Lecture 6: Application of GAN

Respond at **PollEv.com/ronisen**

Text **RONISEN** to **22333** once to join, then text your message

Feel free to share your questions...

Today's class

- Unconditional Image generation
  - DC-GAN
  - Wasserstein GAN
  - Progressive GAN
  - StyleGAN

- Conditional Image generation
  - Class conditional (Big GAN)
  - Paired (Pix2Pix)
  - Unpaired (CycleGAN)

Today's class

- Unconditional Image generation
  - DC-GAN
  - Wasserstein GAN
  - Progressive GAN
  - StyleGAN
- Conditional Image generation
  - Class conditional (Big GAN)
  - Paired (Pix2Pix)
  - Unpaired (CycleGAN)

UNSUPERVISED REPRESENTATION LEARNING WITH DEEP CONVOLUTIONAL GENERATIVE ADVERSARIAL NETWORKS

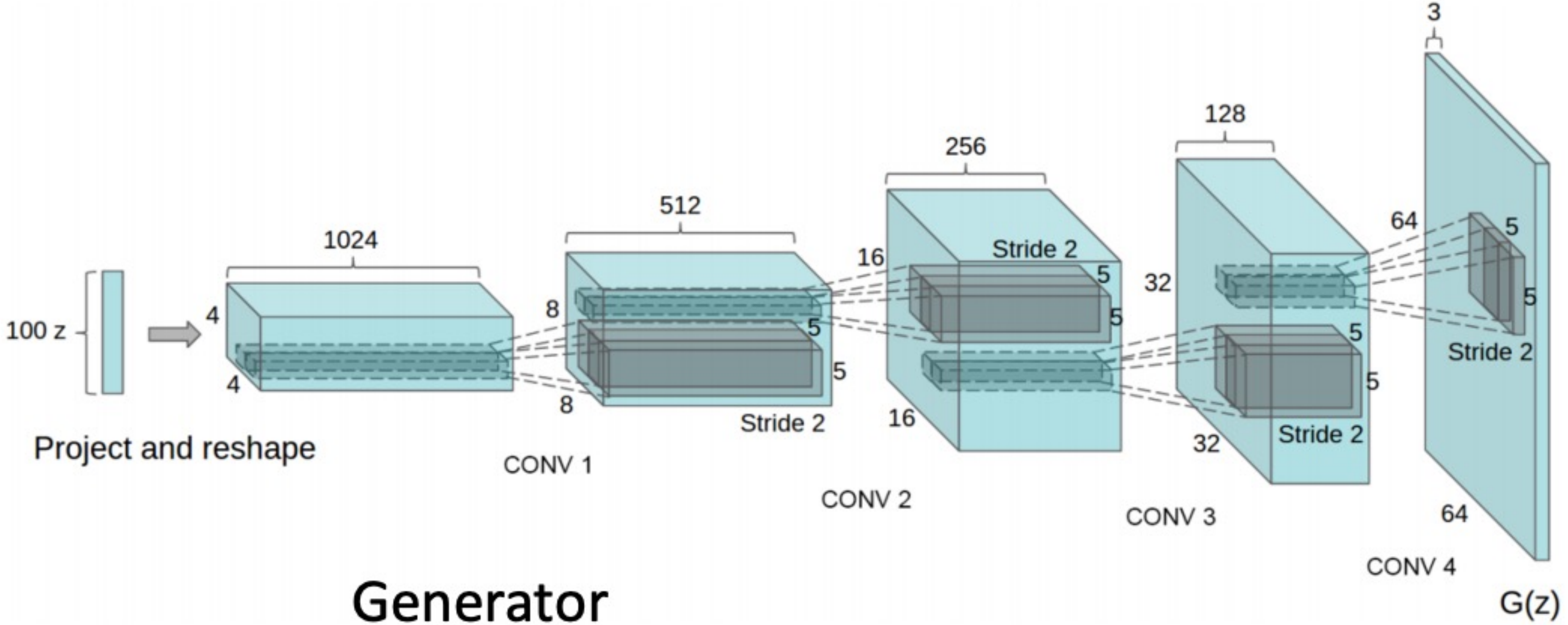
Alec Radford & Luke Metz
indico Research
Boston, MA
{alec,luke}@indico.io

Soumith Chintala
Facebook AI Research
New York, NY
soumith@fb.com

ABSTRACT

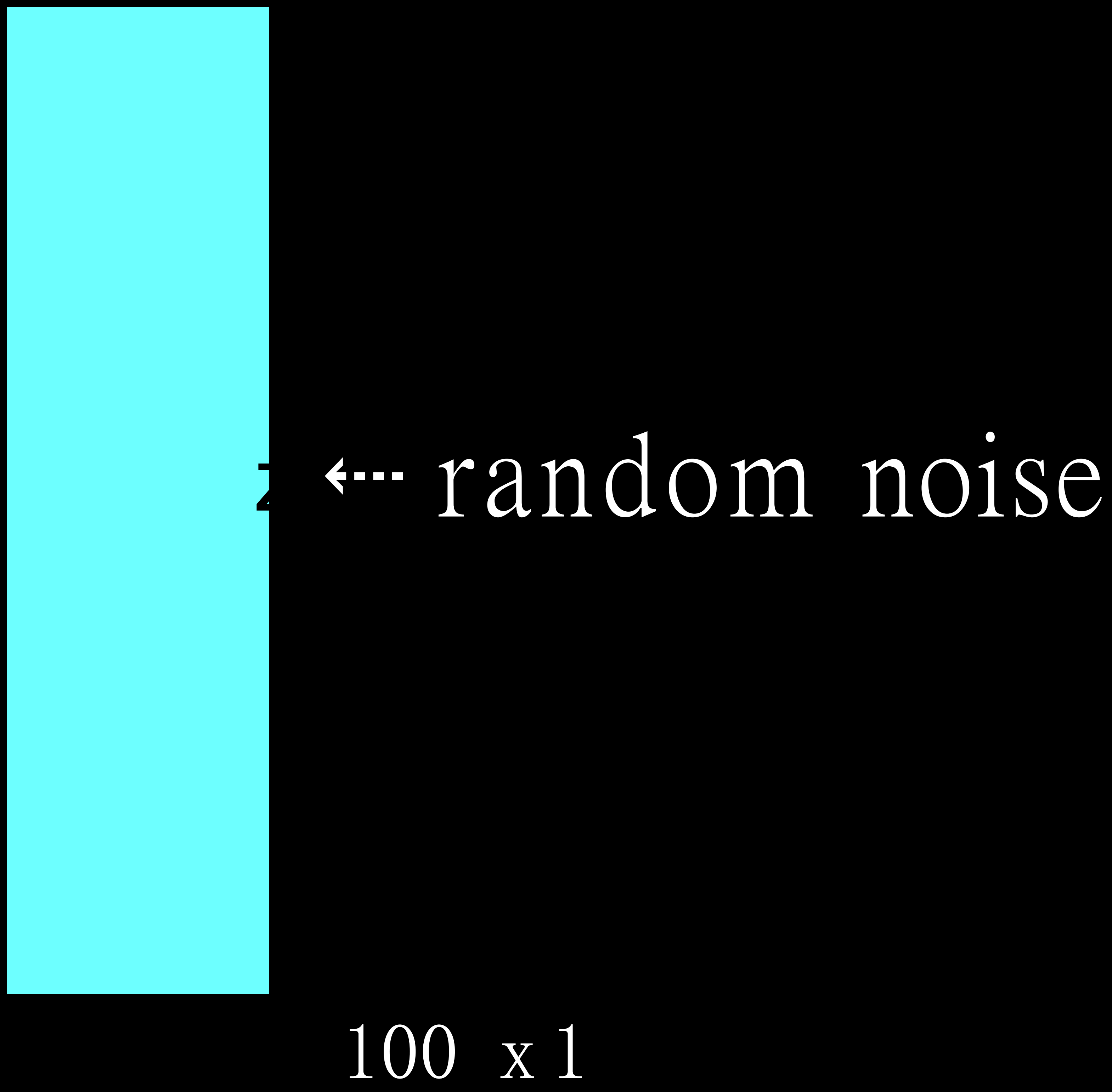
In recent years, supervised learning with convolutional networks (CNNs) has seen huge adoption in computer vision applications. Comparatively, unsupervised learning with CNNs has received less attention. In this work we hope to help

