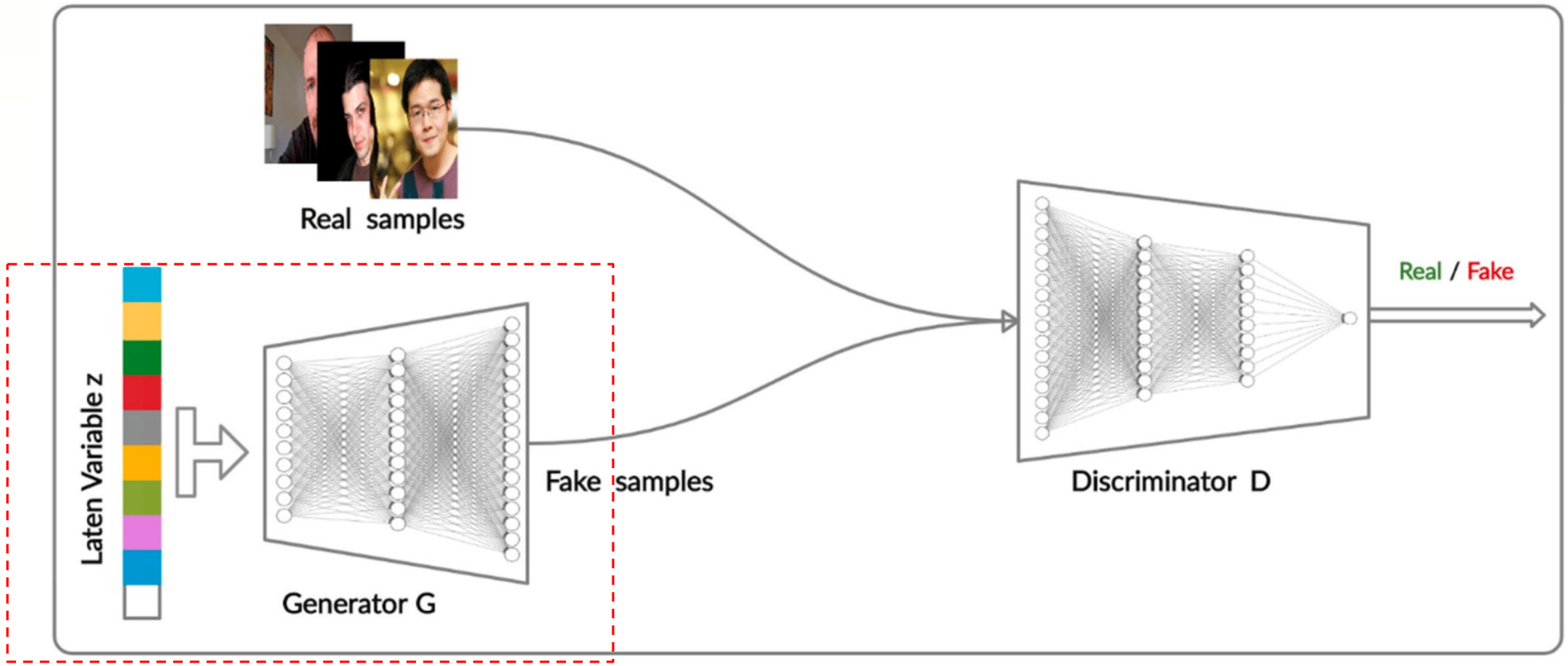
$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right]$$

Generator Minimizes:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)$$

Generative Adversarial Networks, Ian Goodfellow, Yoshua Bengio

Generative Adversarial Network (GAN)



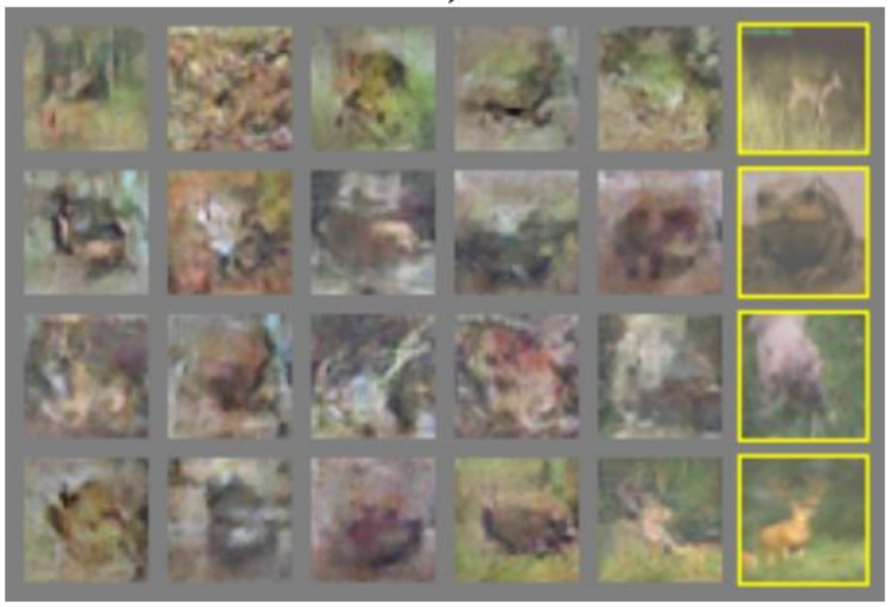
[Generative Adversarial Networks]

Lightricks

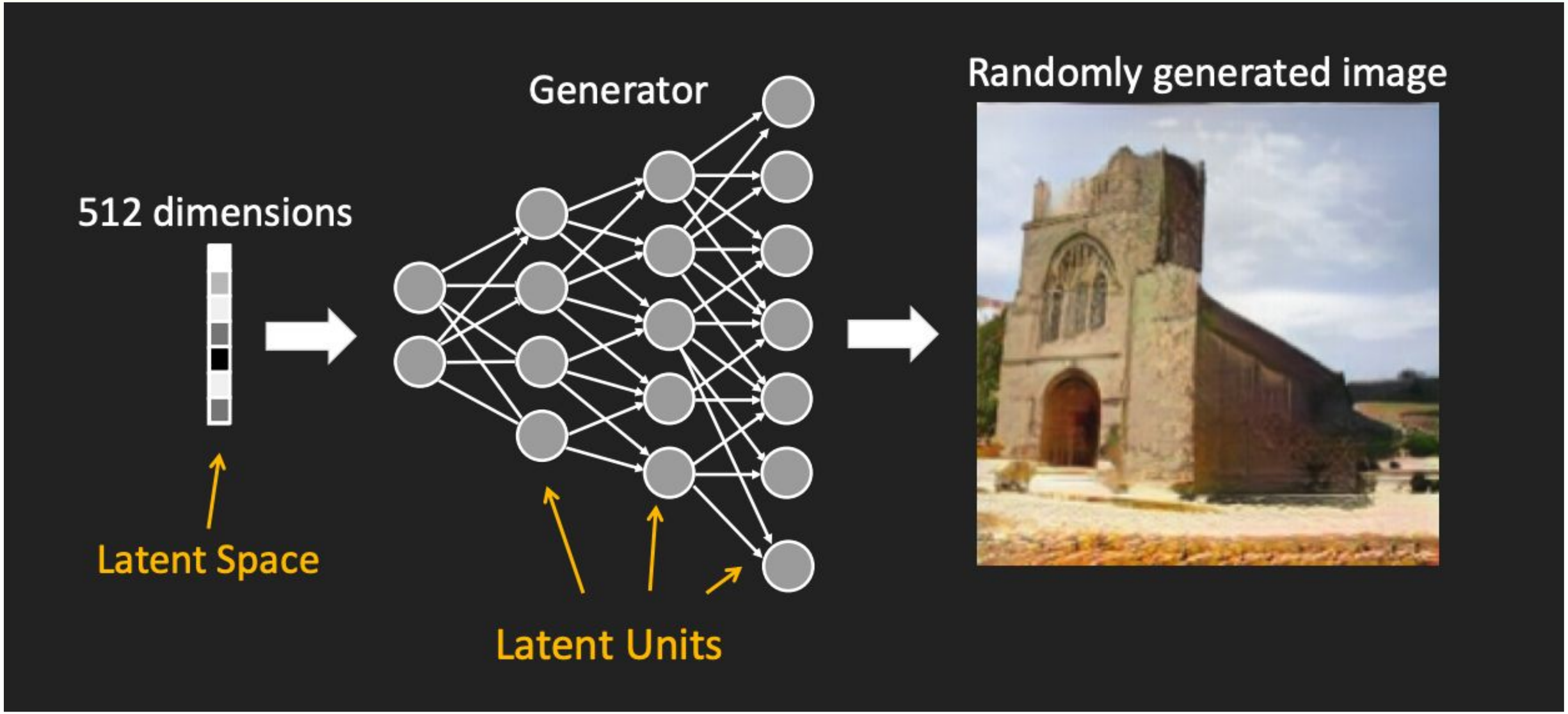
a)

b)

c)

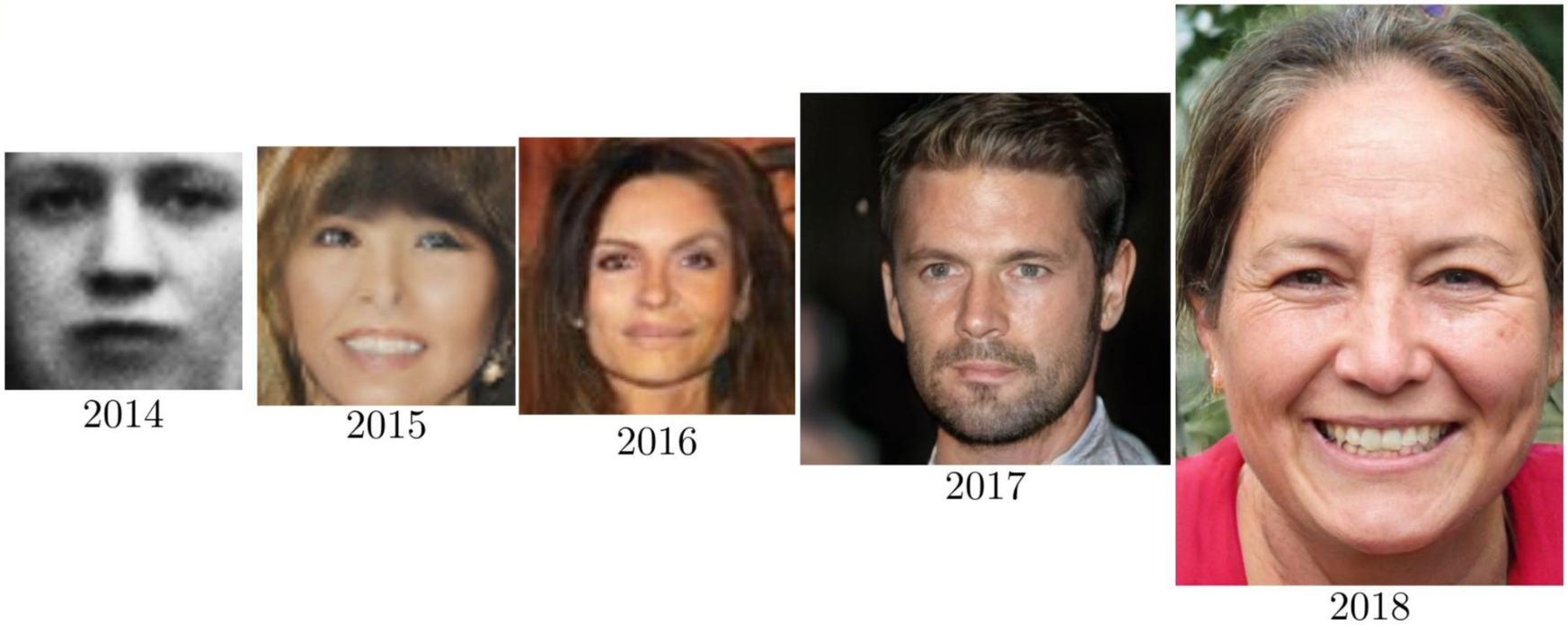


d)



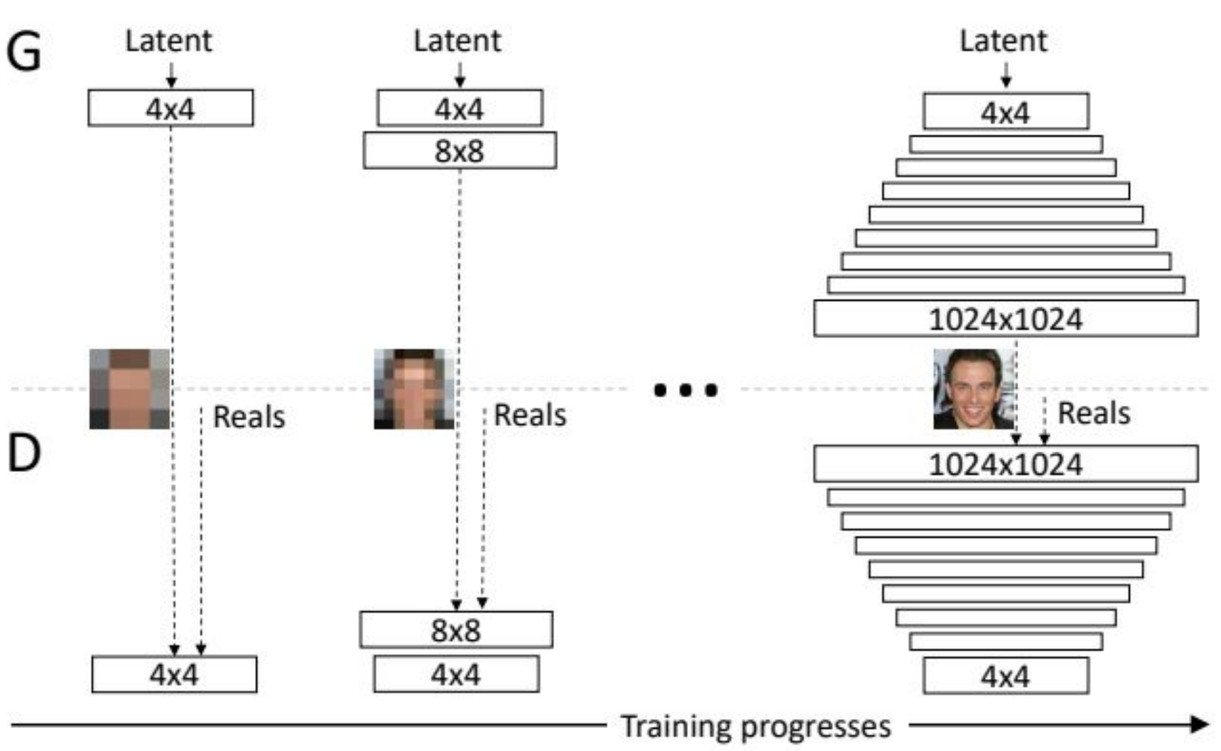
*Input and output

GANs Resolution Progress



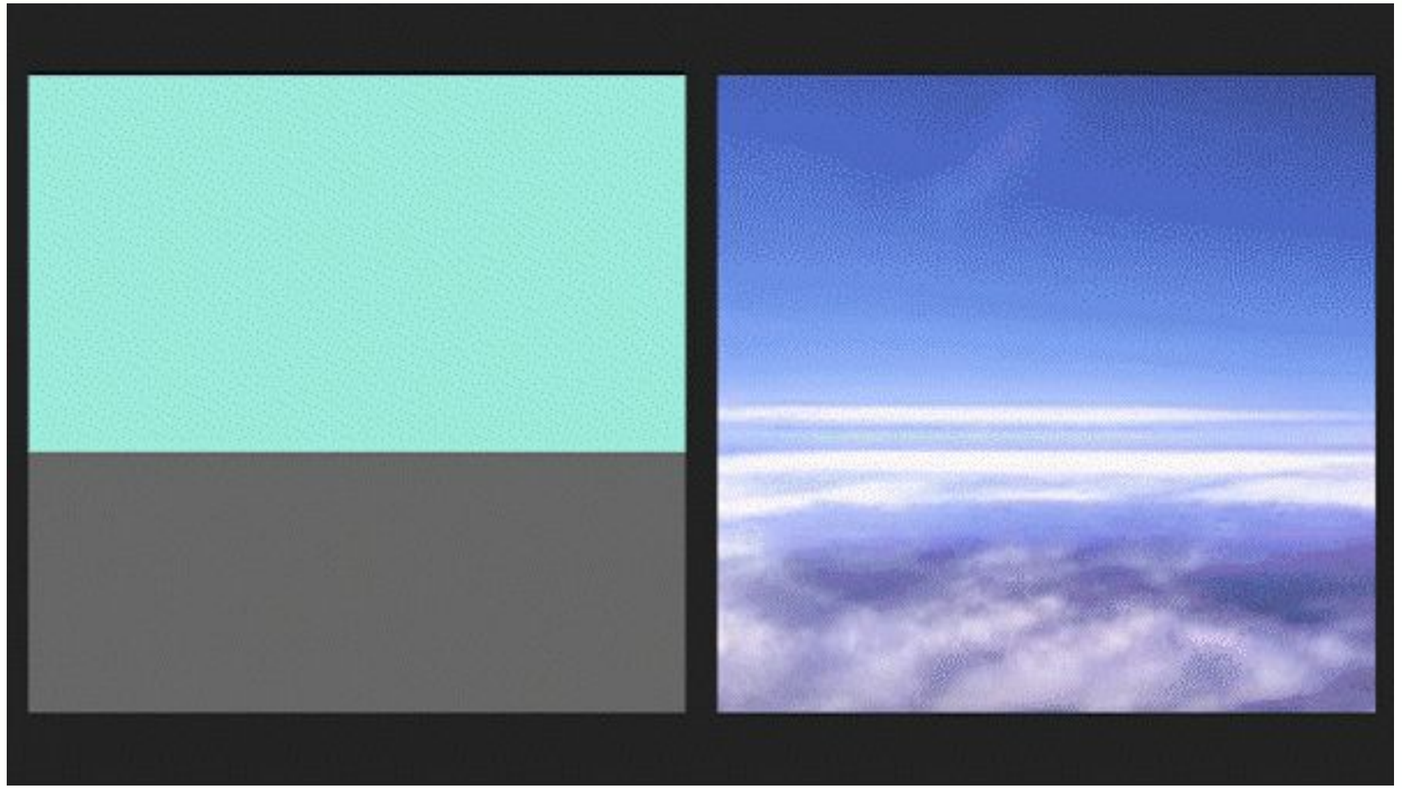
Ian Goodfellow

Progressive Growing



[PROGRESSIVE GROWING OF GANS FOR IMPROVED QUALITY, STABILITY, AND VARIATION, Nvidia]

Now Lets Try to Add Control



Pix2Pix

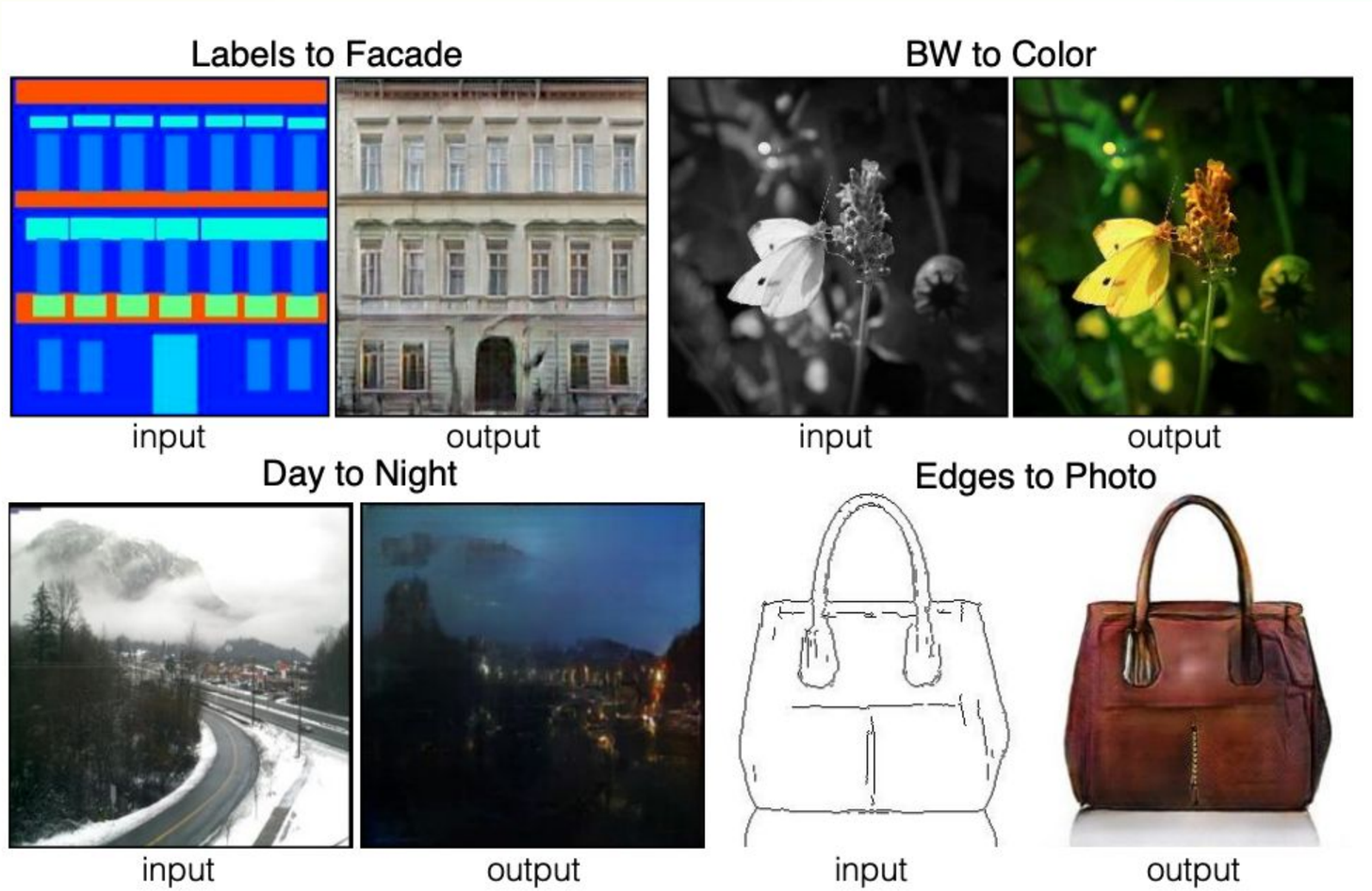
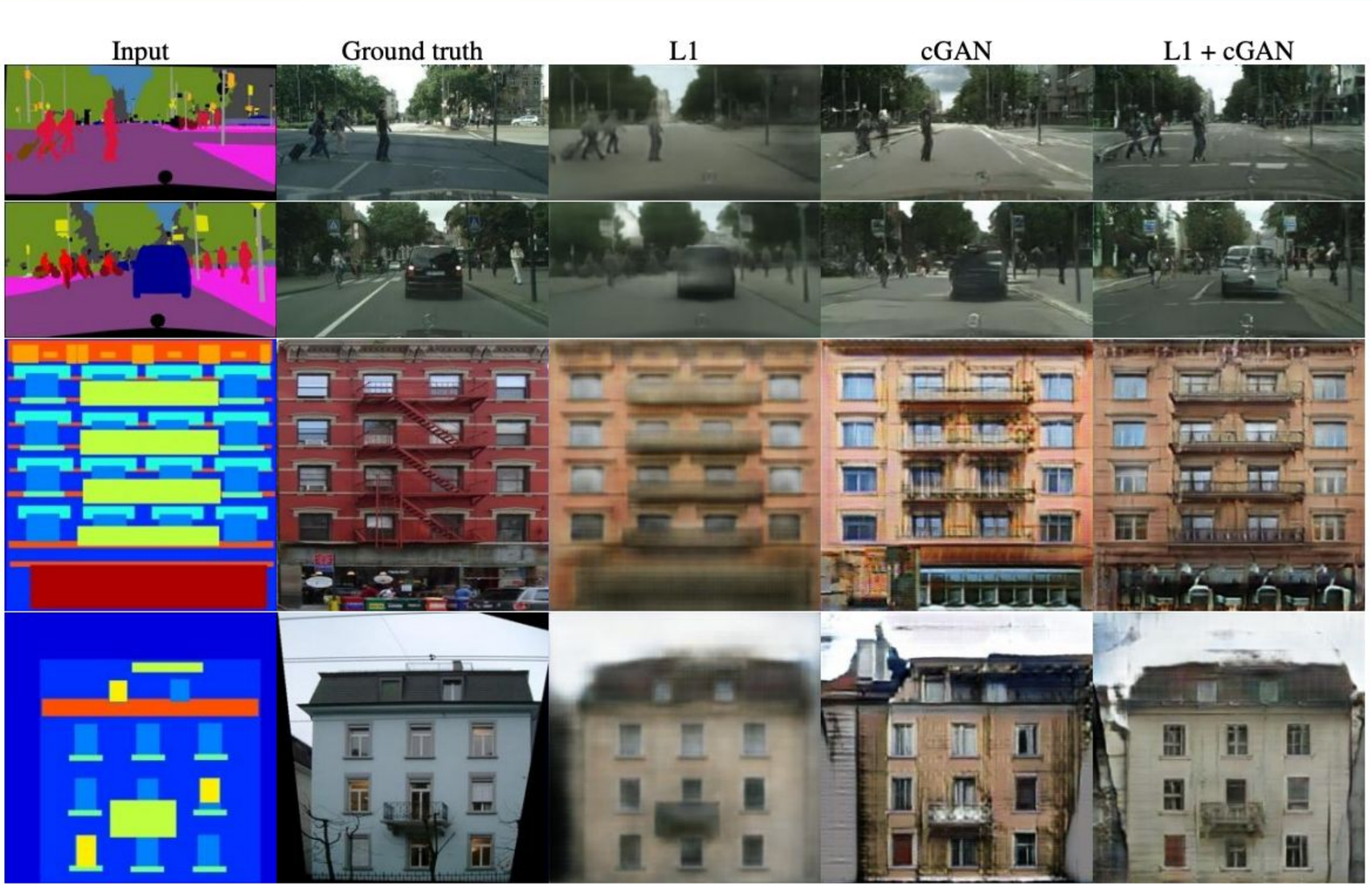


Image-to-Image Translation with Conditional Adversarial Networks
Phillip Isola, Alexei A. Efros, Berkeley

Why not Train Image 2 Image Unet?



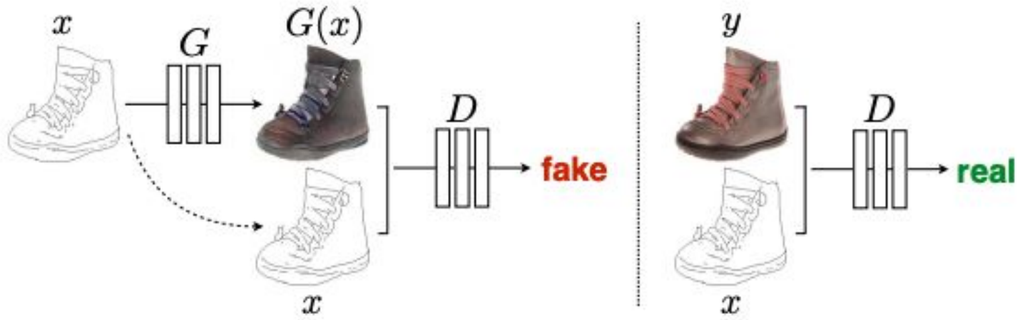
Pix2Pix

The objective: Generator minimizes
Discriminator maximizes

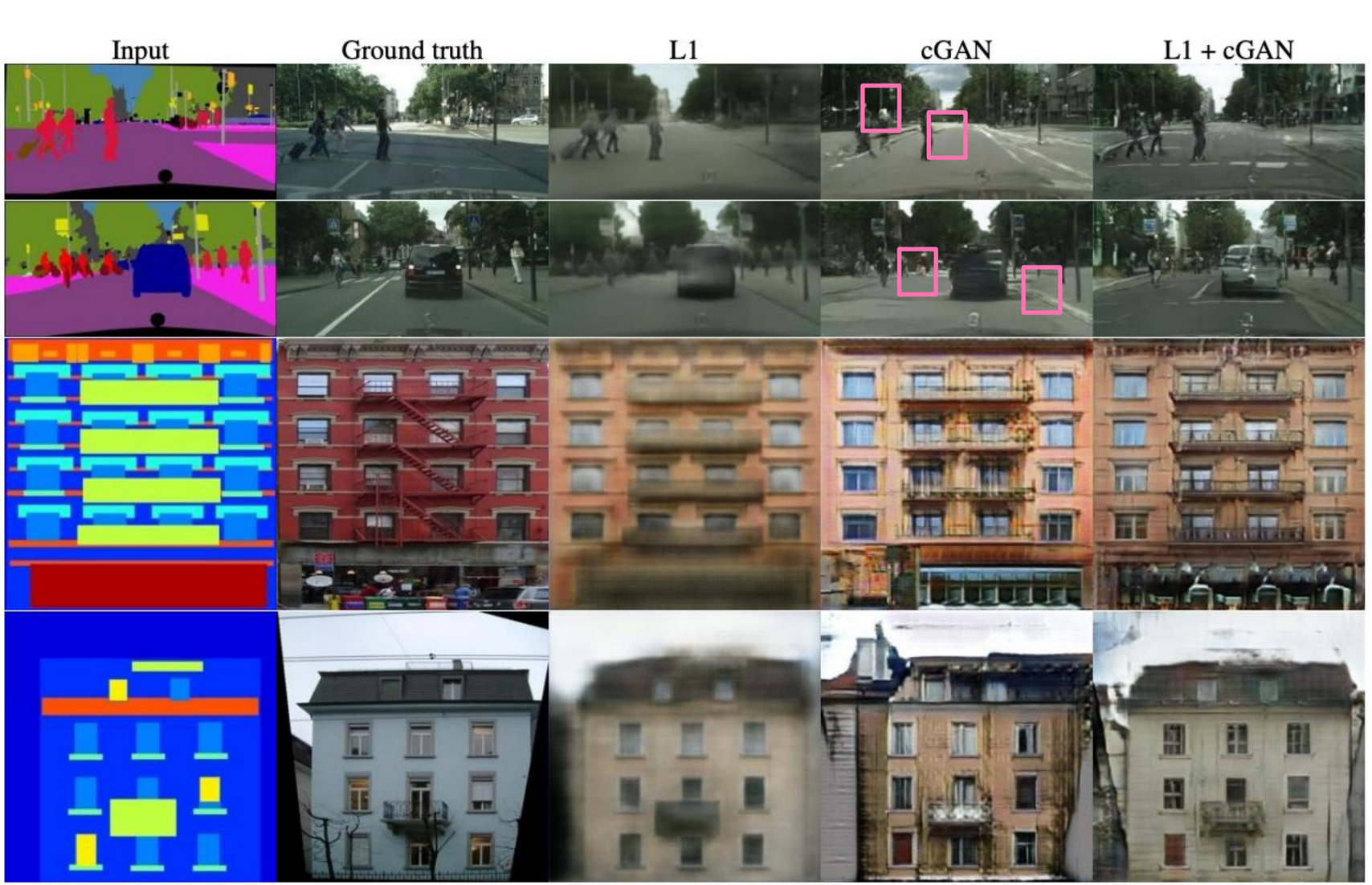
$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x,z))],$$

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G)$$

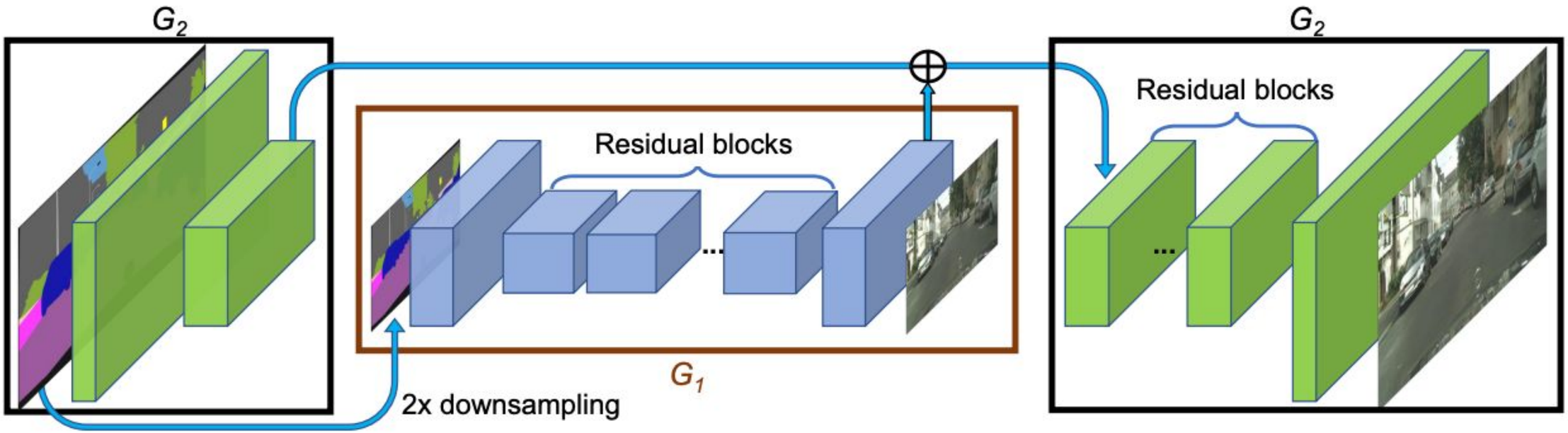
Architecture



PatchGan



Pix2PixHD - Can we improve the resolution?



