Image Synthesis – COMP4423 Computer Vision

Xiaoyong Wei (魏驍勇)

x1wei@polyu.edu.hk

Department of Computing
電子計算學系

The Hong Kong Polytechnic University
香港理工大學

Opening Minds • Shaping the Future
啟迪思維 • 成就未來

A ChatGPT point of view

>1. Introduction to Generative Models

>2. Generative Adversarial Networks (GANs)

>3. Variational Autoencoders (VAEs)

>4. Autoregressive Models

>5. Recent Trends: Energy-Based GANs (EBGANs), Adversarial VAEs, etc. / image super-resolution, style transfer, etc. / open problems and future directions for research

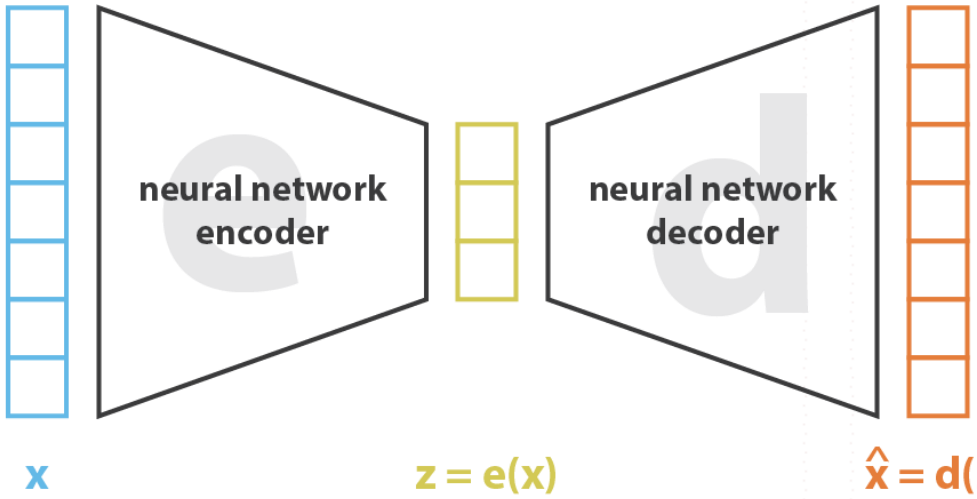
>6. Conclusion

This is just FYI.

The training data cutoff date of ChatGPT and its variants / competitors is Sep. 2021.

Let's briefly review the "history" of image synthesis (on VAEs and GANs) and move to its recent trend (on diffusion models).

Variational Auto-Encoder(VAE)



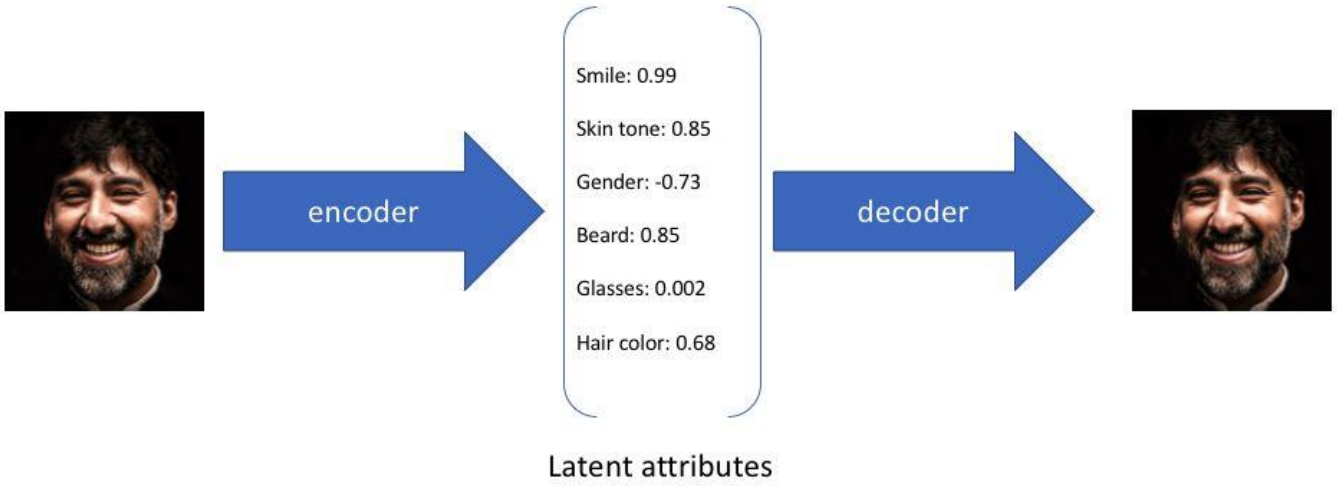
$$\text{loss} = \| x - \hat{x} \|^2 = \| x - d(z) \|^2 = \| x - d(e(x)) \|^2$$

In the training step, the input data is compressed into a vector of low dimensional normal distribution by the encoder, and the vector is reconstructed into the input data by the decoder.

In the generation step, a vector of normal distribution is randomly sampled and input into the decoder to generate data.

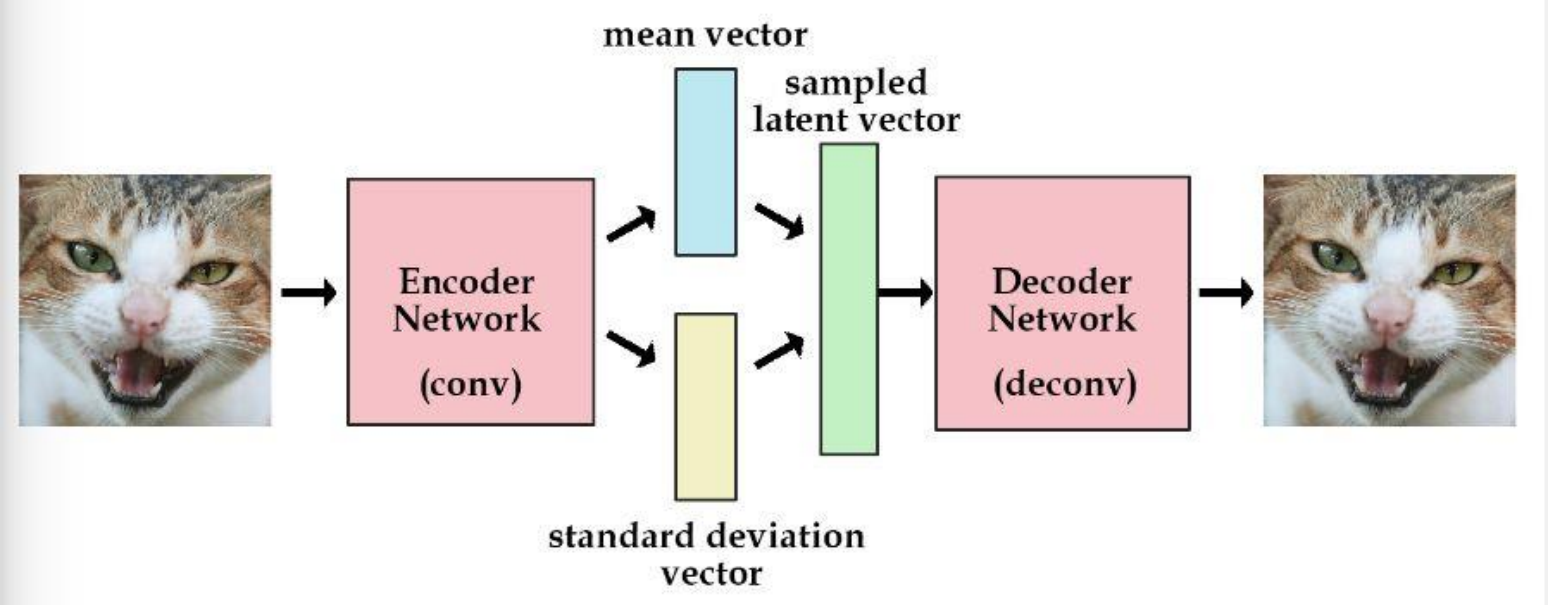
KINGMA D P, WELLING M. Auto-encoding variational bayes[Z]. 2014.
https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Variational Auto-Encoder(VAE)



KINGMA D P, WELLING M. Auto-encoding variational bayes[Z]. 2014.

Variational Auto-Encoder(VAE)



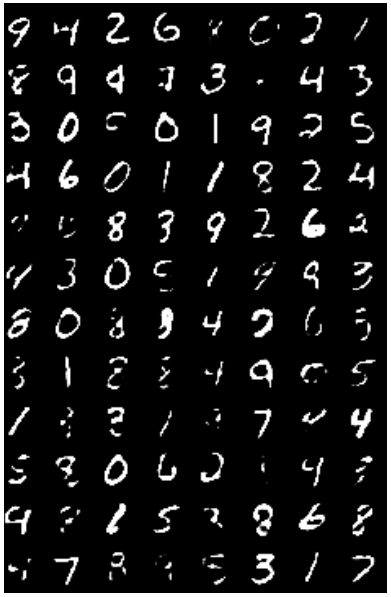
How to make the vector generated by the encoder follow the normal distribution is the core of VAE.

In fact, the output of VAE encoder is not a vector, but two values, which represent the mean and variance of a normal distribution, respectively. Once learned, the parameters can be used to modify a vector E that follows the standard normal distribution $N(0,1)$ by adding the mean values and multiplying by the variance. The modified vector is then used as the input of the decoder for generation.

This process is called reparameterization, which avoids the situation that the sampling is non differentiable (we cannot calculate the gradients in this case).

KINGMA D P, WELLING M. Auto-encoding variational bayes[Z]. 2014.

Variational Auto-Encoder(VAE)



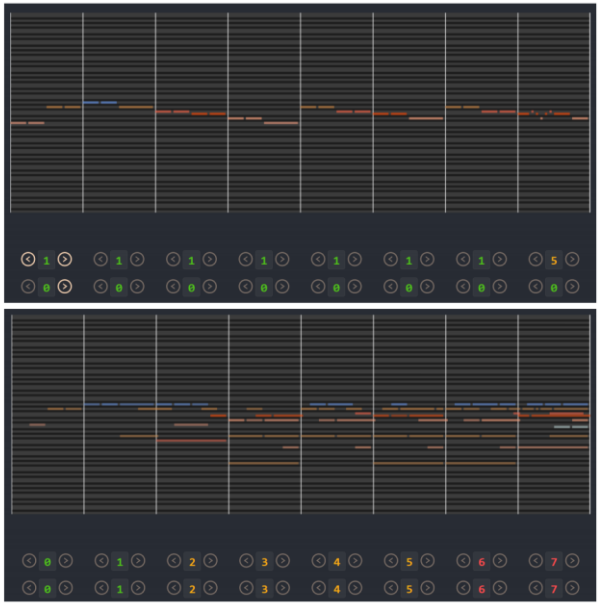


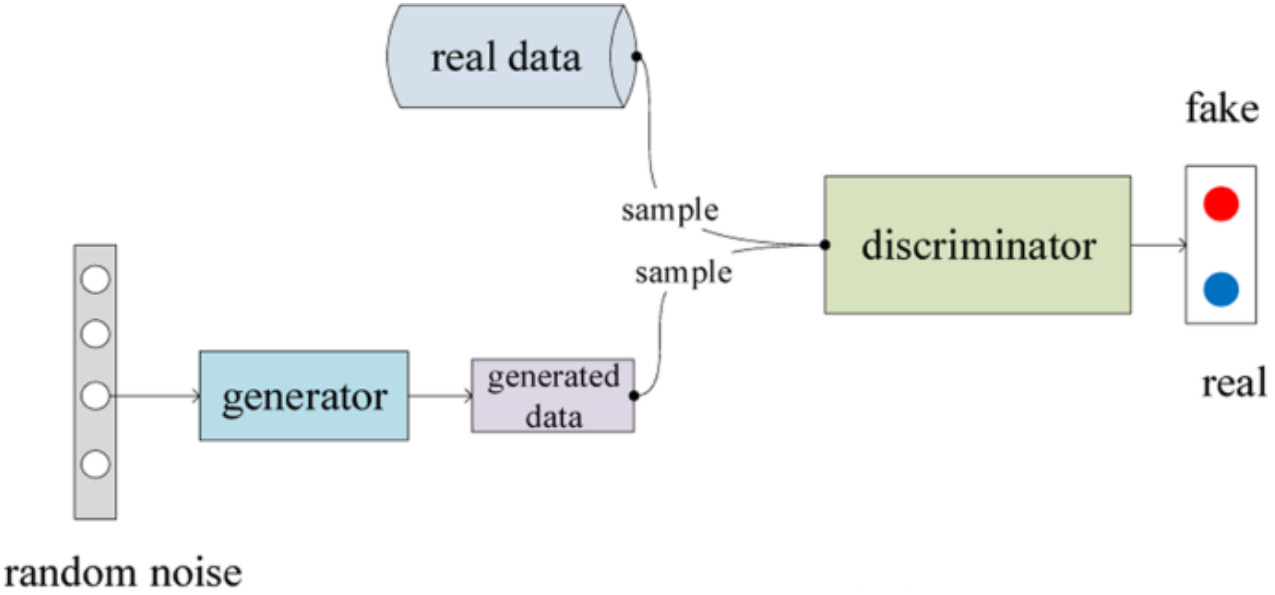
Image generation

Symbolic music generation

https://blog.csdn.net/ppp8300885/article/details/80070723
Muse Morphose: Full-Song and Fine-Grained Music Style Transfer with One Transformer VAE

VAE is not the only choice. Another alternative is GAN.