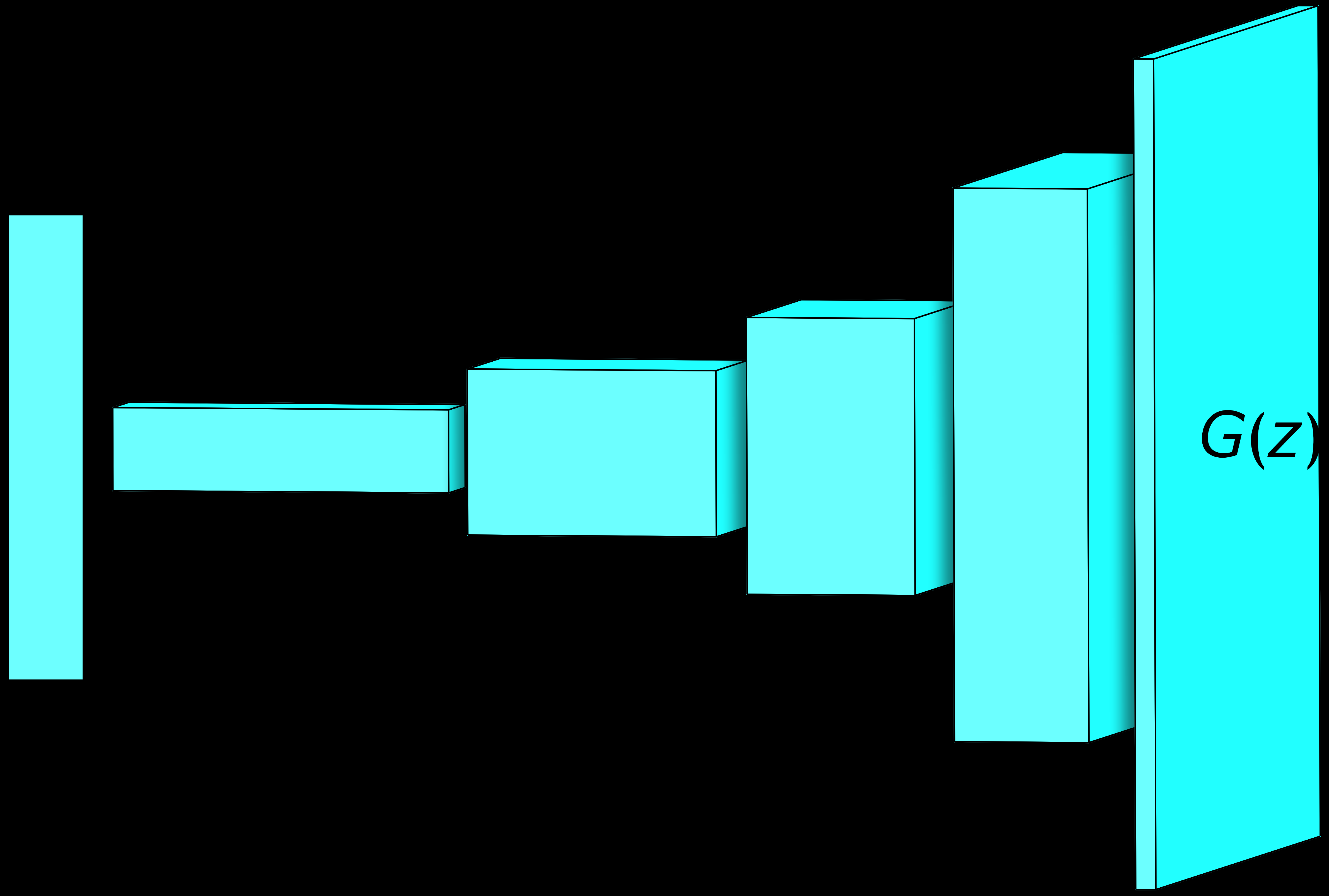
Generative Adversarial Networks: DC-GAN

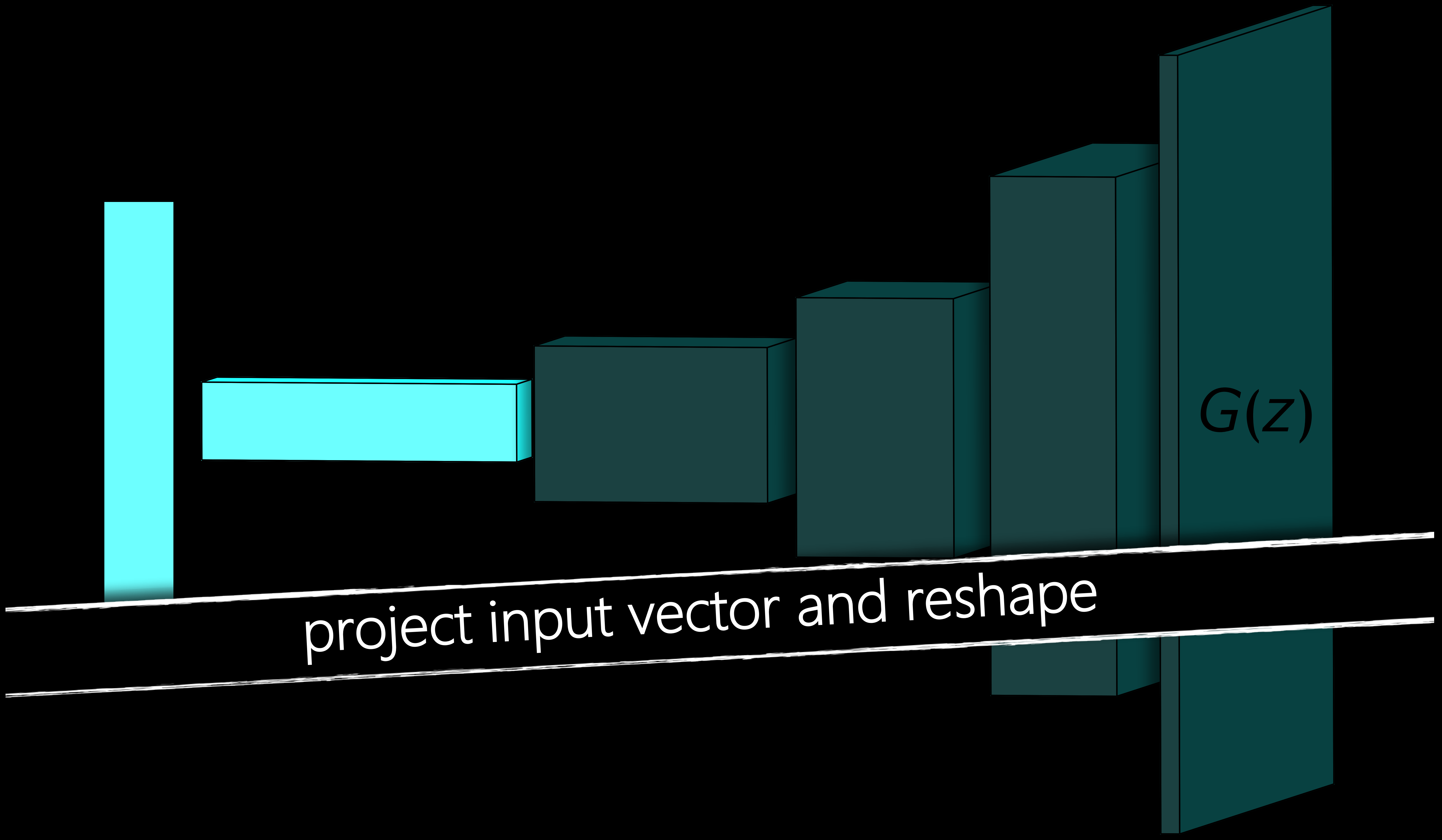


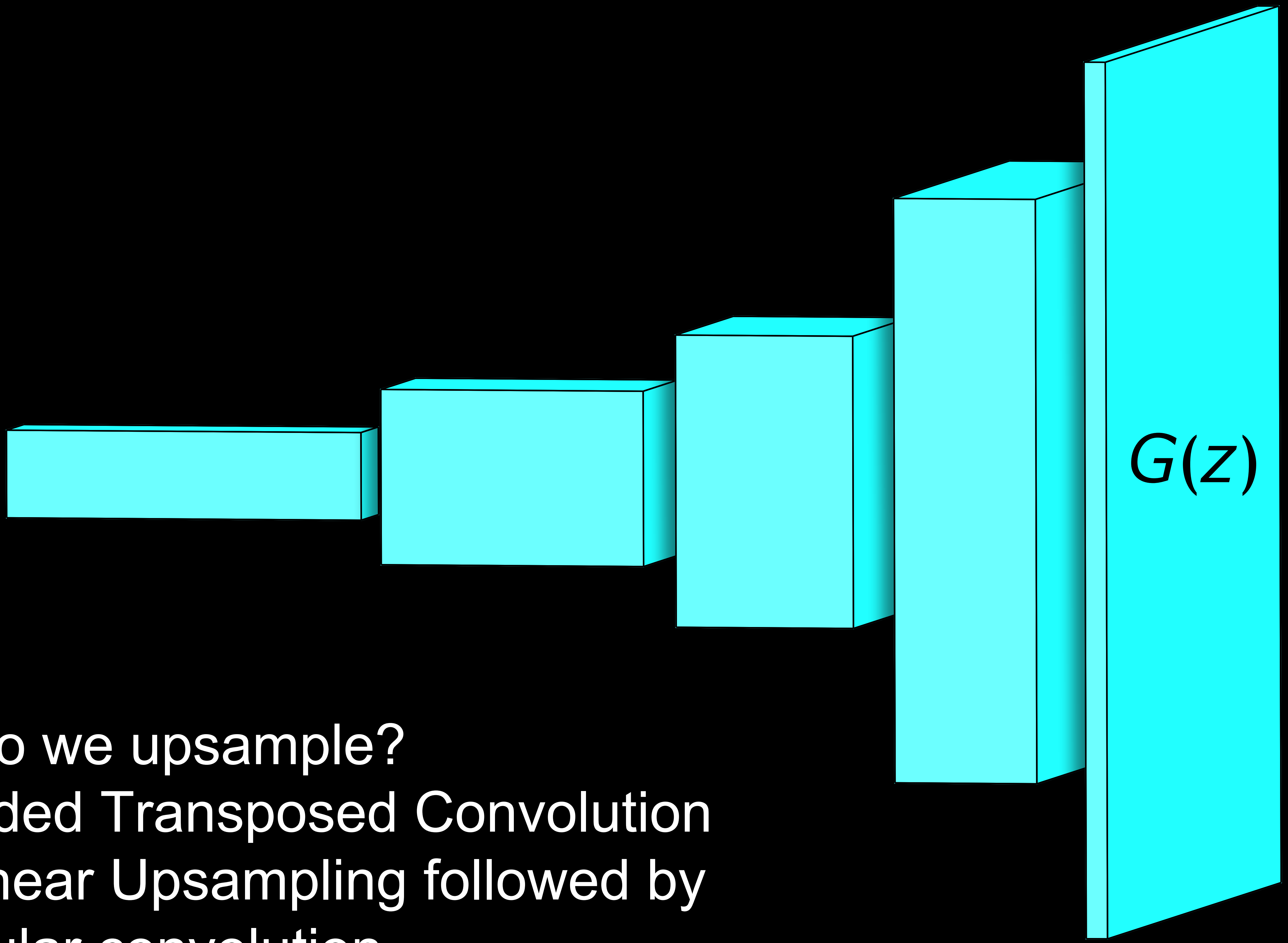
Generator

Radford et al, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", ICLR 2016


z
random noise
100 x 1


G(z)


G(z)
project input vector and reshape

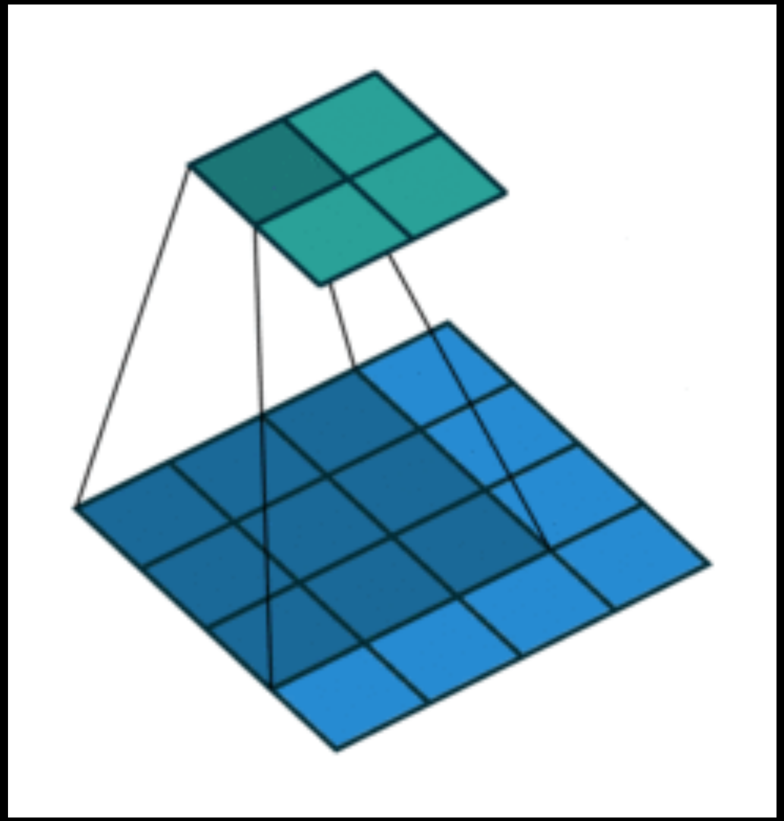


How do we upsample?