High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs, Berkeley and Nvidia

GauGAN - Spade

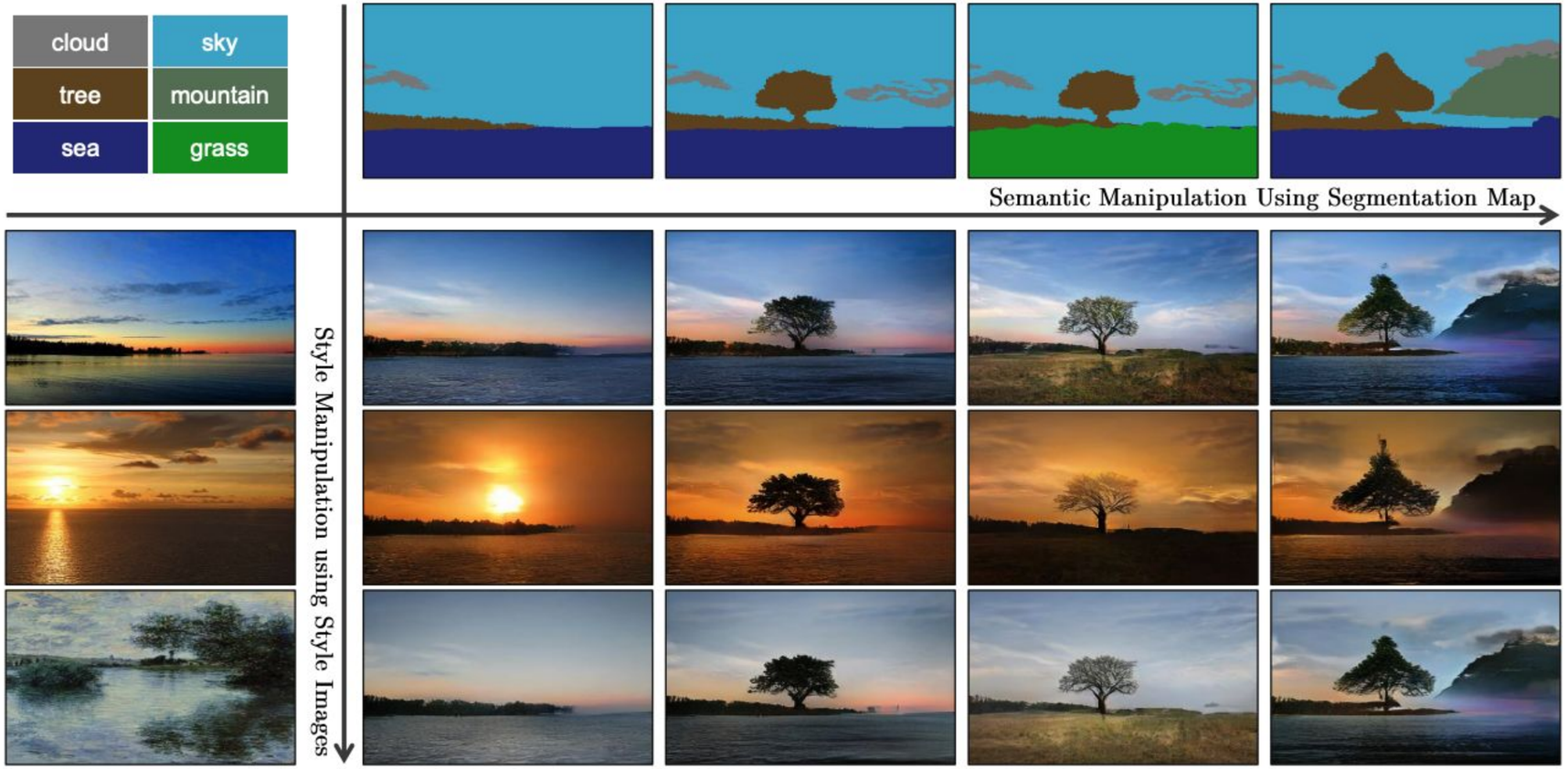
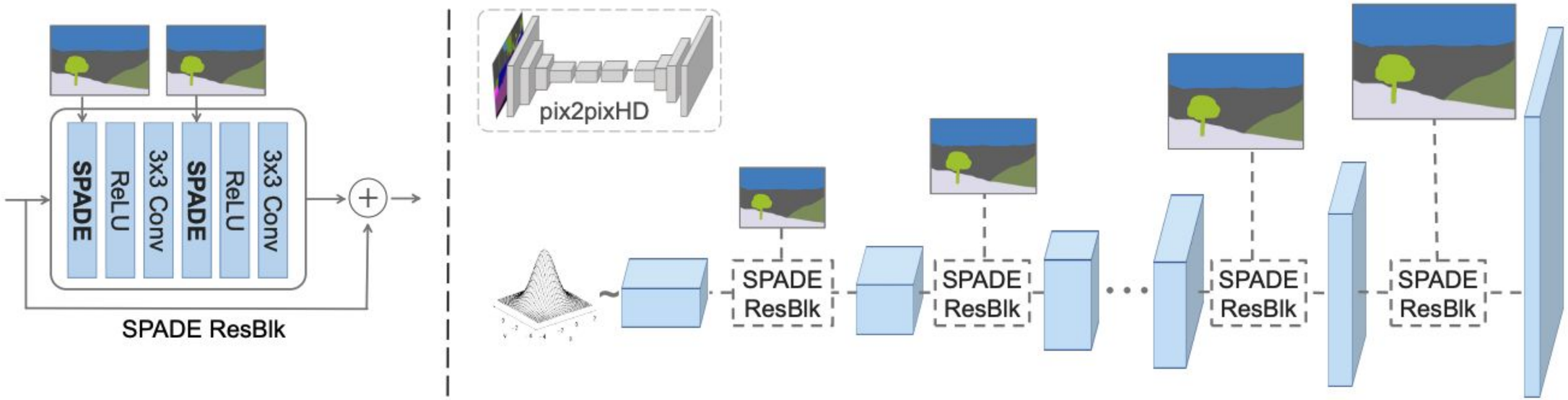


Figure 1: Our model allows user control over both semantic and style as synthesizing an image. The semantic (e.g., the existence of a tree) is controlled via a label map (the top row), while the style is controlled via the reference style image (the leftmost column). Please visit our website for interactive image synthesis demos.

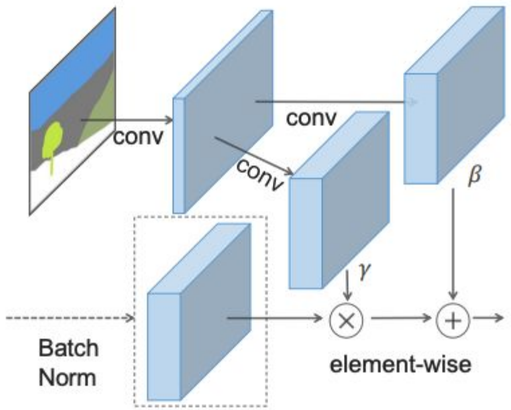
Semantic Image Synthesis with Spatially-Adaptive Normalization

GauGAN - Spade



Semantic Image Synthesis with Spatially-Adaptive Normalization

GauGAN - Spade



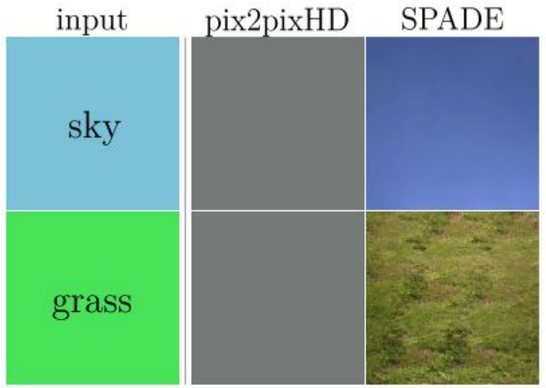
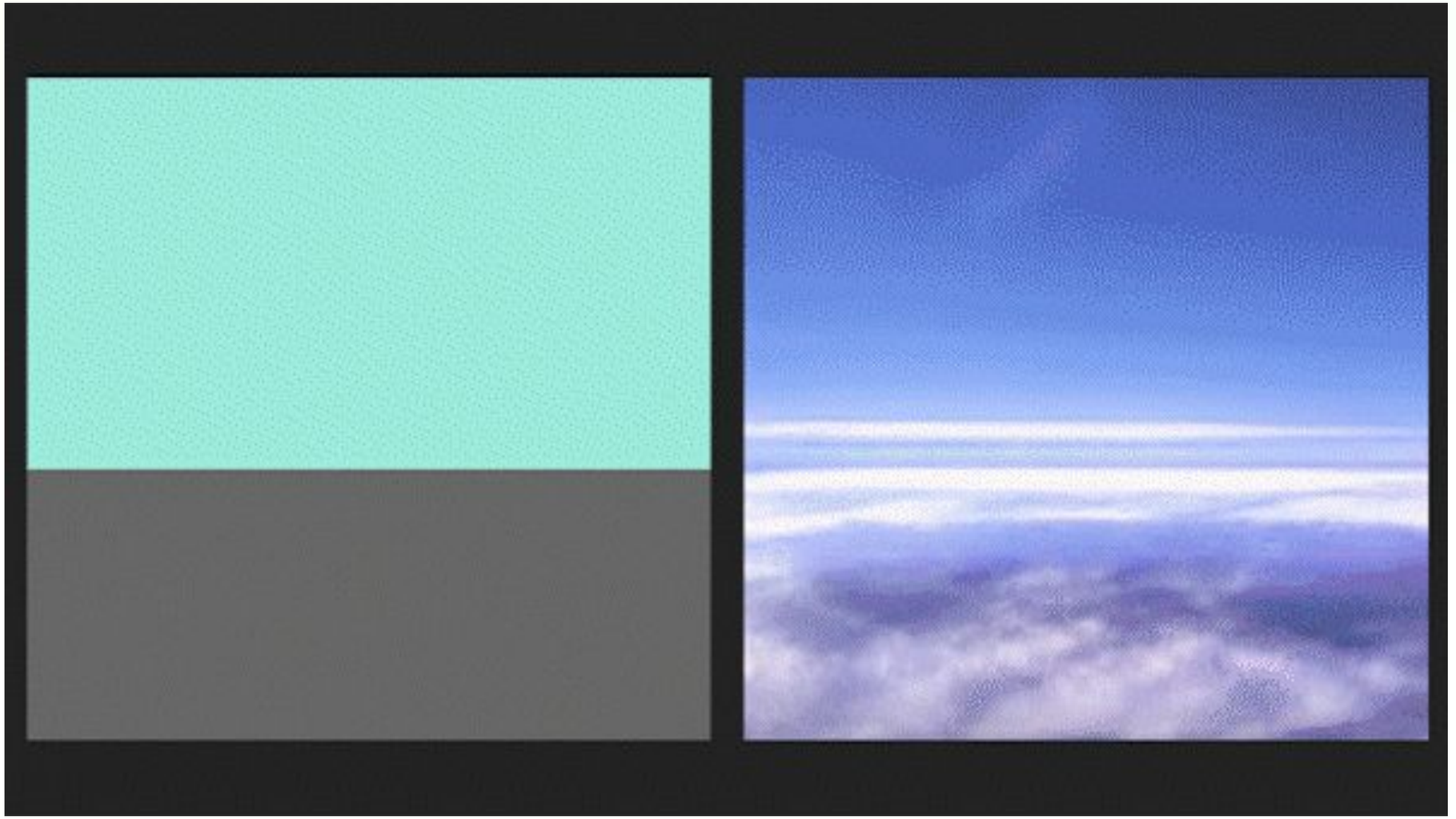
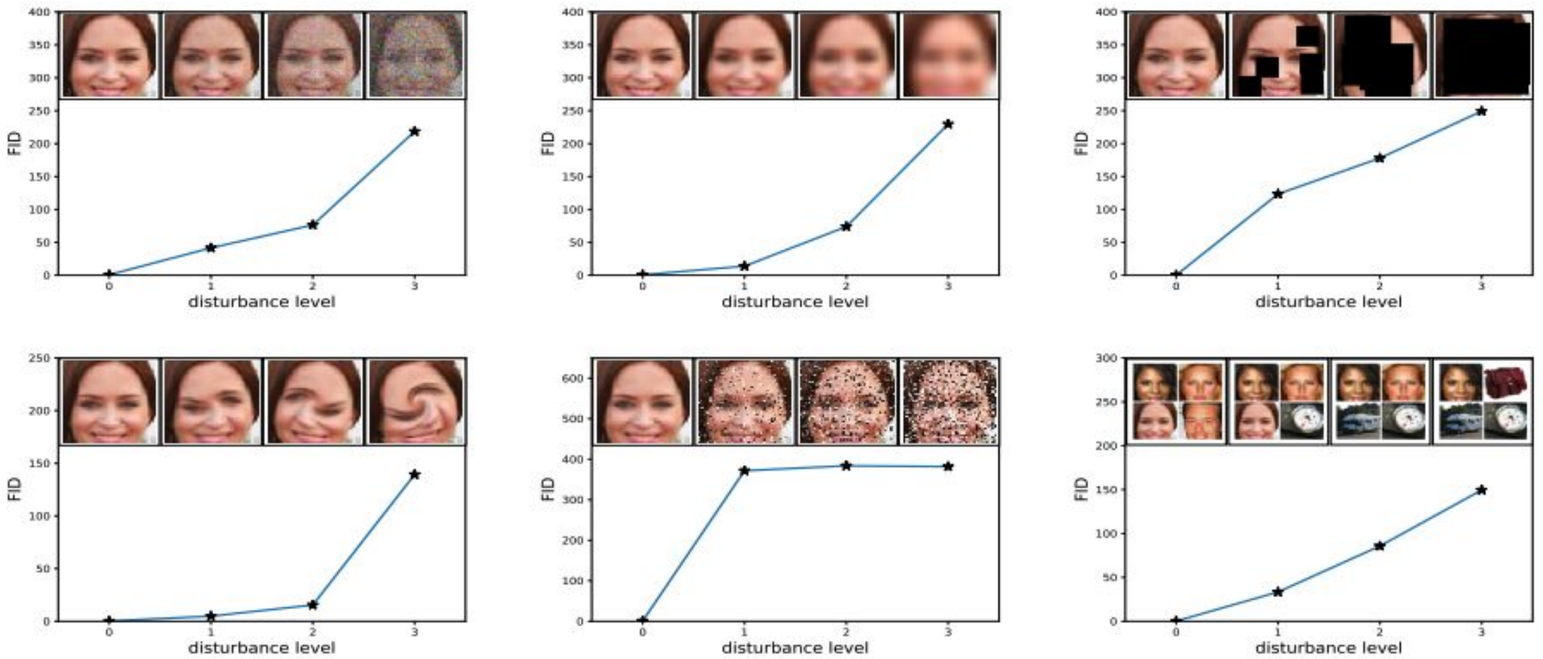


Figure 3: Comparing results given uniform segmentation maps: while the SPADE generator produces plausible textures, the pix2pixHD generator [48] produces two identical outputs due to the loss of the semantic information after the normalization layer.

Semantic Image Synthesis with Spatially-Adaptive Normalization

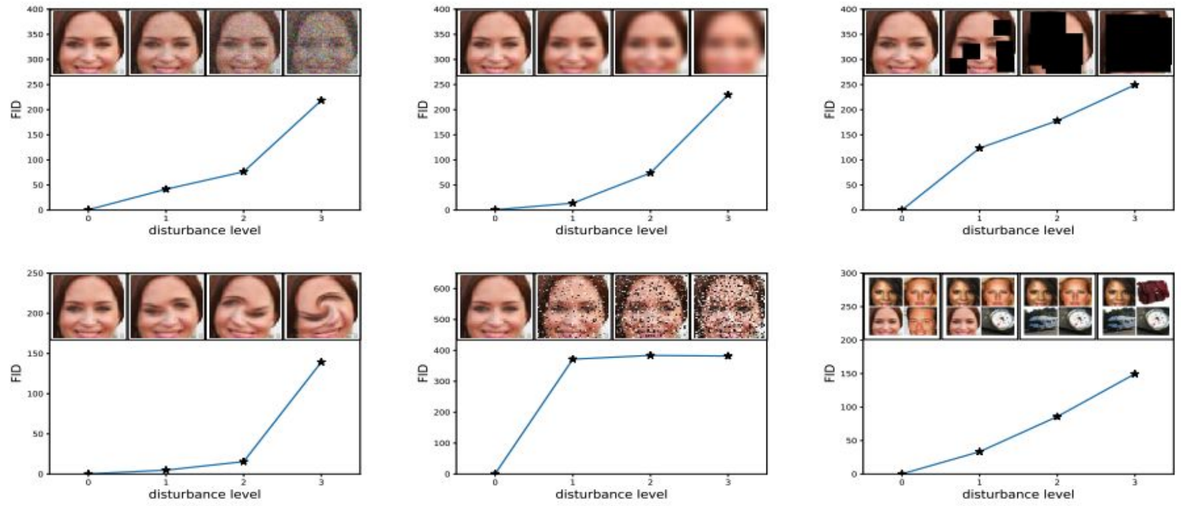


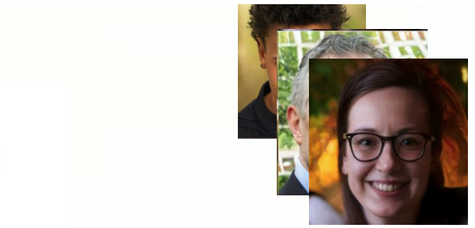
Generative Models Evaluation: FID - Fréchet Inception Distance



[GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium]

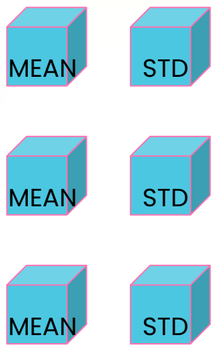
Generative Models Evaluation: FID - Fréchet Inception Distance

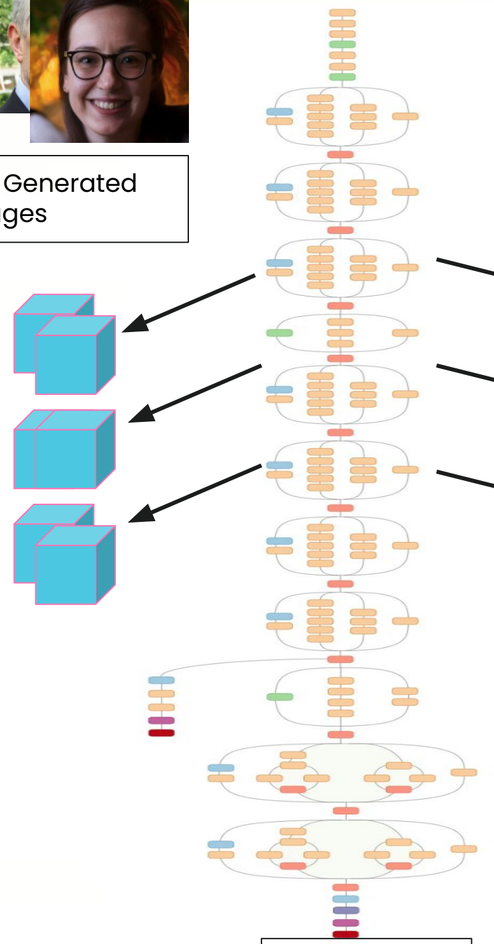


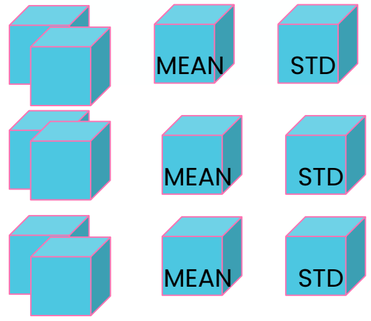


30k Generated Images

30k Real Images







Inception Model

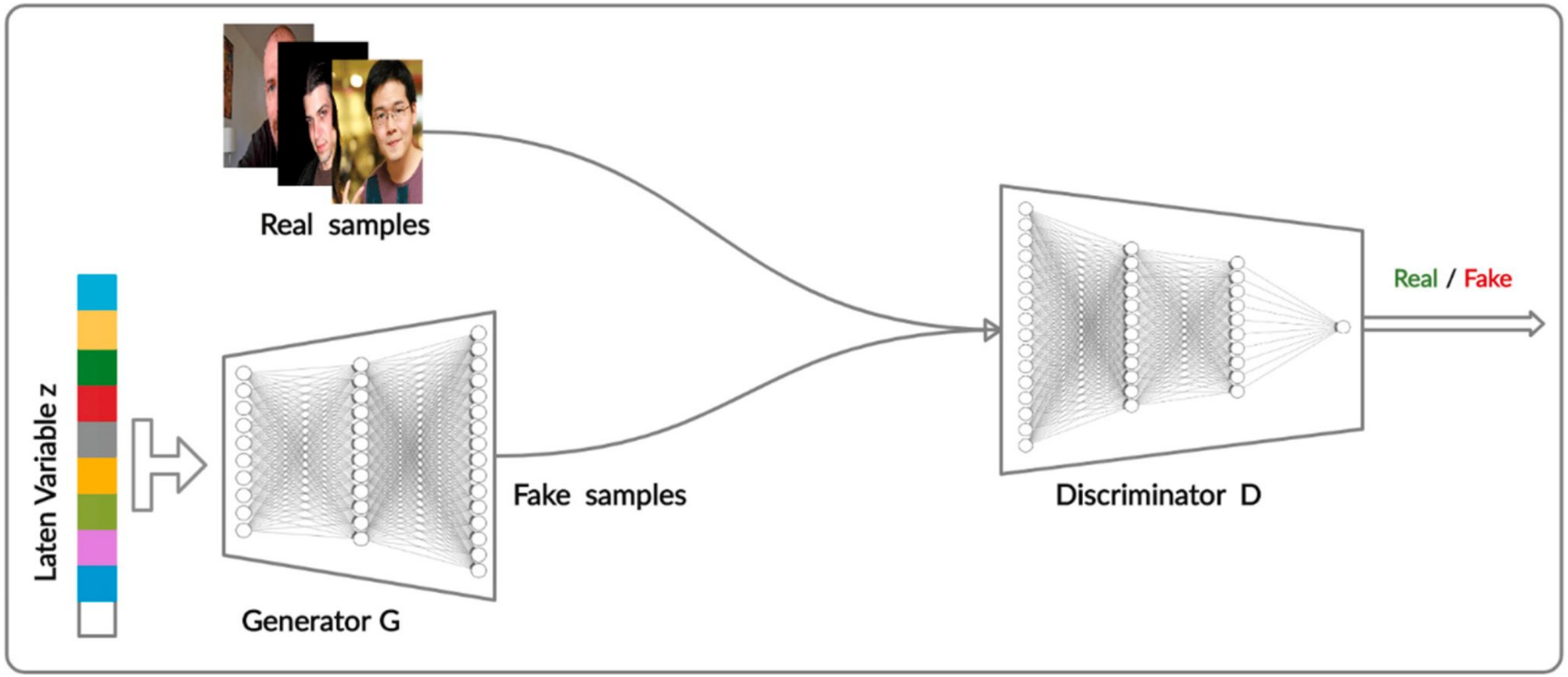
[GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium]

StyleGAN

Nvidia, 2018

[Analyzing and Improving the Image Quality of StyleGAN]

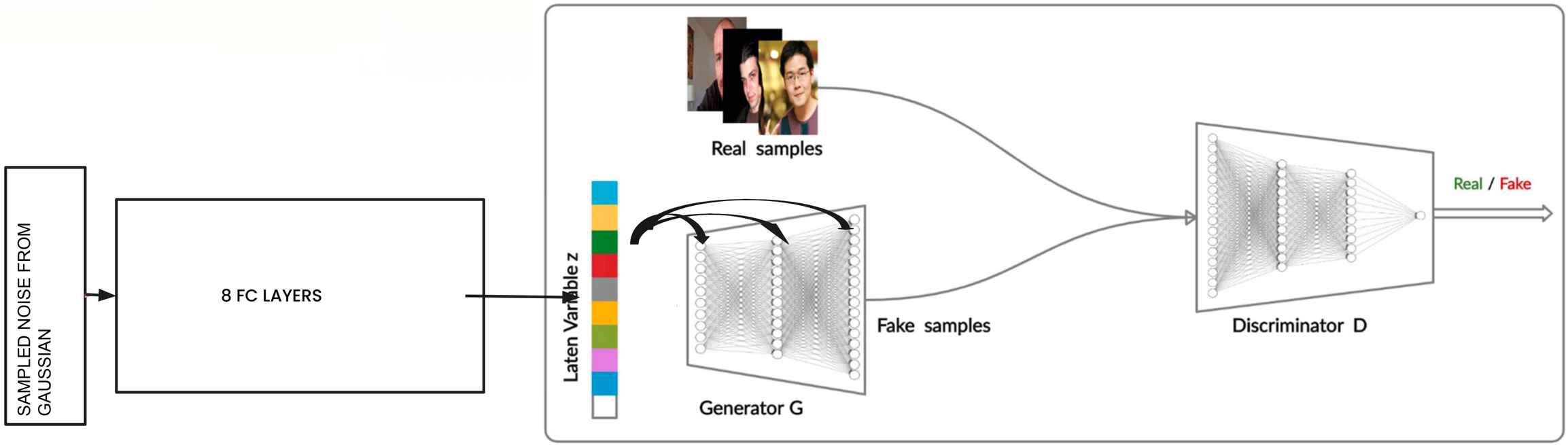
GAN



Generative Adversarial Networks, Ian Goodfellow, Yoshua Bengio

StyleGAN

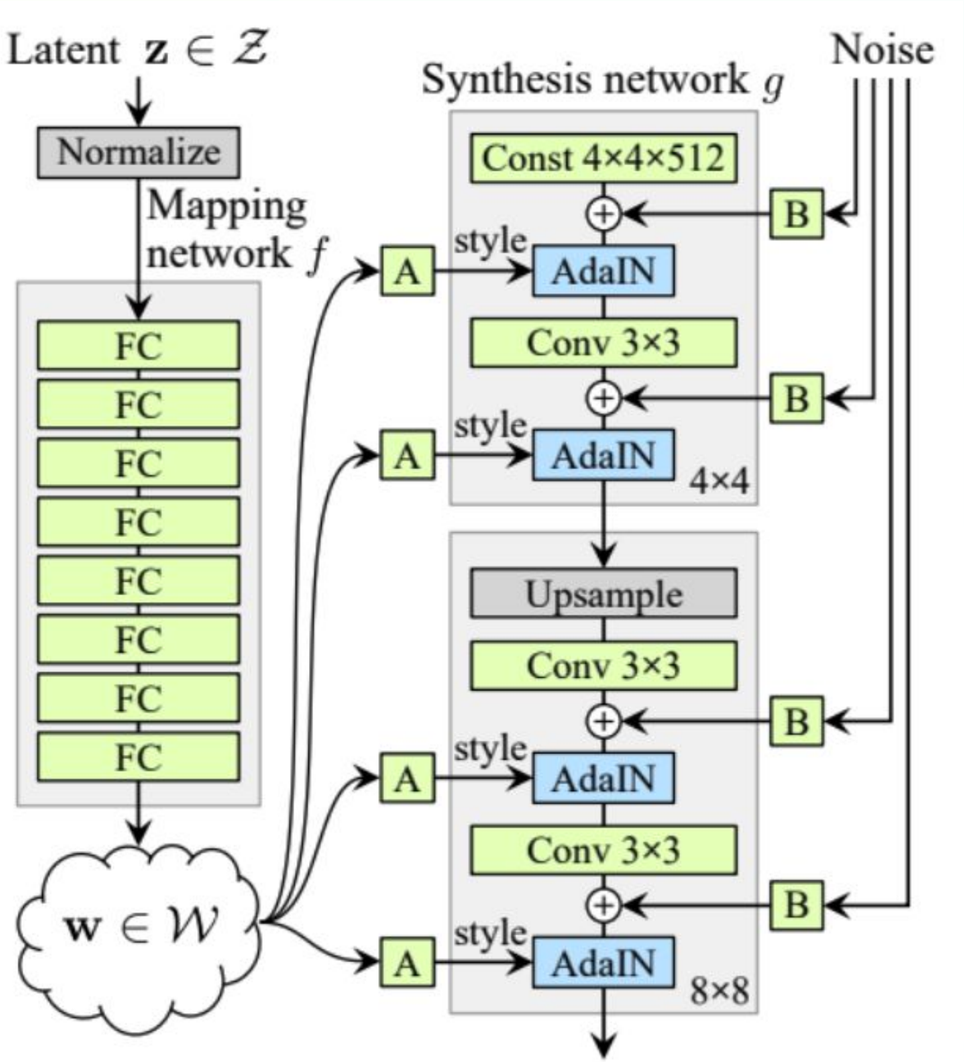
Nvidia, 2018



[Analyzing and Improving the Image Quality of StyleGAN]

StyleGAN

Nvidia, 2018



[Analyzing and Improving the Image Quality of StyleGAN]

AdaIN Block - Style Transfer

content image

Ancient city of Persepolis

+

style image

The Starry Night (Van Gogh)