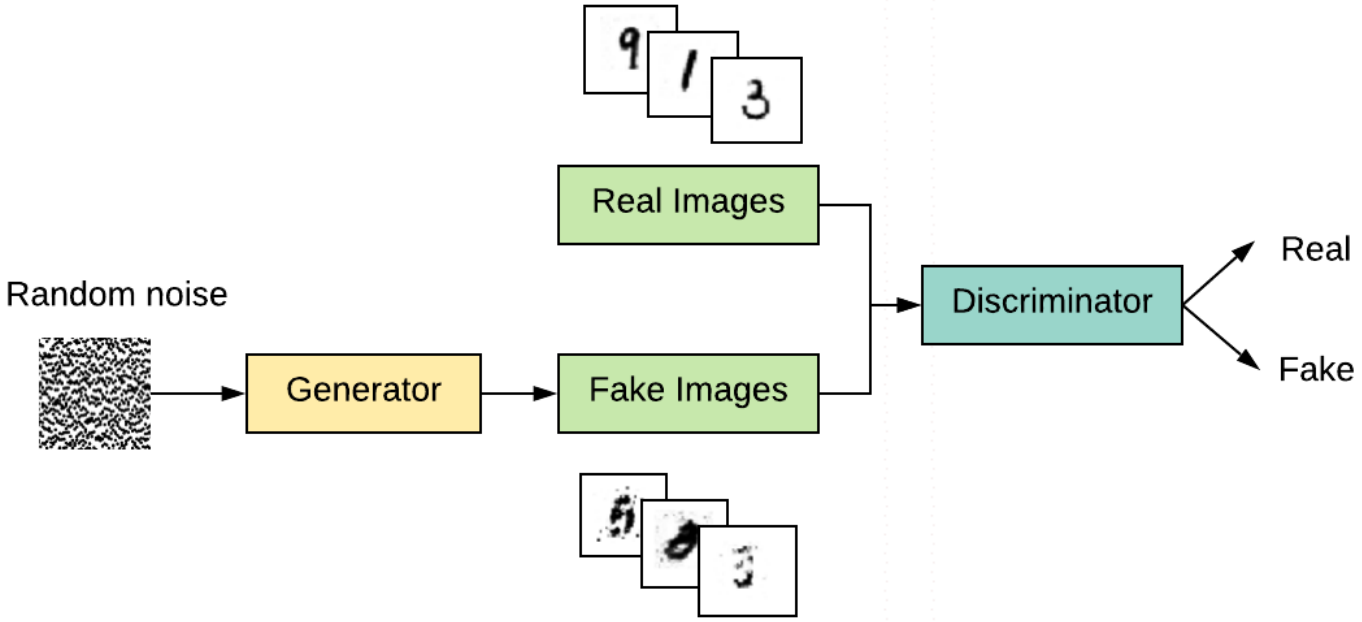
Generative Adversarial Networks(GAN)



Source: Electrocardiogram generation with a bidirectional LSTM-CNN generative adversarial network

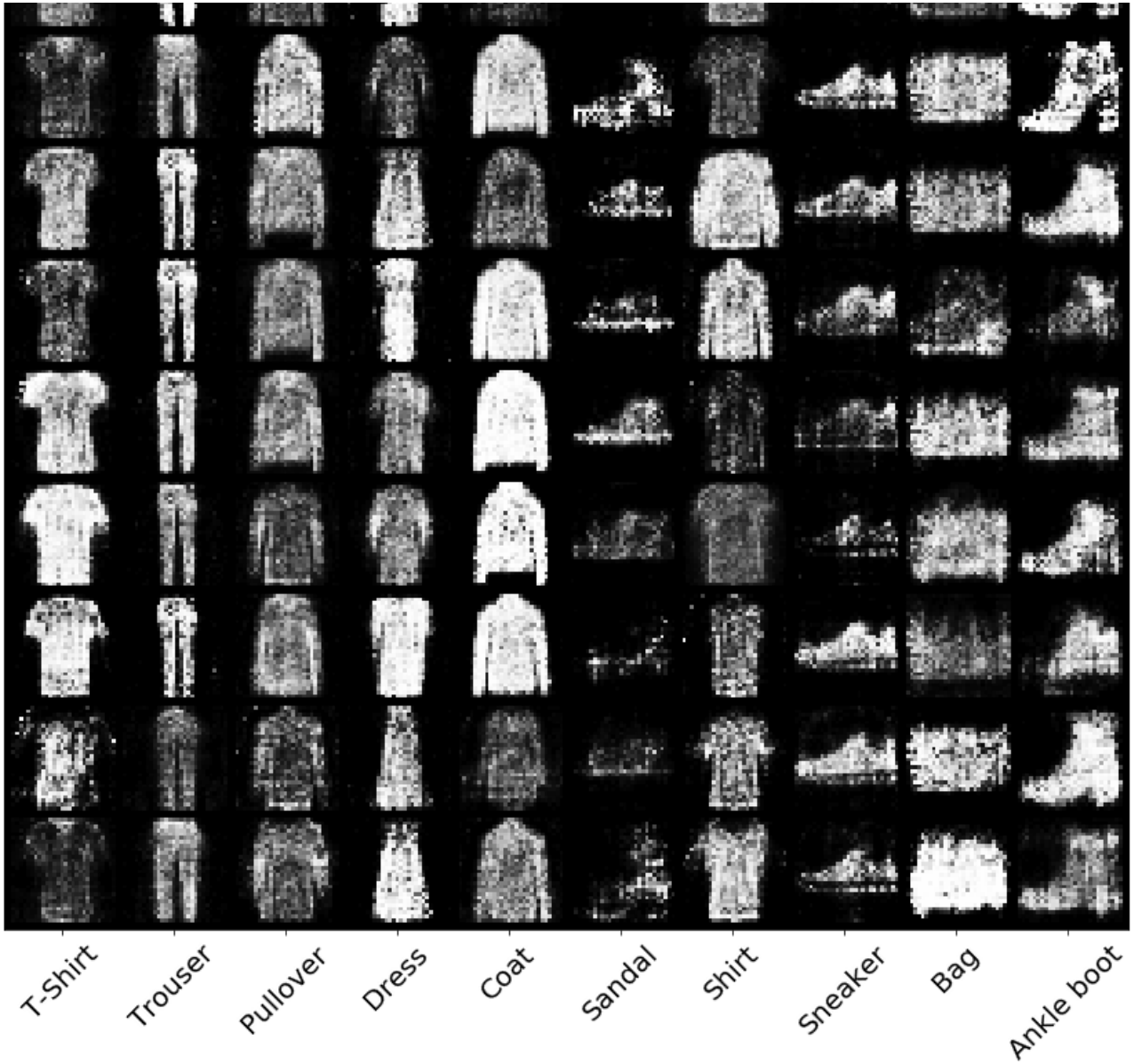
> GAN is an unsupervised generation model. It consists of two parts of generator and discriminator. The two parts are “opponents” to each other, in the way the generator tries its best to generate the “fake” data to “fool” the discriminator, while the discriminator tries its best to avoid being fooled. They make each other stronger through fighting.

Generative Adversarial Networks(GAN)



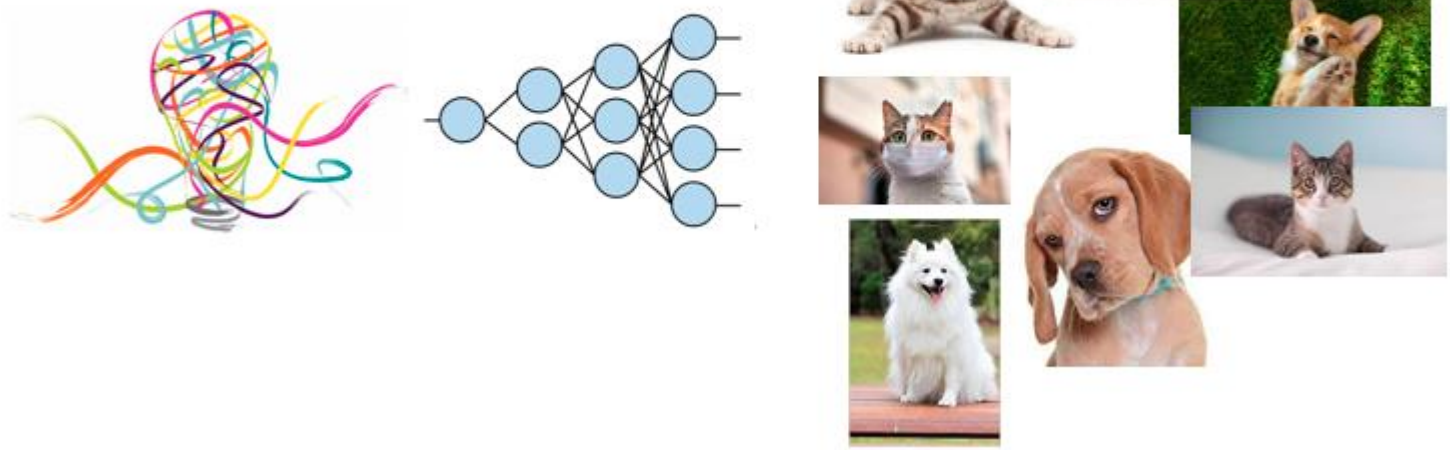
Taking MNIST dataset as an example. We can input a noise into G to generate a fake image, and then D judges whether the image is fake or not. If it's being recognized as a fake image, the fooling failed, which means G is not good enough and its parameters need to be updated. Otherwise, D's is not smart enough and the parameters of D need to be updated.

Project


T-Shirt
Trouser
Pullover
Dress
Coat
Sandal
Shirt
Sneaker
Bag
Ankle boot

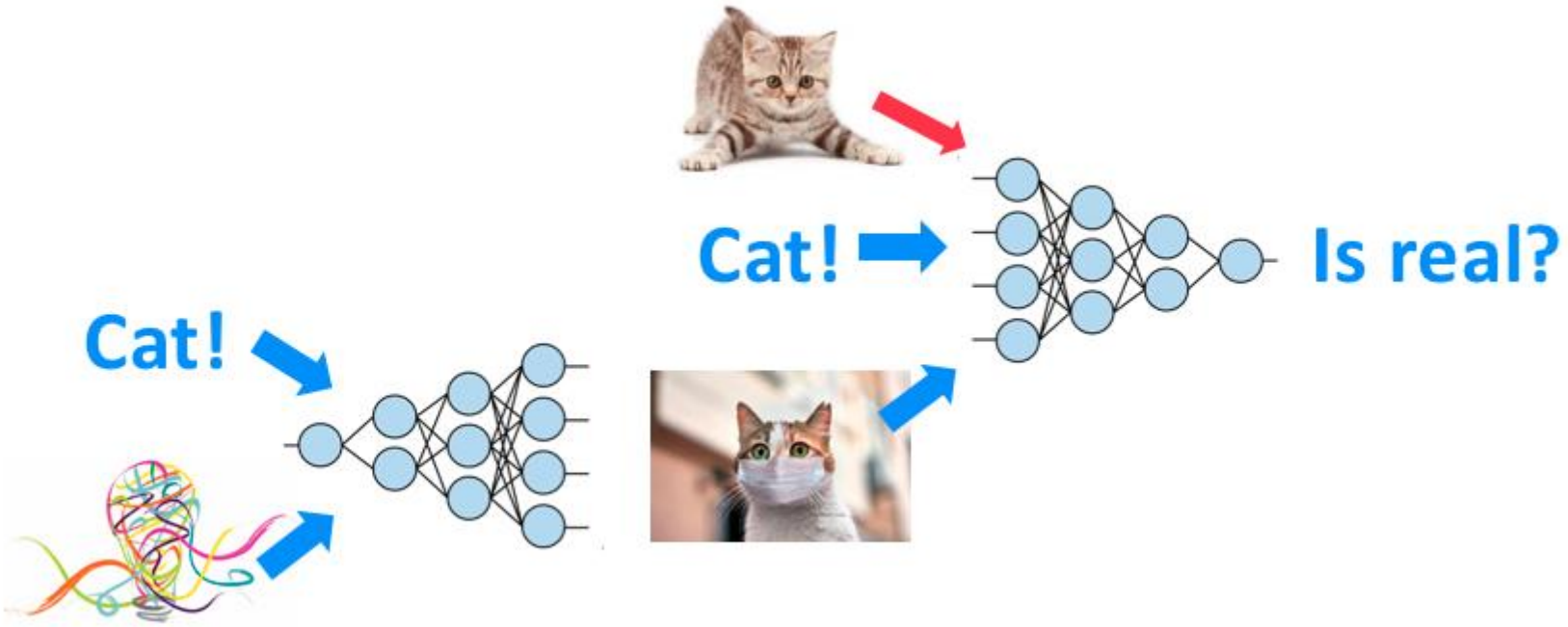
We need to control the generation.