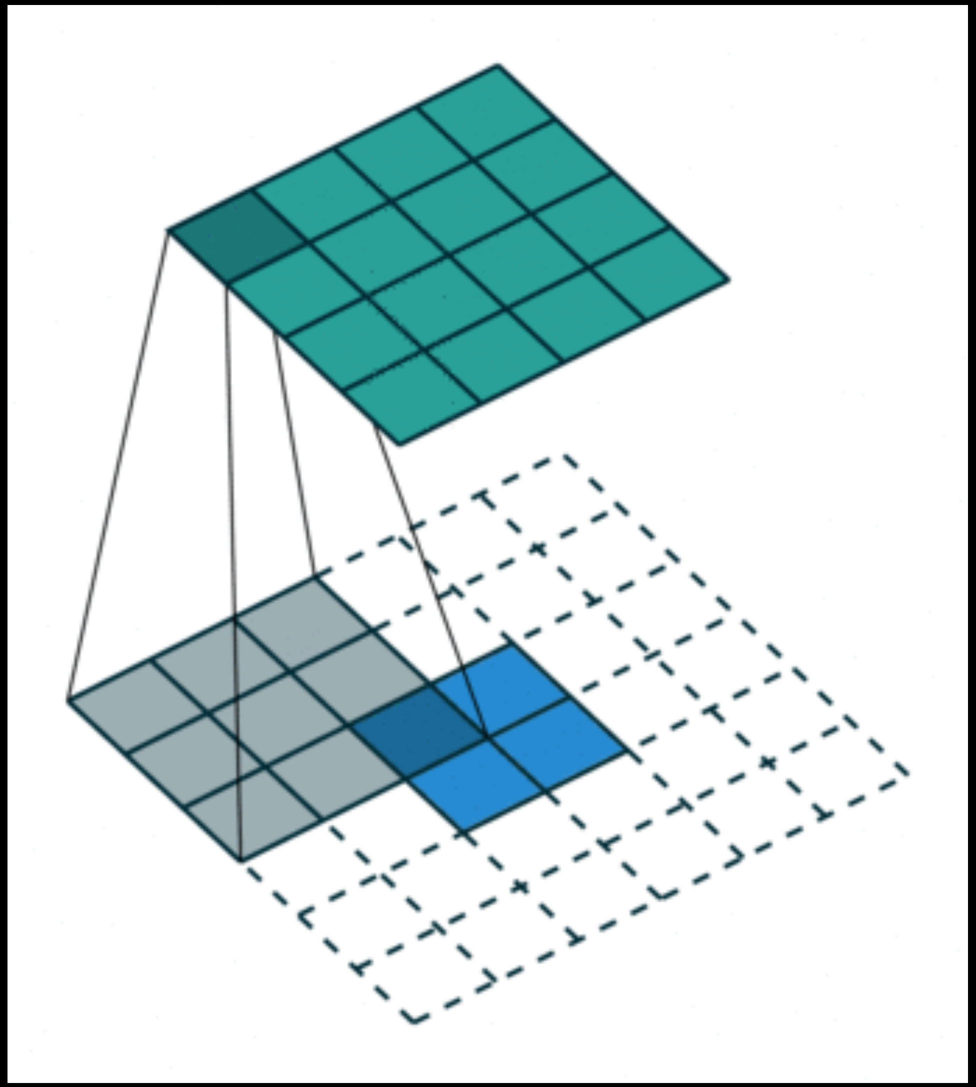
- Strided Transposed Convolution
- Bilinear Upsampling followed by regular convolution.

Regular vs Transposed Convolution

Filter size is 3x3



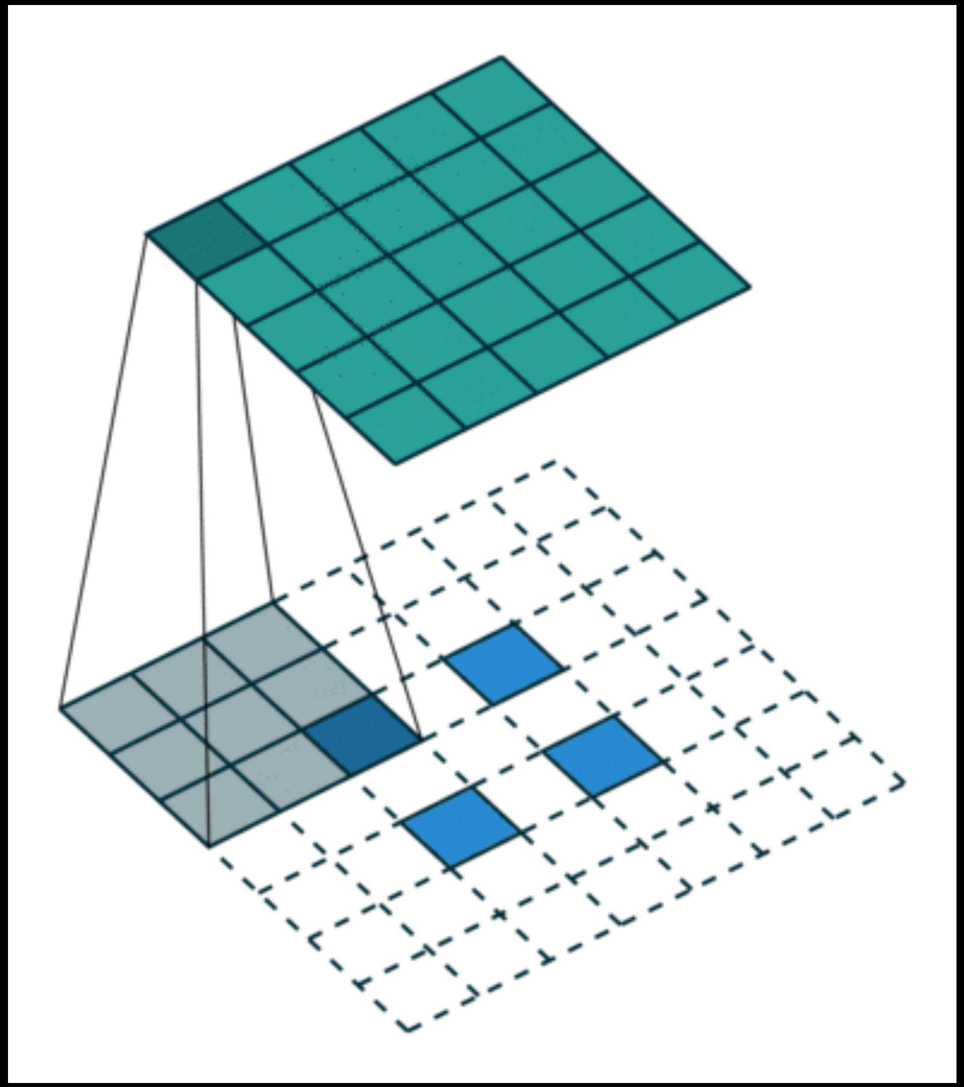
Regular Convolution reduces feature size



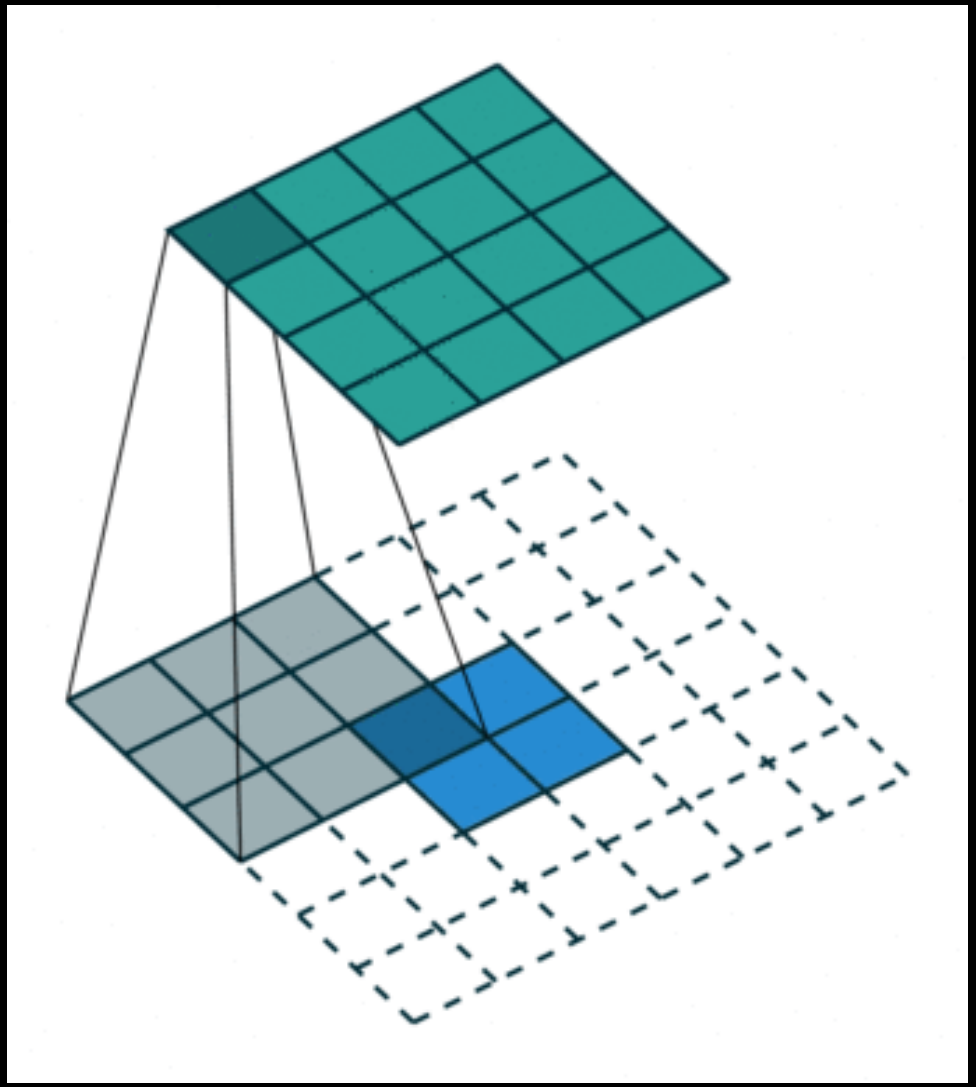
Transposed convolution increases feature size

Strided Transposed Convolution

Filter size is 3x3



With stride



Without stride

Bilinear Upsampling follow by regular convolution

Let the current feature map be 16x16x512.