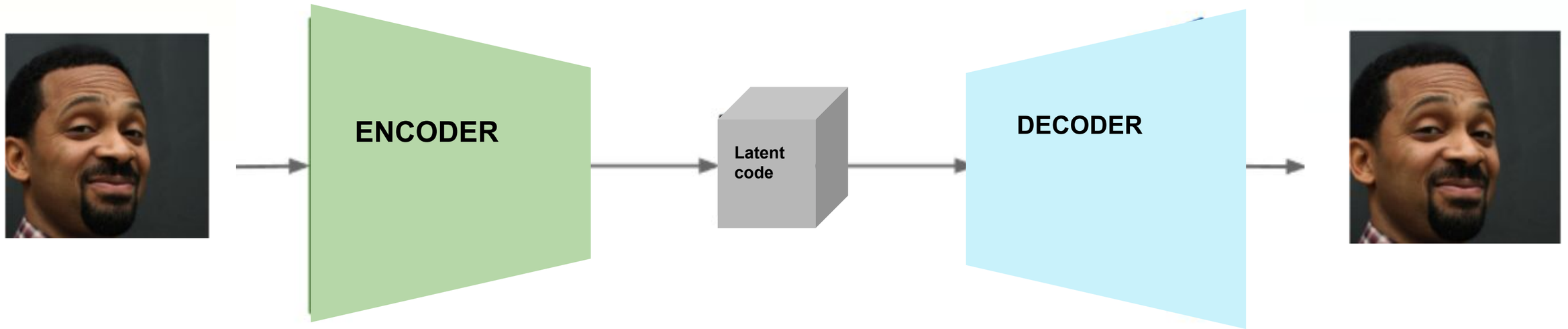
=

generated image

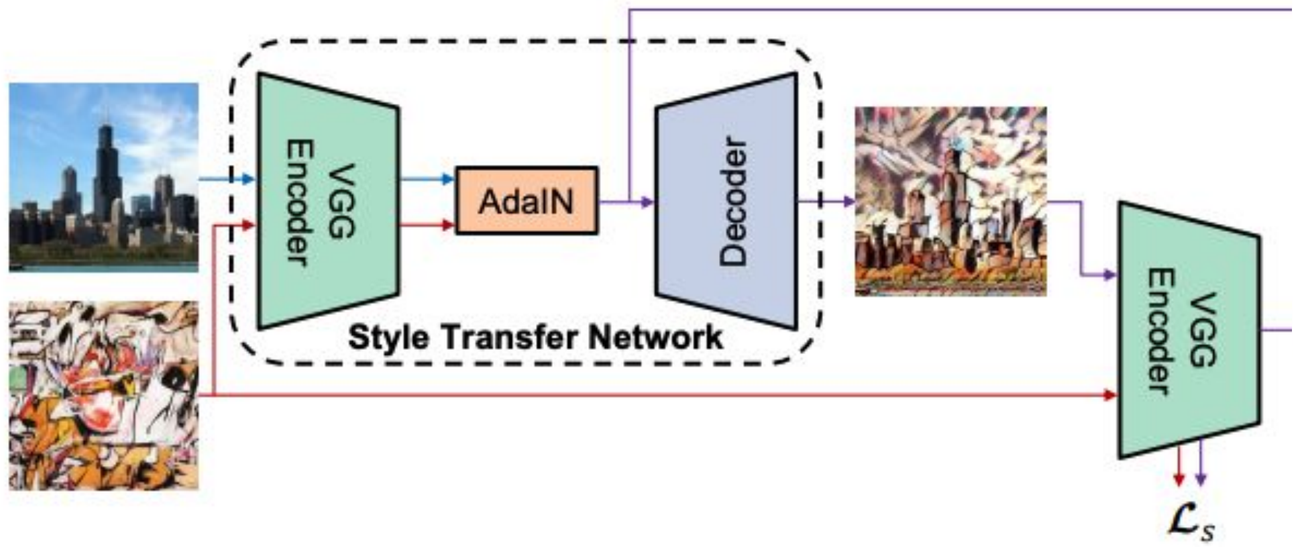
Persepolis
in Van Gogh style

Lightricks

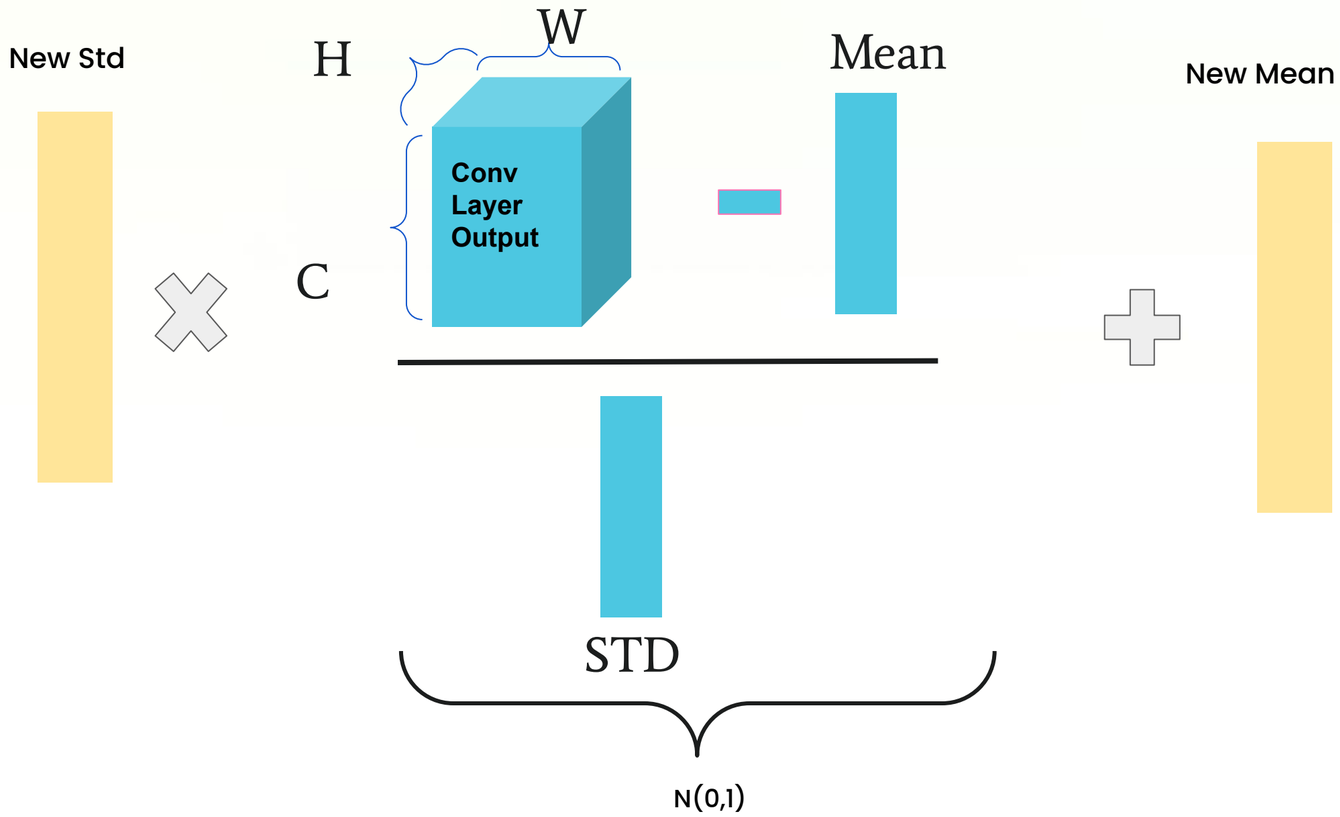
AutoEncoder

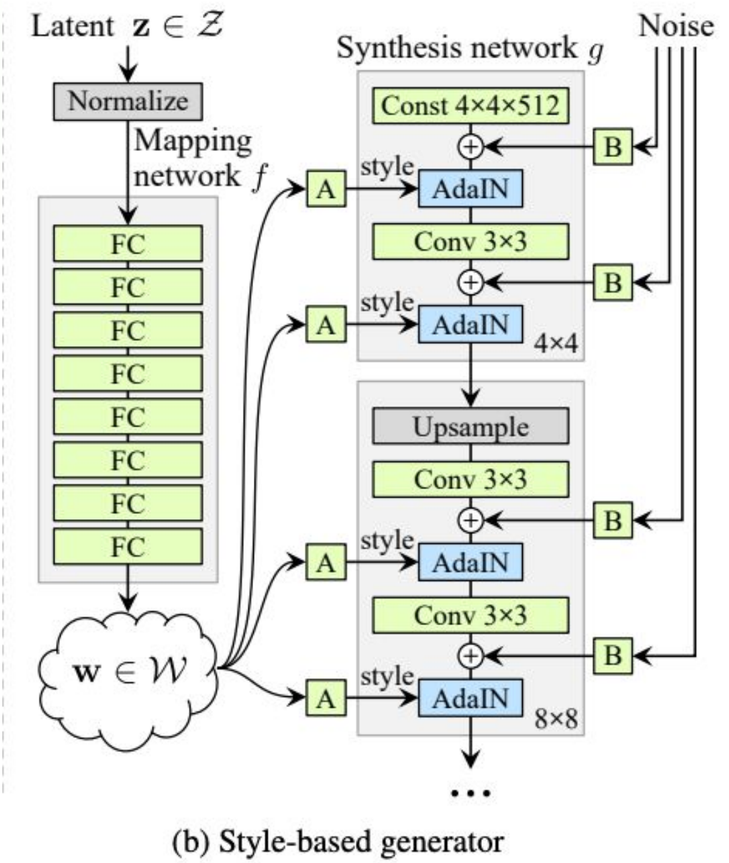


Introducing AdaIN Block



X. Huang, S. Belongie, Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. ICCV 2017

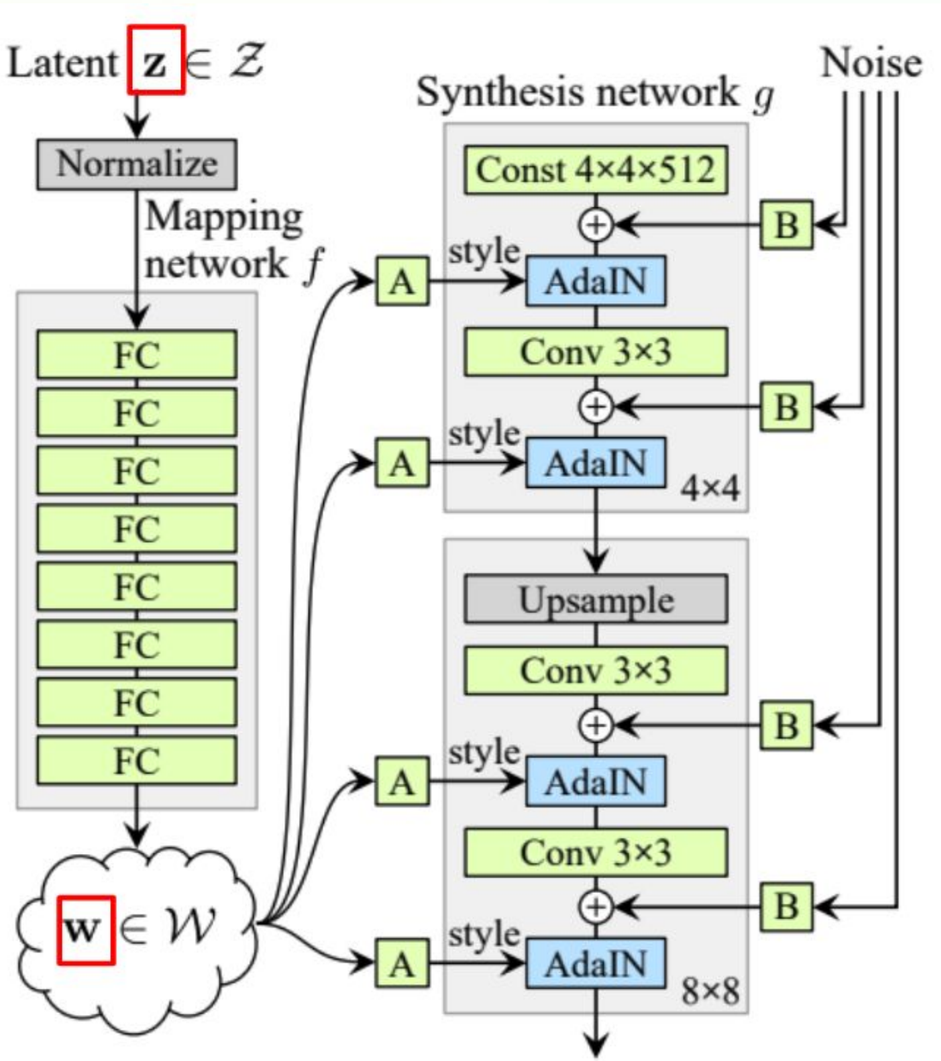


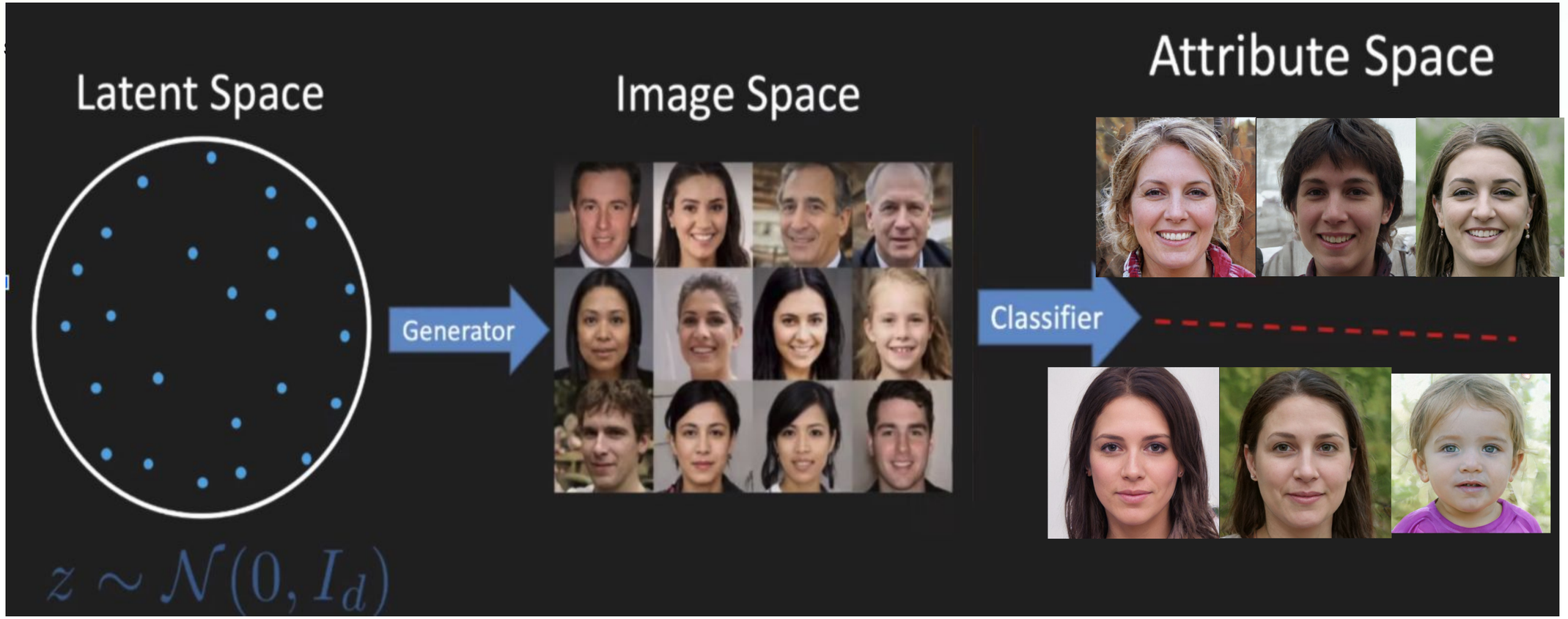


(b) Style-based generator

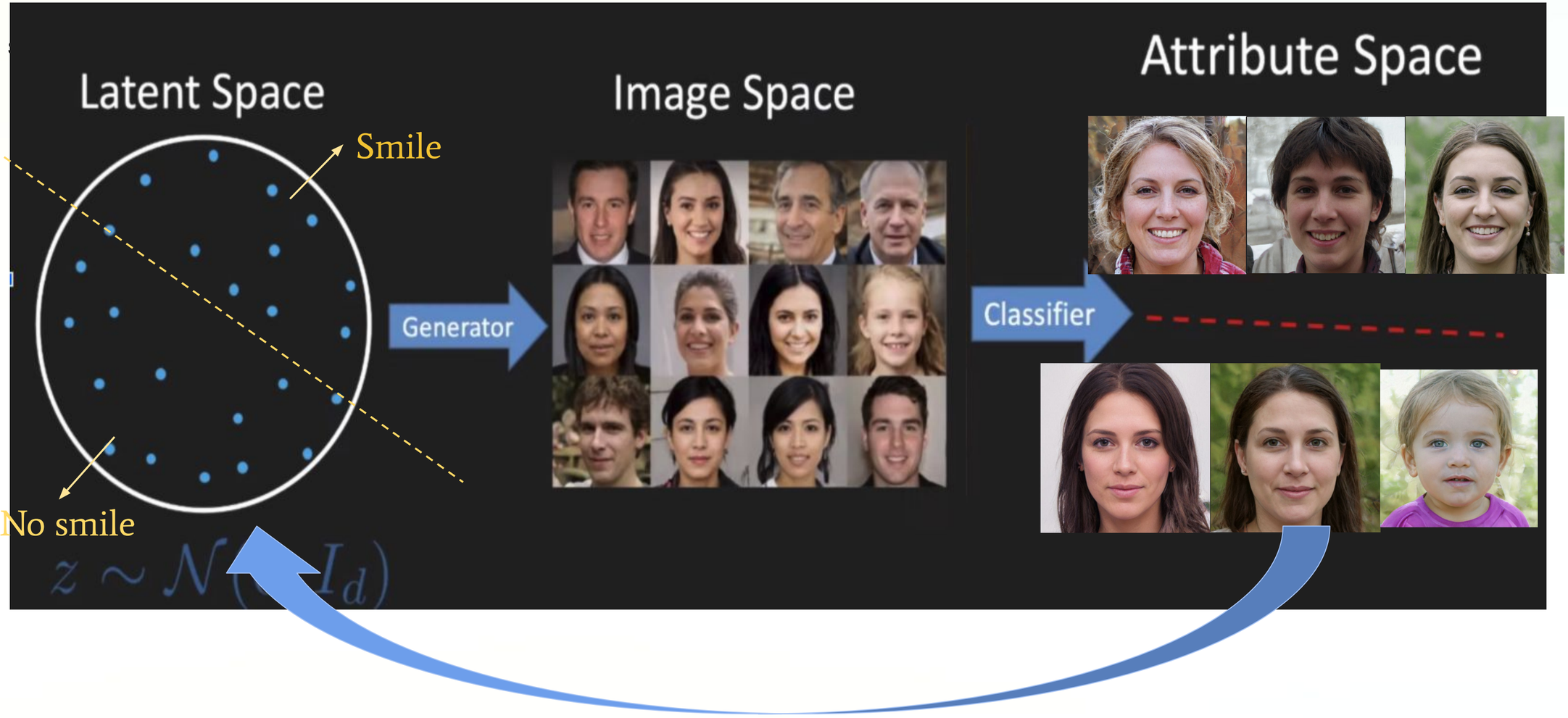
$$\mathrm{AdaIN}(\mathbf{x}_i, \mathbf{y}) = \mathbf{y}_{s,i} \frac{\mathbf{x}_i - \mu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)} + \mathbf{y}_{b,i},$$

StyleGAN Control





Yujun Shen, Jinjin Gu, Xiaoou Tang, Bolei Zhou, Interpreting the Latent Space of GANs for Semantic Face Editing CVPR 2020

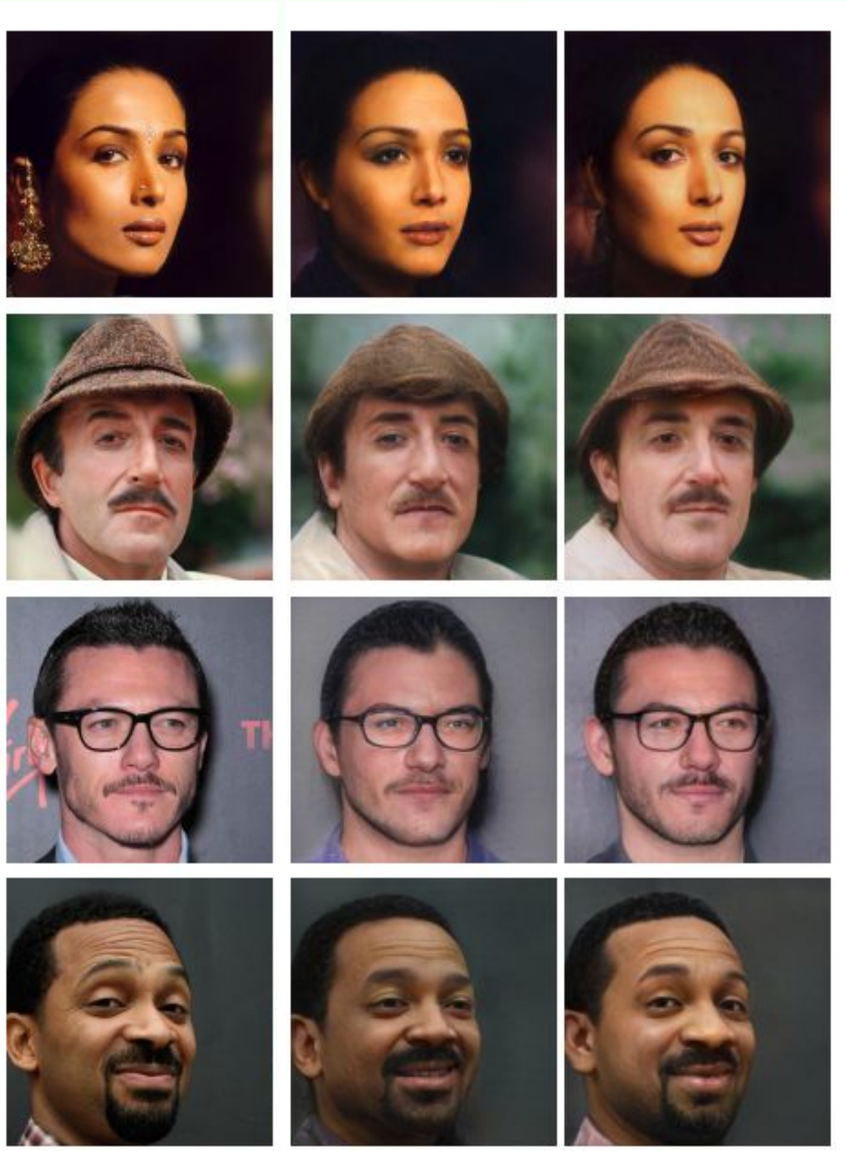


Yujun Shen, Jinjin Gu, Xiaoou Tang, Bolei Zhou, Interpreting the Latent Space of GANs for Semantic Face Editing CVPR 2020

Lightricks

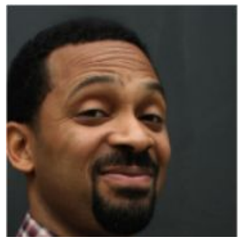
Lightricks

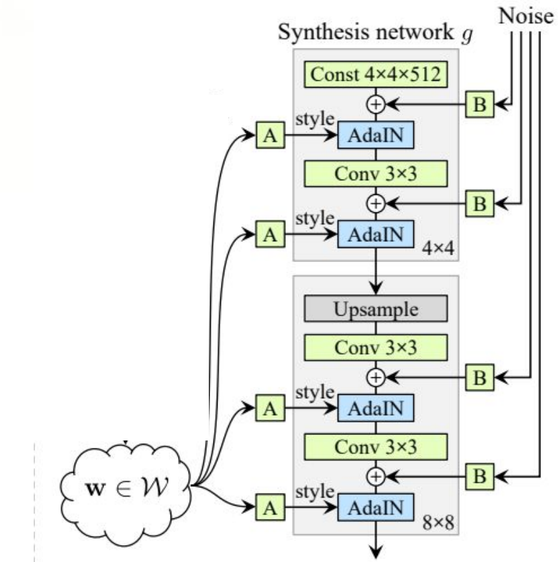
StyleGAN Inversion
Direct / Encoder



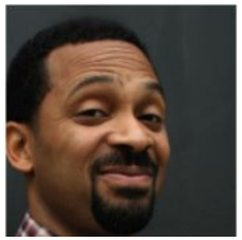
Lightricks

StyleGAN Inversion
Direct / Encoder



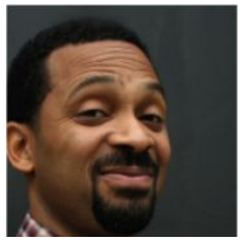


Direct optimization

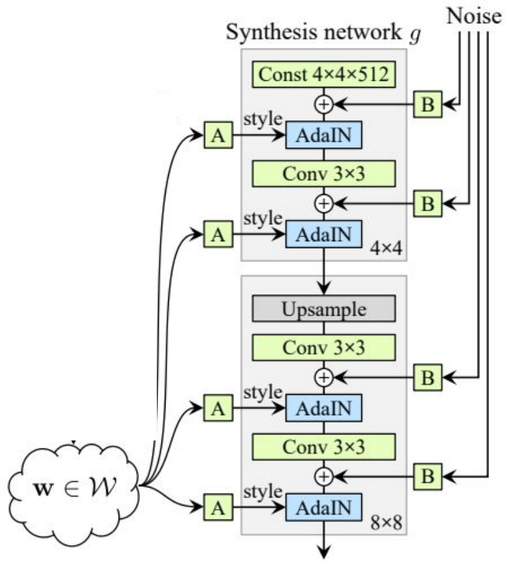


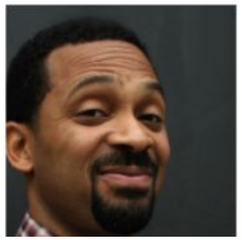
Lightricks

StyleGAN Inversion
Direct / Encoder



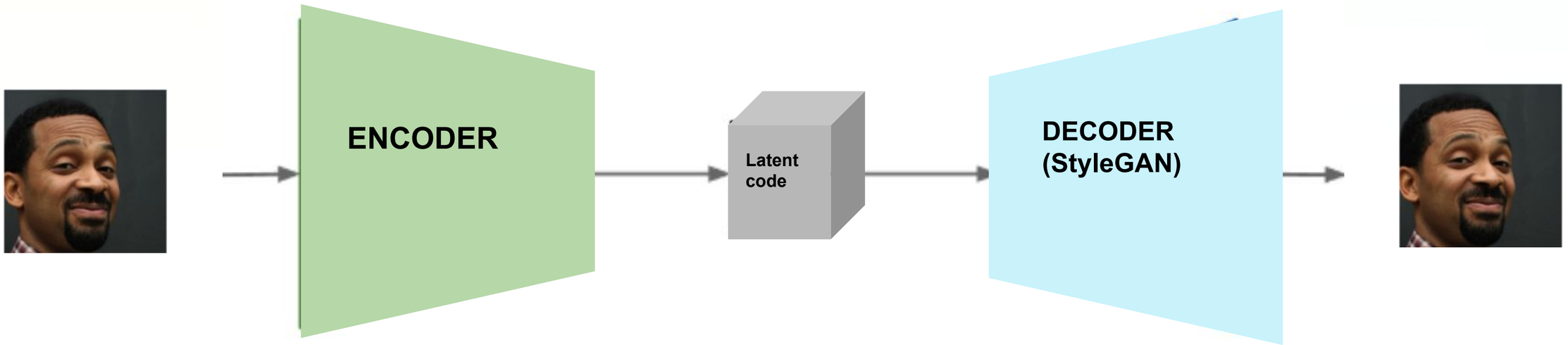
Encoder





Lightricks

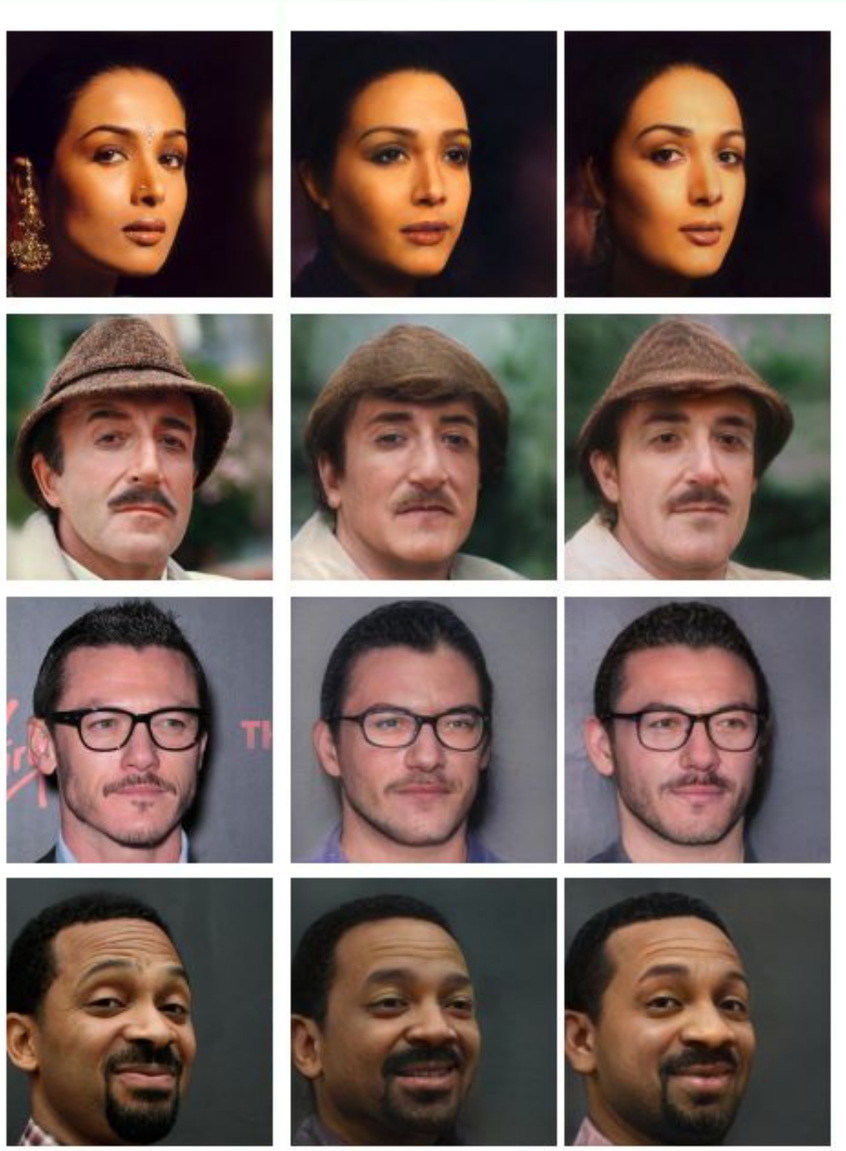
AutoEncoder



Lightricks

StyleGAN Inversion
Direct / Encoder

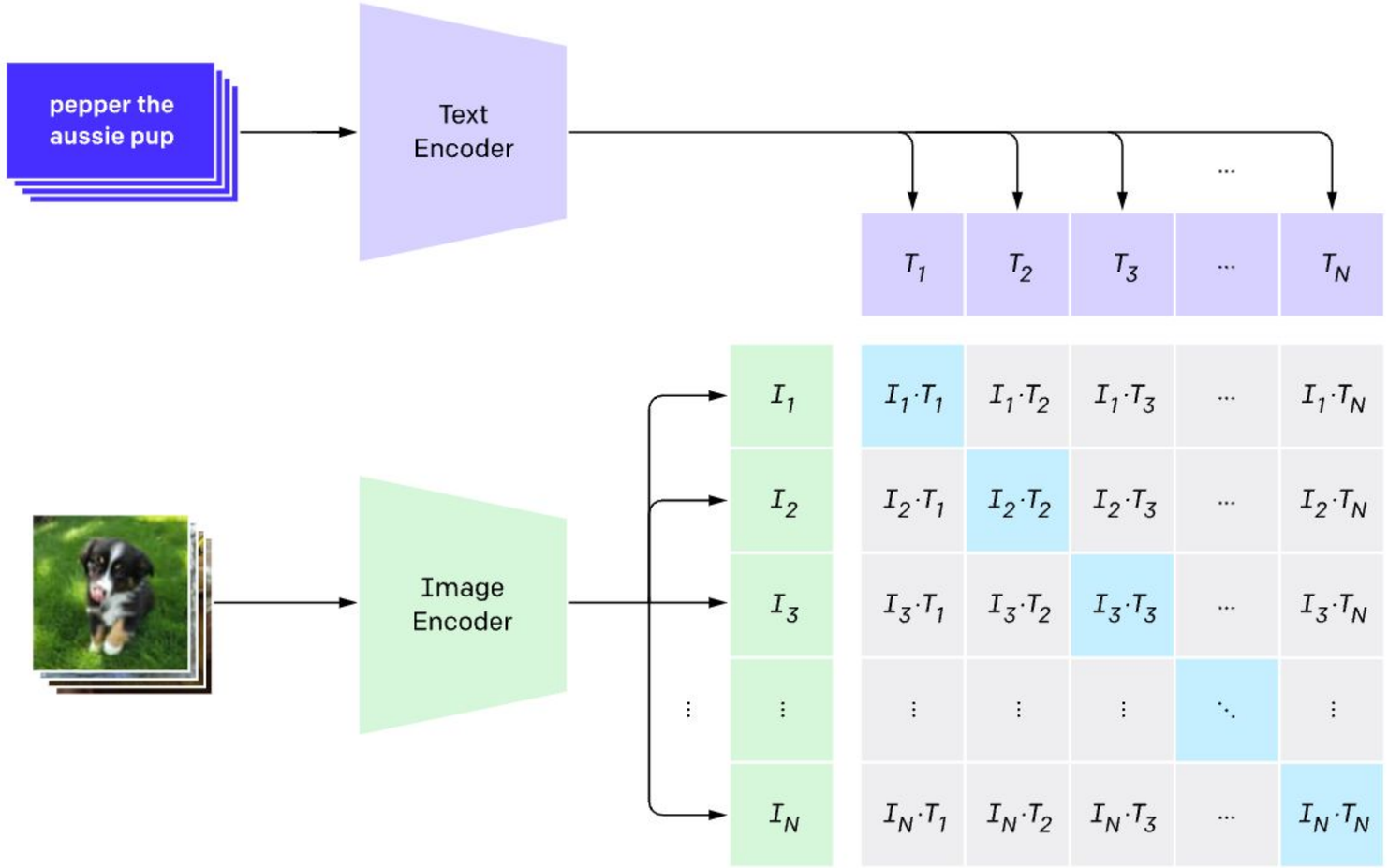
Which one is better?