GAN

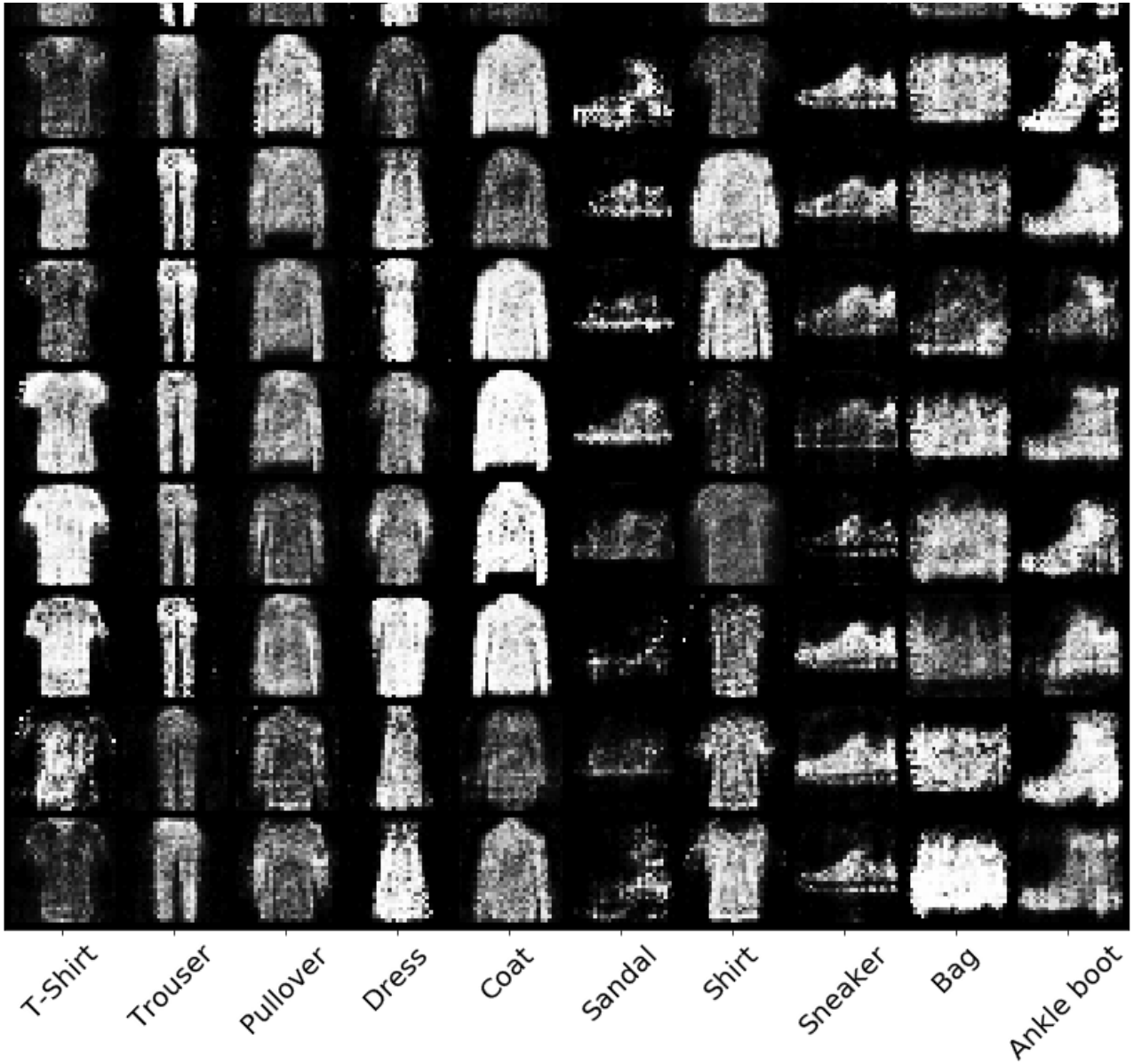


https://mofanpy.com/tutorials/machine-learning/gan/cgan

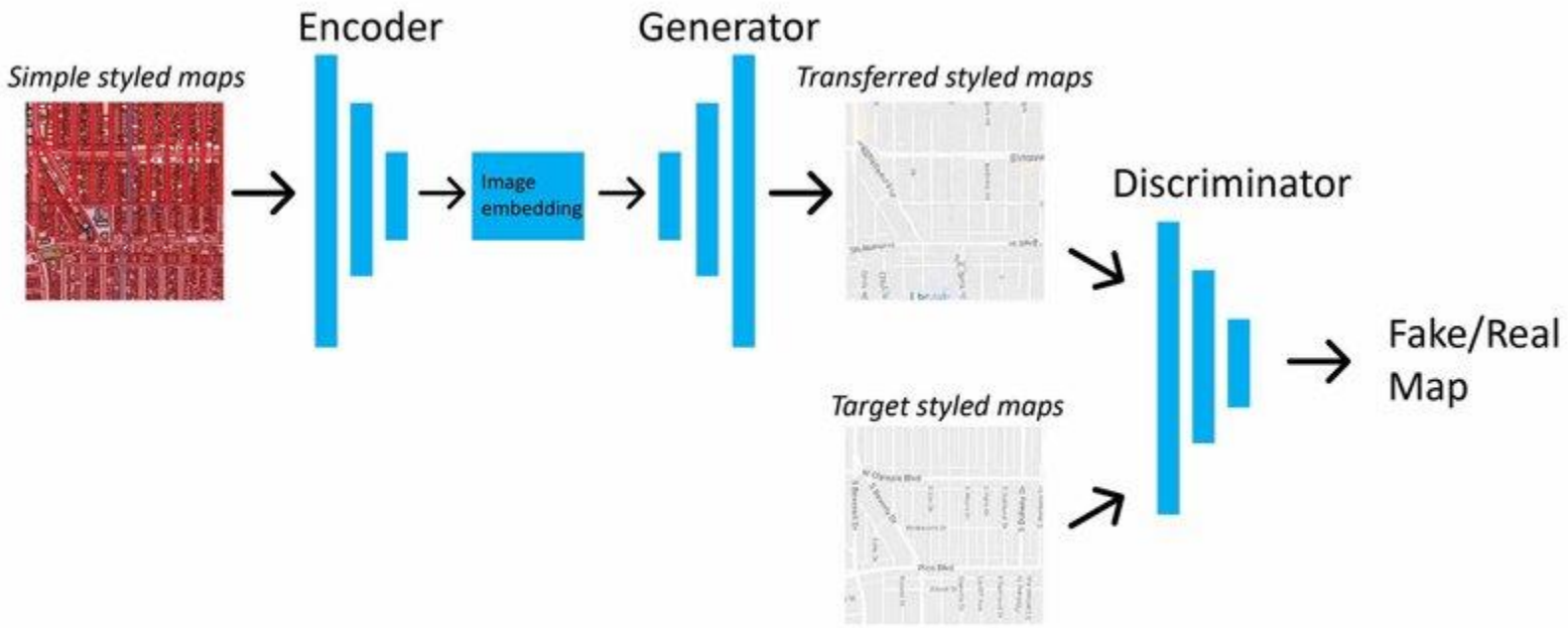
Conditional GAN (CGAN)



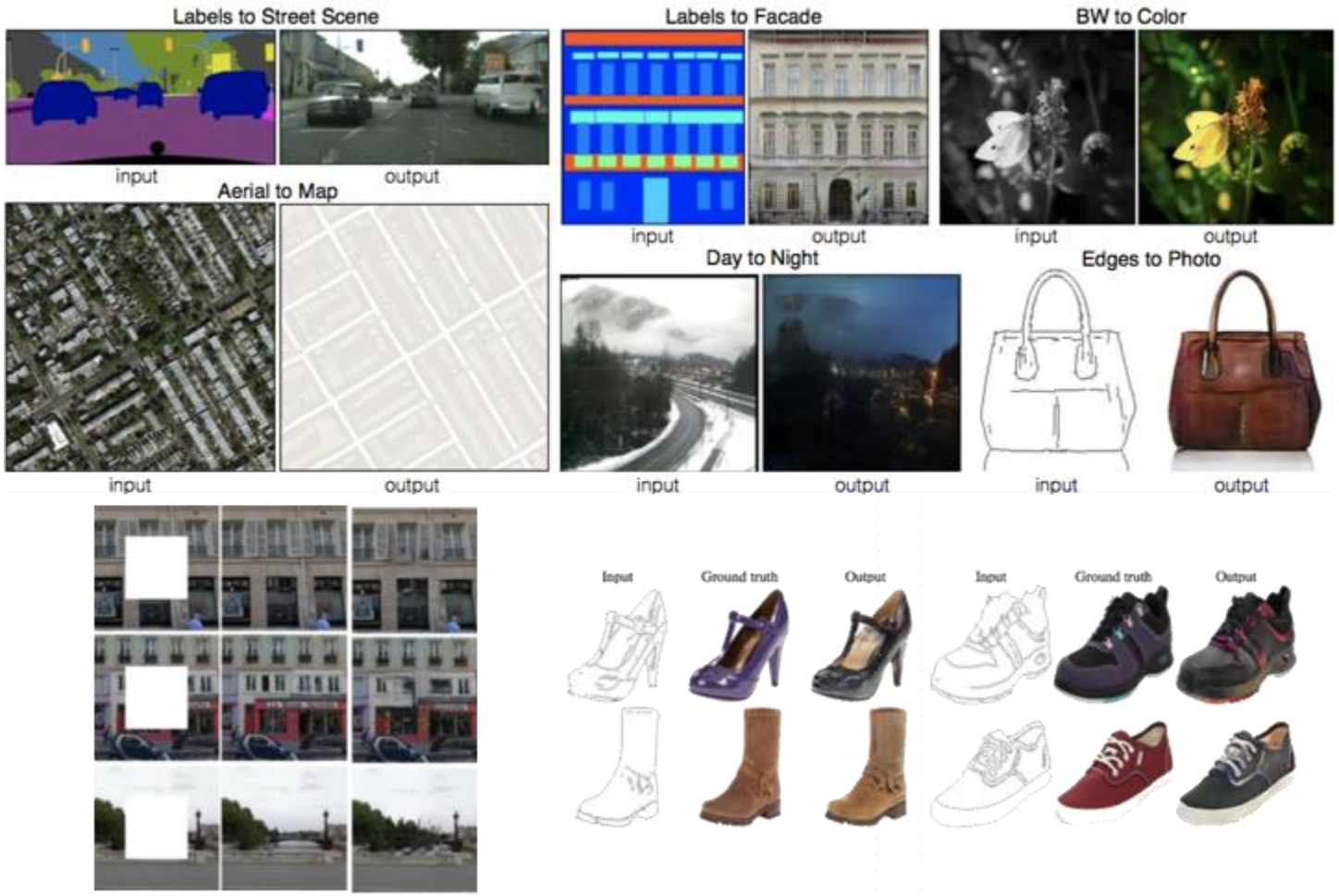
https://mofanpy.com/tutorials/machine-learning/gan/cgan


T-Shirt
Trouser
Pullover
Dress
Coat
Sandal
Shirt
Sneaker
Bag
Ankle boot

Pix2pix

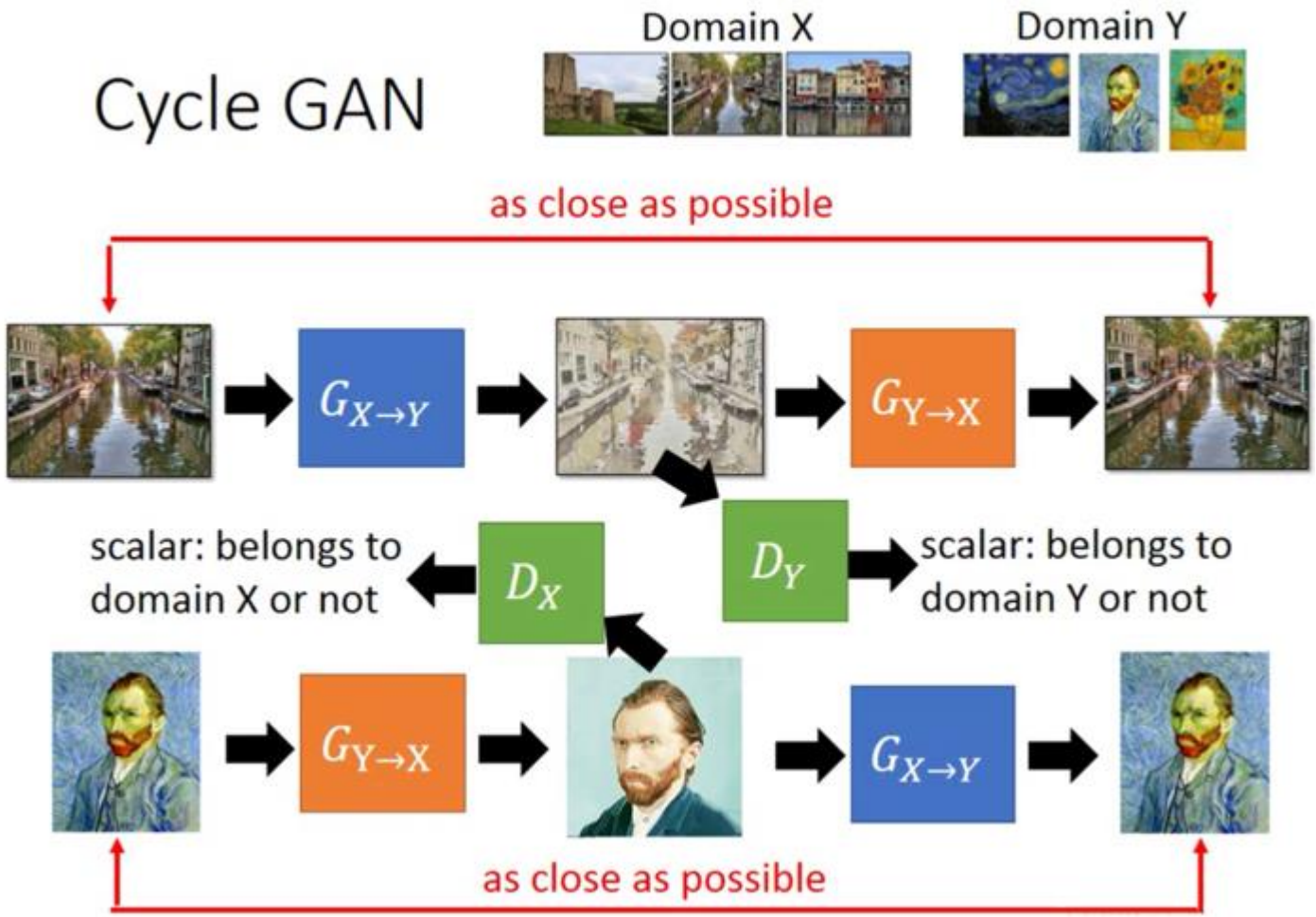

Simple styled maps
Encoder
Image embedding
Generator
Transferred styled maps
Discriminator
Target styled maps
Fake/Real Map

Pix2pix


Labels to Street Scene
input
output
Labels to Facade
input
output
BW to Color
input
output
Aerial to Map
input
output
Day to Night
input
output
Edges to Photo
input
output
Input
Ground truth
Output
Input
Ground truth
Output

Looks cool! But one disadvantage is that we need paired data for training.