Suppose, we want to create the next stage of feature map at 32x32x256.

We first upsample the current feature map from 16x16x512 to 32x32x512 using regular bi-linear upsampling.

Then we apply 256 3x3 convolutional filters of stride 1 and padding 1 to make sure the same image resolution of 32x32 is maintained.

What is stride, what is padding? How do I figure out these parameters?

Padding: extra rows and columns you add around the input (default is 0)
Stride: while applying convolution how many pixels you shift the convolution filters at a time. (default is 1)

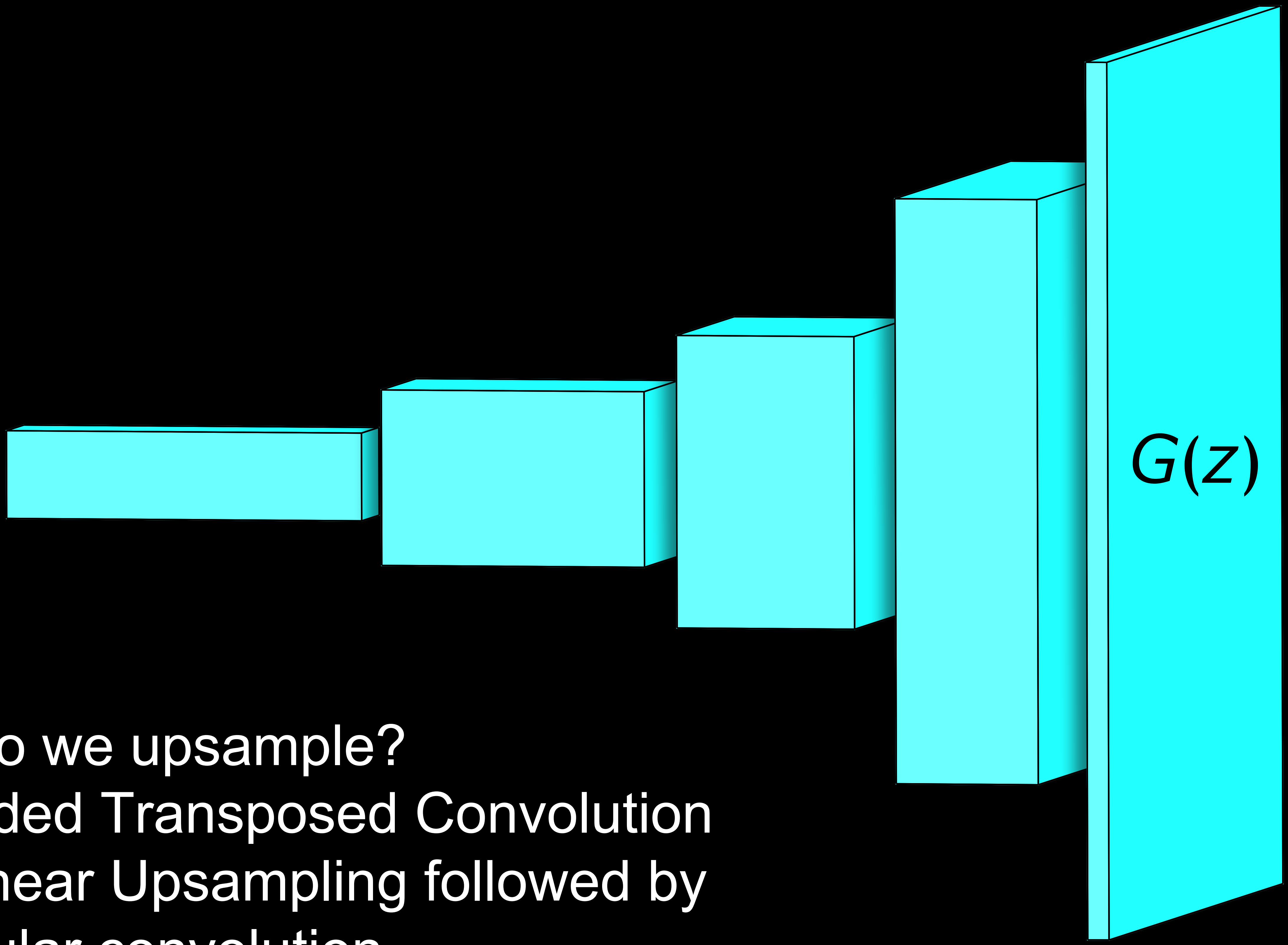
Shape:

- Input: $(N, C_{in}, H_{in}, W_{in})$ or (C_{in}, H_{in}, W_{in})
- Output: $(N, C_{out}, H_{out}, W_{out})$ or $(C_{out}, H_{out}, W_{out})$, where

$$H_{out} = \left\lfloor \frac{H_{in} + 2 \times \text{padding}[0] - \text{dilation}[0] \times (\text{kernel\_size}[0] - 1) - 1}{\text{stride}[0]} + 1 \right\rfloor$$

$$W_{out} = \left\lfloor \frac{W_{in} + 2 \times \text{padding}[1] - \text{dilation}[1] \times (\text{kernel\_size}[1] - 1) - 1}{\text{stride}[1]} + 1 \right\rfloor$$

Check out PyTorch nn.conv2d page for more details!



How do we upsample?

- Strided Transposed Convolution
- Bilinear Upsampling followed by regular convolution.

Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks

Alec Radford, Luke Metz and Soumith Chintala

Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks

Alec Radford, Luke Metz and Soumith Chintala

How do we edit images in DC-GAN?