Lightricks

CLIP- Text and Image encoder

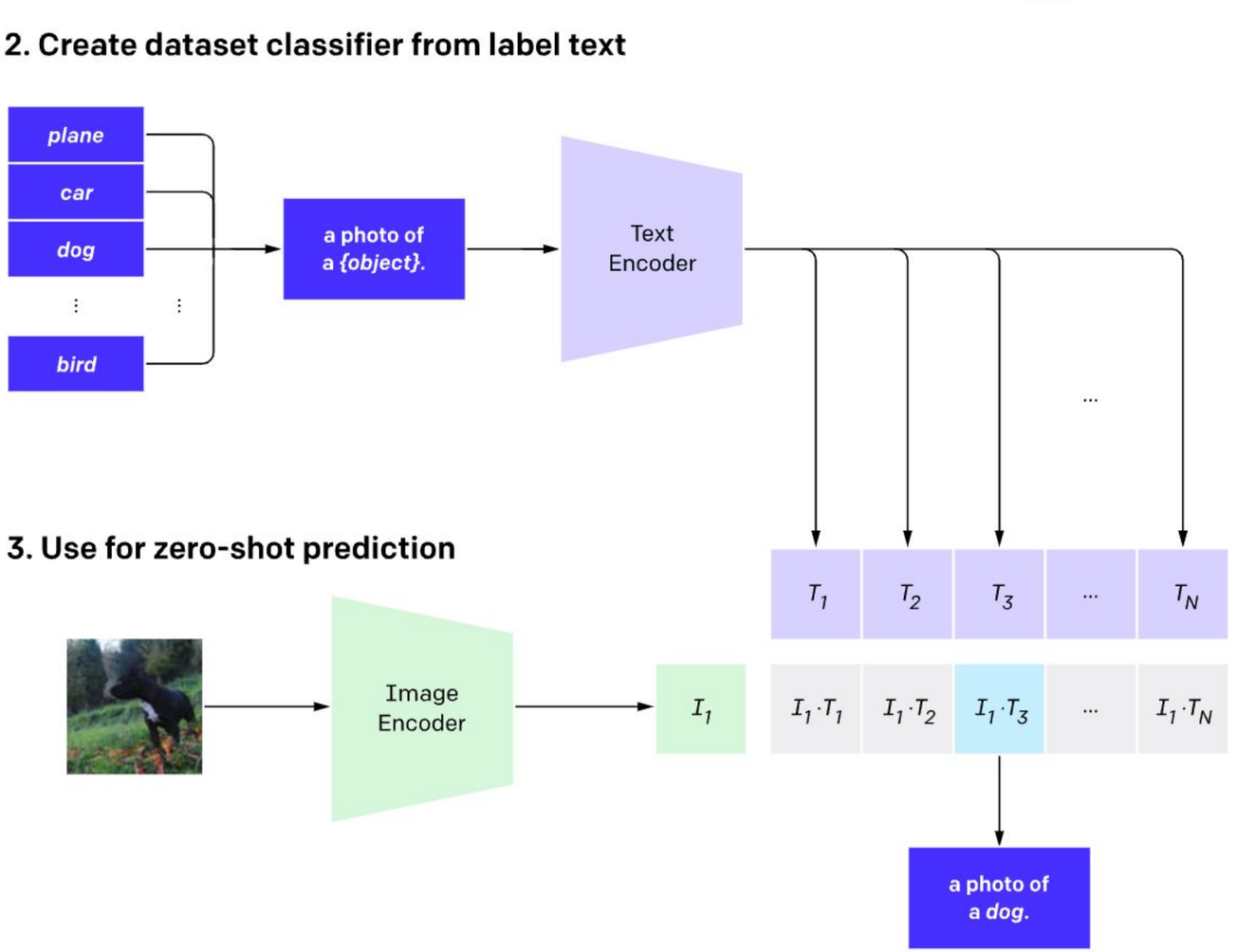
OpenAI, Jan 2021



400 M scraped pairs of image and caption

CLIP- Text and Image encoder

OpenAI, Jan 2021

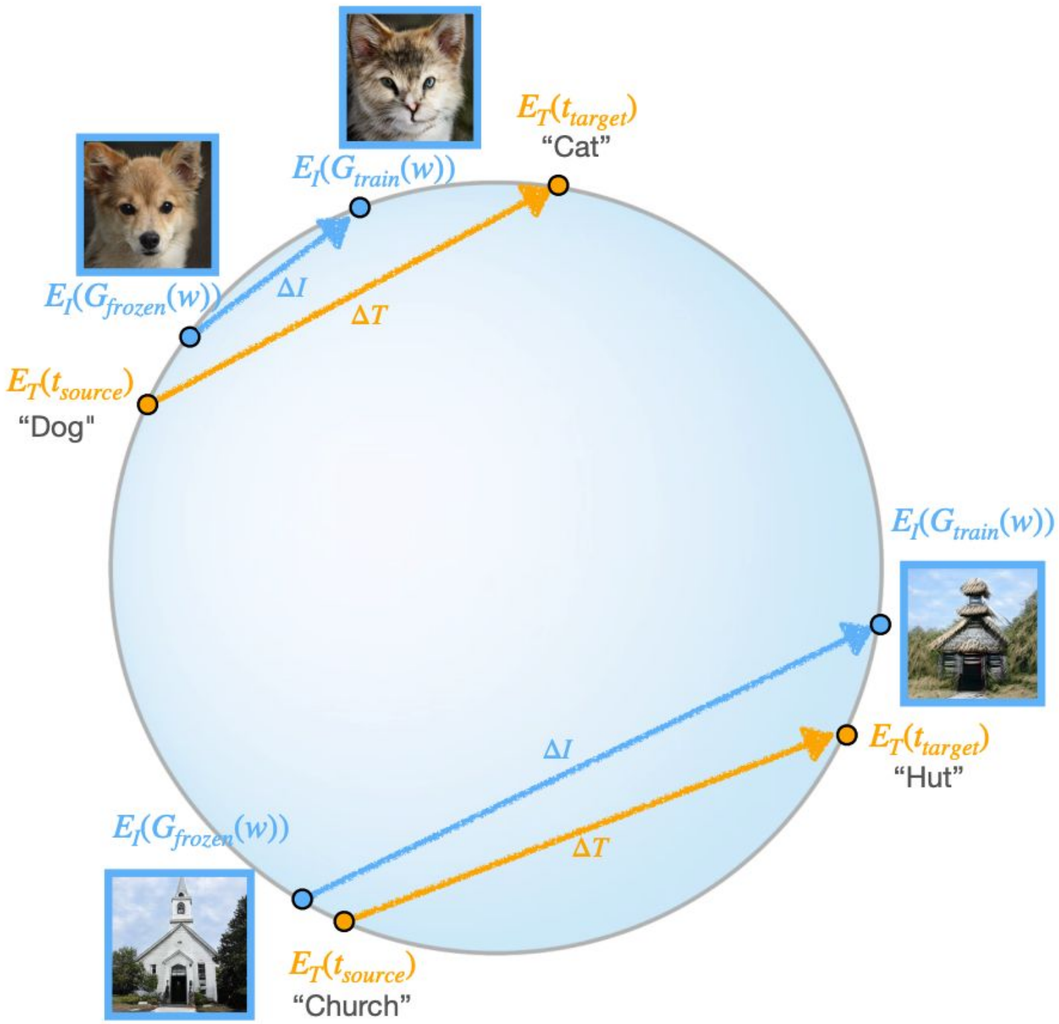


Lightricks

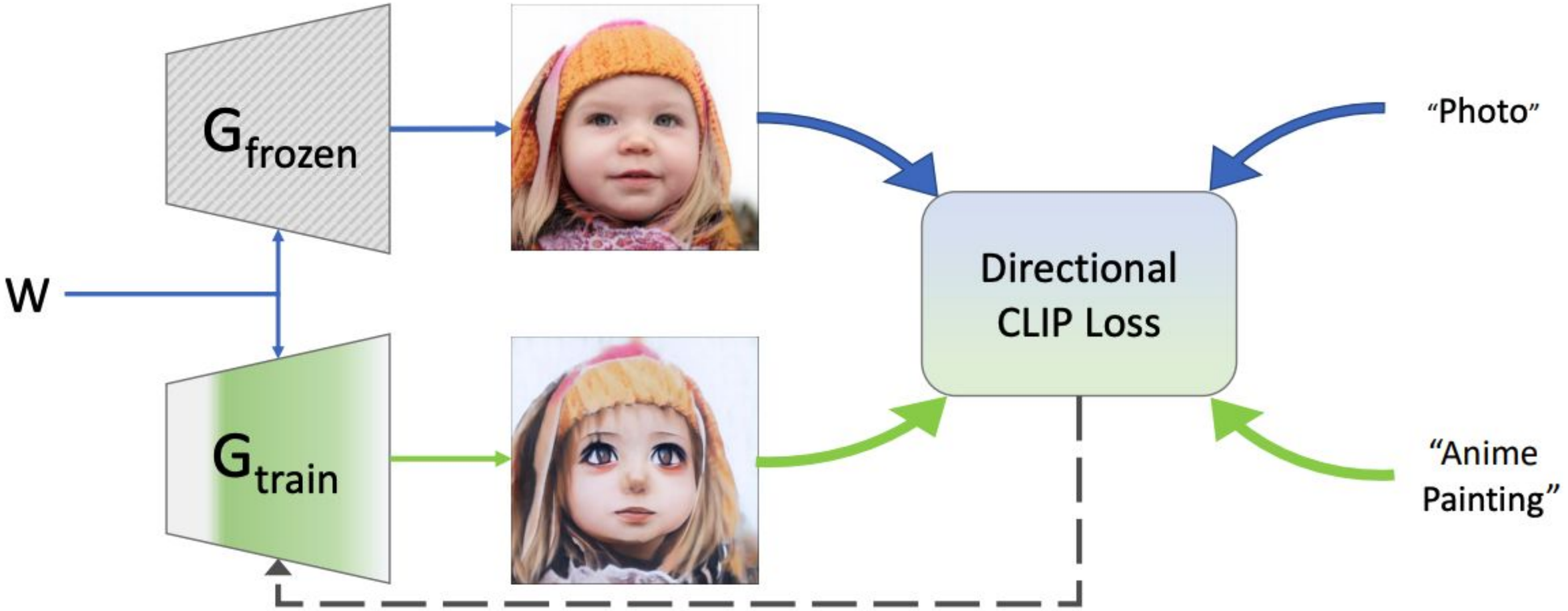
StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery

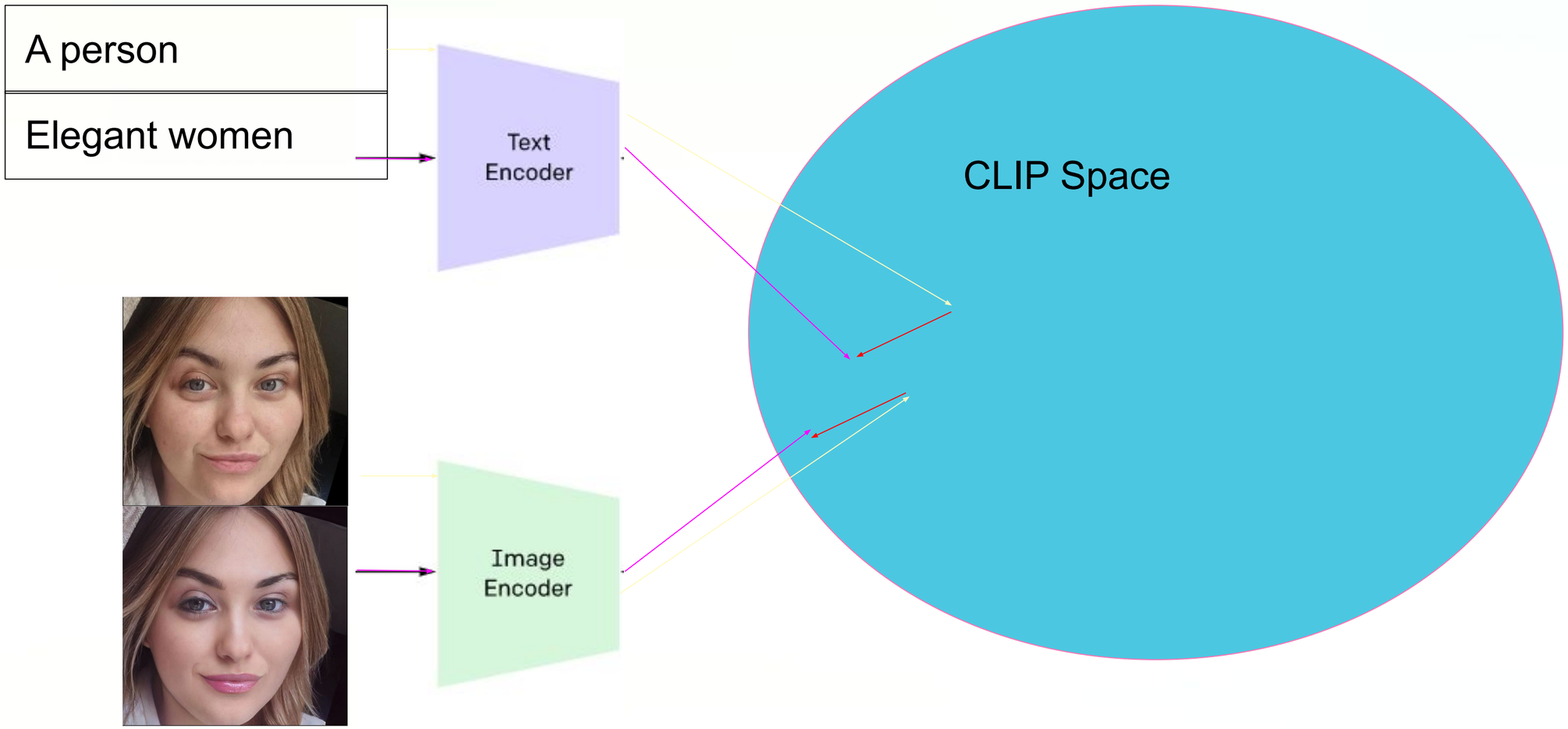
StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators

https://arxiv.org/pdf/2103.17249.pdf, Or Patashnik, Rinon Gal, Dani Cohen Or, et al



$$\Delta T = E_T\left(t_{target}\right) - E_T\left(t_{source}\right) ,$$
$$\Delta I = E_I\left(G_{train}\left(w\right)\right) - E_I\left(G_{frozen}\left(w\right)\right) ,$$
$$\mathcal{L}_{direction} = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I|\,|\Delta T|} ,$$


G frozen
W
G train
Directional
CLIP Loss
"Photo"
"Anime
Painting"


A person
Elegant women
Text Encoder
CLIP Space
Image Encoder

Lightricks

Face GANs

Before After

“Elegant business”

Before After

“Bold Lips”

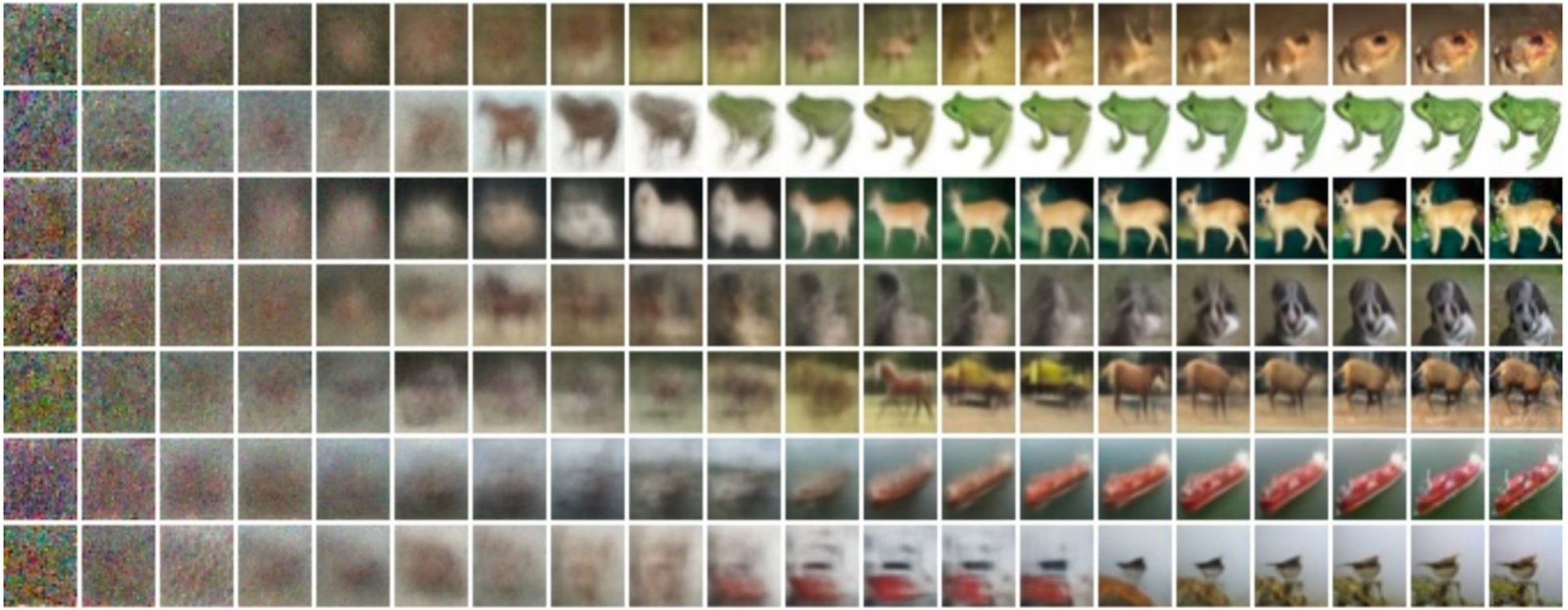
Part 2

Gans Drawbacks

- Unstable Training
- Mode collapse

Diffusion models

Guided image generation, *UC Berkeley, June 2020, OpenAI, June 2021, Google Brain, July 2021*

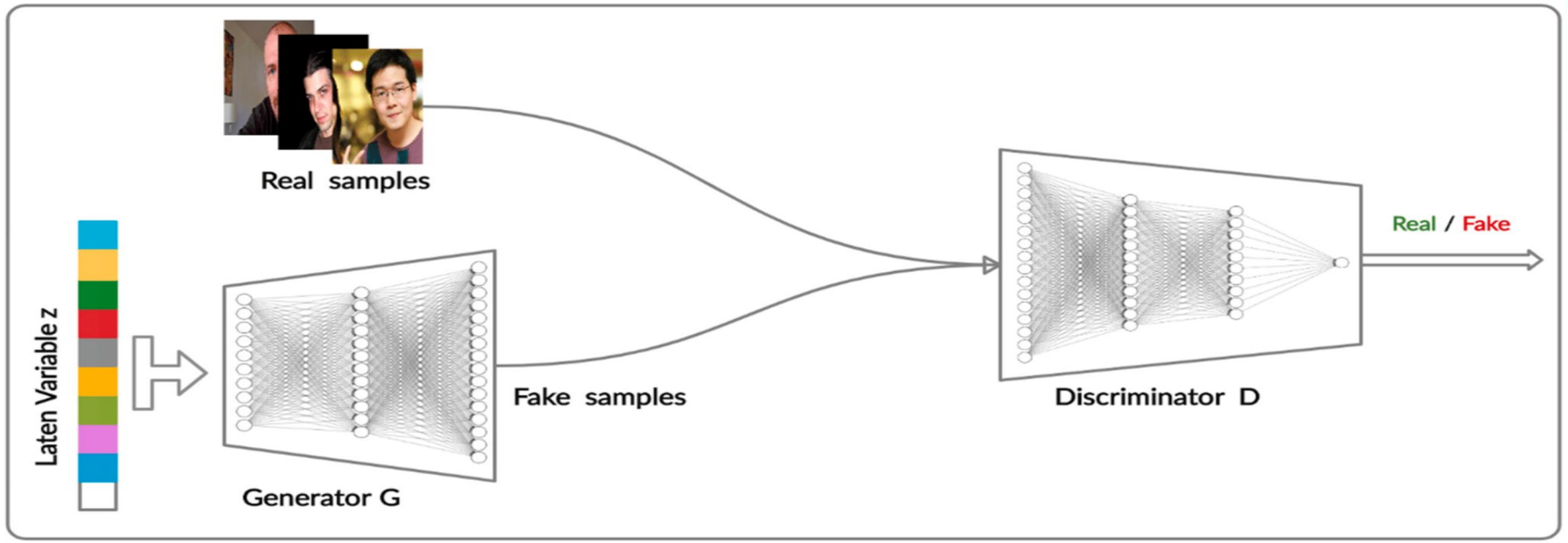


Diffusion model

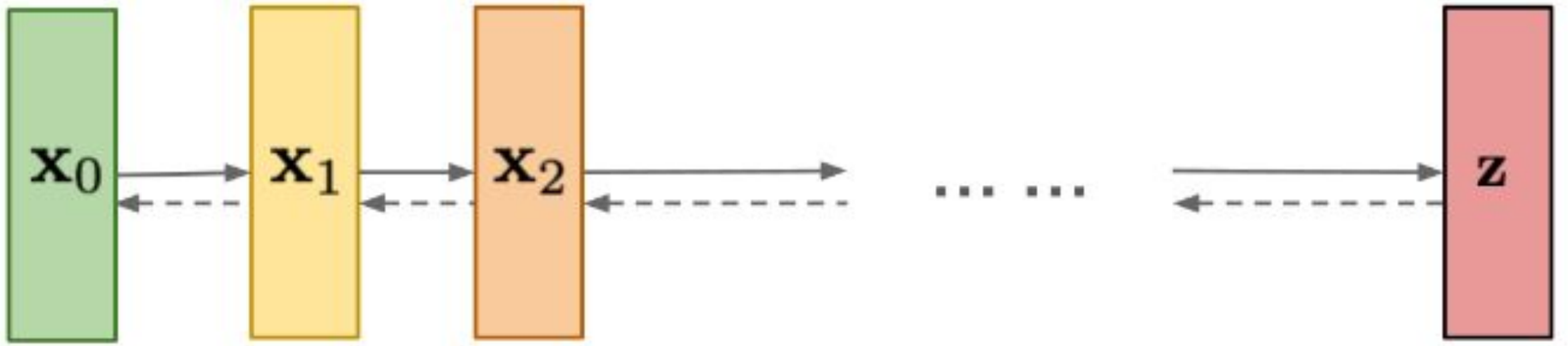
Input: random noise tensor (dimensions - 64x64x6)

Output: tensor (dimensions 64x64x3)

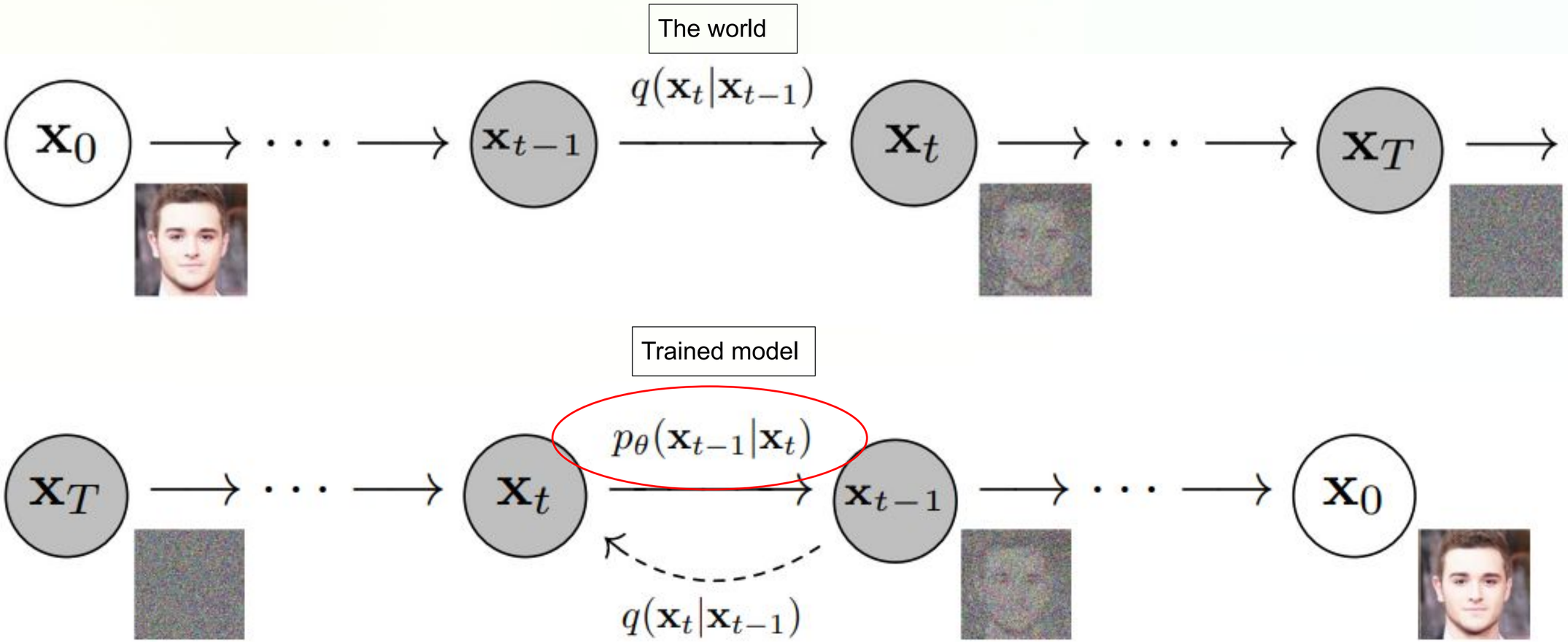
GAN: Adversarial training



Diffusion models: Gradually add Gaussian noise and then reverse

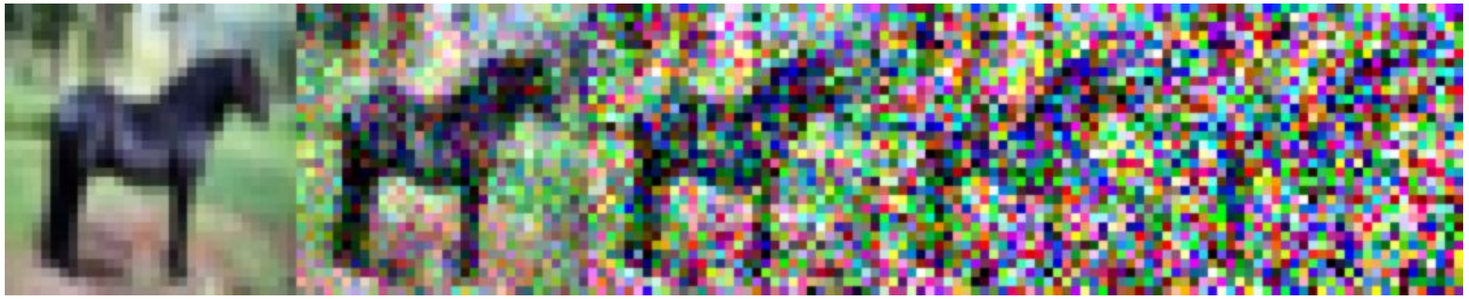


Lightricks


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

Creating Noised Images

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$$



sample an arbitrary step of the noised latents directly conditioned on the input x0

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \text{ where } \bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i \qquad \alpha_t := 1-\beta_t$$

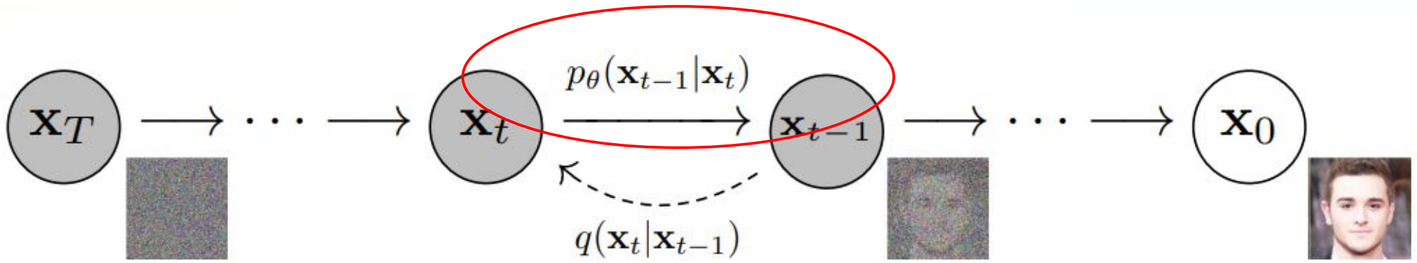
$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

Lightricks

Training a Diffusion Model - the Reverse Step

Diffusion model can be trained to predict $\mathbf{x}_{t-1}$ given $\mathbf{x}_t$, t

But, In practice those models are trained to predict the added noise



$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

Loss: L2($noise_\theta(\mathbf{x}_t), \epsilon$)

Iteration step: $\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_t}(\mathbf{x}_t - noise_\theta(\mathbf{x}_t)) + \sqrt{(1-\bar{\alpha}_t)}noise_\theta(\mathbf{x}_t)$

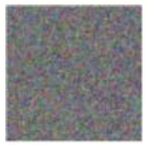
Dimensions:

$\bar{\alpha}_t$ - scalar

$\mathbf{x}_t$ Tensor 64x64x3

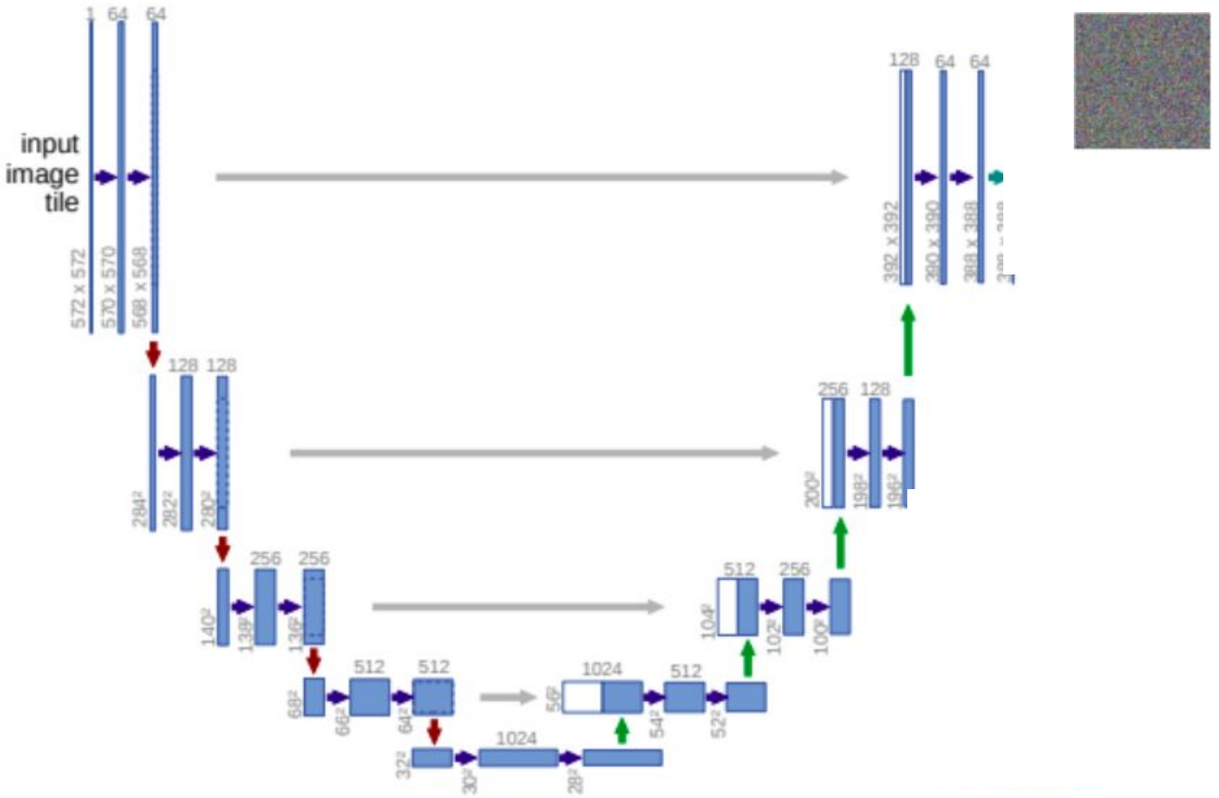
Training process

L2 loss

$\sqrt{\bar{\alpha}_t}$  + $\sqrt{1-\bar{\alpha}_t}$  =  , t

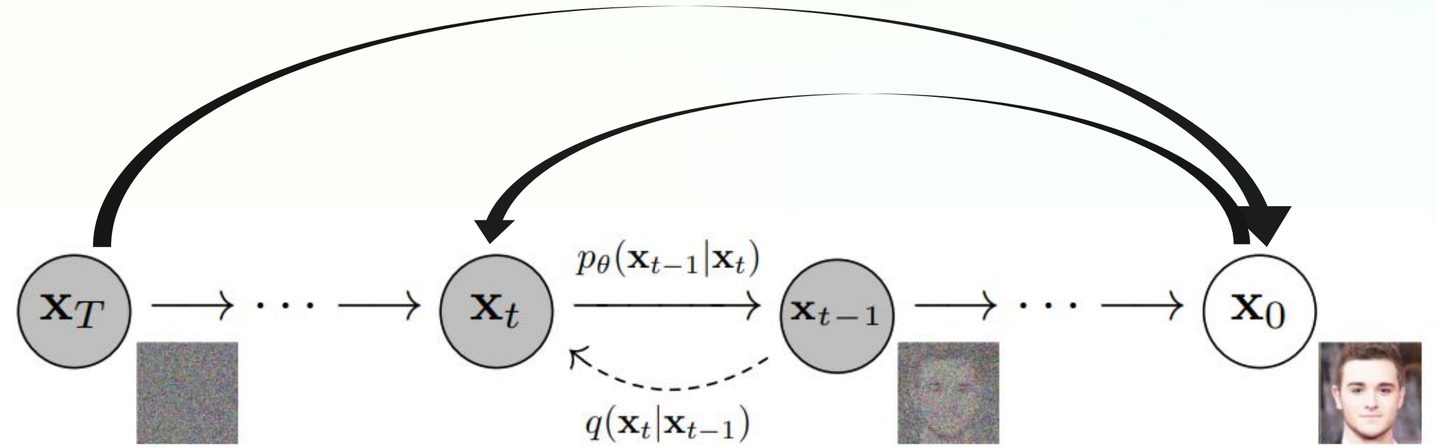
$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

t in (0,1000)

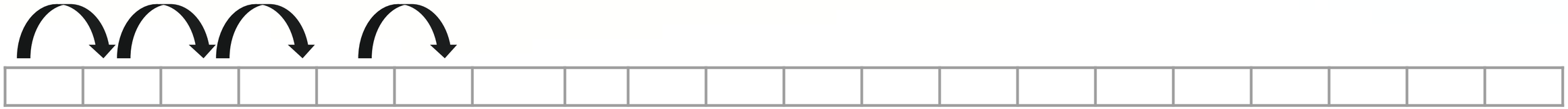


Unet 2015

Inference



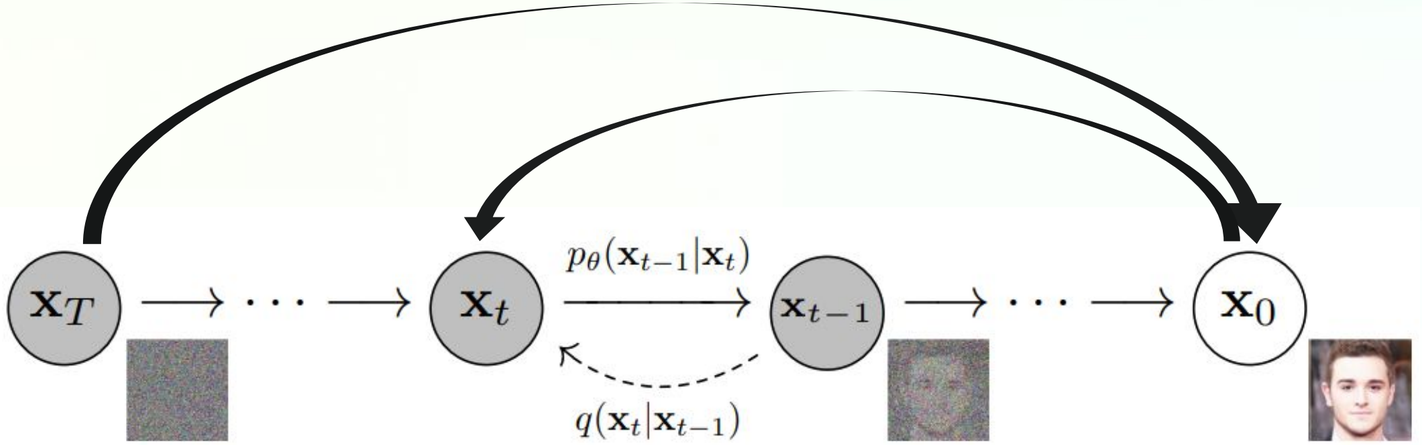
1000 steps??