CycleGAN


Cycle GAN
Domain X
Domain Y
as close as possible
$G_{X \to Y}$
$G_{Y \to X}$
scalar: belongs to domain X or not
D_X
D_Y
scalar: belongs to domain Y or not
$G_{Y \to X}$
$G_{X \to Y}$
as close as possible

CycleGAN

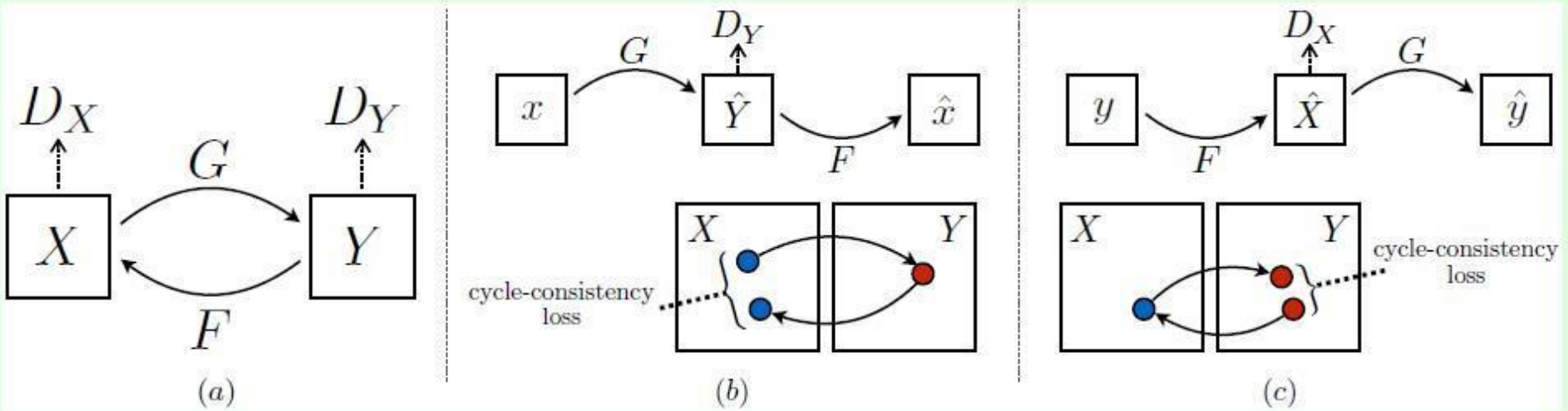
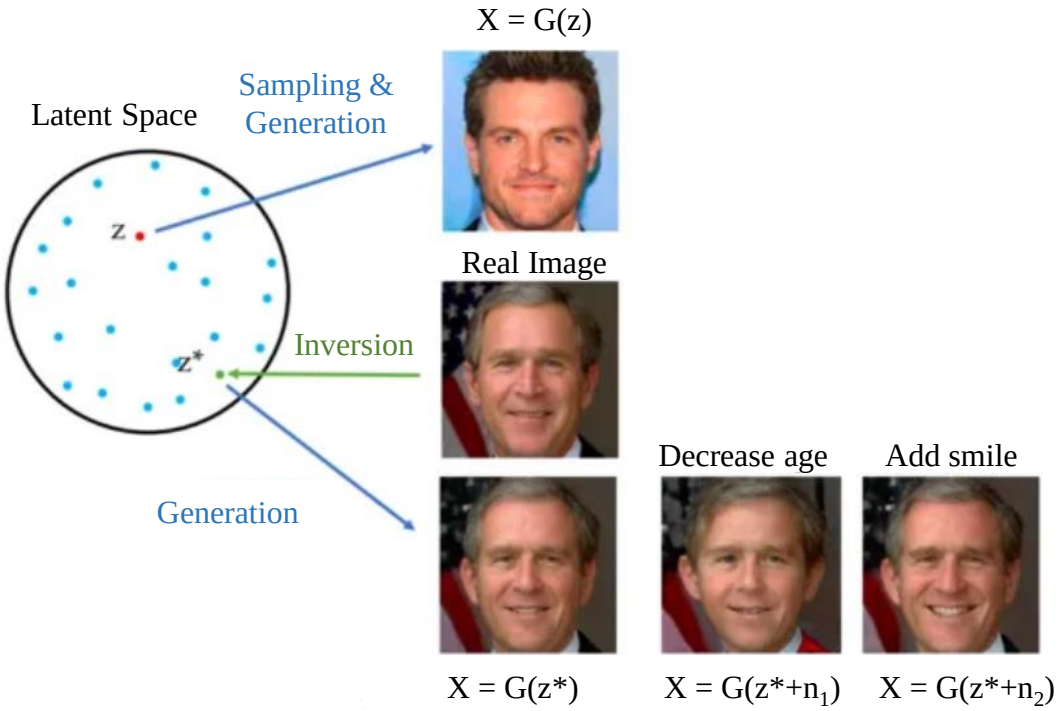


Figure 3: (a) Our model contains two mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and associated adversarial discriminators D_Y and D_X. D_Y encourages G to translate X into outputs indistinguishable from domain Y, and vice versa for D_X and F. To further regularize the mappings, we introduce two *cycle consistency losses* that capture the intuition that if we translate from one domain to the other and back again we should arrive at where we started: (b) forward cycle-consistency loss: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, and (c) backward cycle-consistency loss: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$

Other Applications

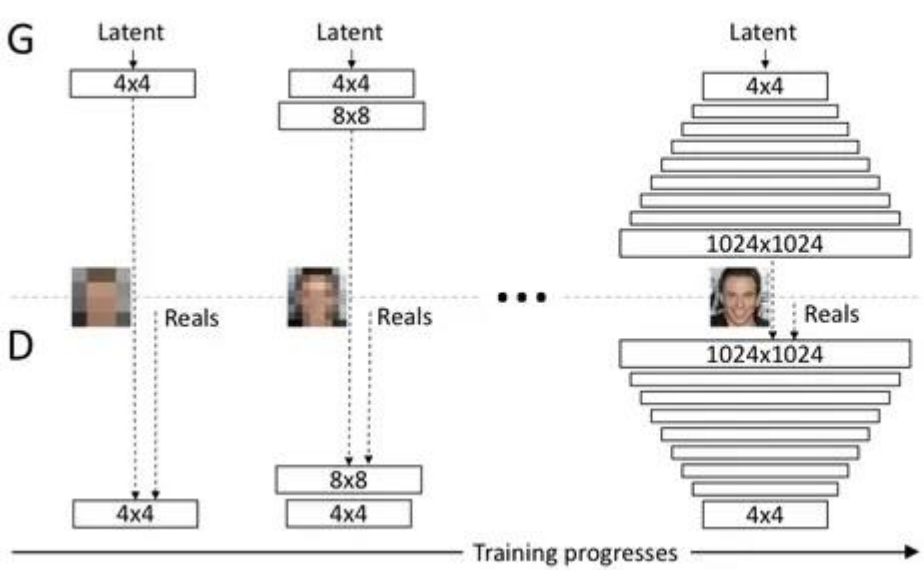
GAN Inversion

GAN inversion aims to **obtain the latent vector for any given image**, such that when passed through the generator, it generates an image close to the real image. Obtaining the latent vector provides more flexibility to perform image manipulation on any image instead of being constrained to GAN-generated images obtained from random sampling and generation.



Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive Growing of GANs for Improved Quality, Stability, and Variation. ArXiv, abs/1710.10196.

Super Resolution



ProGAN is a high-resolution image generation model proposed by NVIDIA, and highly influential models such as StyleGAN are based on ProGAN improvements.

The network starts from a very low resolution (e.g. 4×4), and gradually train and increase the image resolution until the generator is able to generate images with a resolution up to the target high resolution (e.g. 1024×1024).

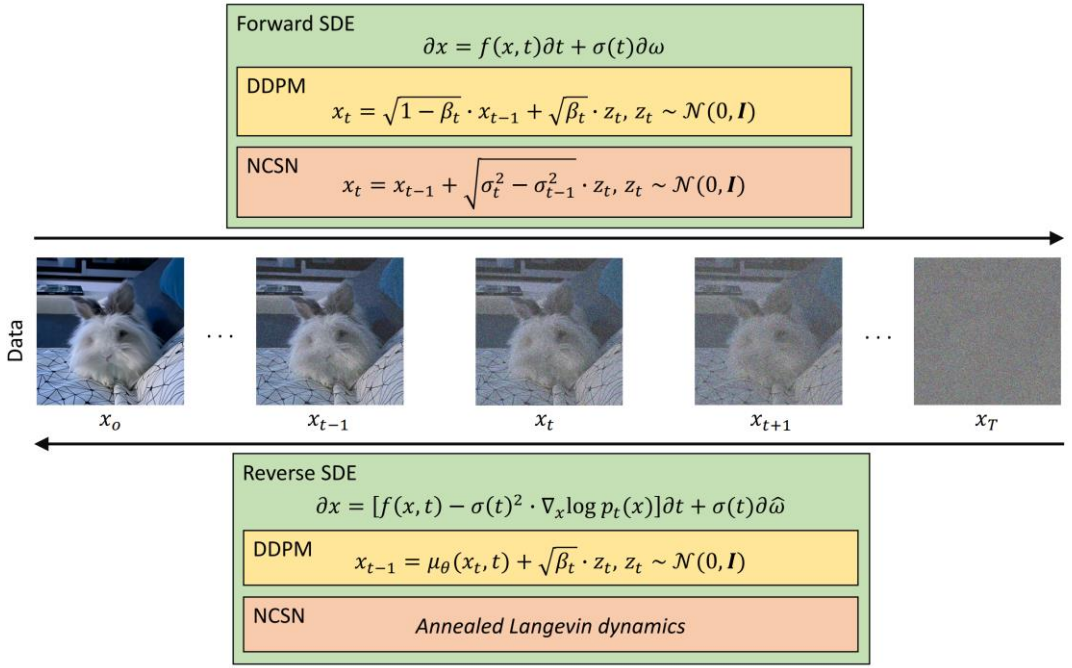
Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive Growing of GANs for Improved Quality, Stability, and Variation. ArXiv, abs/1710.10196.