Generative Adversarial Networks: Vector Math

Samples from the model

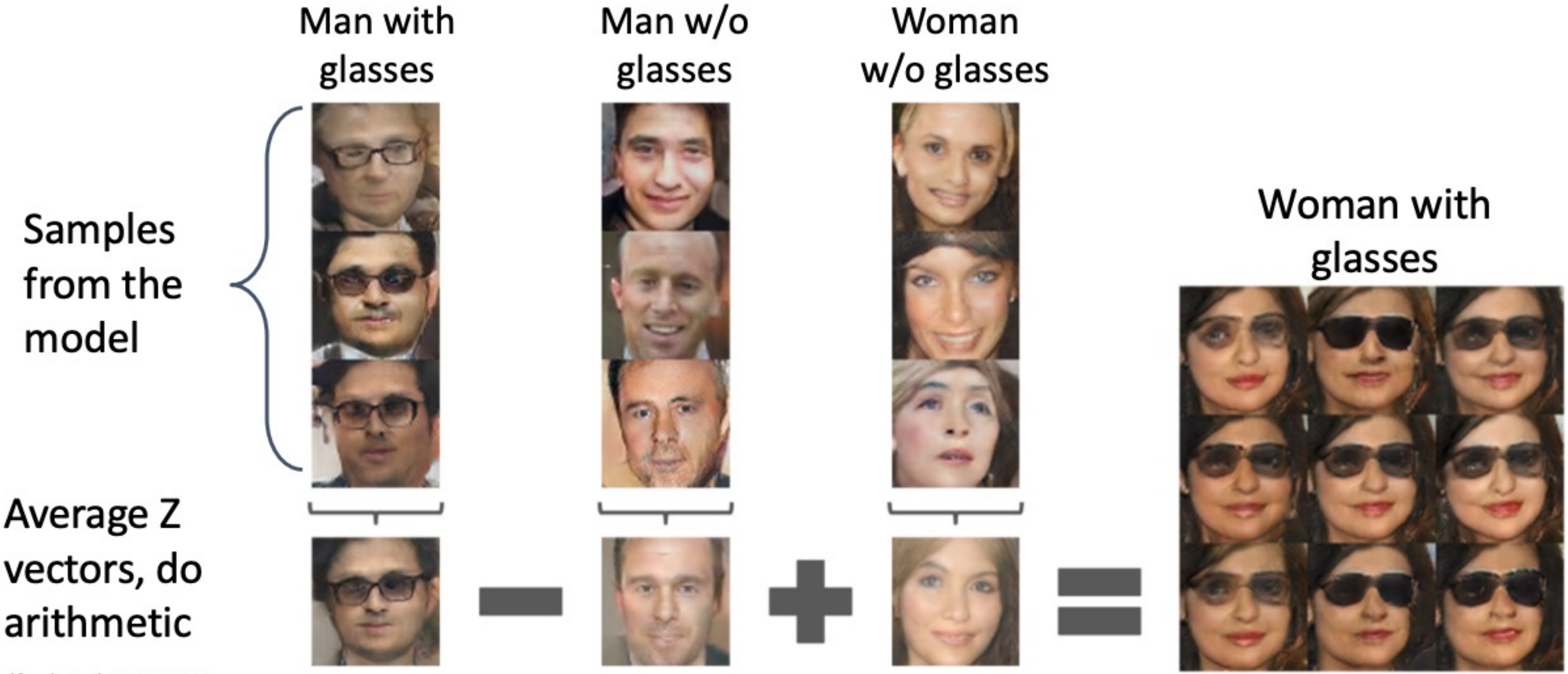
Smiling woman

Neutral woman

Neutral man

Smiling Man

Generative Adversarial Networks: Vector Math



Radford et al, ICLR 2016

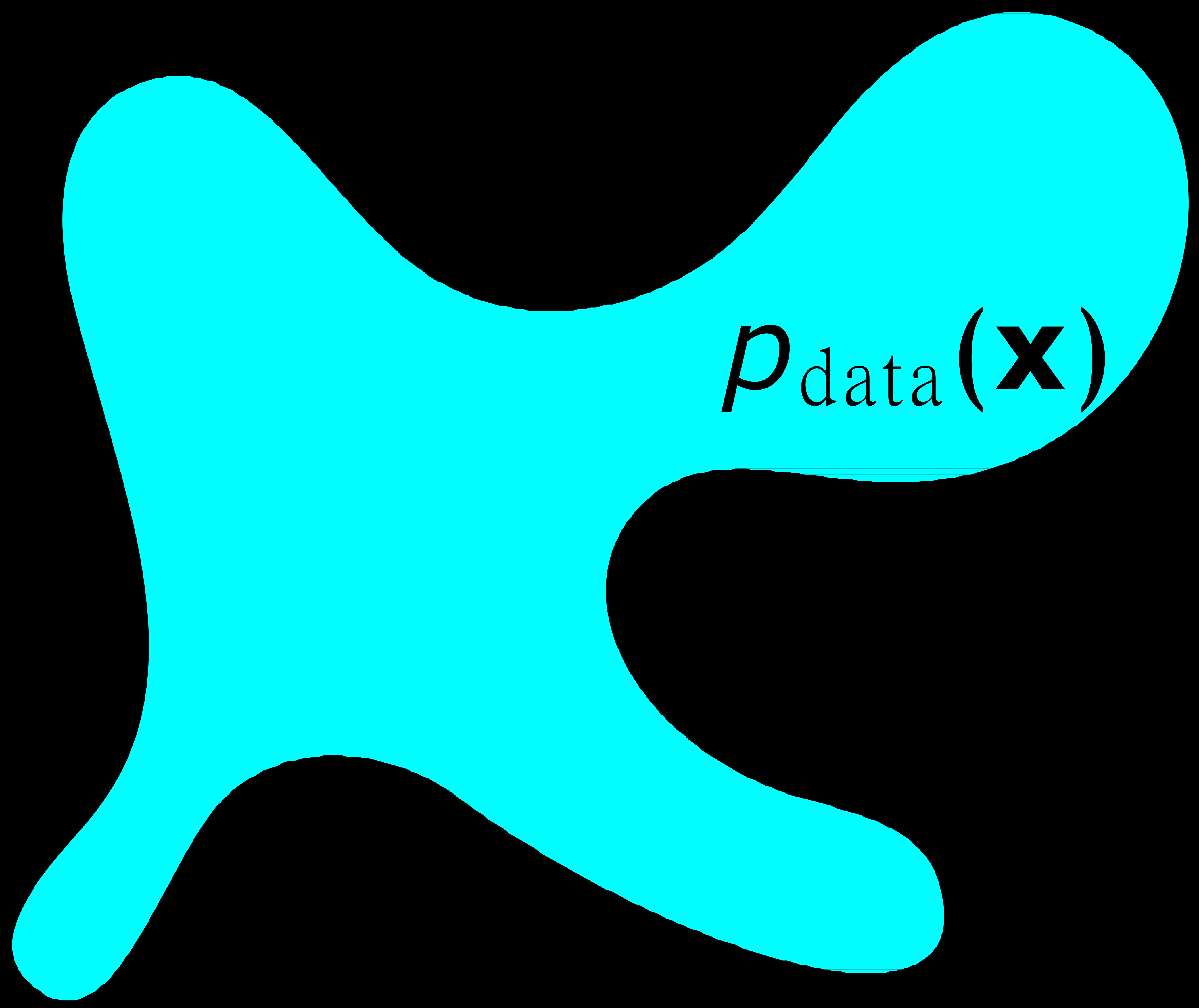
Today's class

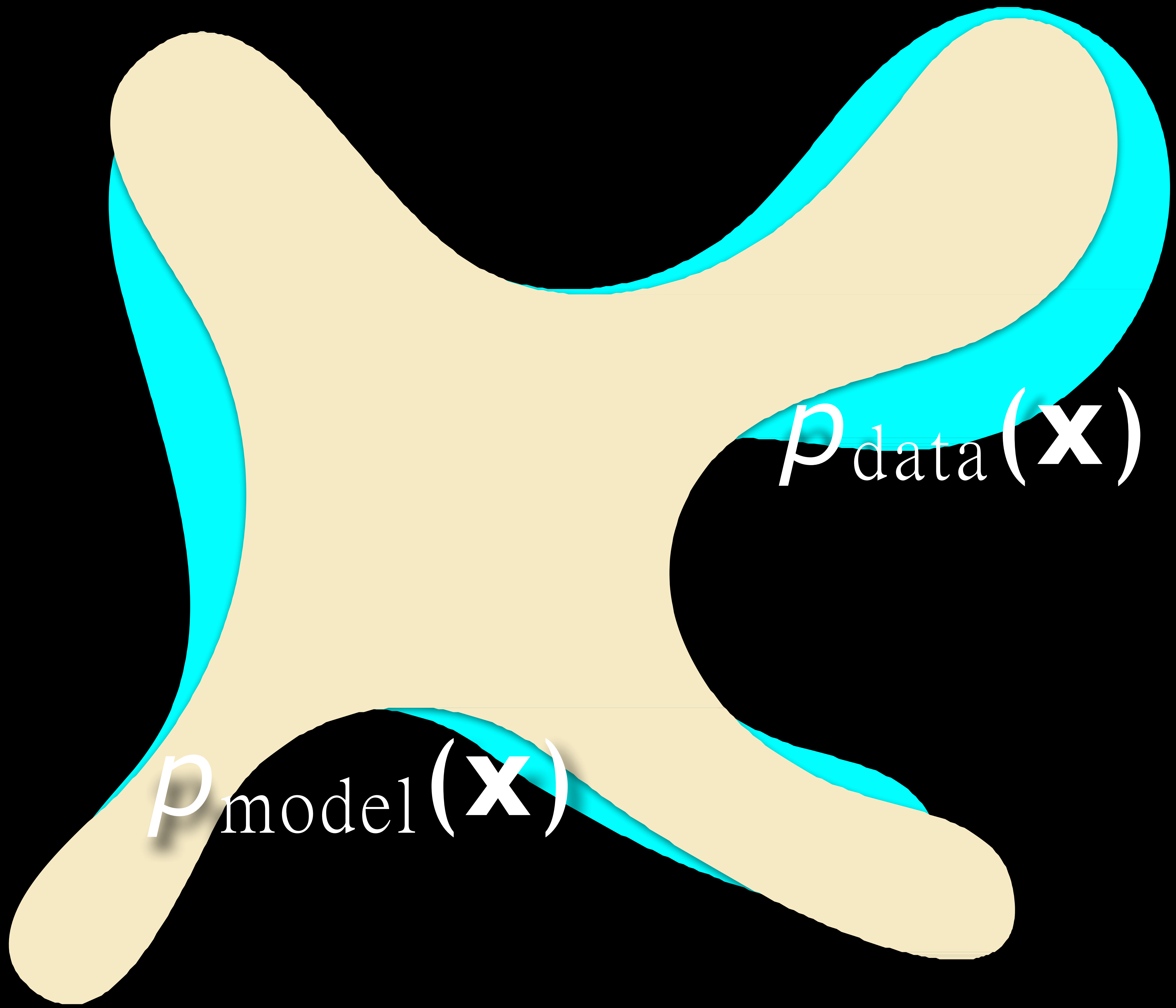
- Unconditional Image generation
    - DC-GAN
    - Wasserstein GAN
    - Progressive GAN
    - StyleGAN

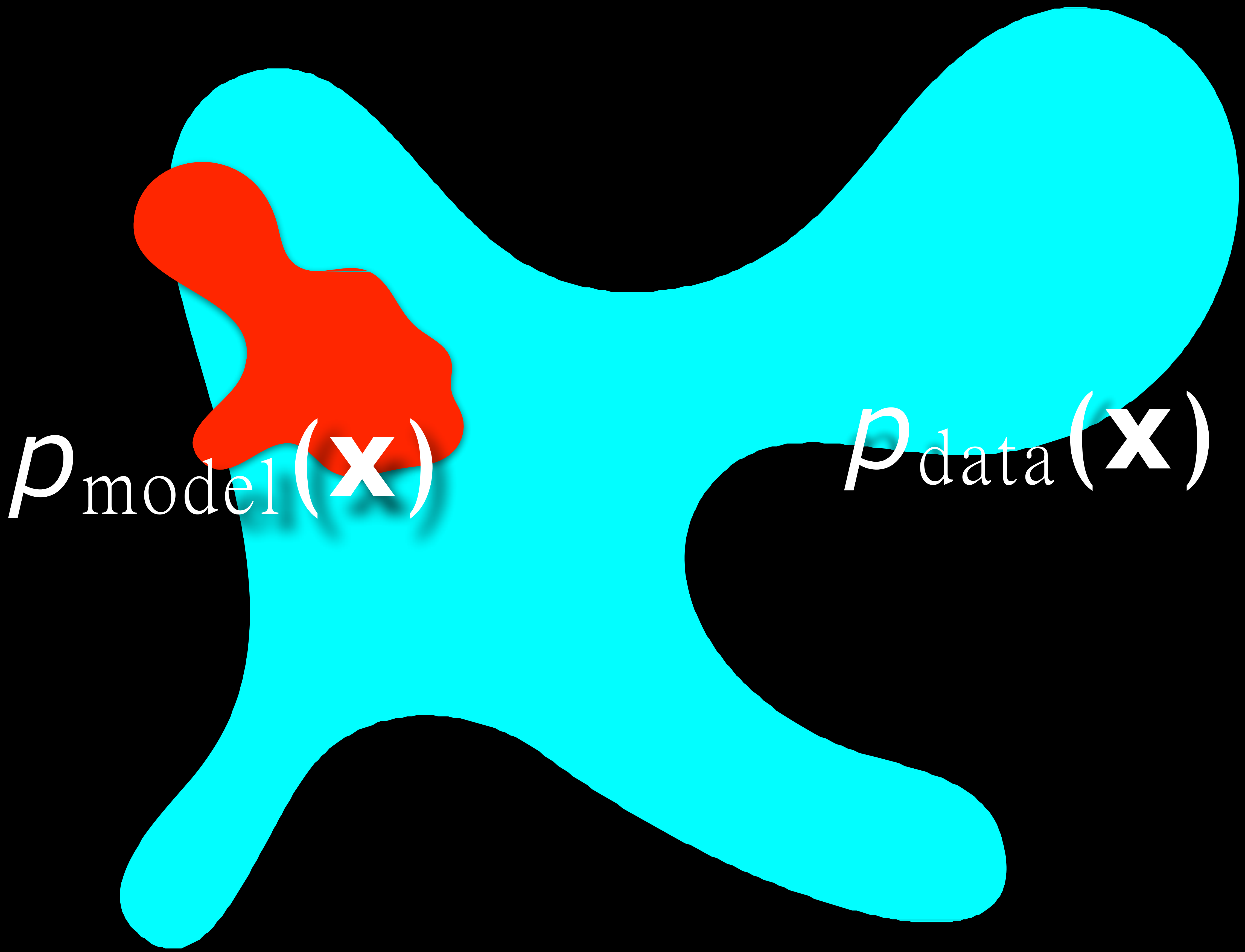
- Conditional Image generation
    - Class conditional (Big GAN)
    - Paired (Pix2Pix)
    - Unpaired (CycleGAN)