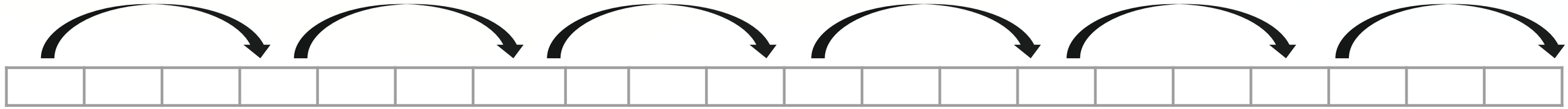


$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$$

Inference

1000 steps??

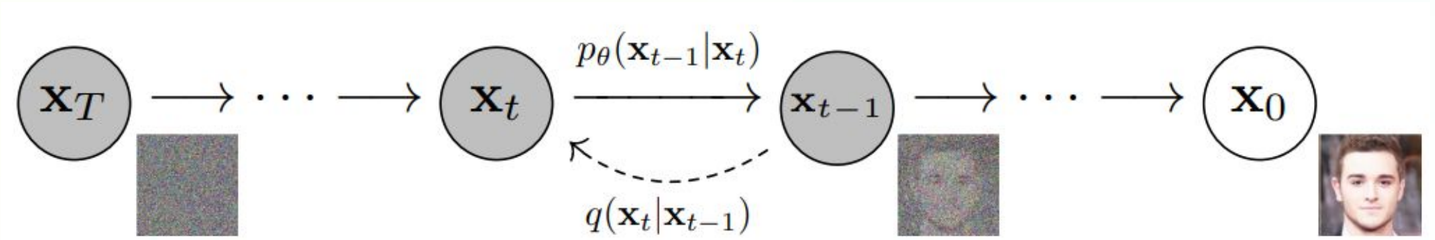


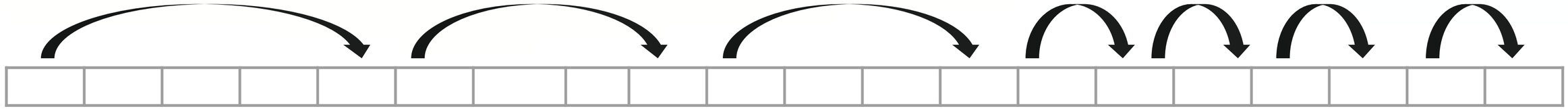


$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$$

Inference

1000 steps??





$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$$

GLIDE

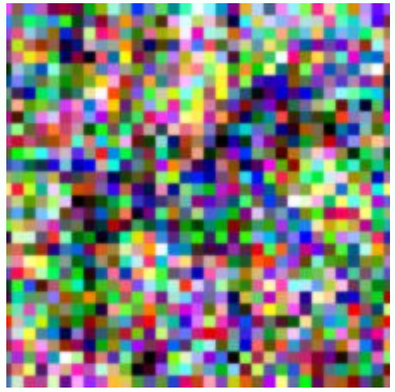
OpenAI, Dec 2021

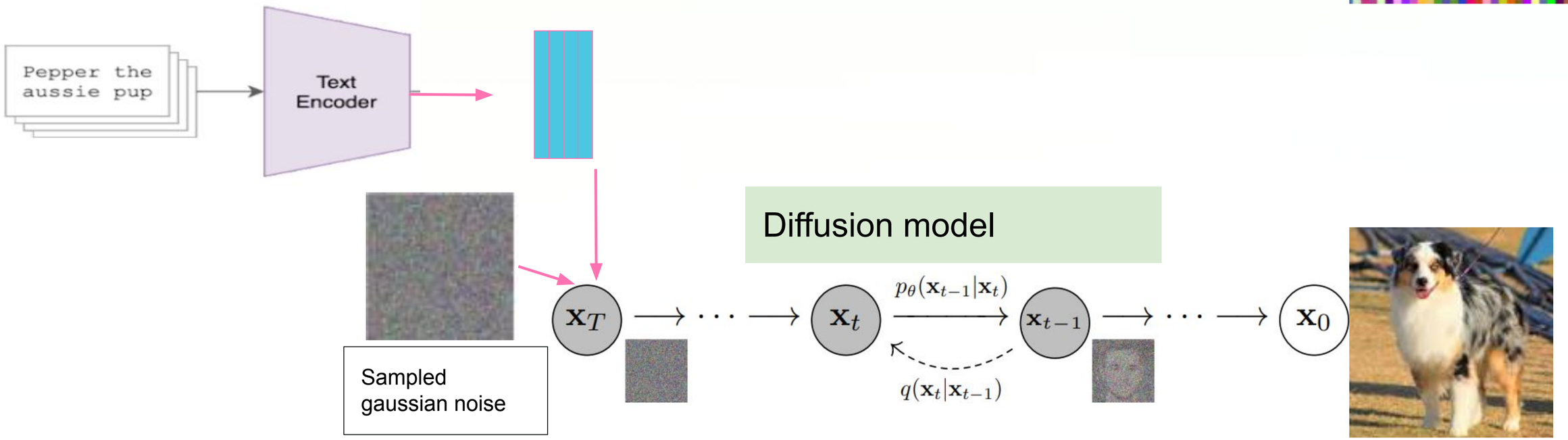
Given a text prompt + random noise image, generates a related image using a text transformer model based on the model of “Attention is all you need” (Bert) and a conditioned diffusion model.

“a boat in the canals of venice”

“a painting of a fox in the style of starry night”

Glide: Text 2 Image Architecture



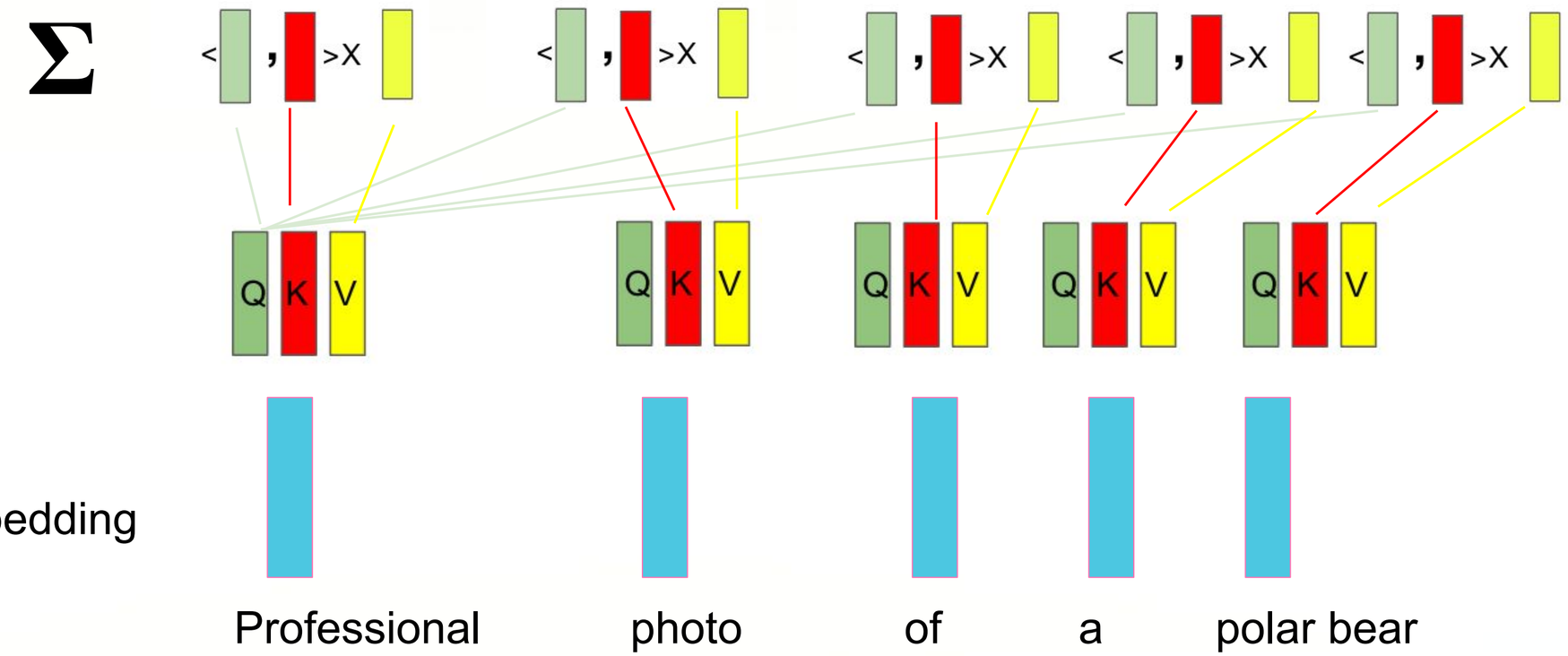


Lightricks

Transformers

Google, Dec 2017

Attention Layer



[Attention is all you need]

Lightricks

Transformers Training

“Deep learning is a subfield of artificial intelligence (AI) that focuses on training computer systems to learn and make intelligent decisions by analyzing large amounts of data. It is inspired by the structure and function of the human brain.”

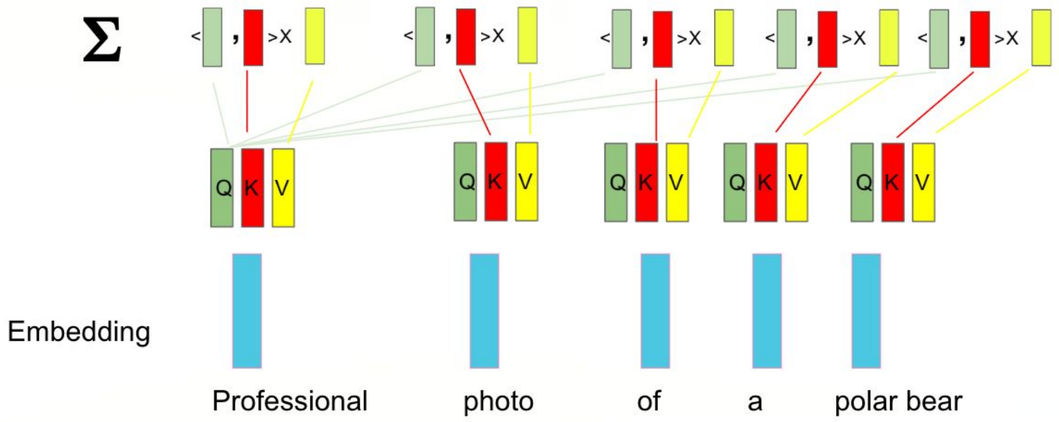
Lightricks

Transformers Training

“Deep learning is a subfield of artificial intelligence (AI) that focuses on training computer systems to learn and make intelligent decisions by analyzing large amounts of data. It is inspired by the structure and function of the human brain.”

Lightricks

Positional Embedding



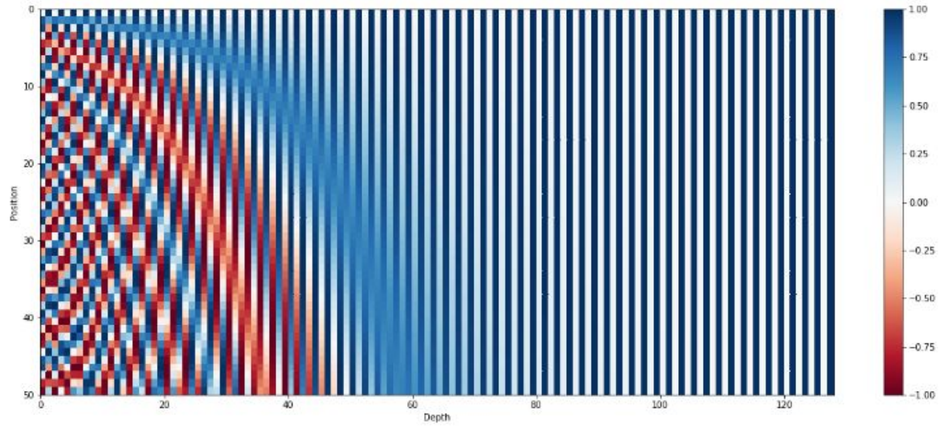
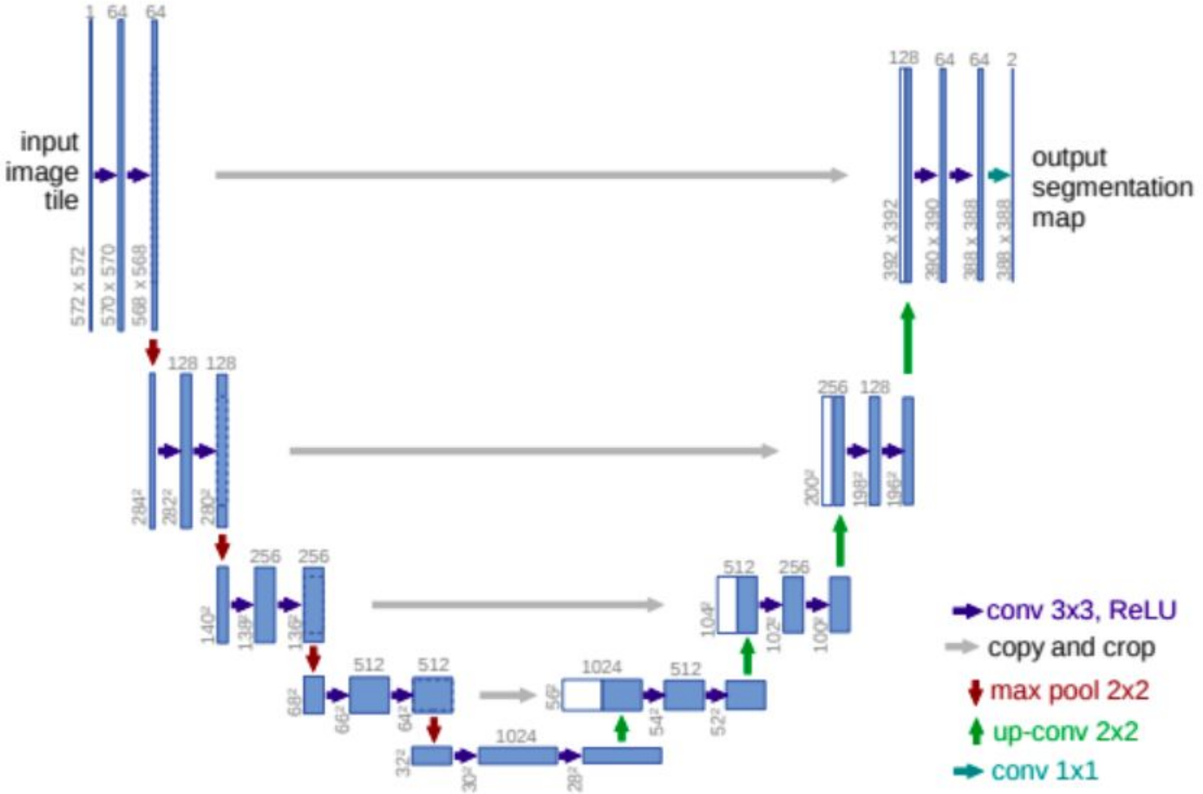


Figure 2 - The 128-dimensional positonal encoding for a sentence with the maximum lenght of 50. Each row represents the embedding vector $\overrightarrow{p_t}$

Lightricks

Global Context in the Vision Domain

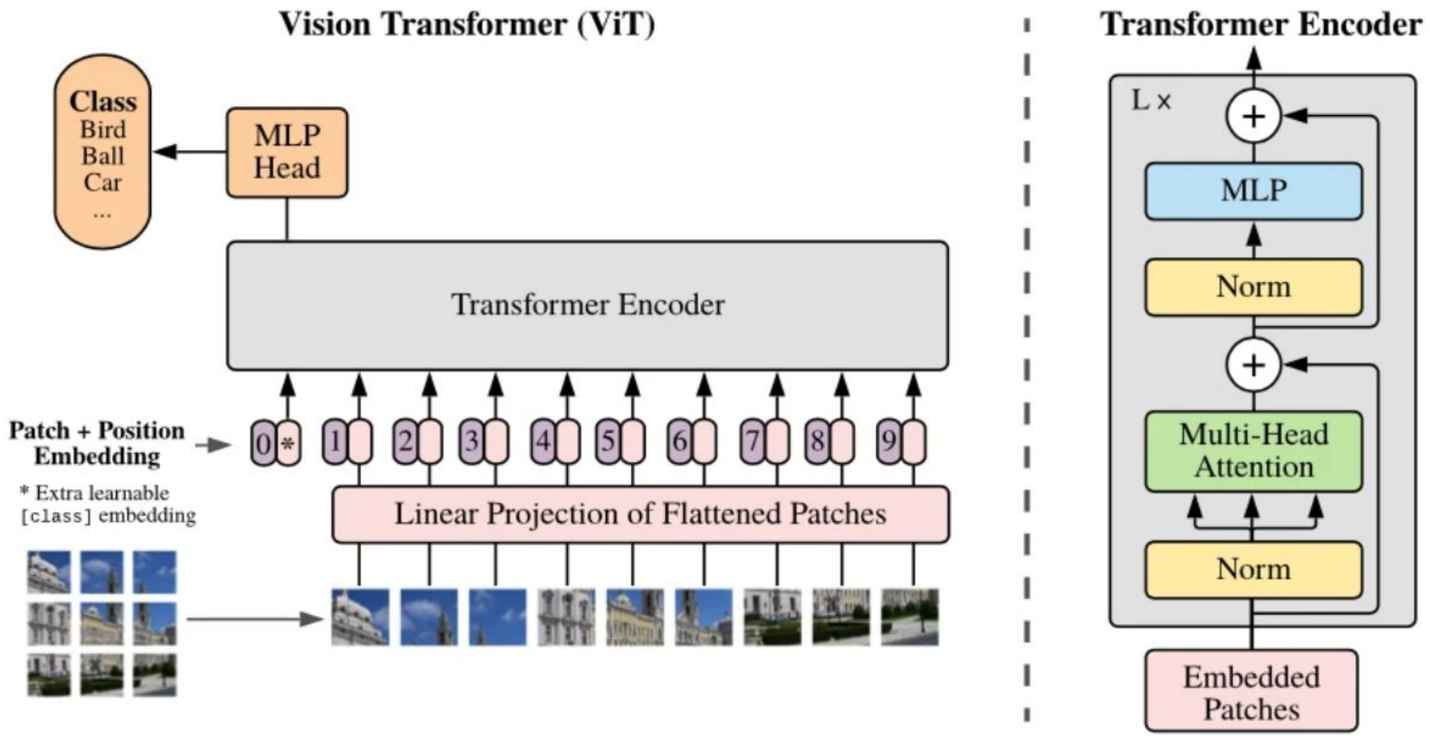


Global context

Unet 2015

Lightricks

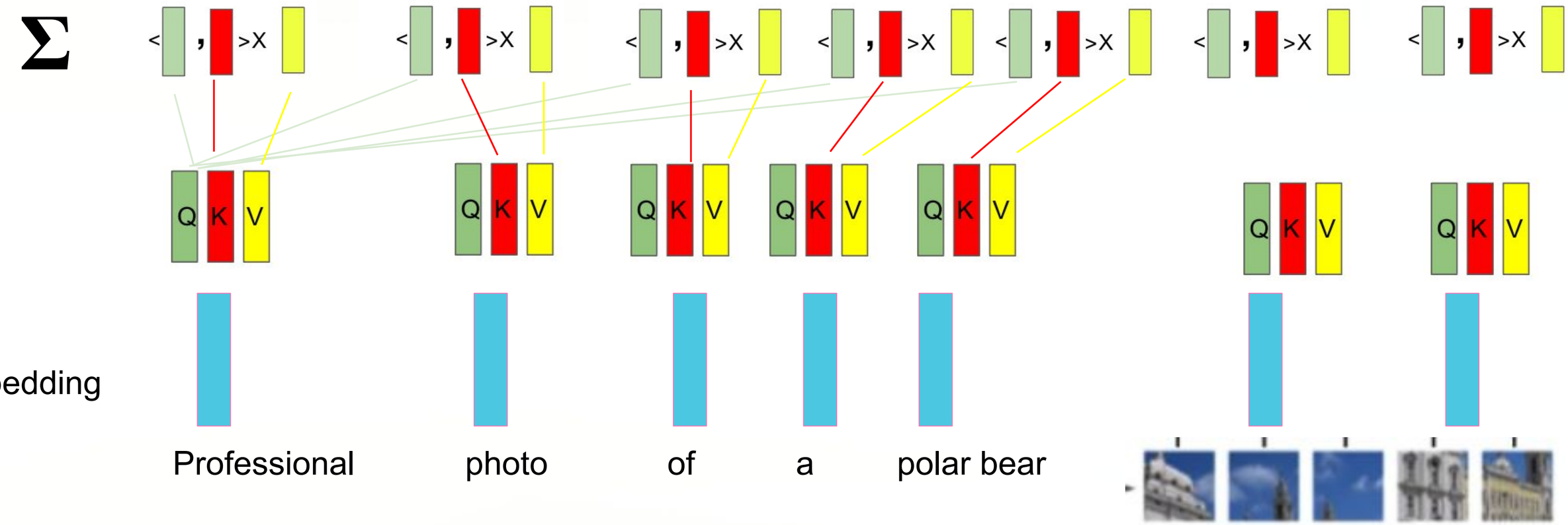
Visual Transformers



[An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale]

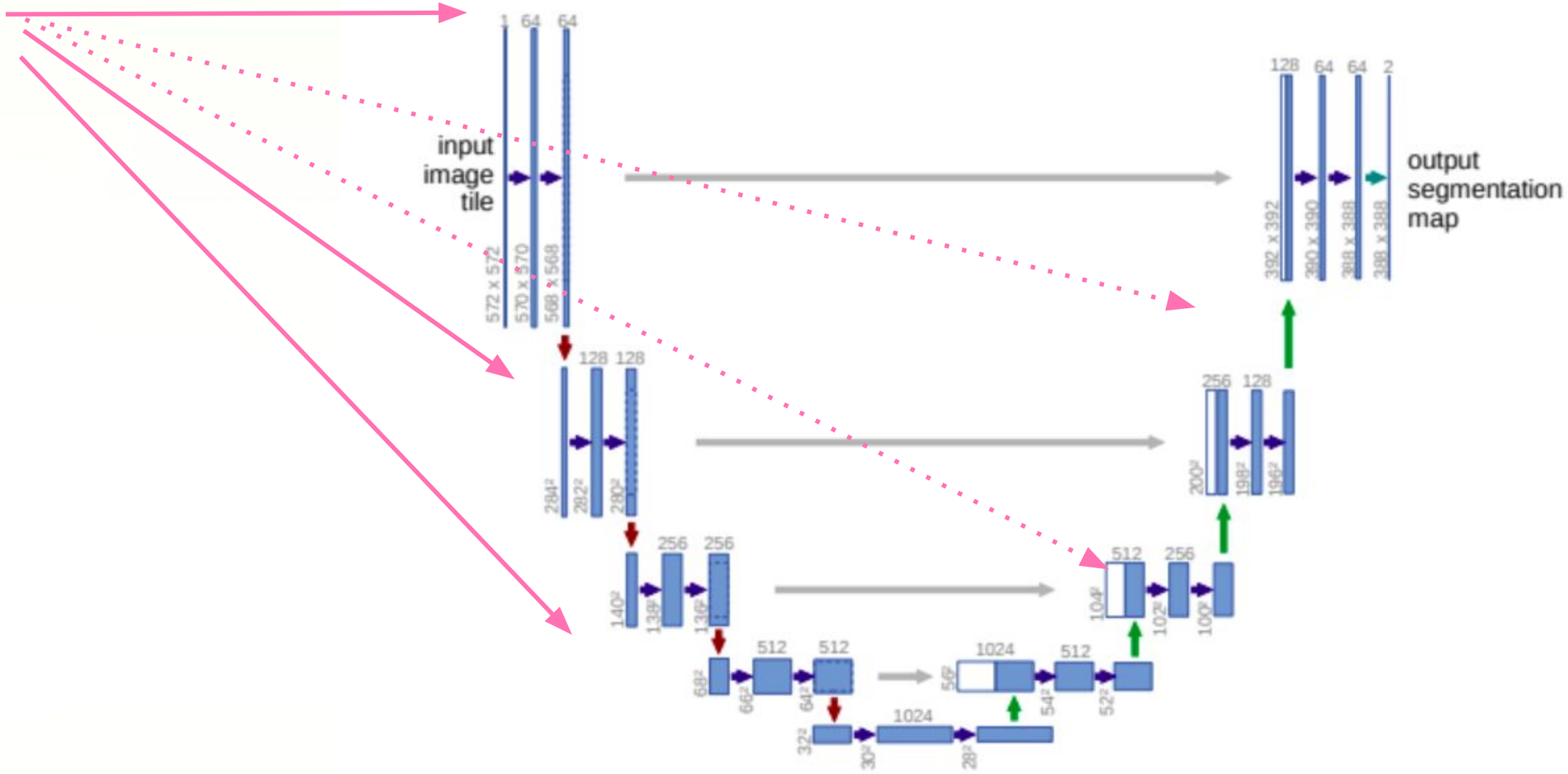
Lightricks

Cross Attention



Lightricks

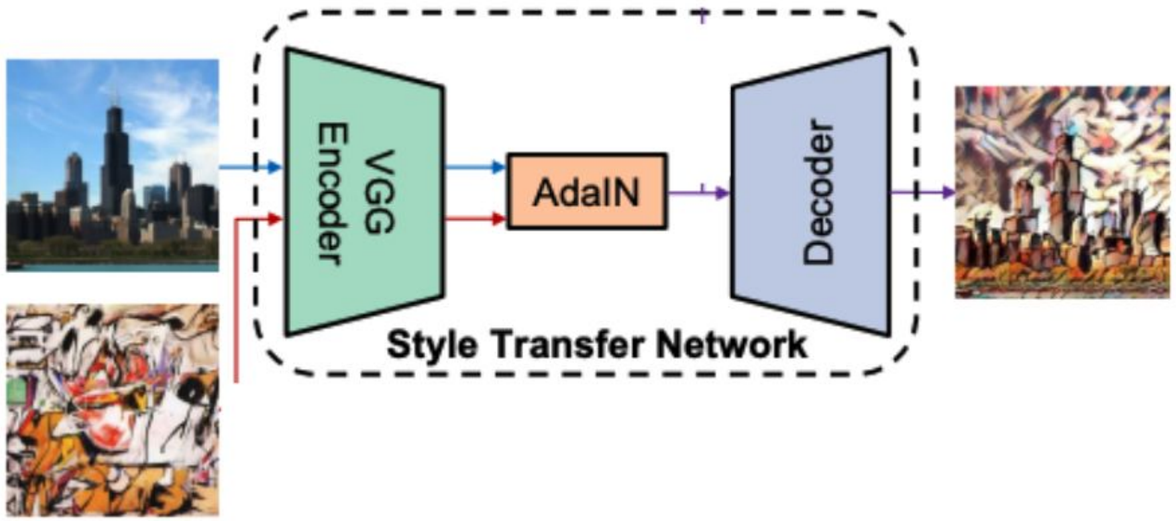
Adding Textual Input into the Diffusion Model