Diffusion Models

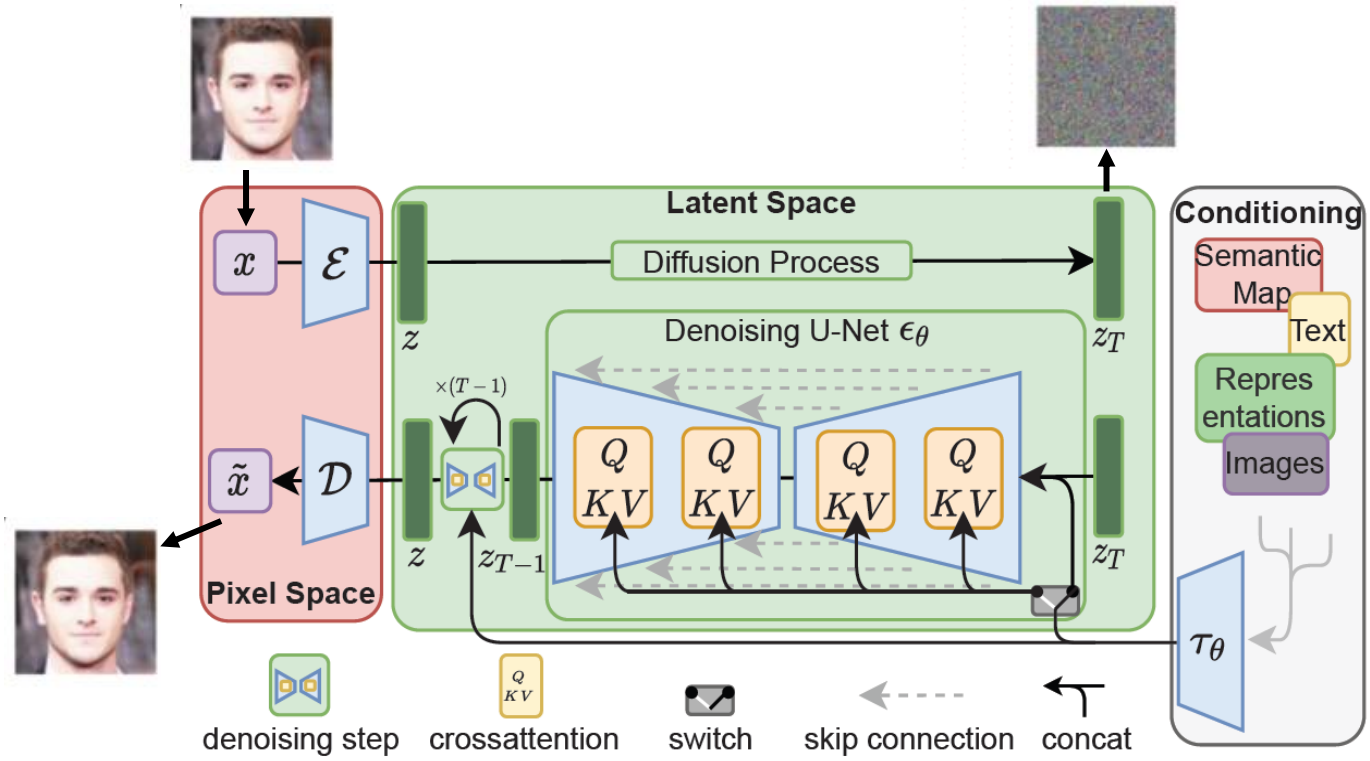
Diffusion Models

A generic framework composing three alternative formulations of diffusion models based on: denoising diffusion probabilistic models (DDPMs), noise conditioned score networks (NCSNs), and stochastic differential equations (SDEs). The formulation based on SDEs is a generalization of the other two.

In the forward process, Gaussian noise is gradually added to the input X_0 over T steps. In the reverse process, a model learns to restore the original input by gradually removing the noise.



Stable Diffusion

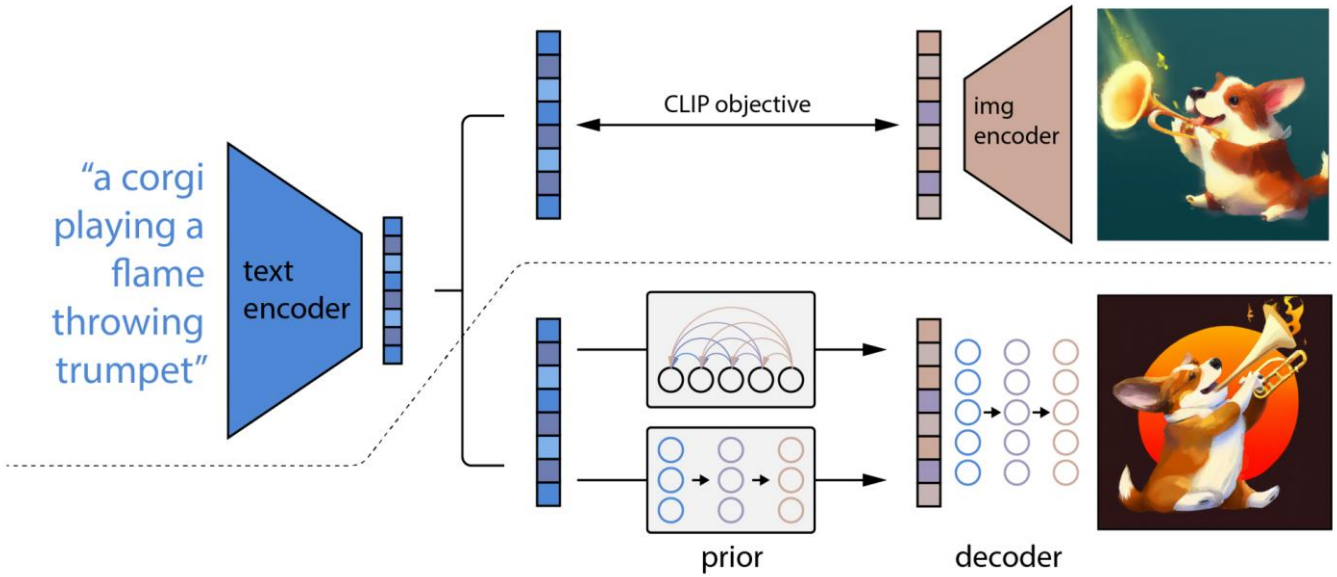


Stable Diffusion: Large-scale training, stable and effective, and open-sourced.

The Stable Diffusion is based on the Latent Diffusion Model, in which the diffusion is conducted in the latent space rather than in the image space.

Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

DALL-E2



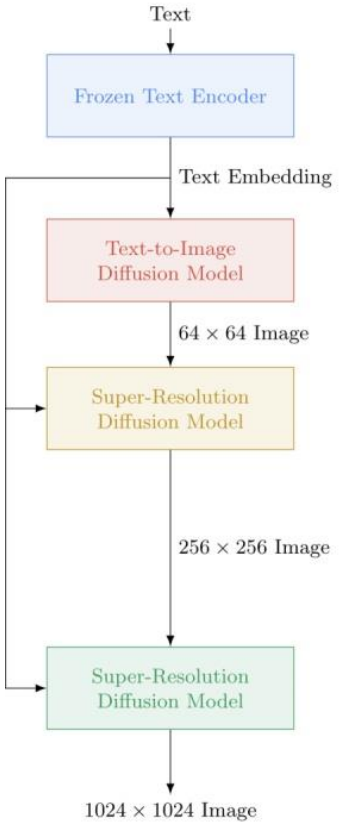
The Text-to-Image model DALLE2 from OpenAI is a diffusion model-based approach. DALLE2 consists of a Text Encoder, a feature vector Diffusion Model and a Image Decoder.

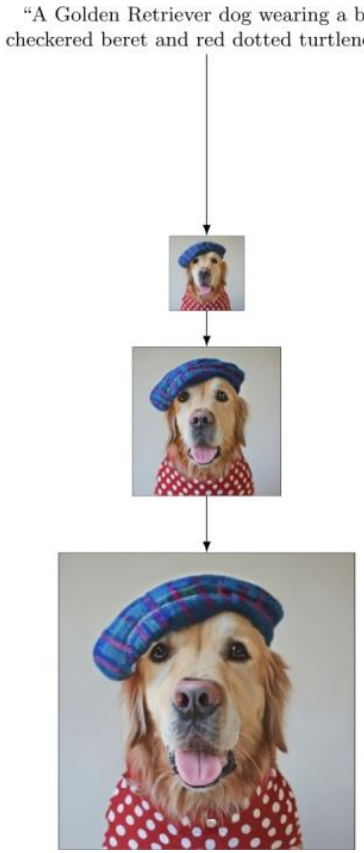
Imagen

Imagen from Google is also a text-to-image model based on the diffusion.

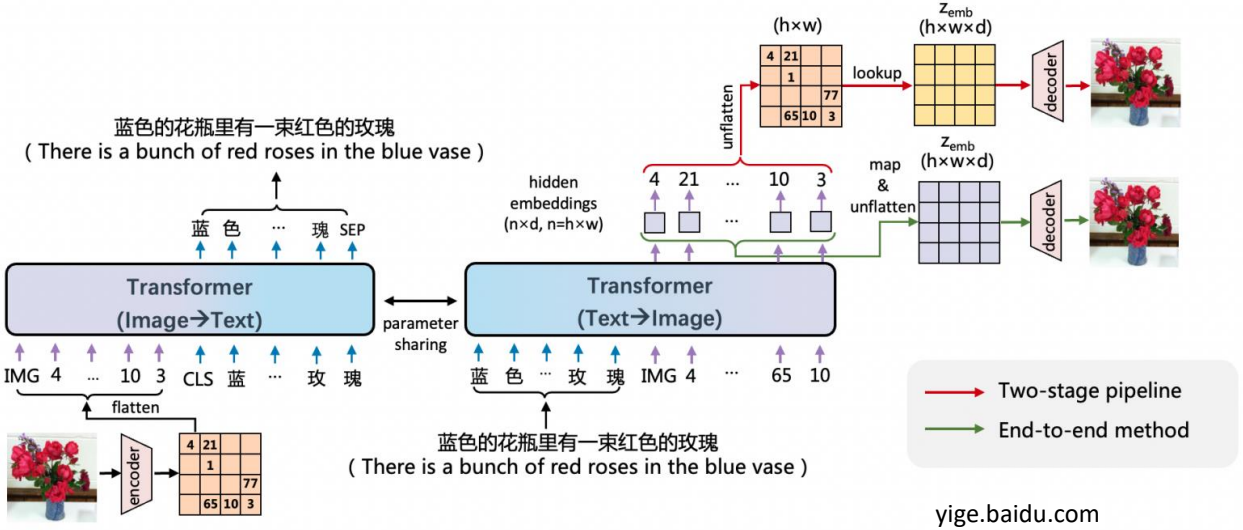
The novelty of Imagen is that it uses multiple super-resolution diffusion models to gradually increase the resolution of the image in order to generate a high-resolution images.

imagen.research.google





ERNIE-ViLG (WENXIN YIGE)



yige.baidu.com

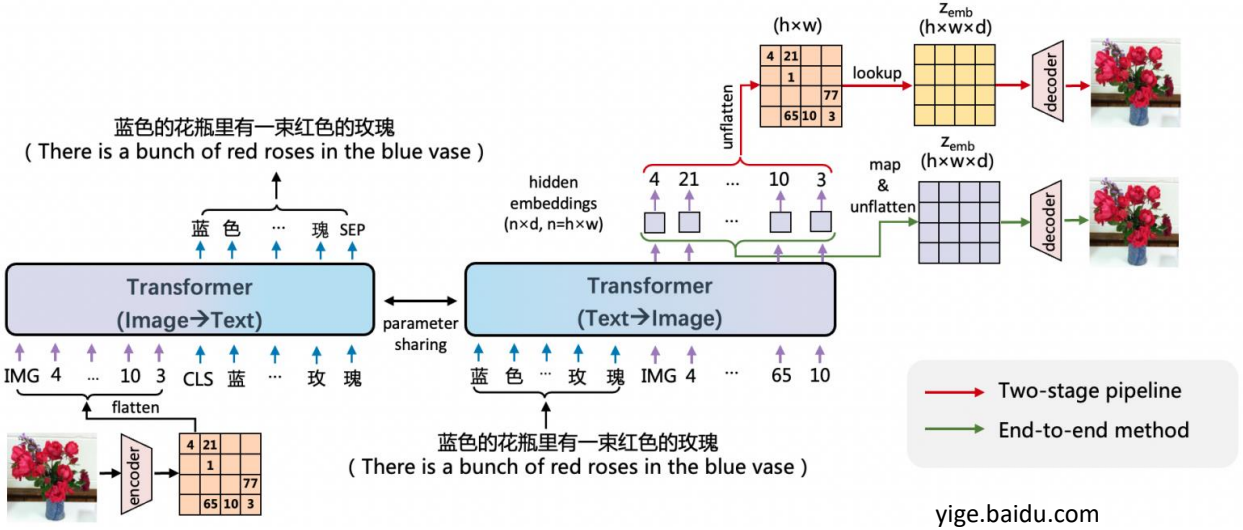
While most text-to-image models use the diffusion, Wenxin Yige from Baidu uses a model based on VQ (Vector Quantization) and Transformer.

First, a Transformer-based VQ-AutoEncoder is trained, i.e., Encoder compresses the image into discrete Tokens, and Decoder decompresses the discrete Tokens into images.