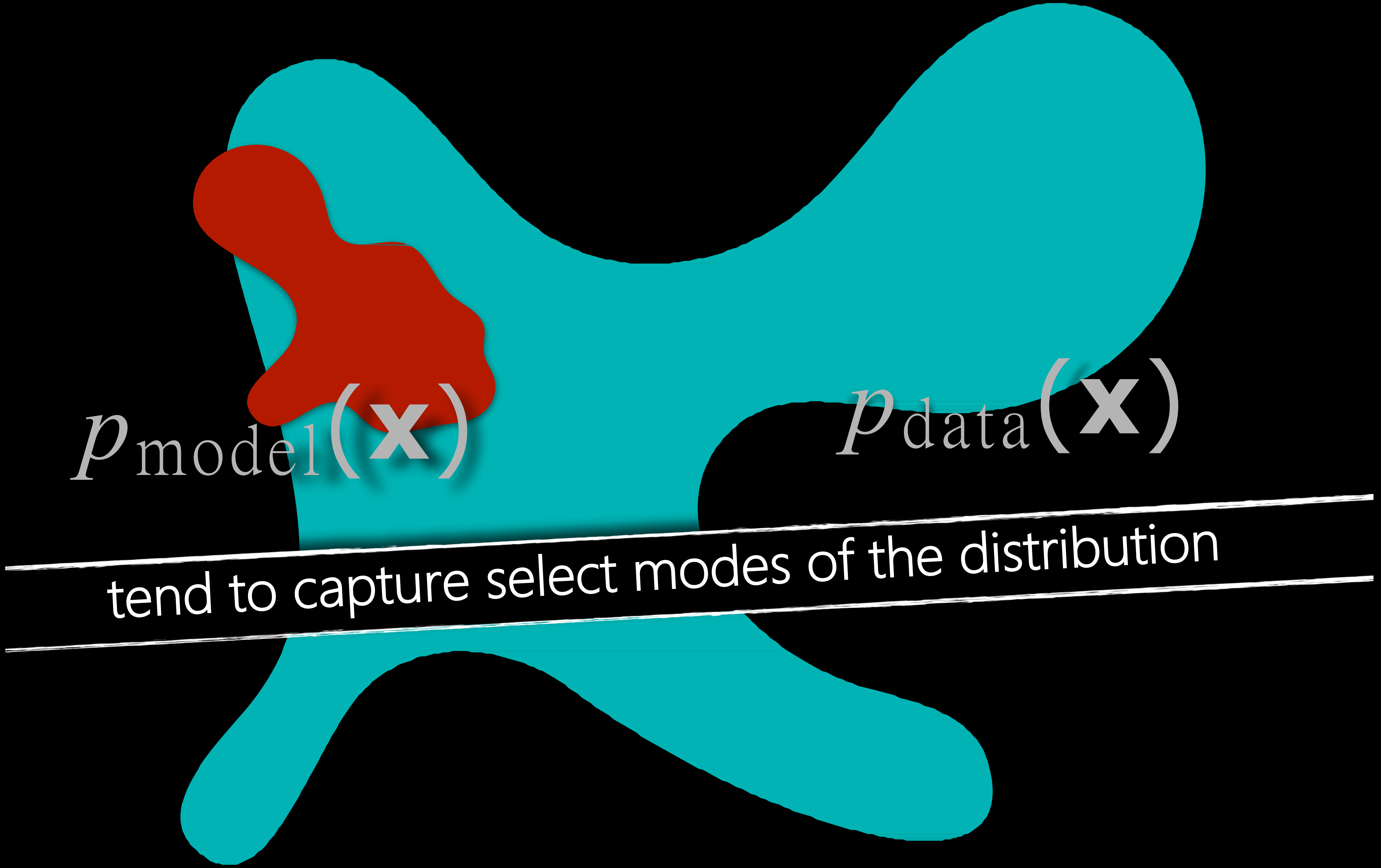
GAN training can be

UNSTABLE


$p_{\text{data}}(\mathbf{x})$


$p_{\text{data}}(\mathbf{x})$
$p_{\text{model}}(\mathbf{x})$

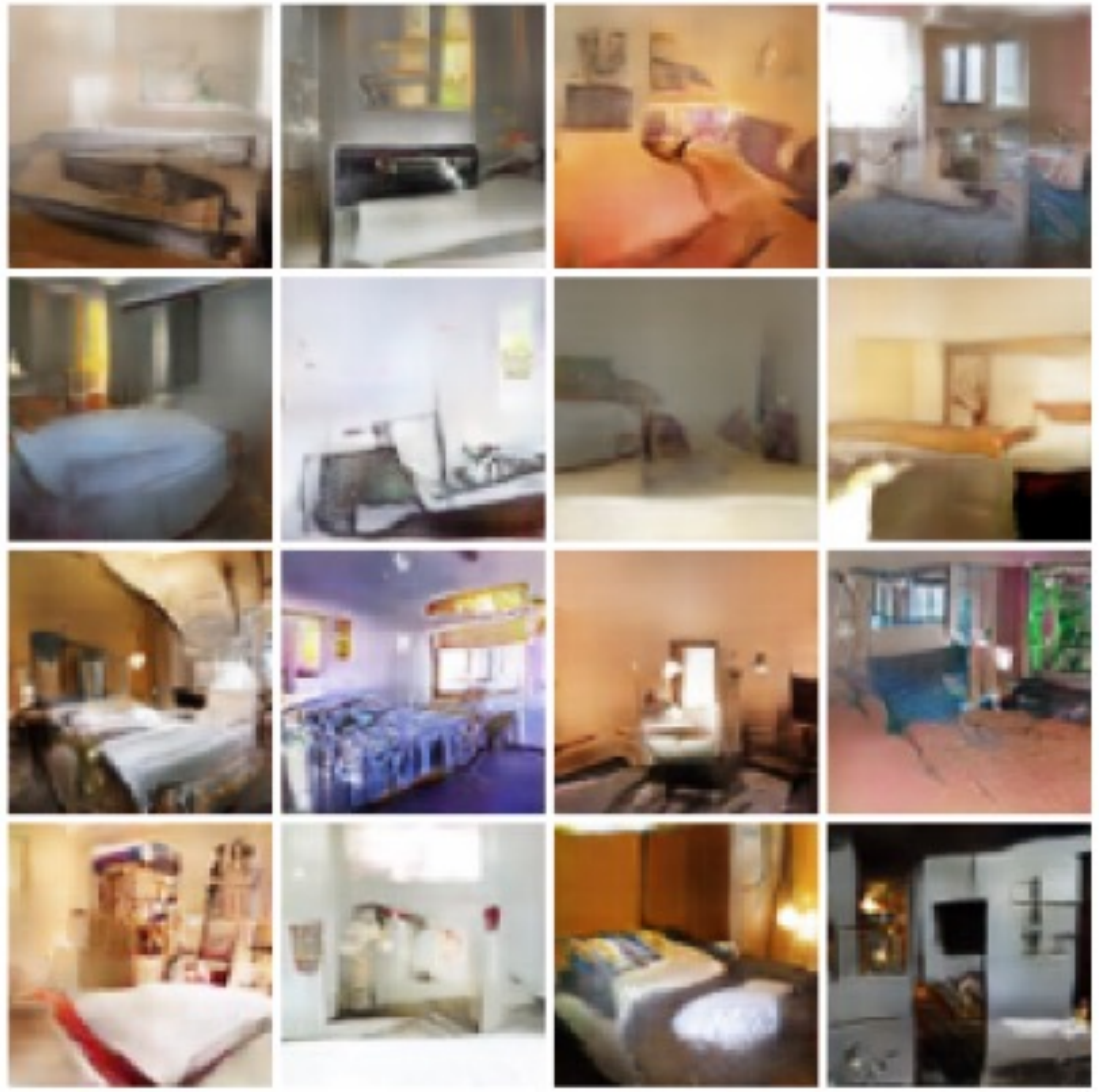

$p_{\text{model}}(\mathbf{x})$
$p_{\text{data}}(\mathbf{x})$



$p_{\text{model}}(\mathbf{x})$
$p_{\text{data}}(\mathbf{x})$
tend to capture select modes of the distribution

GAN Improvements: Improved Loss Functions

Wasserstein GAN (WGAN)



Arjovsky, Chintala, and Bouttou, "Wasserstein GAN", 2017

WGAN with Gradient Penalty (WGAN-GP)

Gulrajani et al, "Improved Training of Wasserstein GANs", NeurIPS 2017

	Discriminator/Critic	Generator
GAN	$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)\right]$	$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)$
WGAN	$\nabla_w \frac{1}{m} \sum_{i=1}^{m} \left[f\left(x^{(i)}\right) - f\left(G\left(z^{(i)}\right)\right)\right]$	$\nabla_\theta \frac{1}{m} \sum_{i=1}^{m} f\left(G\left(z^{(i)}\right)\right)$

WGAN:
Instead of classifying fake and real image,
simply minimize the discriminator score for real image
and maximize the discriminator score for fake images.

Instead of classifying all fake images as real,
simply minimize the discriminator score.

(remember we changed the sign of real=1 to
avoid vanishing gradient)

GAN Improvements: Higher Resolution

256 x 256 bedrooms

1024 x 1024 faces

Karras et al, "Progressive Growing of GANs for Improved Quality, Stability, and Variation", ICLR 2018

Today's class

- Unconditional Image generation
  - DC-GAN
  - Wasserstein GAN
  - Progressive GAN
  - StyleGAN

- Conditional Image generation
  - Class conditional (Big GAN)
  - Paired (Pix2Pix)
  - Unpaired (CycleGAN)

How do you generate 1024x1024 images without artifacts?