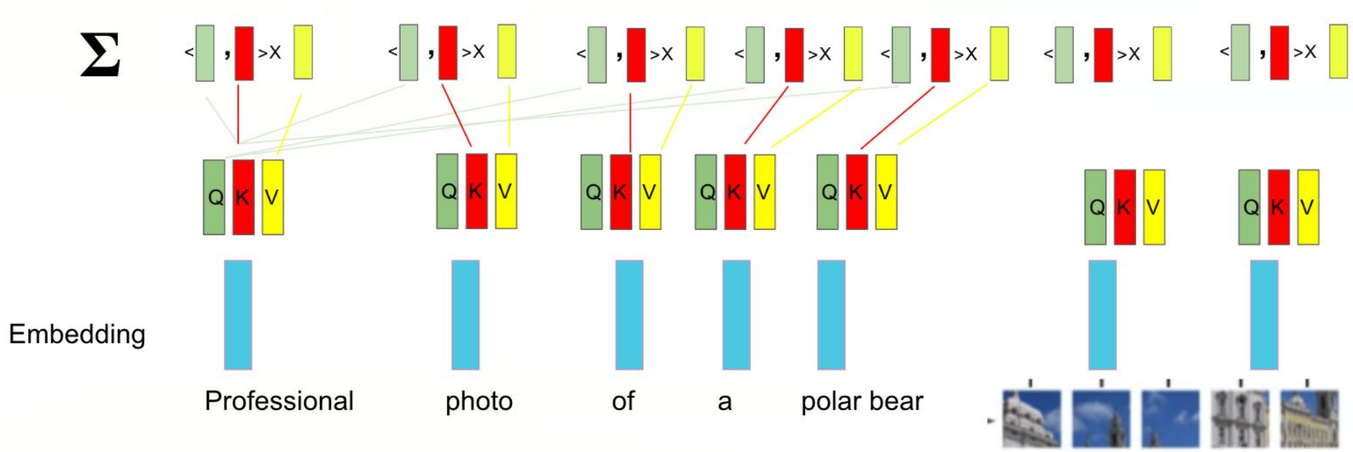


Lightricks

Multimodal Layers

1. Style/Text injection via channel statistics
2. Style/Text injection via Cross attention

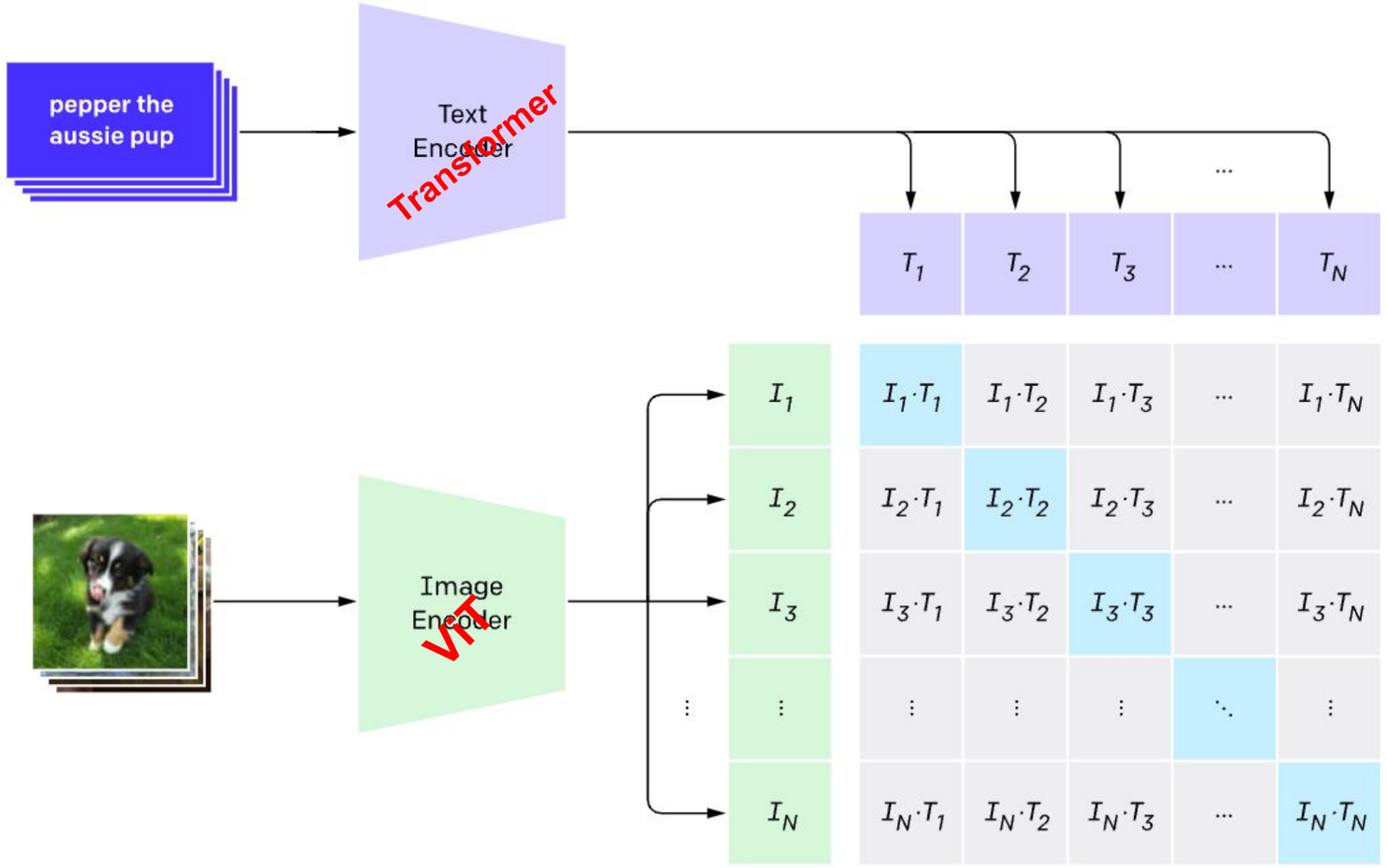




Lightricks

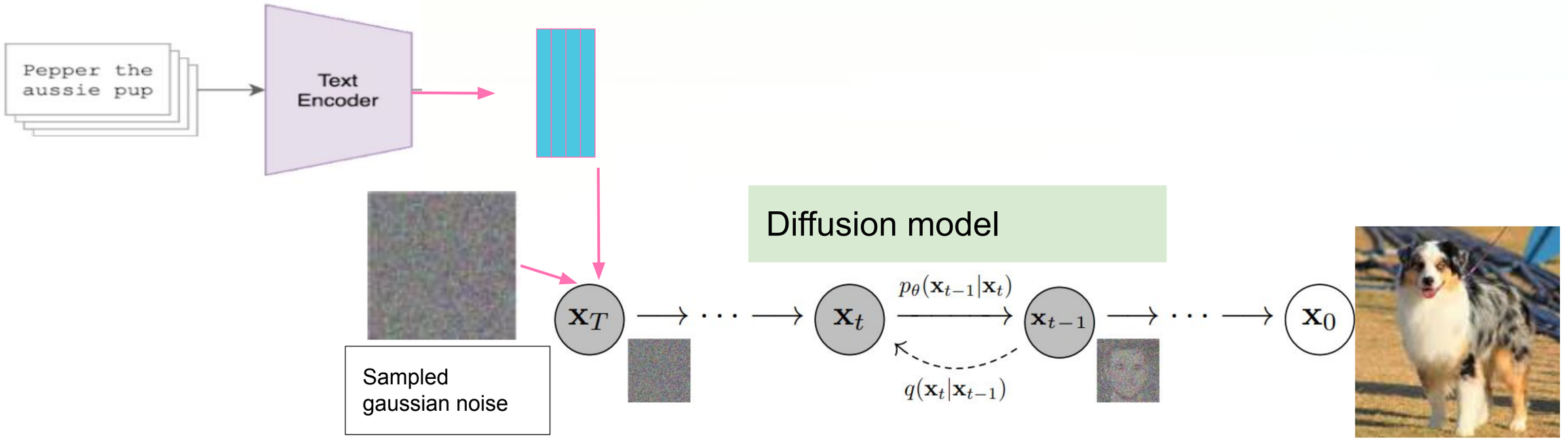
CLIP- Text and Image encoder

OpenAI, Jan 2021



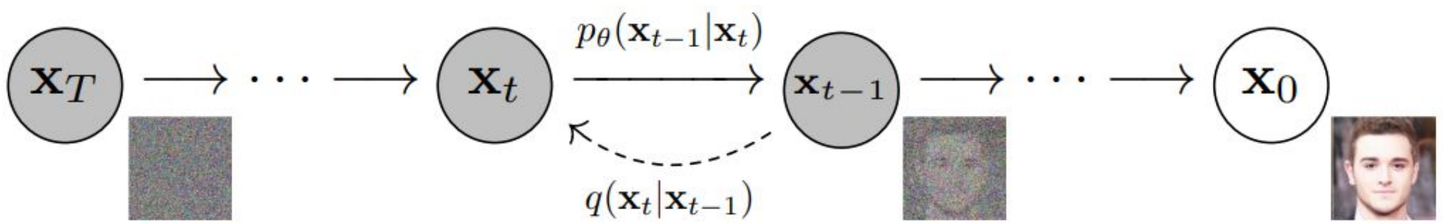
400 M scraped pairs of image and caption

Glide Text 2 Image Architecture



Class Free Guidance -

extrapolation toward the textual concept

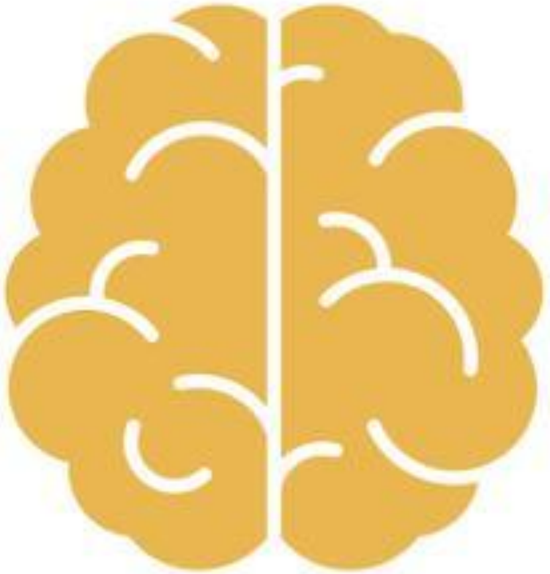


$$x = \mathbf{x}_{empty} + guidancescale * (\mathbf{x}_{prompt} - \mathbf{x}_{empty})$$

$$\mathbf{x}_{empty:t-1} = P(x_t, \text{""})$$

$$\mathbf{x}_{prompt:t-1} = P(x_t, \text{" prompt "})$$

System 1



Fast, intuitive and emotional

System 2

Slow, conscious and effortful

What Caused this Fast Progress?

1. **DATA**
   5 Billion image, text pairs!

2. **Parallel training in scale** -
   2.5M training steps of batch size 2048 in very short time. (256 TPU-v4 chips)

3. **New architectures**
   Multi domian, layers (Bert, GPT3, StyleGAN, ViT)

Lightricks

Imagen

A dragon fruit wearing karate belt in the snow.

An art gallery displaying Monet paintings. The art gallery is flooded. Robots are going around the art gallery using paddle boards.

A marble statue of a Koala DJ in front of a marble statue of a turntable. The Koala is wearing large marble headphones.

Imagen

Google Brain, May 2022

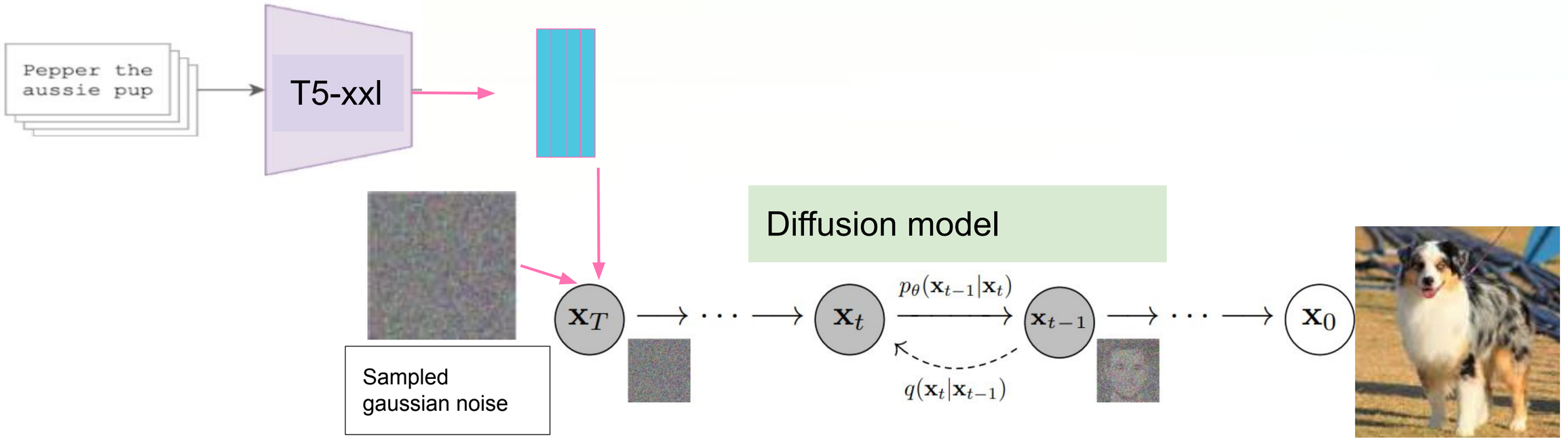
Huge text model (T5-XXL)

Given text embedding generates an image using a guided diffusion model.

Learn two 4x super-resolution steps (16x in total).

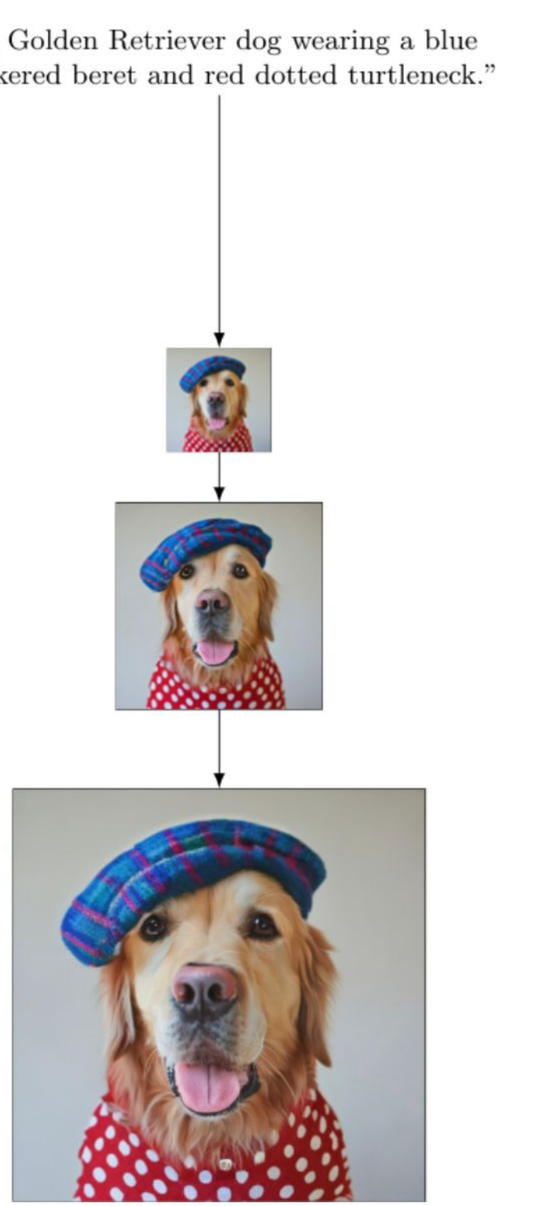
An art gallery displaying Monet paintings. The art gallery is flooded. Robots are going around the art gallery using paddle boards.

Imagen Architecture



Lightricks

Imagen:
Super Resolution models

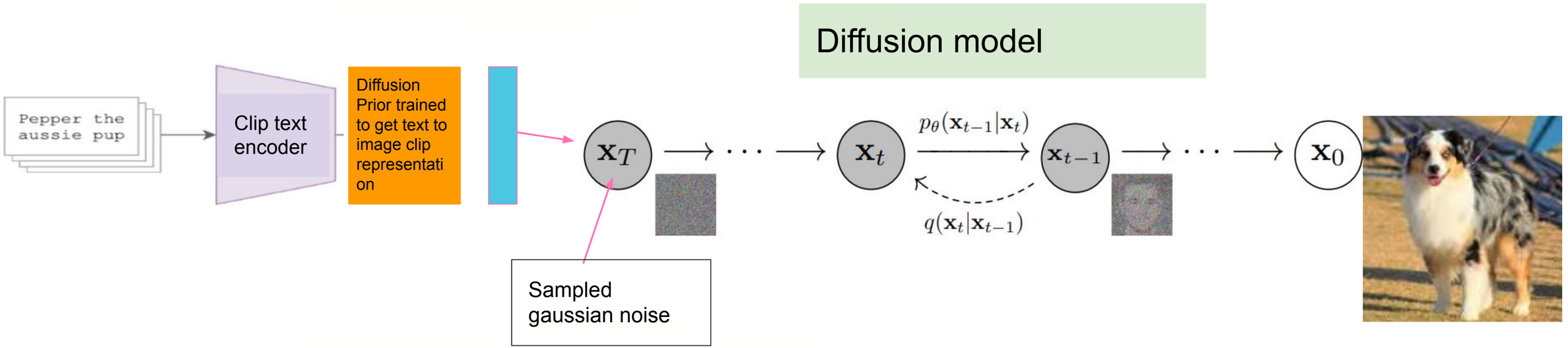


Lightricks

Dalle-2

openAI april 2022

Unclip - from clip image vector back into image



...

Emad
@EMostaque

Delighted to announce the public open source release of #StableDiffusion!

Please see our release post and retweet!
stability.ai/blog/stable-di...

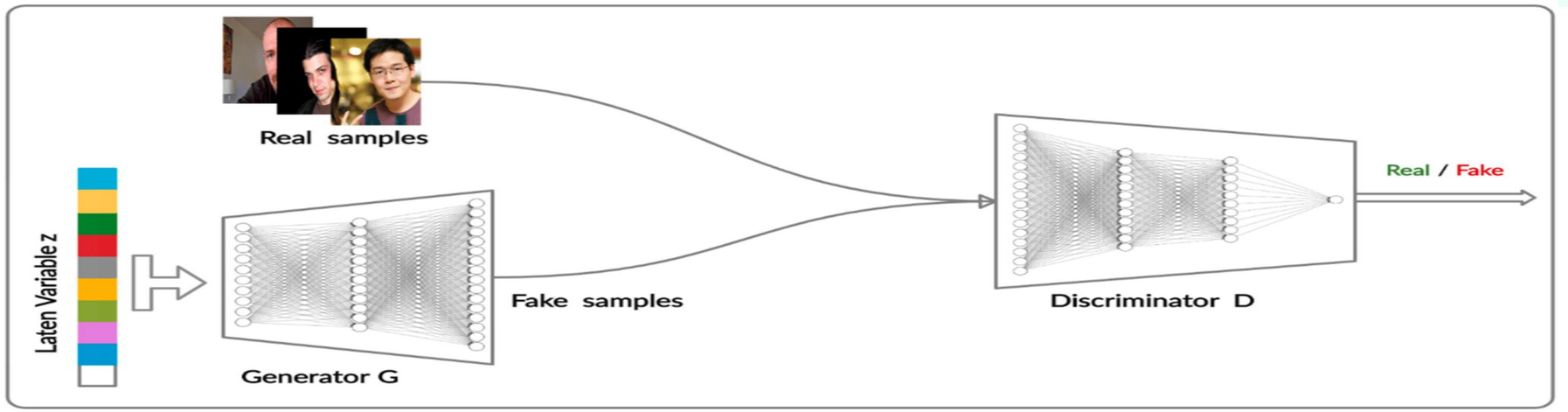
Proud of everyone involved in releasing this tech that is the first of a series of models to activate the creative potential of humanity

תרגם את הציוץ

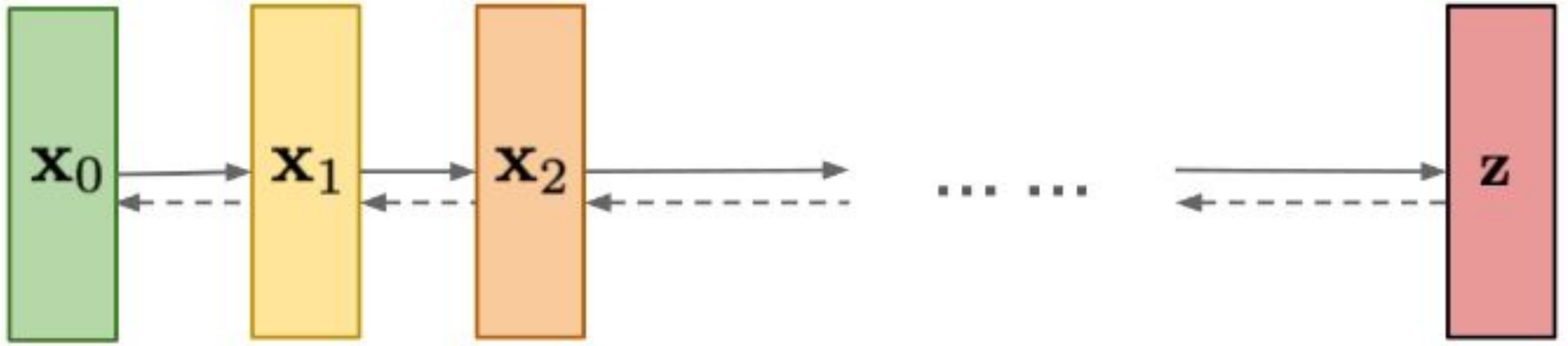
stability.ai
Stable Diffusion Public Release — Stability.Ai
We are delighted to announce the public release of Stable Diffusion and the launch of DreamStudio Lite.

9:07 אחה"צ · 22 באוג' 2022 · Twitter Web App

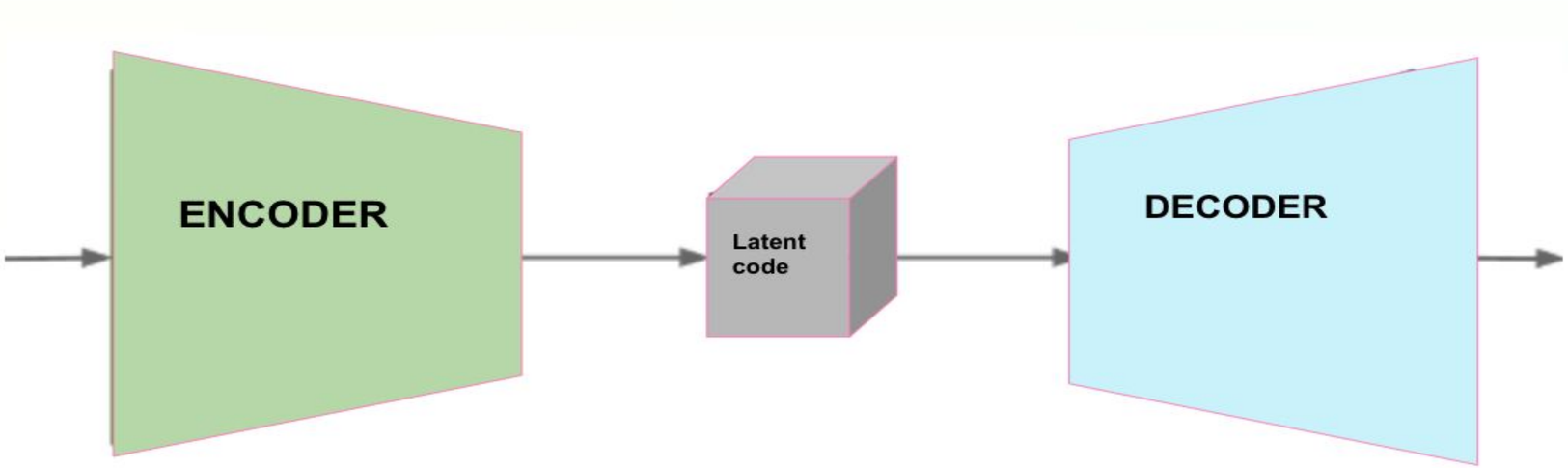
GAN: Adversarial training



Diffusion models: Gradually add Gaussian noise and then reverse

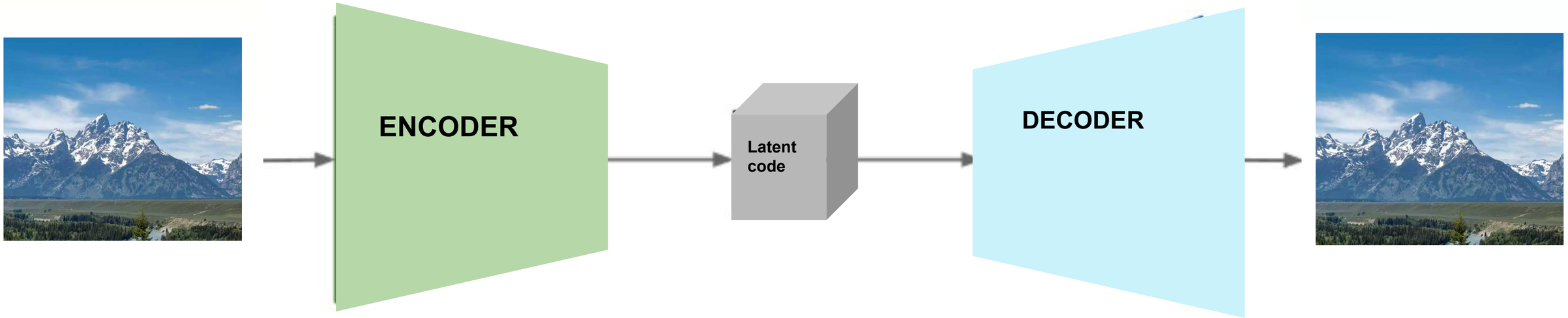


Variational Autoencoder:



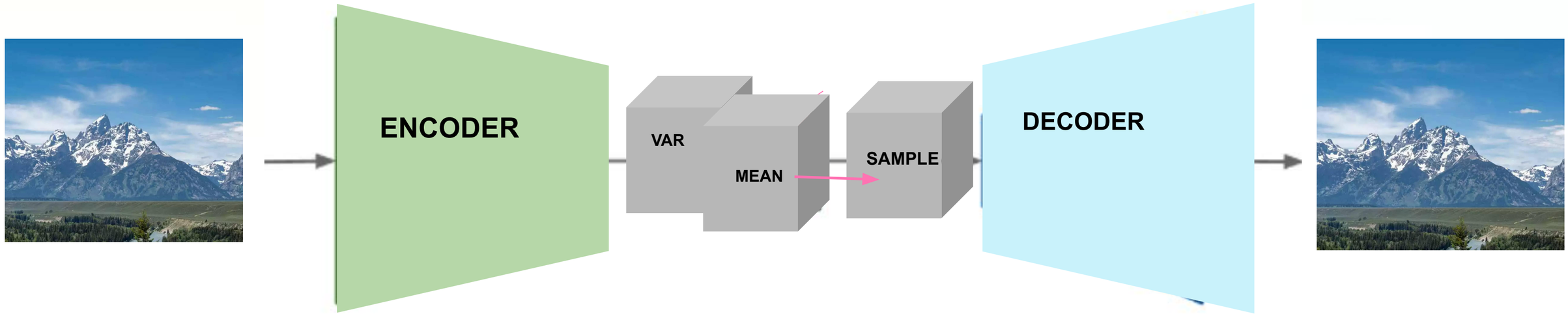
Lightricks

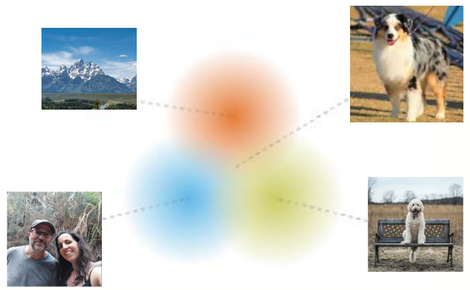
AutoEncoder



Lightricks

Variational AutoEncoder





Lightricks

Taming Transformers for High-Resolution Image Synthesis

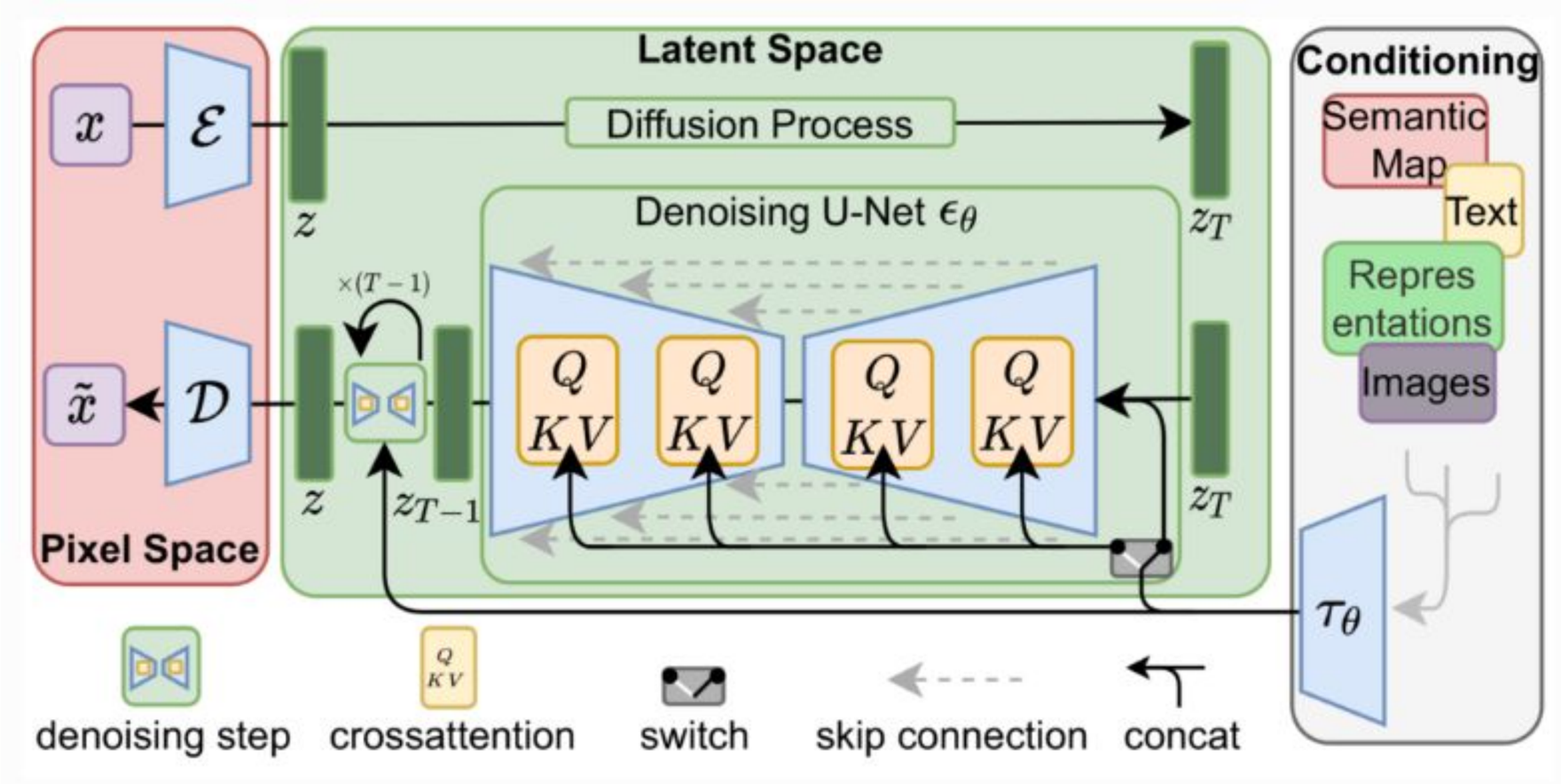
Patrick Esser*, Robin Rombach*, Björn Ommer

https://github.com/CompVis/taming-transformers

Lightricks

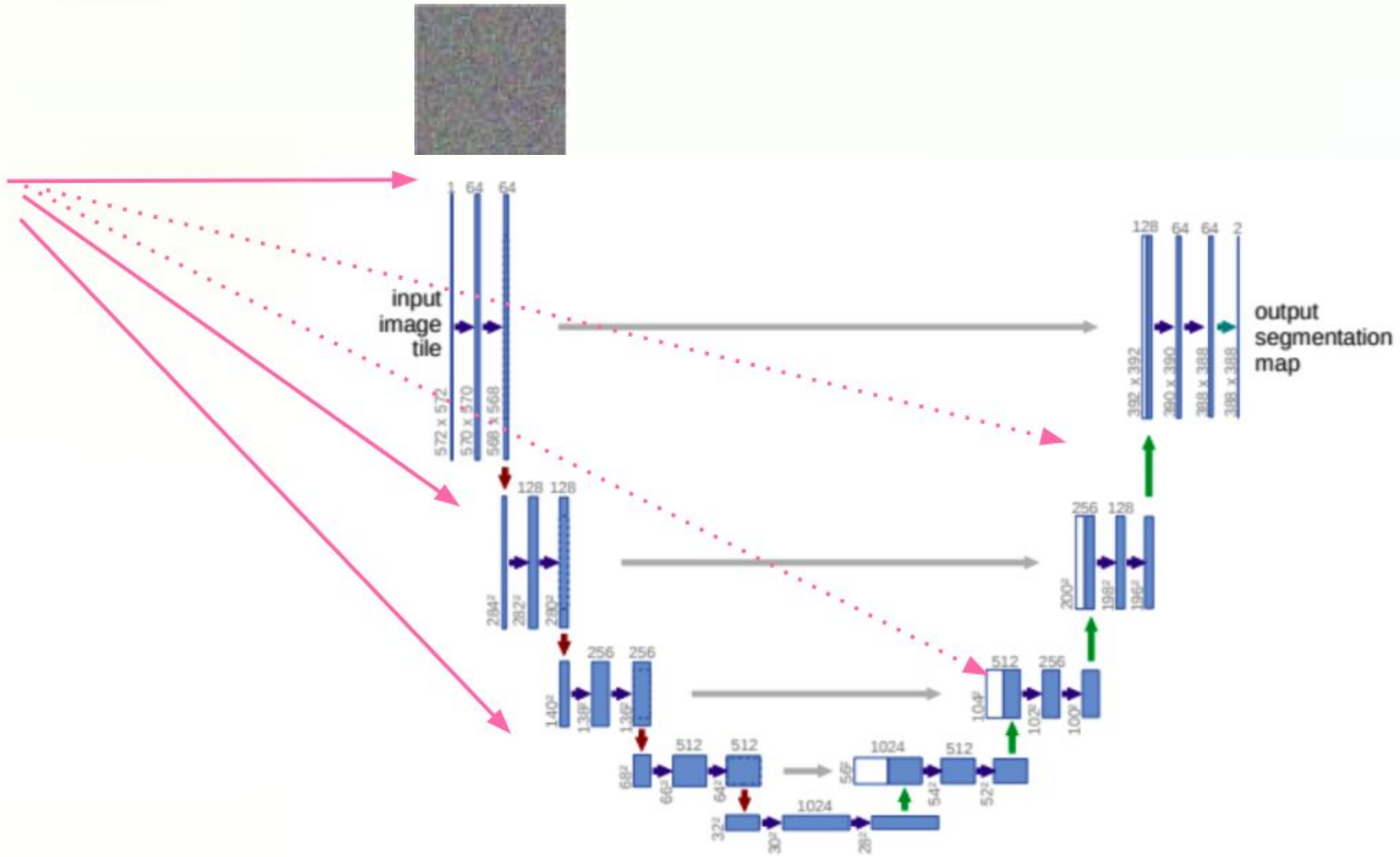
High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer