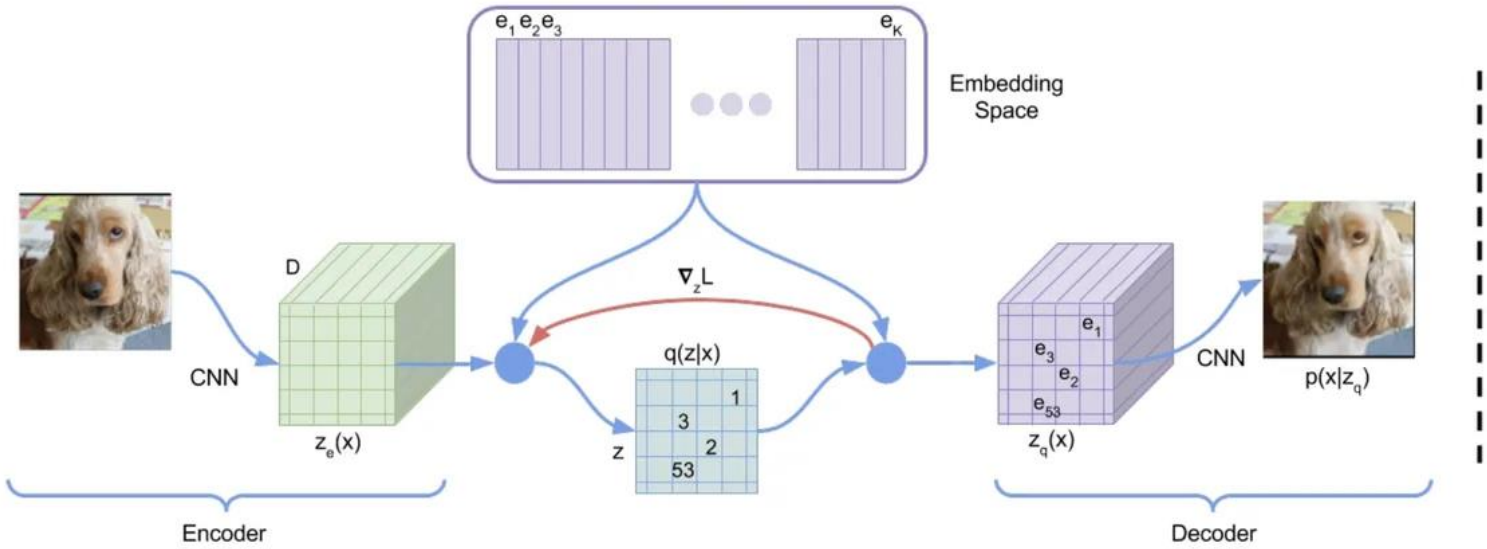
Then, two Transformers with shared parameters are trained to generate image Tokens based on text Tokens and text Tokens based on image Tokens, respectively.

After training, any text can be used as the input to the Transformer to obtain the image Token, and then the image Token is used as the input to the Decoder of VQ-AutoEncoder to decompress it to an image.

ERNIE-ViLG (WENXIN YIGE)

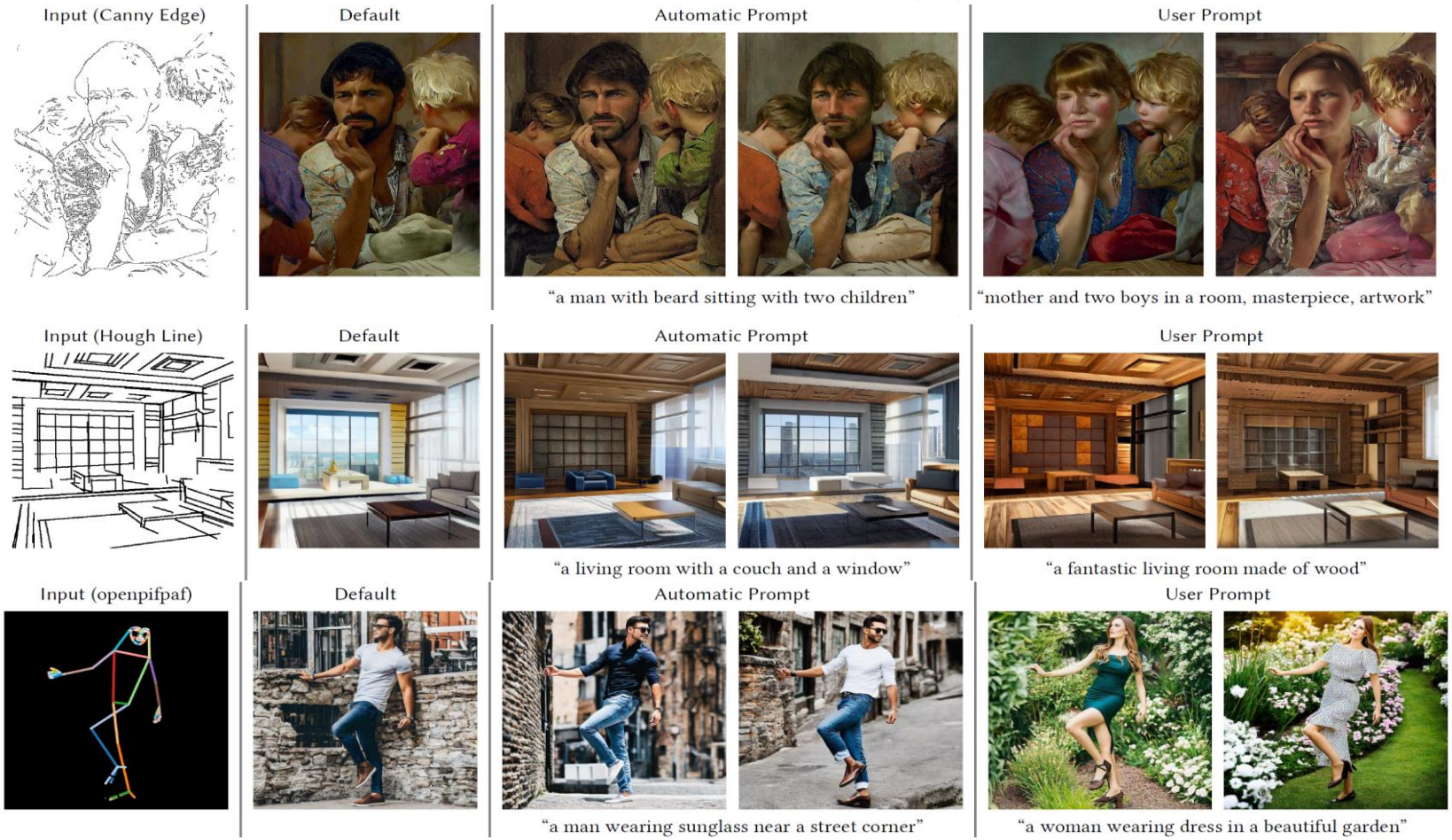


yige.baidu.com



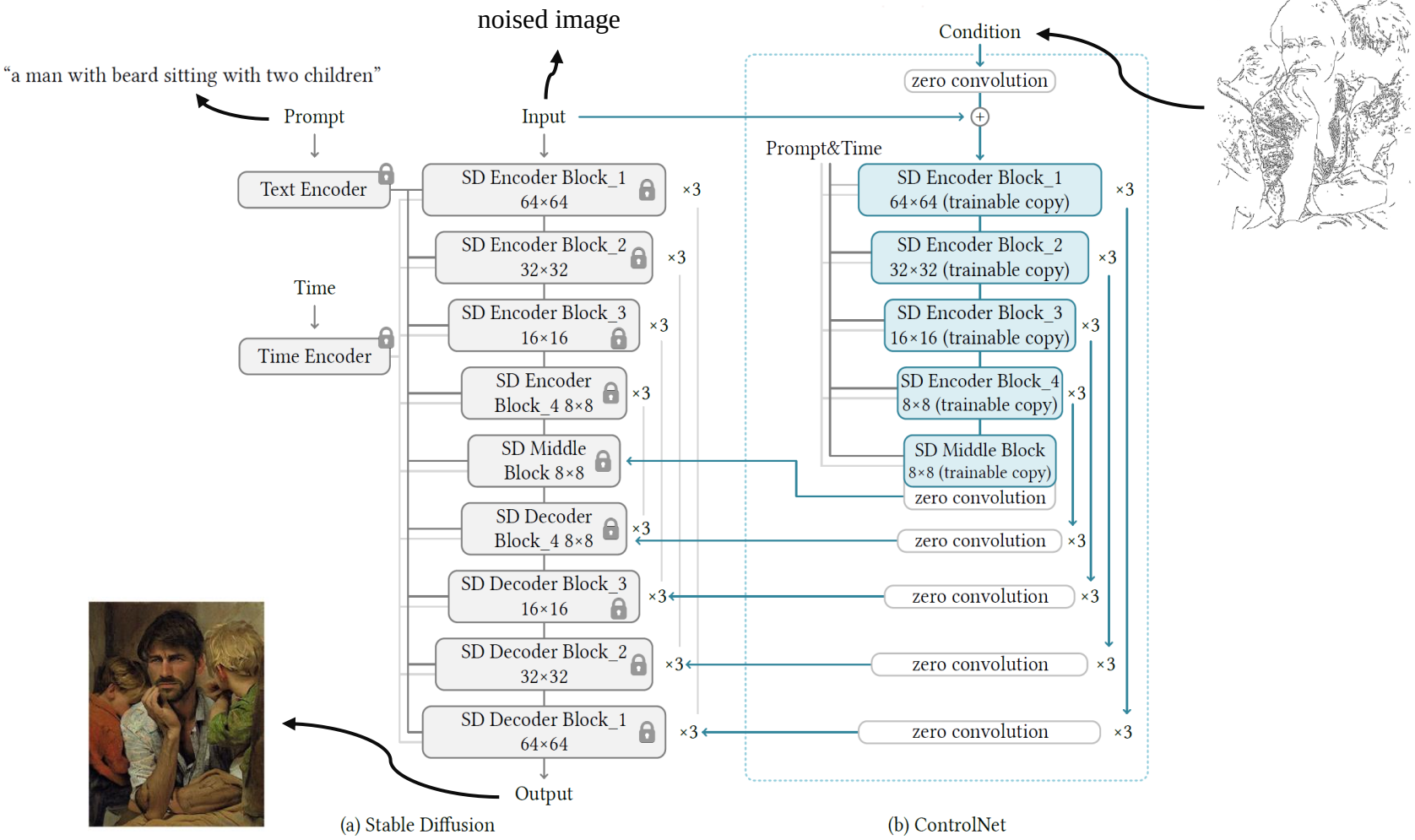
Recent trend to control the synthesis

ControlNet – Task specific conditions



Zhang, Lvmin, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." arXiv preprint arXiv:2302.05543 (2023).

ControlNet



Zhang, Lvmin, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." arXiv preprint arXiv:2302.05543 (2023).

Text-driven Image Editing

GLIDE

During fine-tuning, random regions of training examples are erased, and the remaining portions are fed into the model along with a mask channel as additional conditioning information.

The diffusion architecture has four additional input channels: **a second set of RGB channels, and a mask channel**.

Nichol, Alex, et al. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models." arXiv preprint arXiv:2112.10741 (2021).

"a corgi wearing a bow tie and a birthday hat"

"a fire in the background"

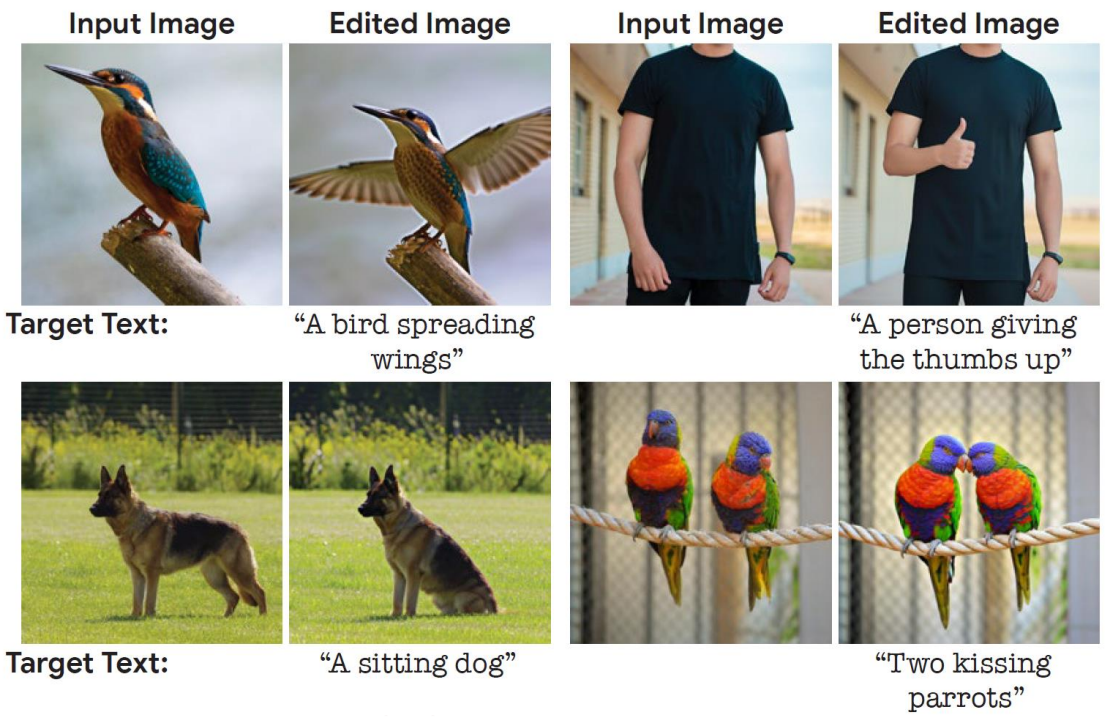
"only one cloud in the sky today"

Text-driven object replacement with mask

Text-driven Image Editing

Imagic

The method can perform various text-based semantic edits on a single real input image, including highly complex non-rigid changes such as posture changes and editing multiple objects.



Kawar, Bahjat, et al. "Imagic: Text-based real image editing with diffusion models." arXiv preprint arXiv:2210.09276 (2022).