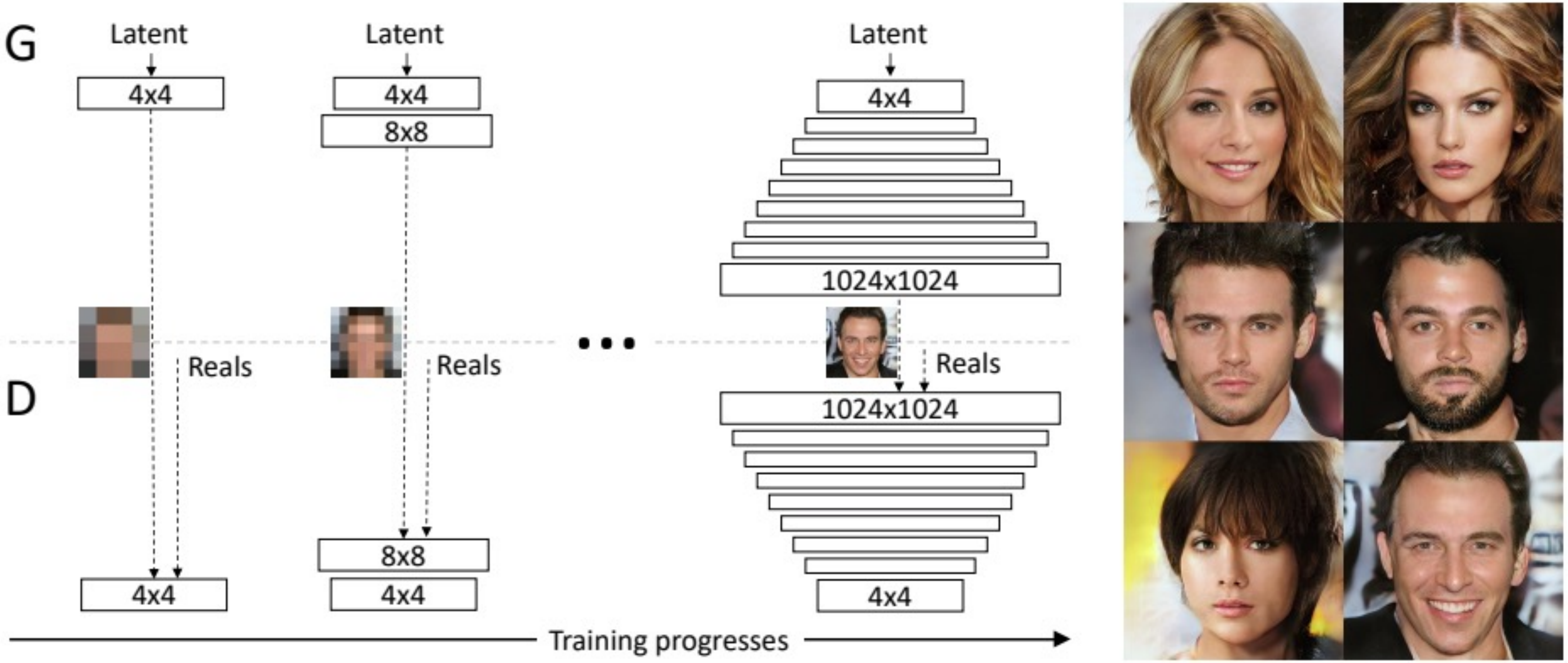
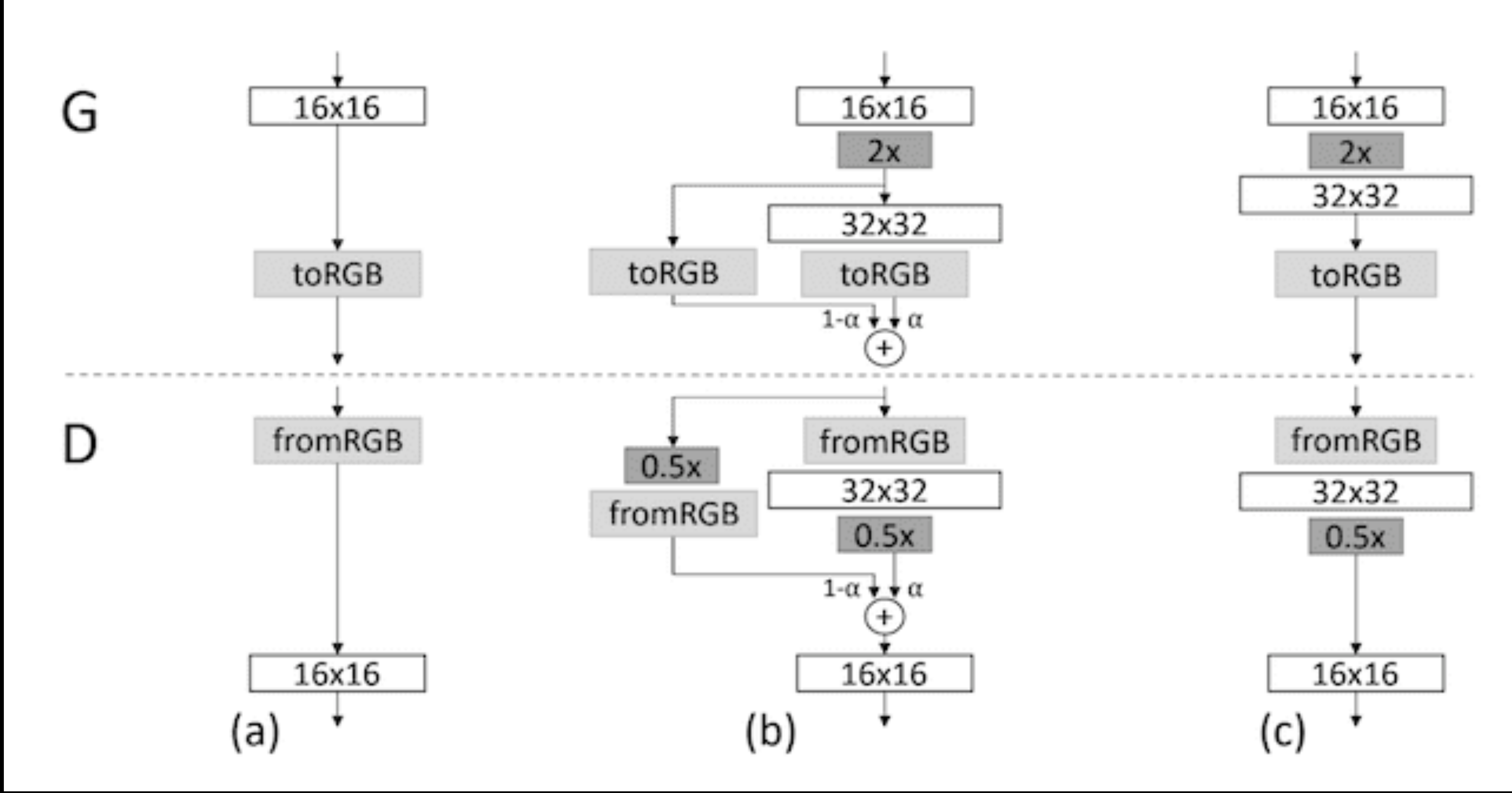


Figure 1: Our training starts with both the generator (G) and discriminator (D) having a low spatial resolution of 4×4 pixels. As the training advances, we incrementally add layers to G and D, thus increasing the spatial resolution of the generated images. All existing layers remain trainable throughout the process. Here $\boxed{N \times N}$ refers to convolutional layers operating on $N \times N$ spatial resolution. This allows stable synthesis in high resolutions and also speeds up training considerably. One the right we show six example images generated using progressive growing at 1024×1024.