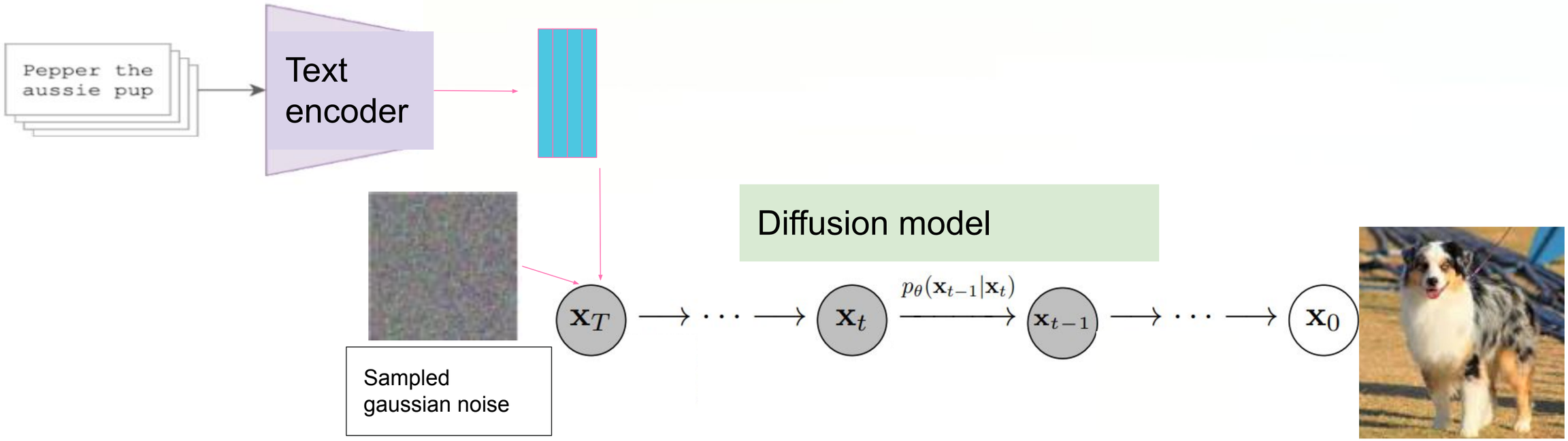


Lightricks

64X64X3 image | 64X64X4 latent representation

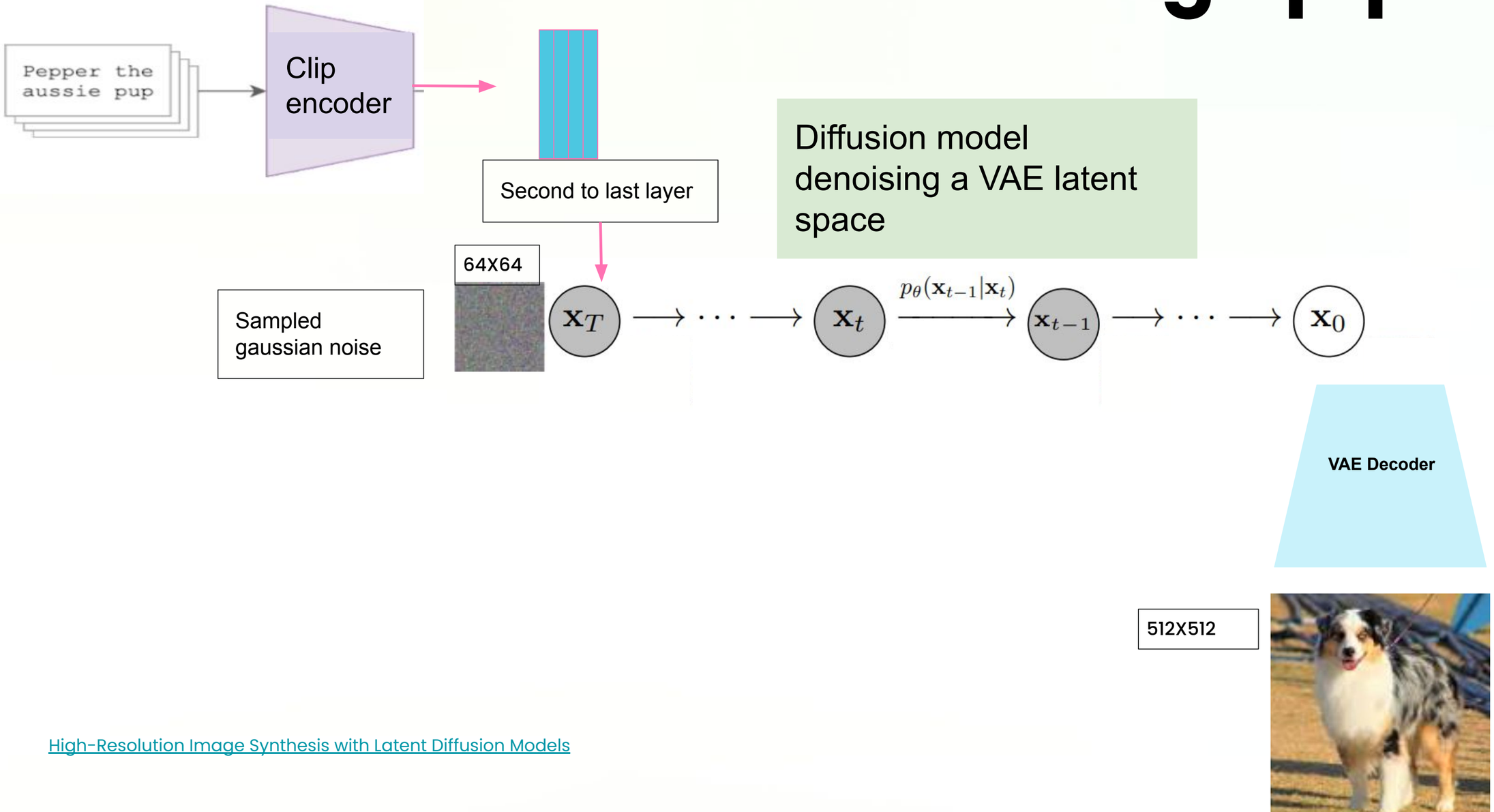


Original Text to Image Pipe



Lightricks

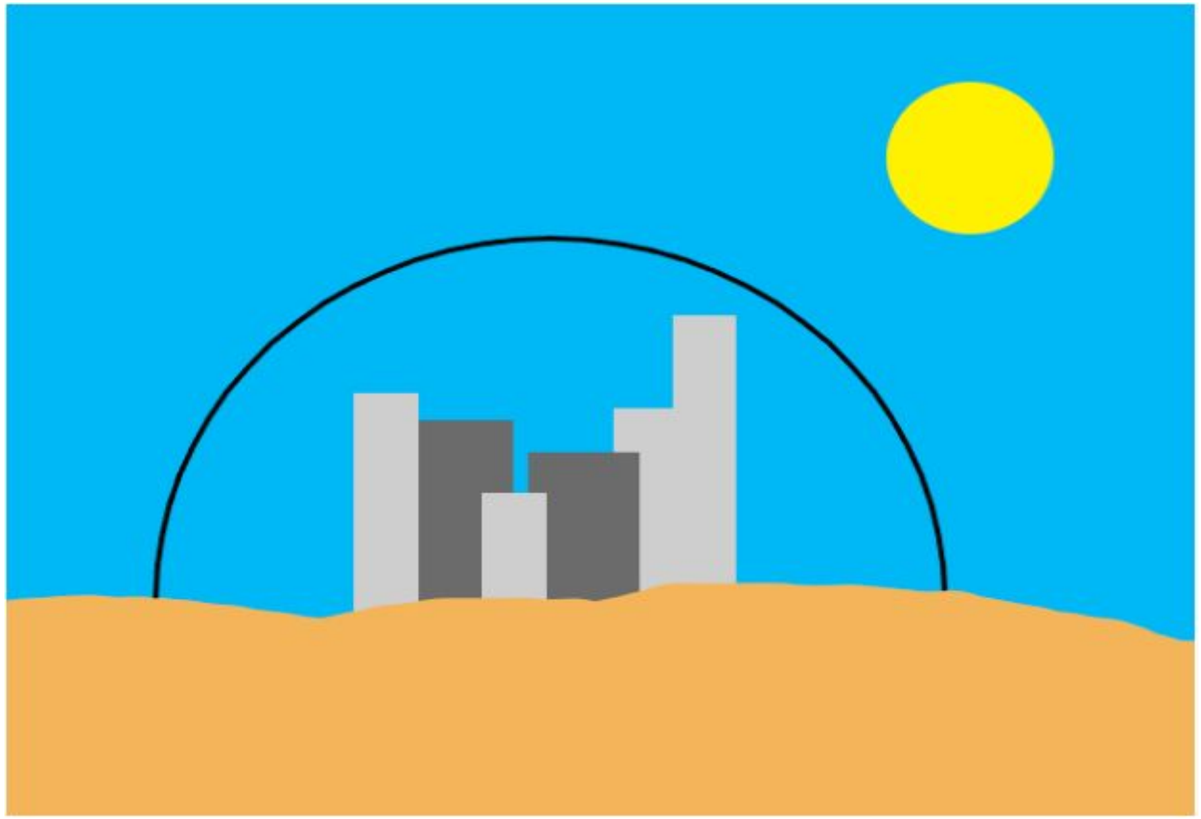
Stable Diffusion Text to Image pipe



[High-Resolution Image Synthesis with Latent Diffusion Models]

Lightricks

Part 3 - Controlling Text V=Condition Diffusion Models

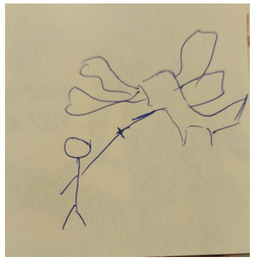


Lightricks

Image to Image pipe

A hero fighting a fire-breathing dragon

Clip encoder

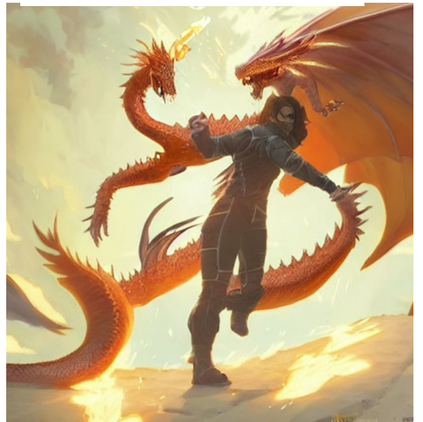


Encoder $q_\phi(\mathbf{z}|\mathbf{x})$

Diffusion model denoising a VAE latent space

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

VAE Decoder



[High-Resolution Image Synthesis with Latent Diffusion Models]

Lightricks

Inpainting pipe (legacy)

A magic castle on the mountains

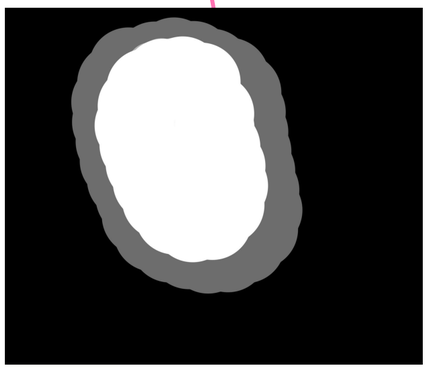
Clip encoder

Diffusion model denoising a VAE latent space

Encoder $q_\phi(\mathbf{z}|\mathbf{x})$

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

VAE Decoder



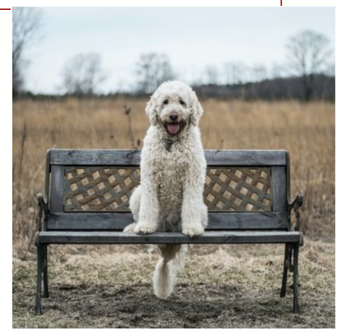
[High-Resolution Image Synthesis with Latent Diffusion Models]

Lightricks

Inpainting pipe

Face of a yellow cat, high resolution, sitting on a park bench

Clip encoder

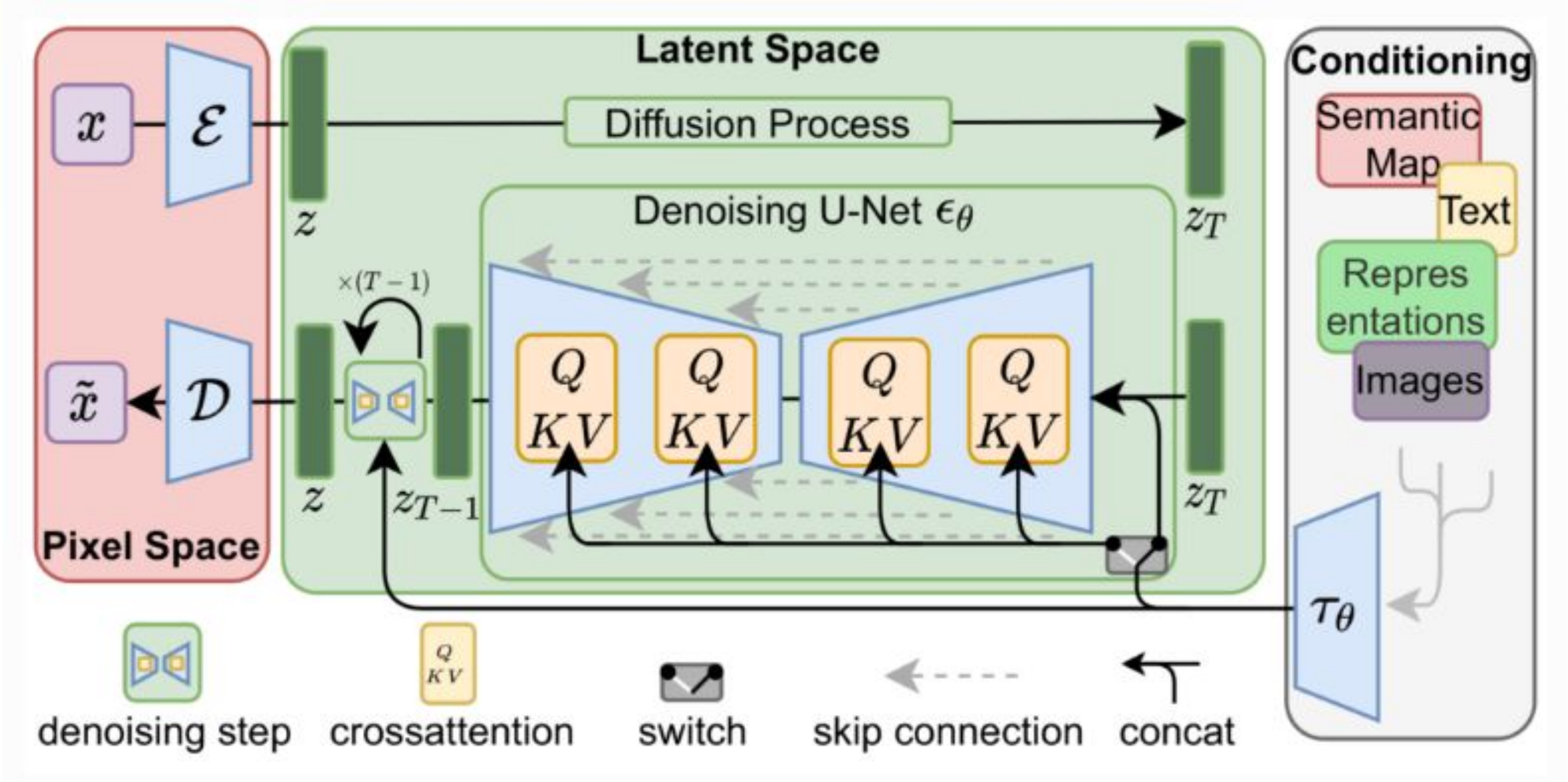


Encoder $q_\phi(\mathbf{z}|\mathbf{x})$

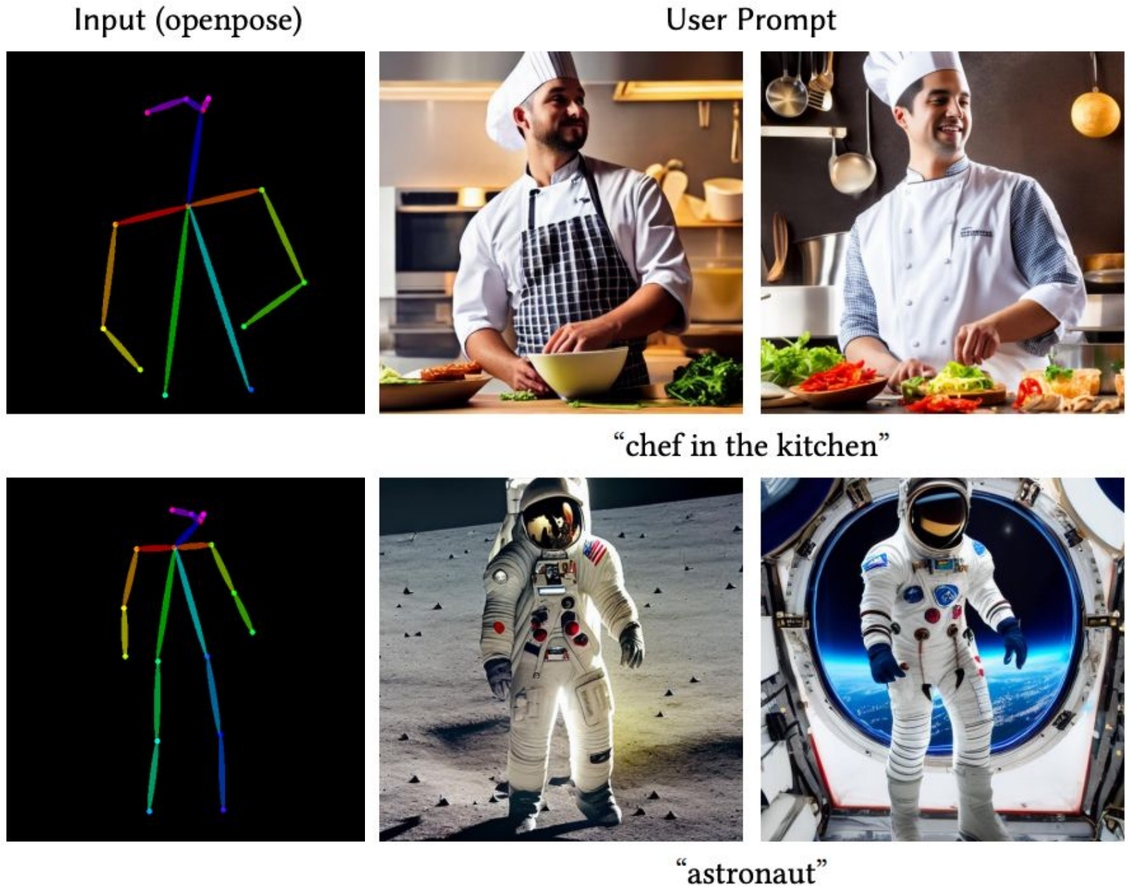
Inpainting finetuned diffusion model denoising a VAE latent space

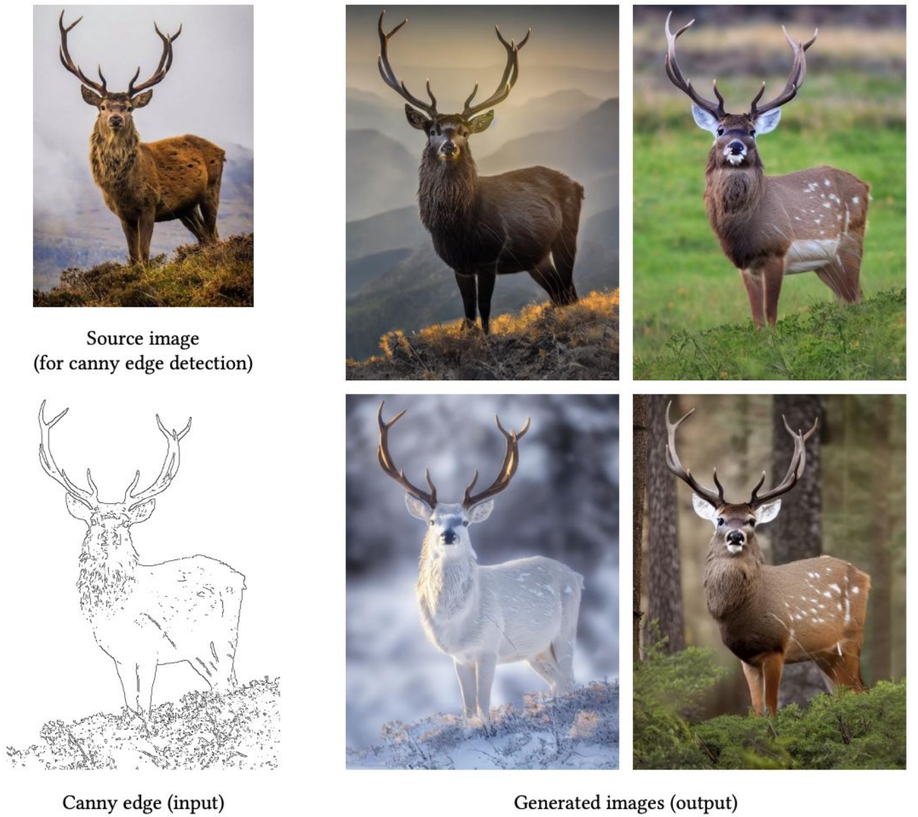
$\mathbf{x}_T \longrightarrow \cdots \longrightarrow \mathbf{x}_t \xrightarrow{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \mathbf{x}_{t-1} \longrightarrow \cdots \longrightarrow \mathbf{x}_0$

VAE Decoder


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

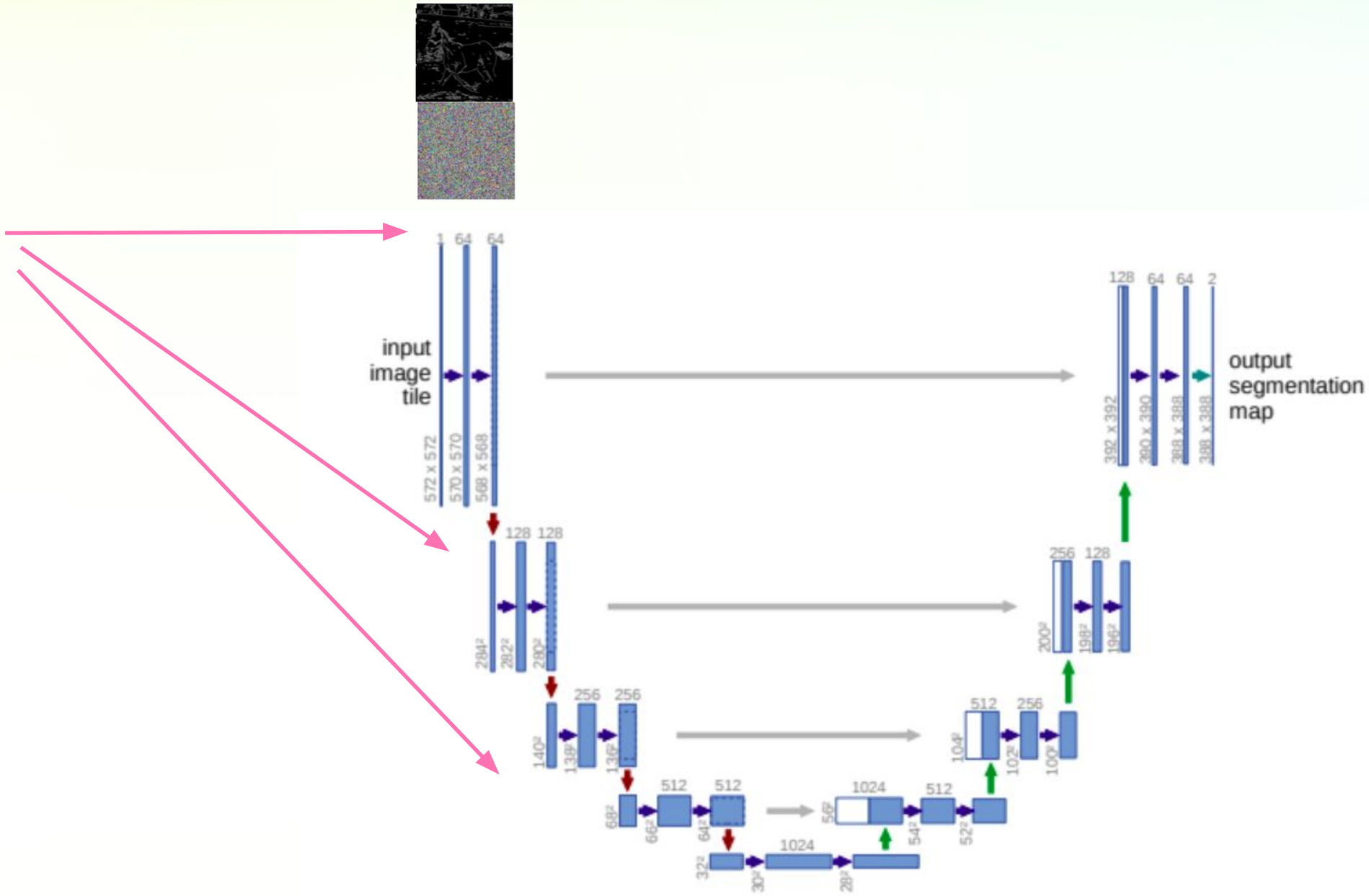
ControlNet Enhance pretrained image diffusion models with task-specific conditions



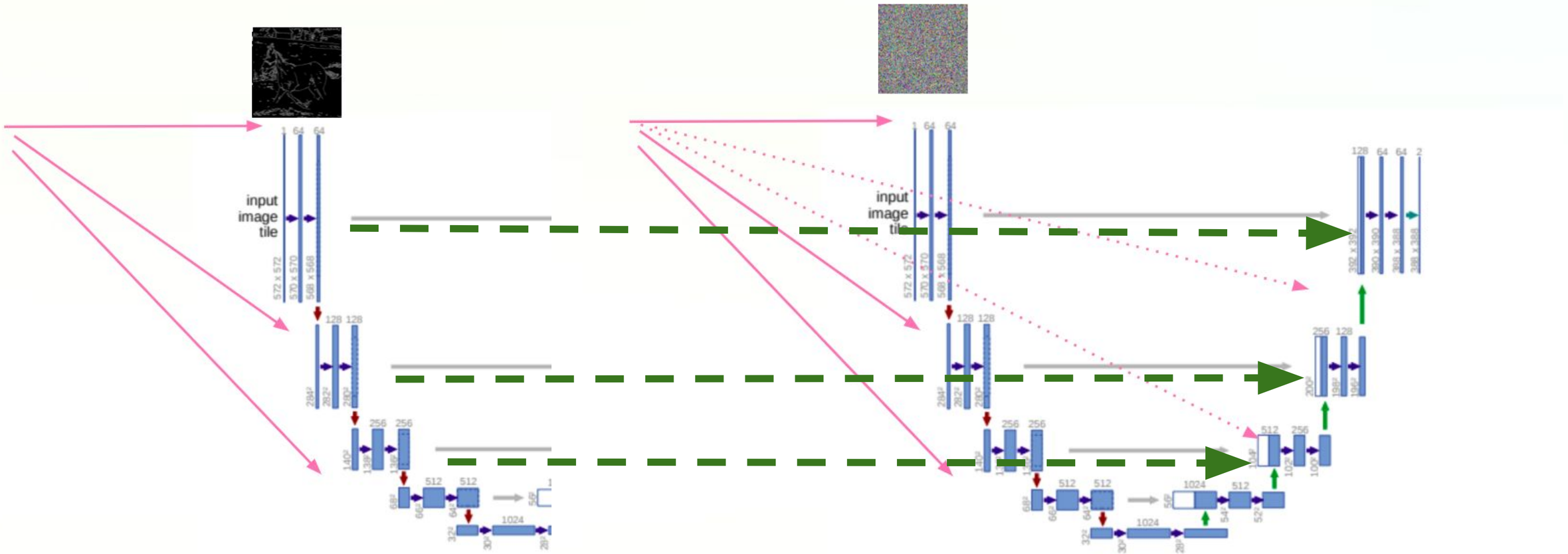


[Adding Conditional Control to Text-to-Image Diffusion Models]