Text-driven object replacement without mask

Text-driven Image Editing

Imagic

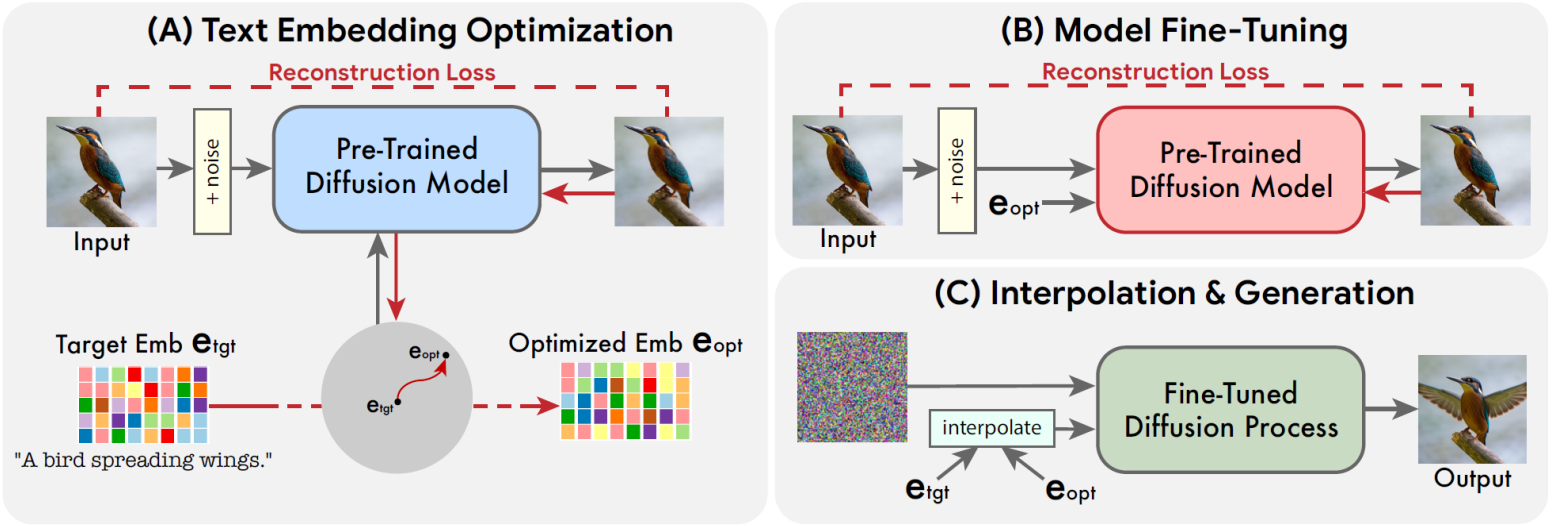


Figure 3. **Schematic description of *Imagic*.** *Given a real image and a target text prompt: (A) We encode the target text and get the initial text embedding* $\mathbf{e}_{tgt}$*, then optimize it to reconstruct the input image, obtaining* $\mathbf{e}_{opt}$*; (B) We then fine-tune the generative model to improve fidelity to the input image while fixing* $\mathbf{e}_{opt}$*; (C) Finally, we interpolate* $\mathbf{e}_{opt}$ *with* $\mathbf{e}_{tgt}$ *to generate the final editing result.*

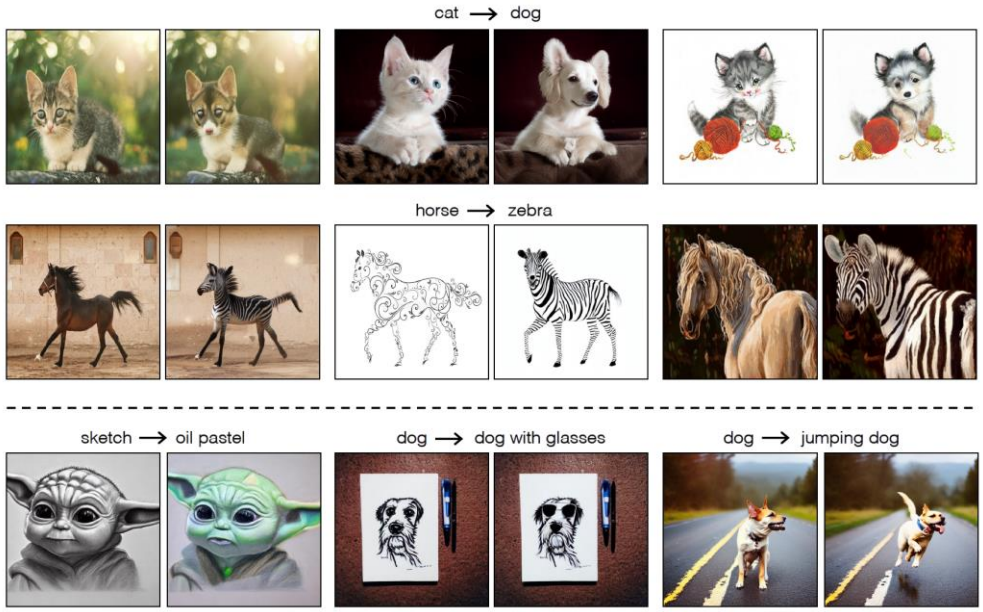
Kawar, Bahjat, et al. "Imagic: Text-based real image editing with diffusion models." arXiv preprint arXiv:2210.09276 (2022).

Text-driven Image Editing

Pix2pix-zero

pix2pix-zero, a diffusion-based image-to-image approach that allows users to **specify the edit direction on-the-fly** (e.g., cat to dog).

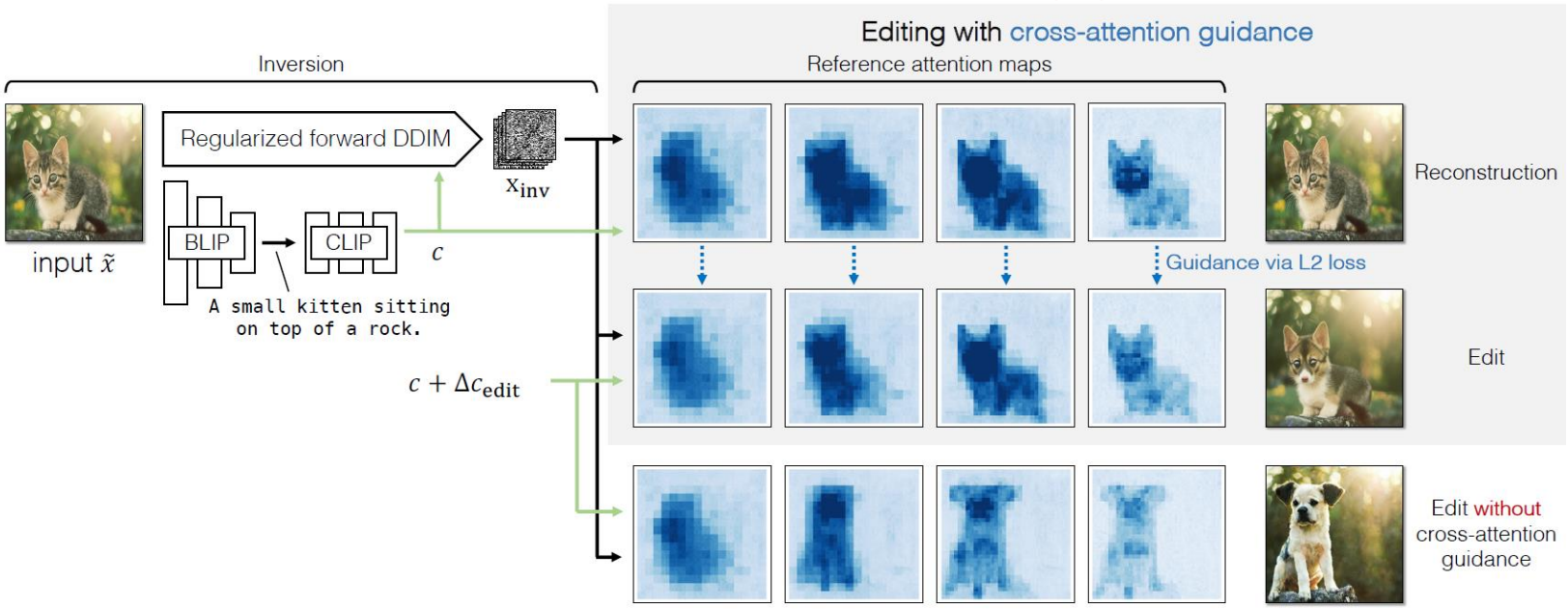
The method can directly use pre-trained text-to-image diffusion models, such as Stable Diffusion, for editing real and synthetic images while preserving the input image's structure.



Parmar, Gaurav, et al. "Zero-shot image-to-image translation." arXiv preprint arXiv:2302.03027 (2023).

Text-driven Image Editing

Pix2pix-zero



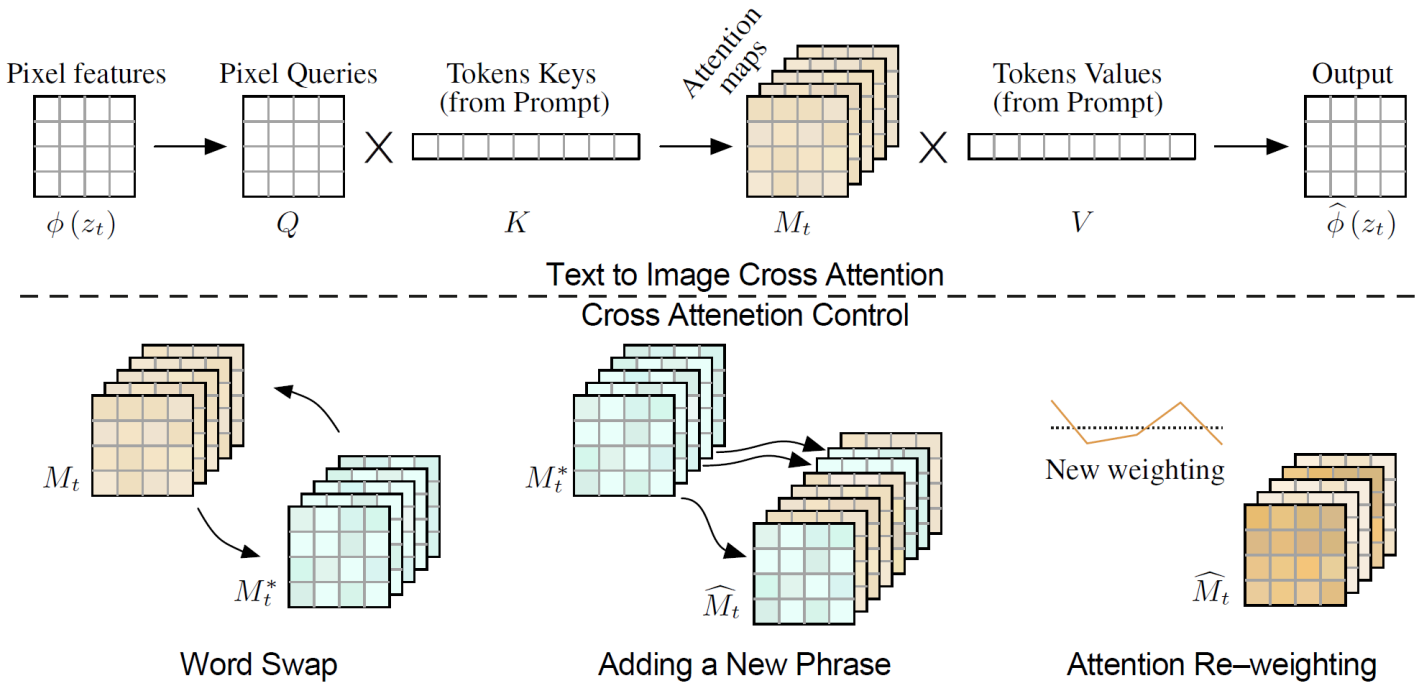
Parmar, Gaurav, et al. "Zero-shot image-to-image translation." arXiv preprint arXiv:2302.03027 (2023).

Prompt-to-prompt Editing

The method provides variety of Prompt-to-Prompt editing capabilities. The user can tune the level of influence of an adjective word, replace items in the image, specify a style for an image, or make further refinements over the generated image. The manipulations are infiltrated through the cross-attention mechanism of the diffusion model without the need for any specifications over the image pixel space.