Lightricks

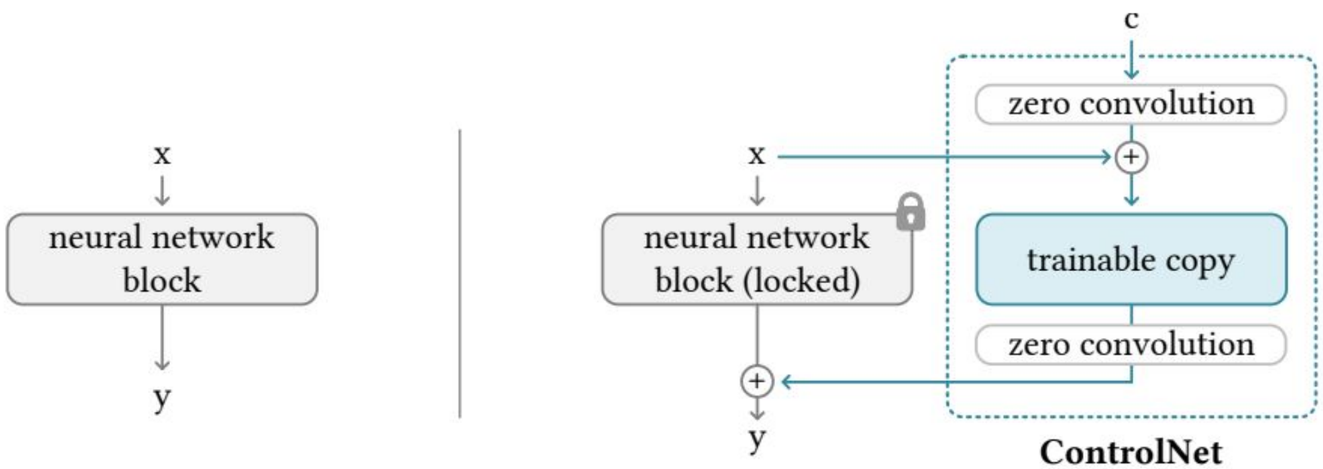


Lightricks


input image tile
input image tile
Trainable copy
Freezed copy

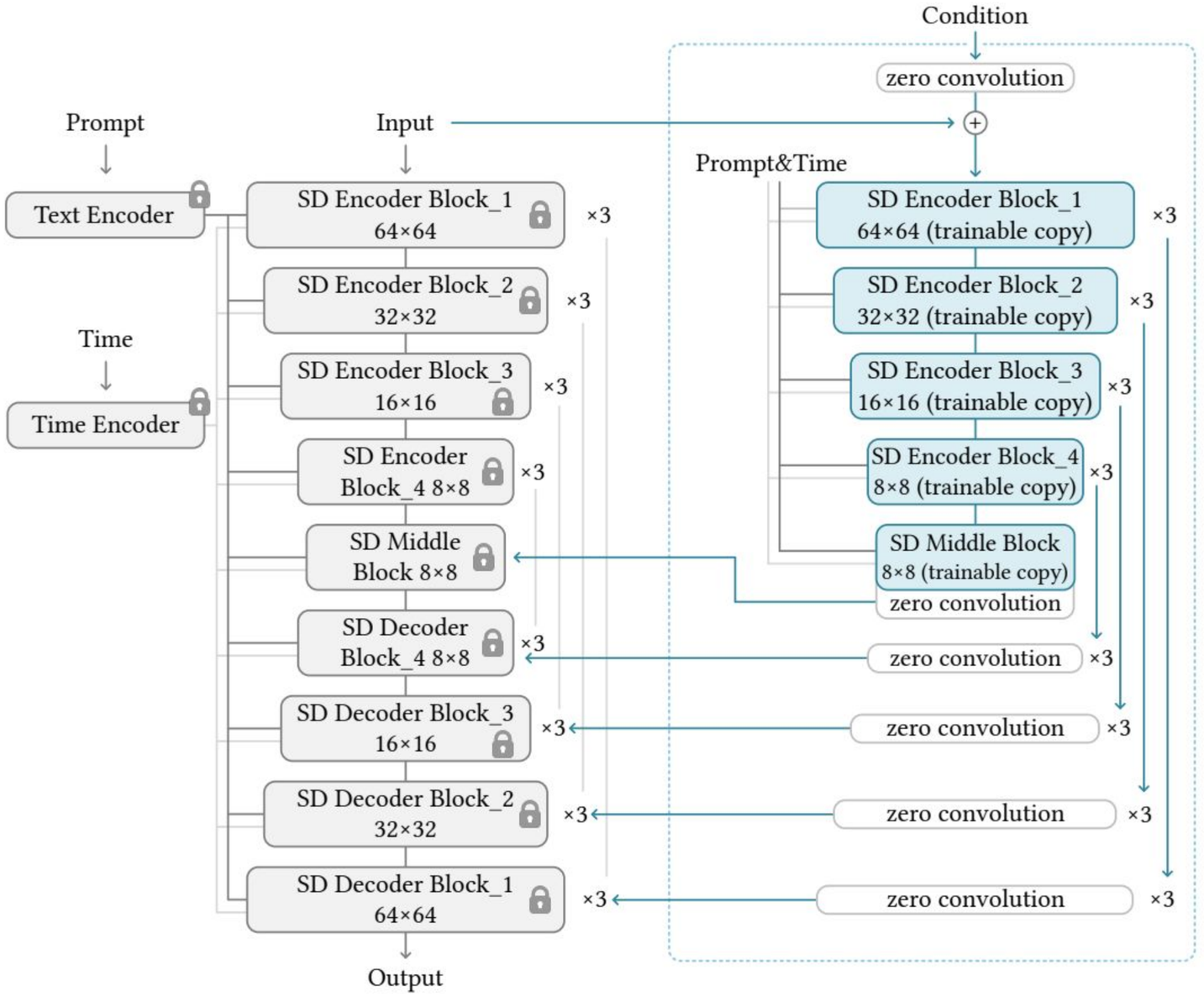
ControlNet

*zero convolution



(a) Before

(b) After



(a) Stable Diffusion (b) ControlNet

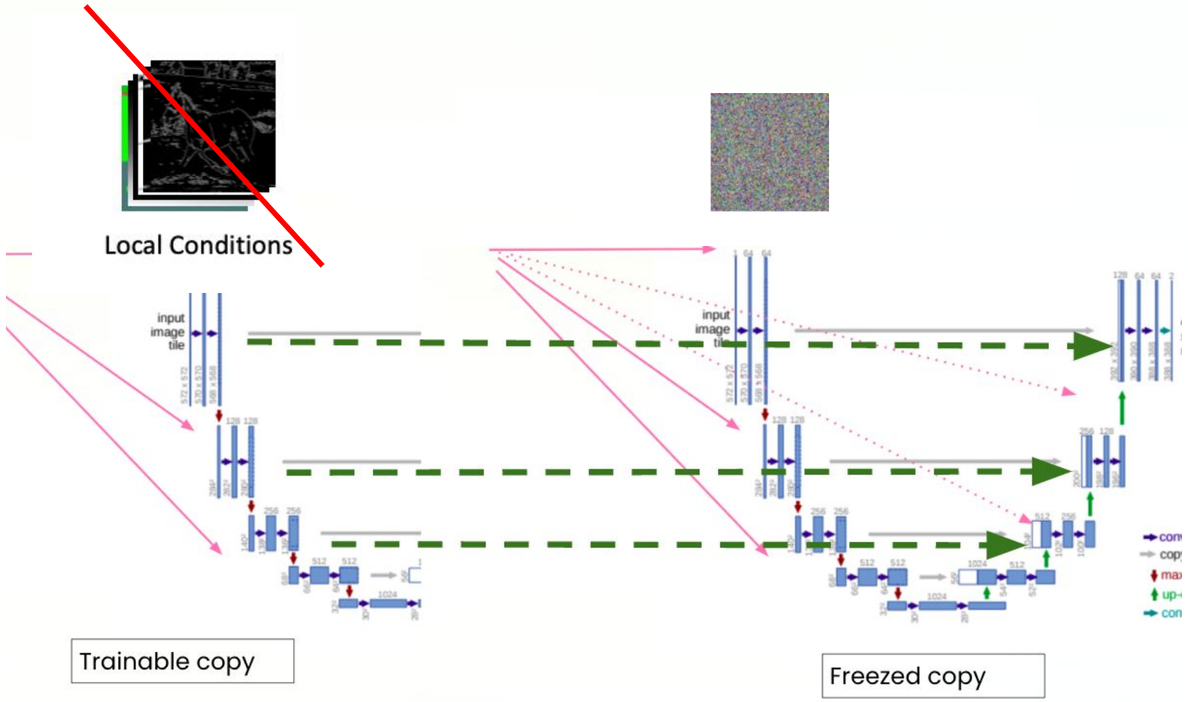
It is all about the control

Uni-ControlNet: All-in-One Control

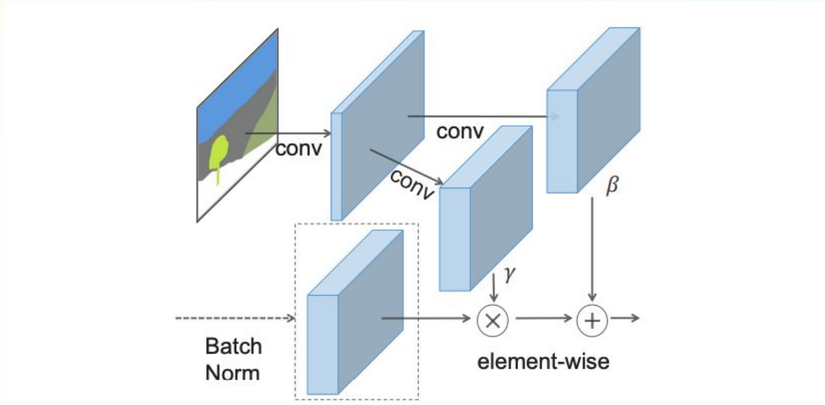
- Local and Global condition adapters
- Global adapter - via text tokens

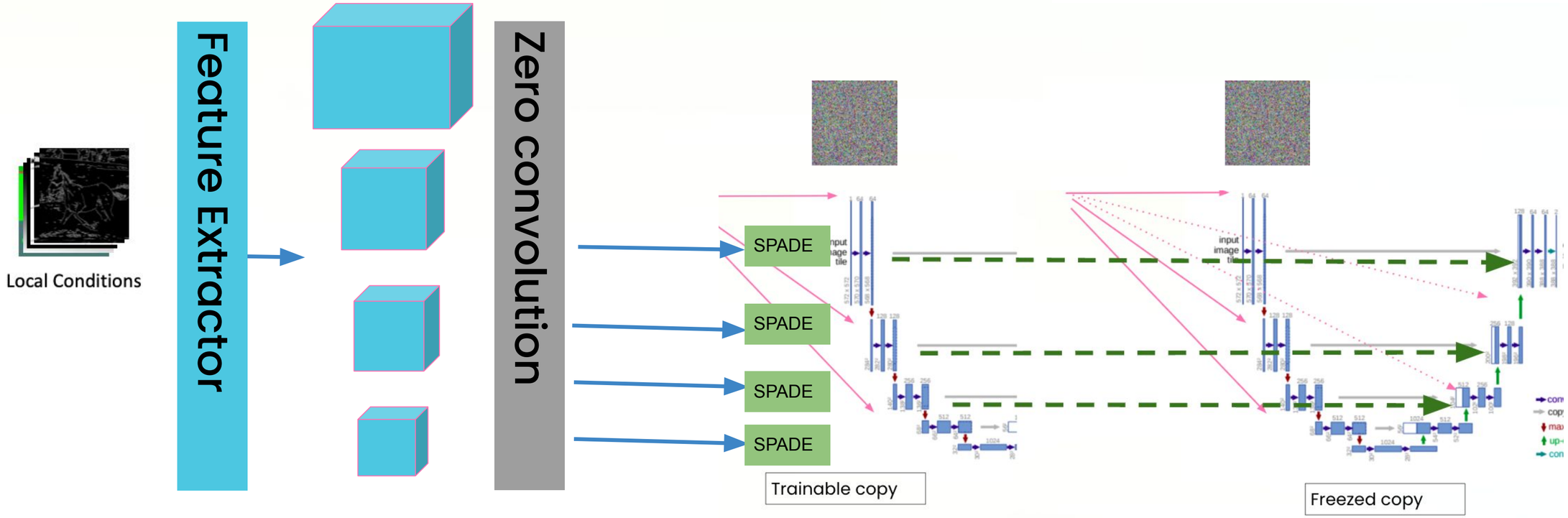
	Fine-tuning	Composable Control	Fine-tuning Cost	Adapter Number
ControlNet	✓	✗	N	N
Uni-ControlNet (Ours)	✓	✓	2	2

Local Control Adapter



Local Control Adapter





Condition-1 Condition-2 Ours

Useful gits to follow

https://github.com/huggingface/diffusers

https://github.com/AUTOMATIC1111/stable-diffusion-webui

Good image 2 image example:

https://colab.research.google.com/github/patil-suraj/Notebooks/blob/master/image_2_image_using_diffusers.ipynb

Lightricks

Questions?

Takeaway

1. Industry will usually push the boundaries of images quality. A lot of expensive big experiments, academy will add the option to get control on the output.

2. Consolidity - today computer vision domain and NLP domain are not separated anymore, each one inherits ideas for the other

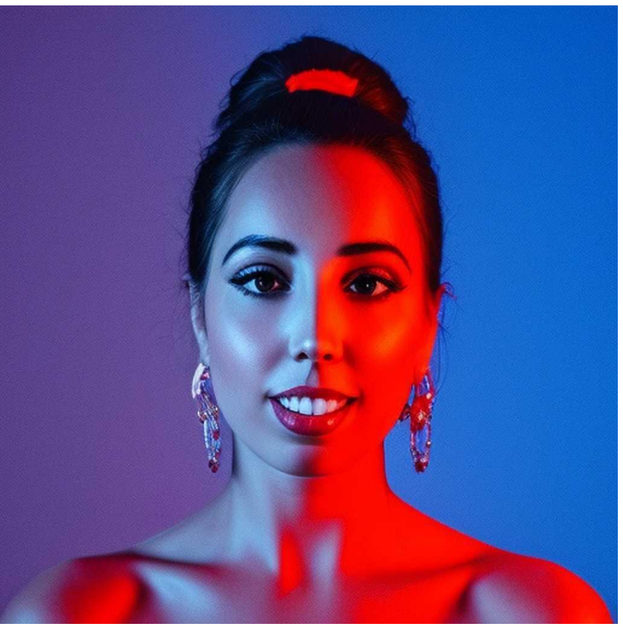
3. We covered a lot of architectures, and technical ideas. Combining those existing ideas leads to great progress

Lightricks

Thank you!

Lightricks

Lets Leverage the control




DreamBooth

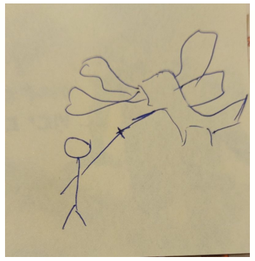
DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

Lightricks

Image to Image pipe

A hero fighting a fire-breathing dragon

Clip encoder

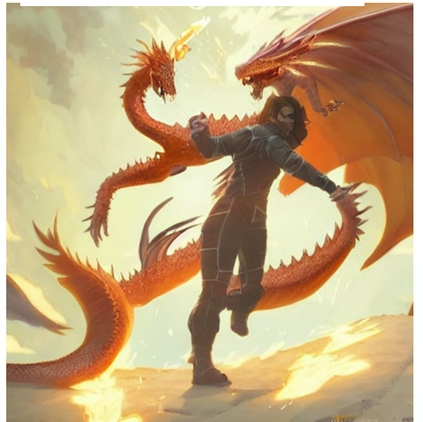


Encoder
$q_\phi(\mathbf{z}|\mathbf{x})$

Diffusion model denoising a VAE latent space

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

VAE Decoder



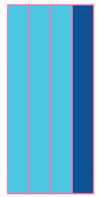
[High-Resolution Image Synthesis with Latent Diffusion Models]

Lightricks

Dreambooth Training

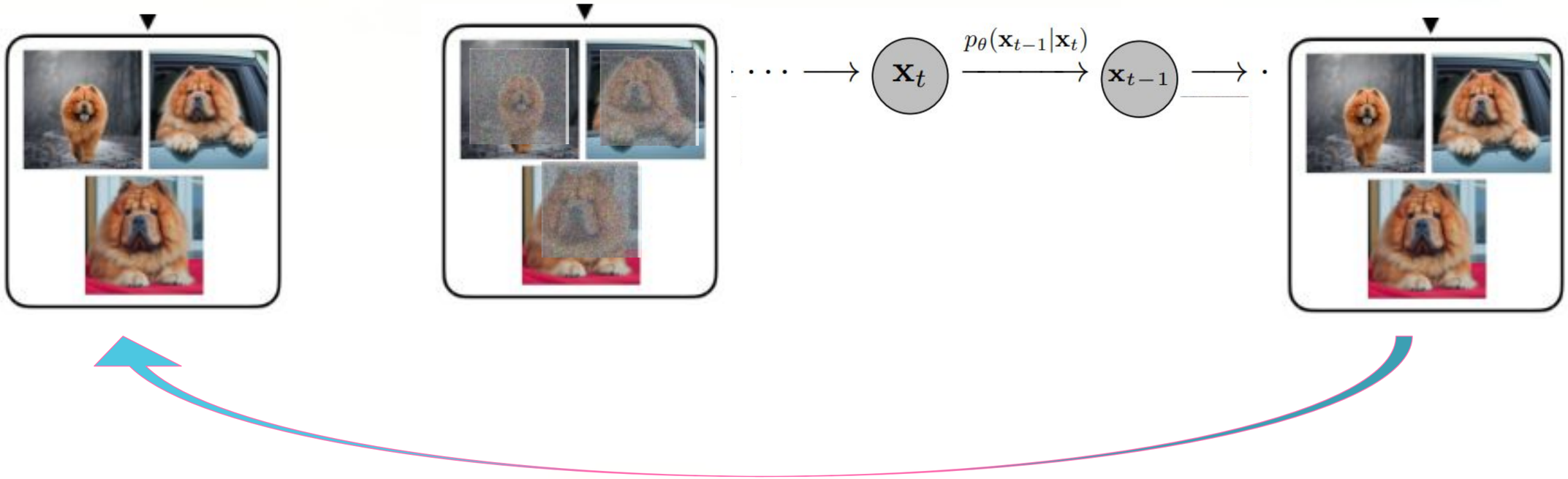
A <V> dog

Clip encoder



1 training step

Predicted start



[High-Resolution Image Synthesis with Latent Diffusion Models]

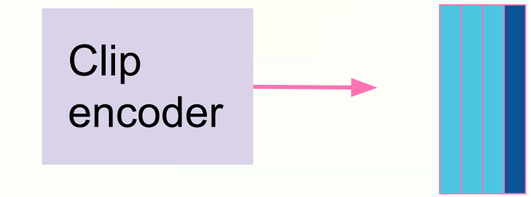
L2 Loss

Lightricks

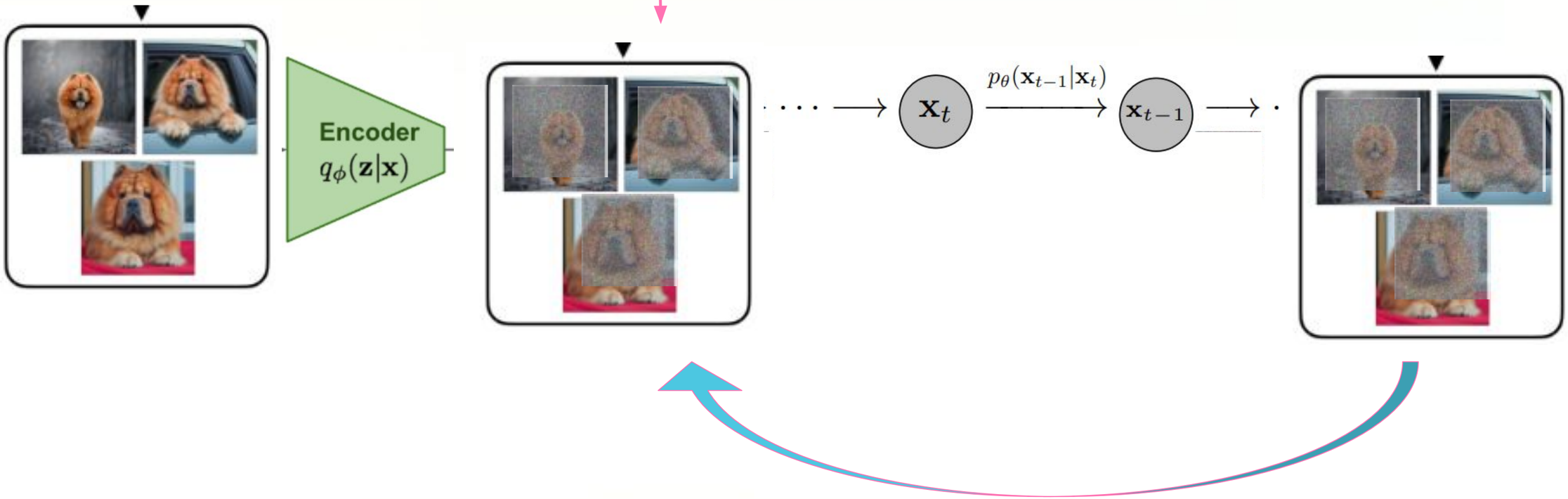
Dreambooth Training

A <V> dog

Clip encoder



1 training step

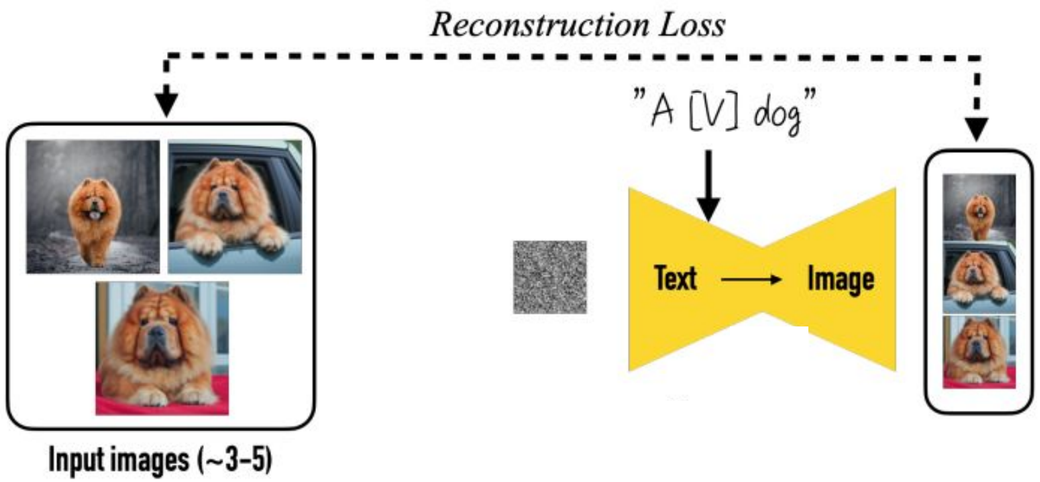


L2 Loss

[High-Resolution Image Synthesis with Latent Diffusion Models]

DreamBooth

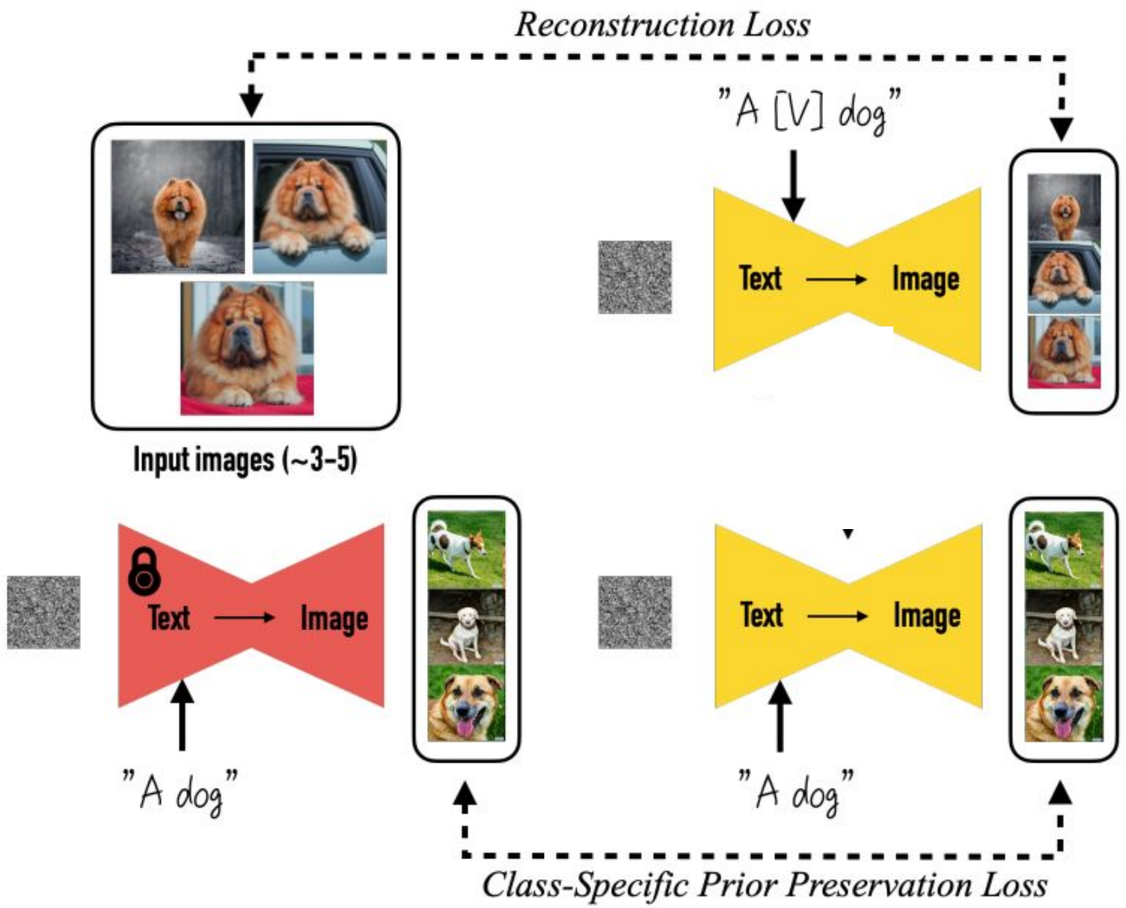
- Rare text token
- Prior Preservation loss



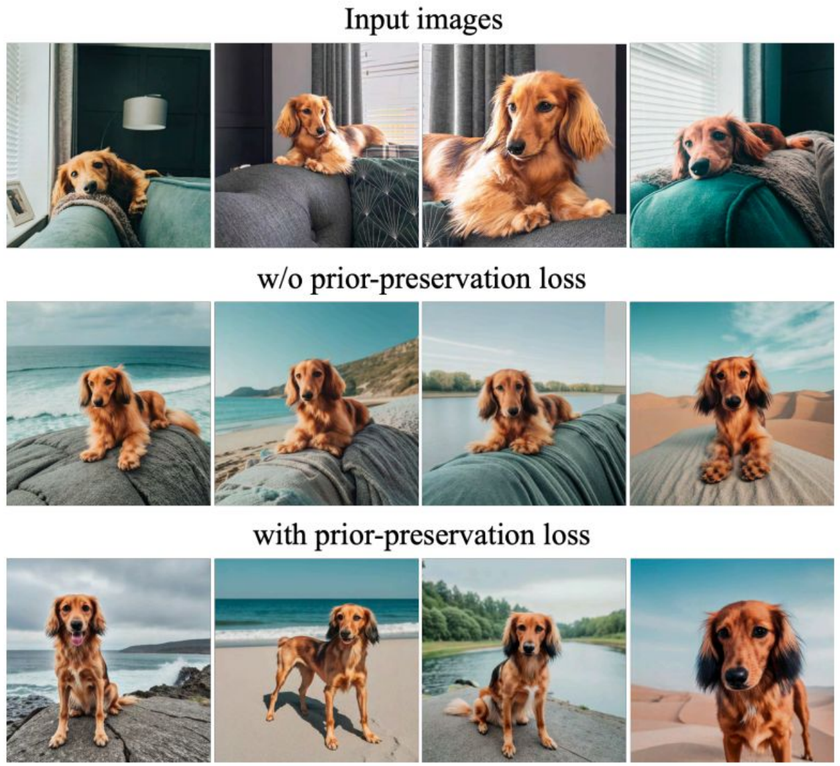
DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

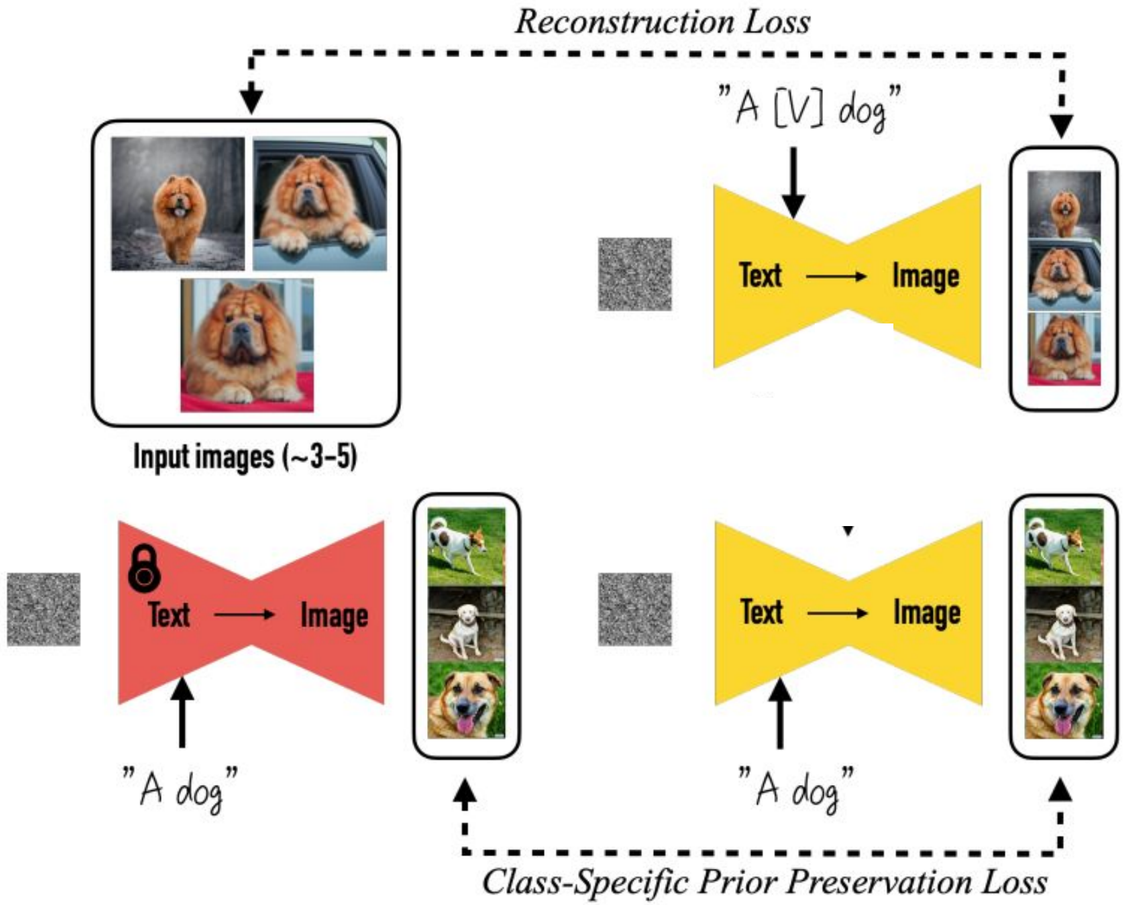
DreamBooth

- Rare text token
- Prior Preservation loss



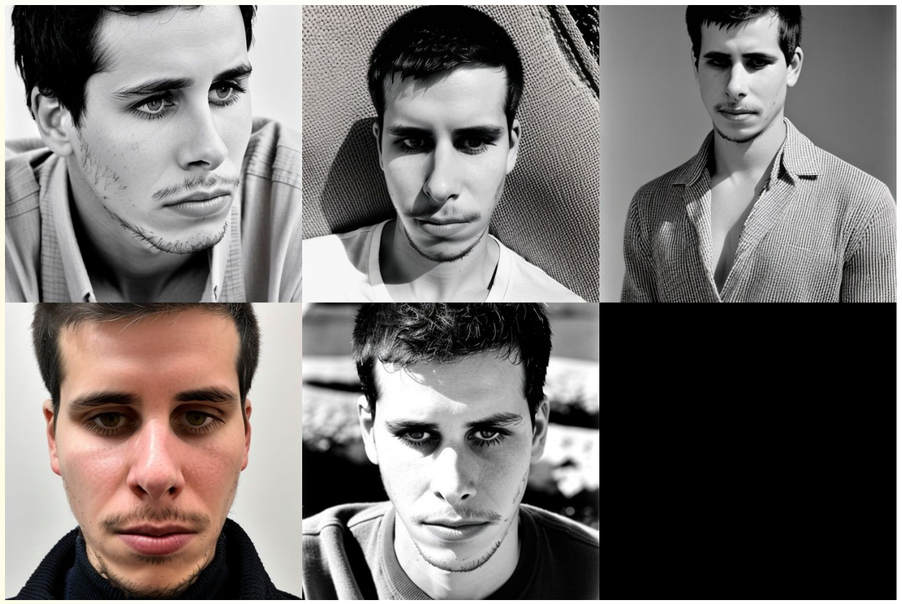
DreamBooth


Input images
w/o prior-preservation loss
with prior-preservation loss


Reconstruction Loss
"A [V] dog"
Text
Image
Input images (~3-5)
Text
Image
"A dog"
Text
Image
"A dog"
Class-Specific Prior Preservation Loss

Lightricks

DreamBooth In practice



Many Conditions in One Model

Uni-ControlNet 10M text-image pairs
Evaluation: fid: 5k

Table 2: FID on different controllable diffusion models. The best results are in **bold**.

	Canny	MLSD	HED	Sketch	Openpose	Depth	Segmentation	Style\Content
ControlNet	18.90	31.36	26.59	22.19	27.84	21.25	**23.08**	31.17
T2I-Adapter	18.98	-	-	**18.83**	29.57	21.35	23.84	28.86
Ours	**17.79**	**26.18**	**17.86**	20.11	**26.61**	**21.20**	23.40	**23.98**

Composer
64X64 (2B) batch size - 1024 1M steps pretrain
Finetune 200K steps on 60M samples
= 1B
256X256 (1.1B) batch size 512 1M steps
1024X1024 (1B) batch size 512 1M steps
Data of 1B
Evaluation:

Composer achieves a zero-shot FID of 9.2 in text-to-image synthesis on the COCO dataset