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

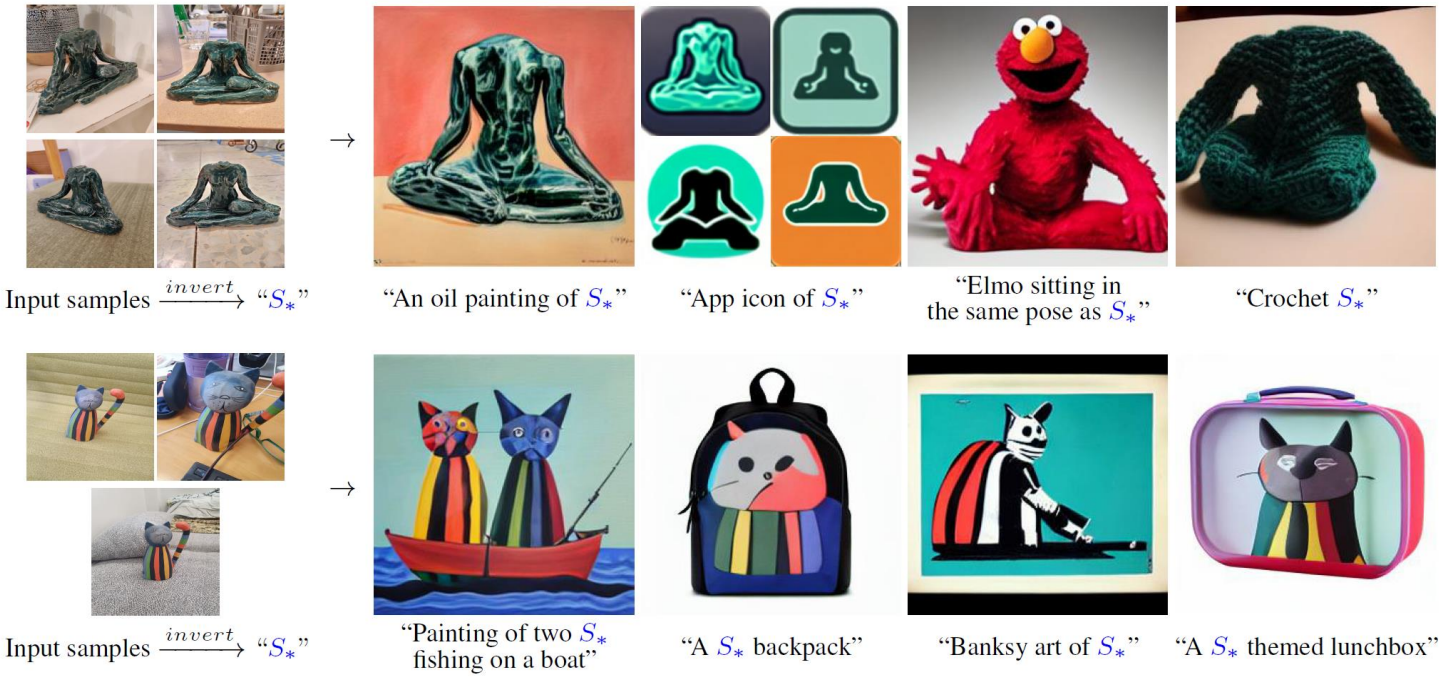
Prompt-to-prompt Editing



Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Personalization

Textual Inversion



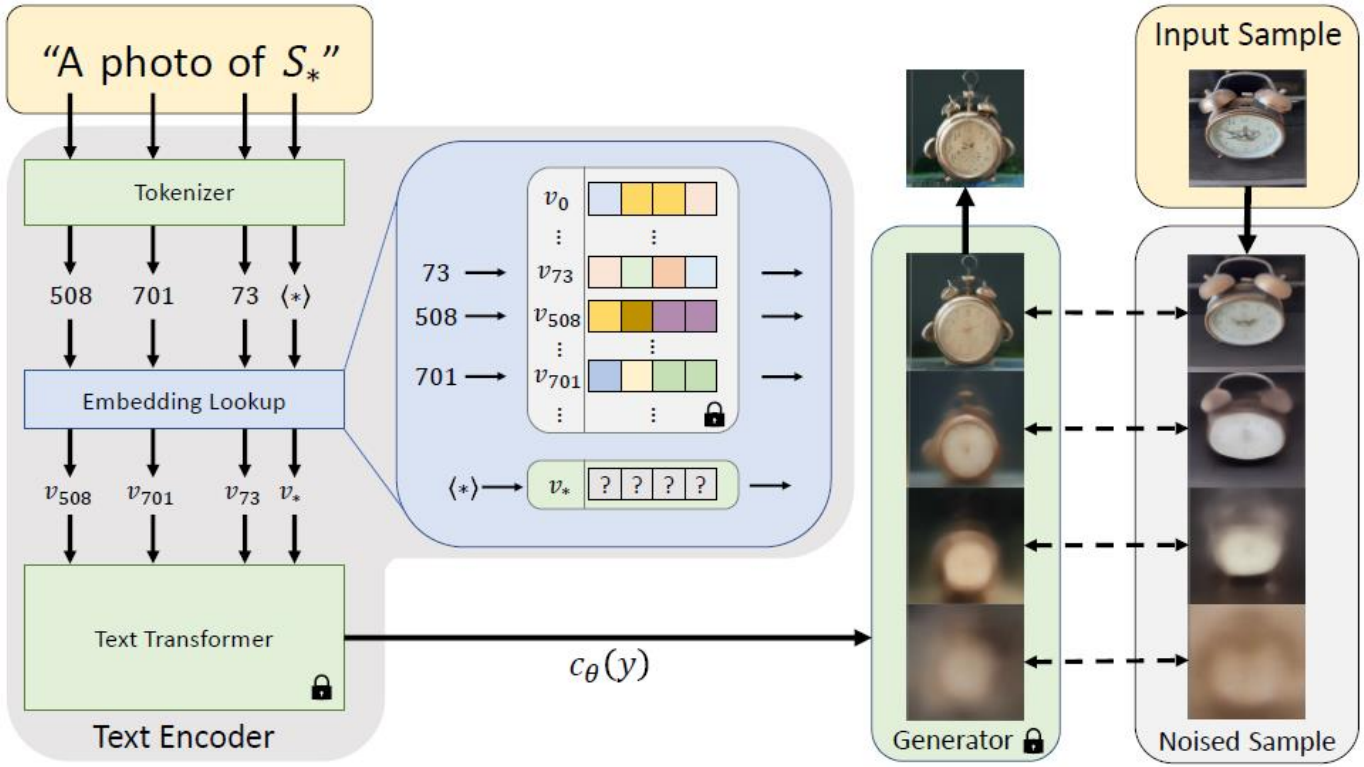
This method finds new pseudo-words in the embedding space of a pre-trained text-to-image model which describe specific concepts.

These pseudo-words can be composed into new sentences, placing the targets in new scenes, changing their style or composition, or ingraining them into new products.

Gal, Rinon, et al. "An image is worth one word: Personalizing text-to-image generation using textual inversion." arXiv preprint arXiv:2208.01618 (2022).

Personalization

Textual Inversion

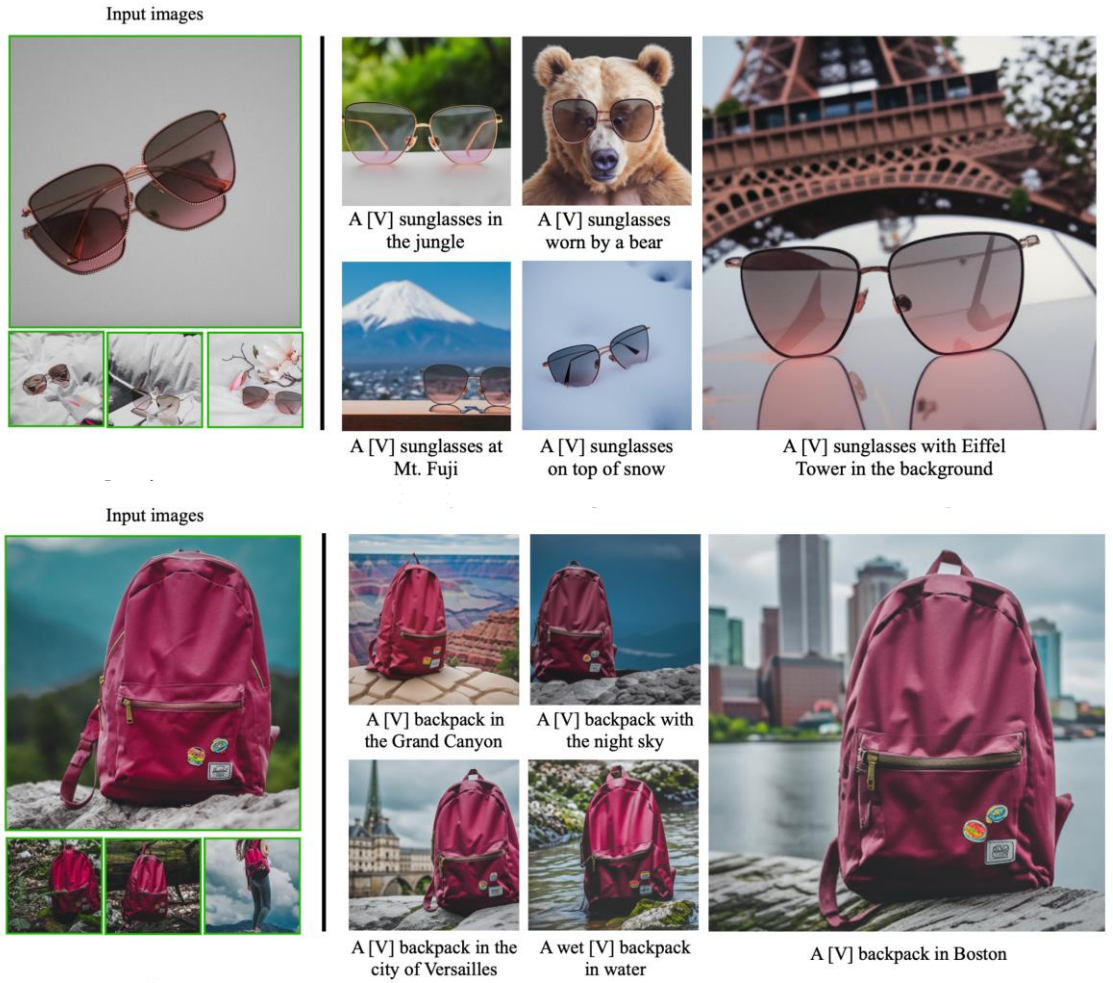


Gal, Rinon, et al. "An image is worth one word: Personalizing text-to-image generation using textual inversion." arXiv preprint arXiv:2208.01618 (2022).

Personalization

DreamBooth

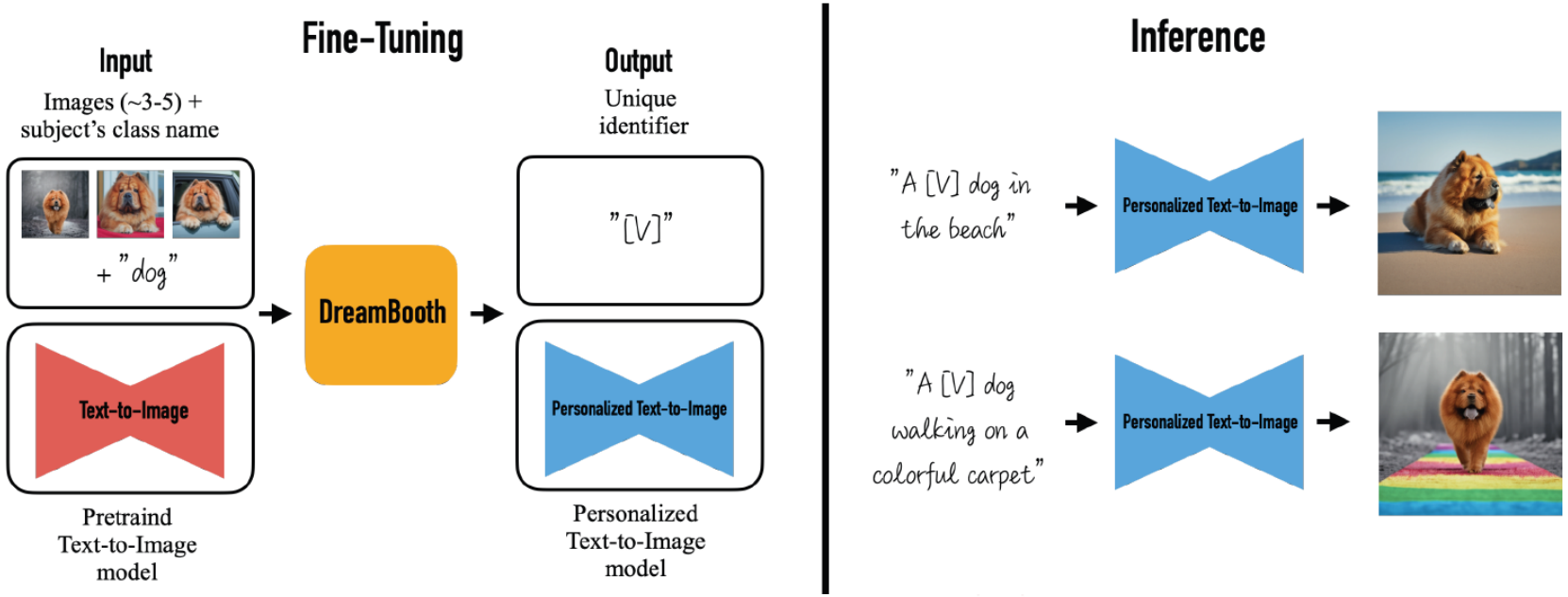
This method is similar with Textual Inversion. Given only a few (3-5) casually captured images of a specific subject, without any textual description, our objective is to generate new images of the subject with high detail fidelity and with variations guided by text prompts.



Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." arXiv preprint arXiv:2208.12242 (2022).

Personalization

DreamBooth



Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." arXiv preprint arXiv:2208.12242 (2022).

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學
Department of Computing
電子計算學系

Thank you!