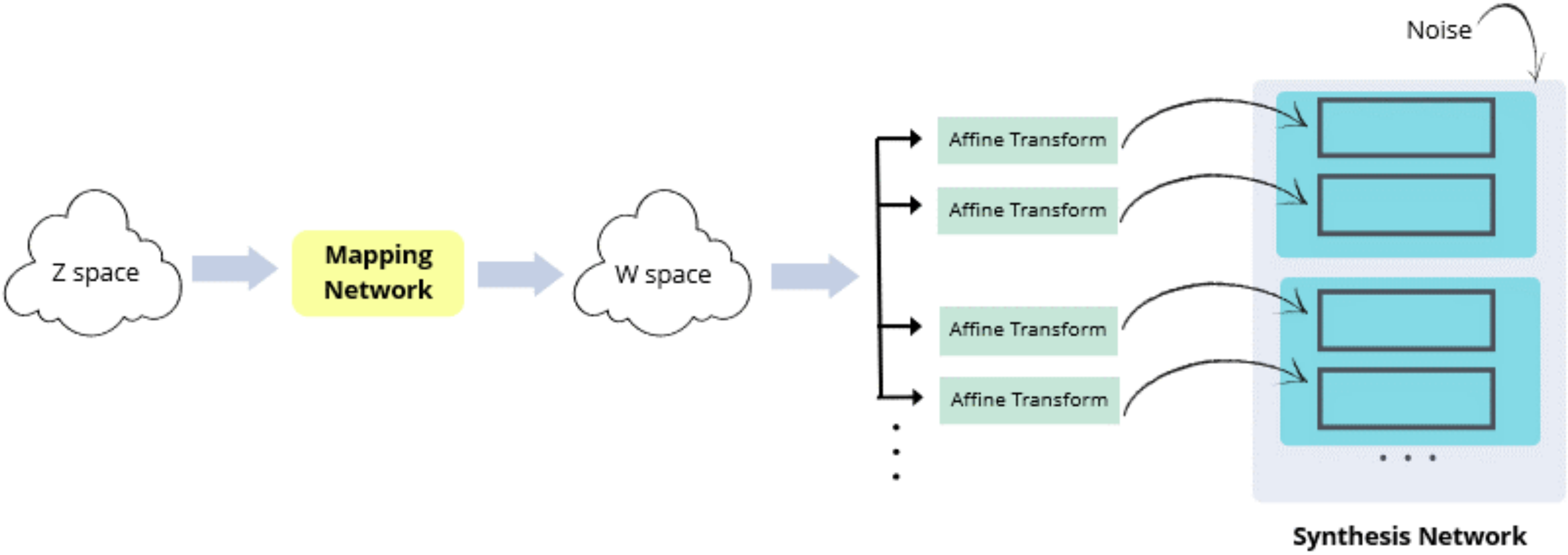


Challenge: Stability

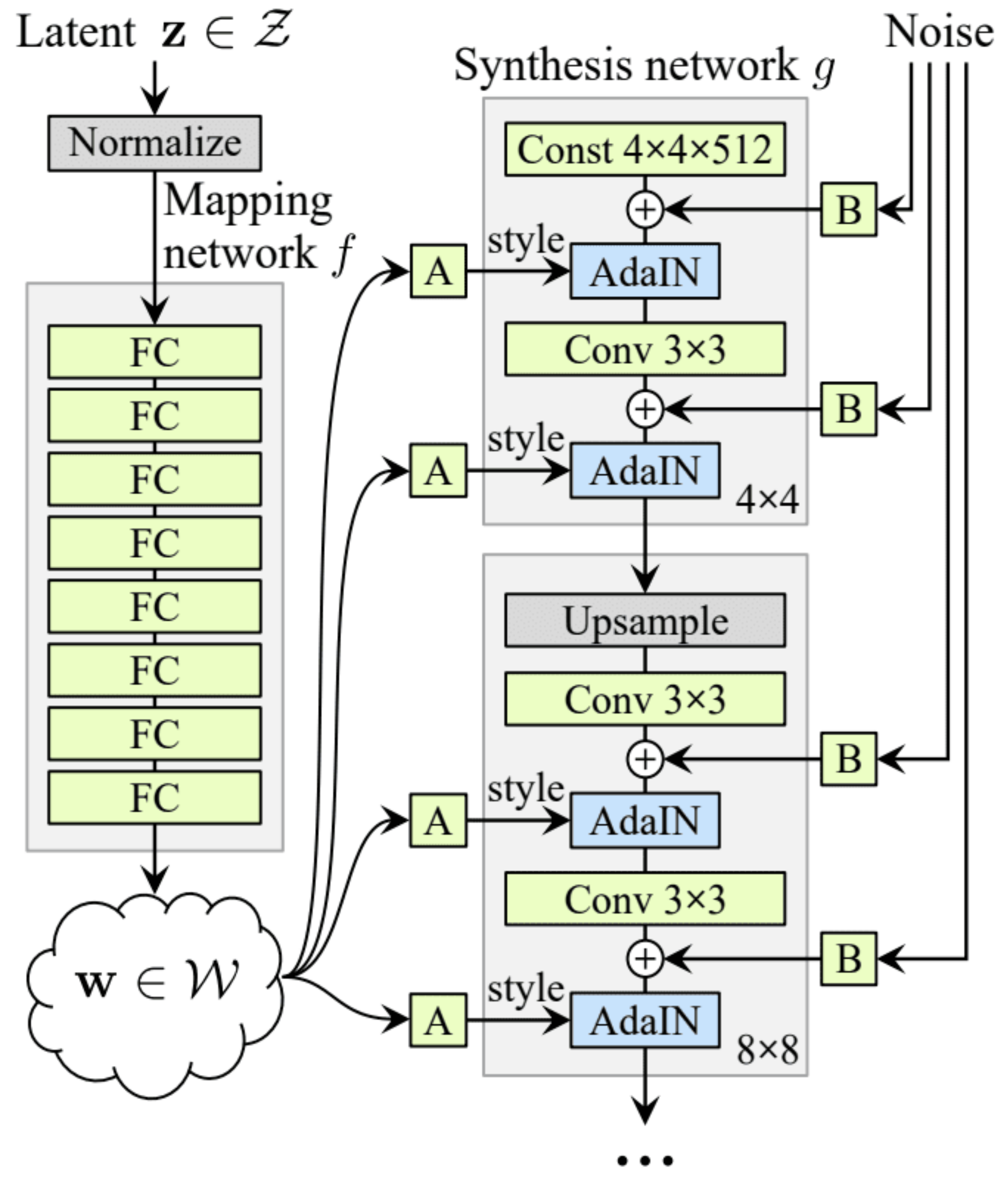
As training progress alpha is linearly changed from 0 to 1

Today's class

- Unconditional Image generation
  - DC-GAN
  - Wasserstein GAN
  - Progressive GAN
  - StyleGAN
- Conditional Image generation
  - Class conditional (Big GAN)
  - Paired (Pix2Pix)
  - Unpaired (CycleGAN)



Goal: Better disentanglement of features in latent space (W space)

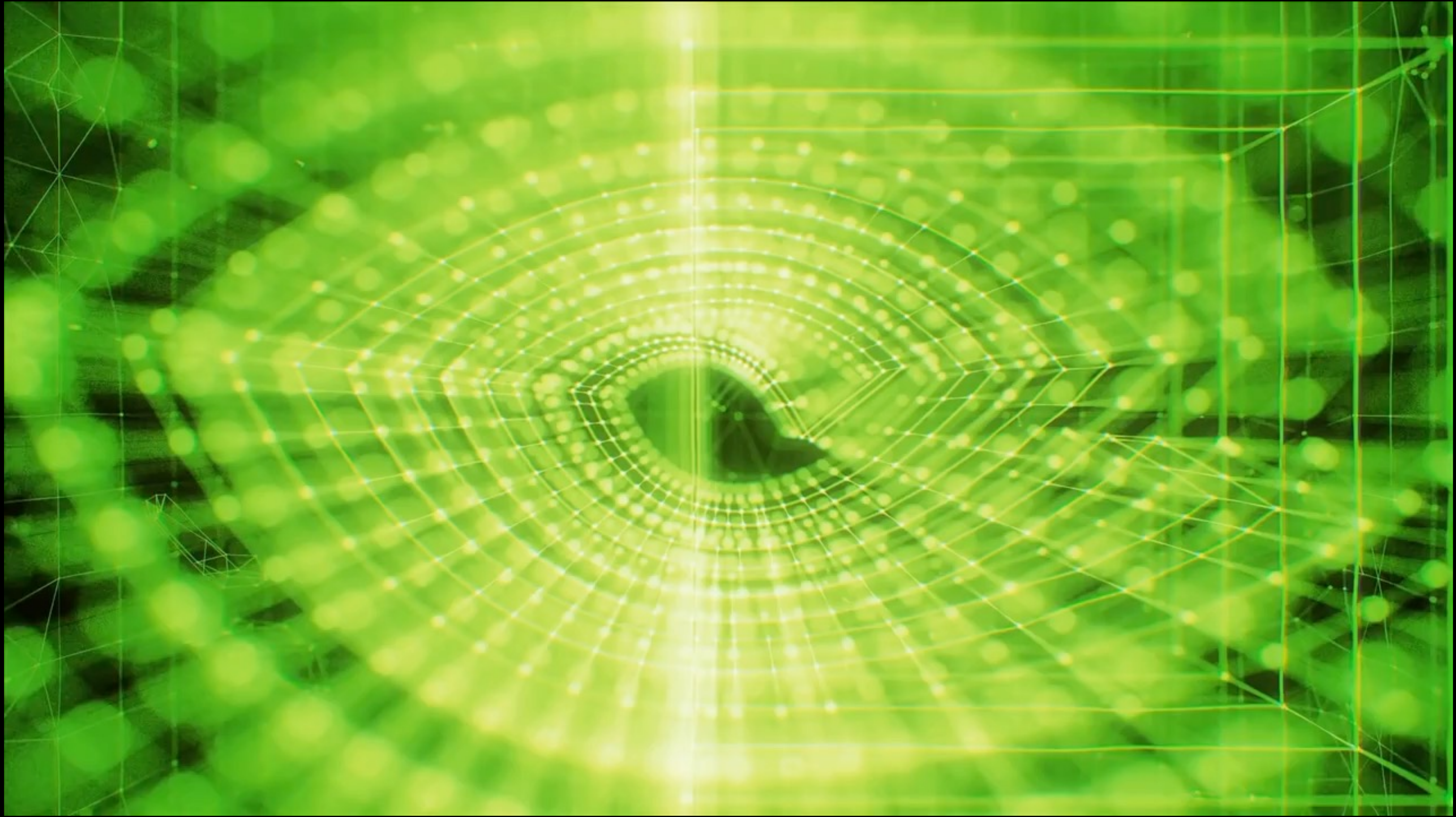


A = learned affine transformation block for AdaIN (predicts y)

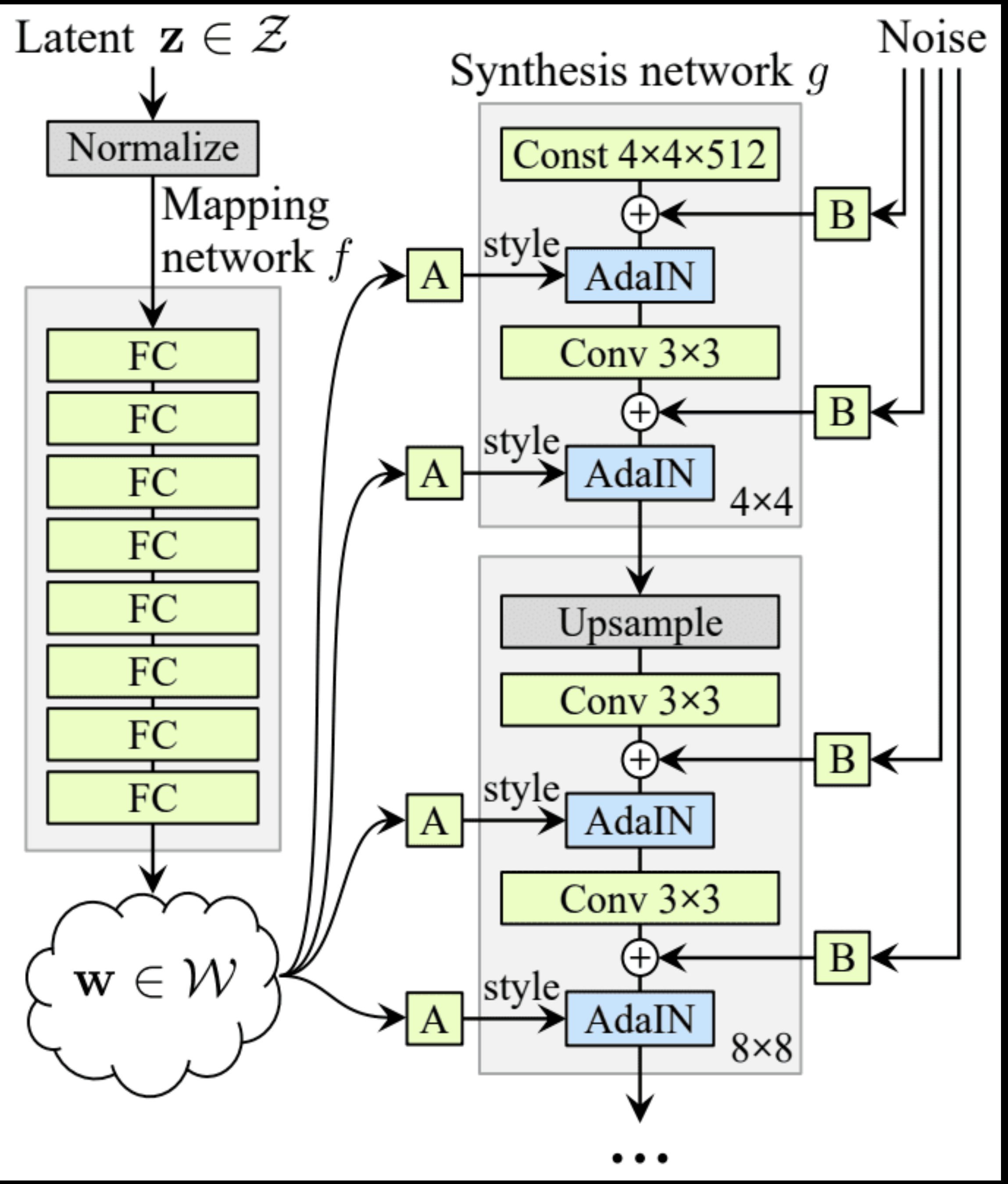
Adaptive Instance Normalization (very effective in controlling styles)

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

B = learned per-channel scaling factor for noise input.



Which latent space to choose for embedding and editing?



- Z : 512 dimensional latent space (not good)
- W : 512 dimensional latent space (better but not perfect)
- W+ : 18x512 dimensional latent space (after affine transformation A has been applied)

- W is better for editing.
- W+ is better for reconstruction or embedding of real images.

Want to know more about embedding in W and W+ space?
Read: Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?

We have 4 papers where we will learn how to embed images in StyleGAN latent space and how to edit these images.

Tue Sept 13	Pivotal Tuning for Latent-based Editing of Real Images.	William Stanford
	Third Time's the Charm? Image and Video Editing with StyleGAN3.	Nurislam Tursynbek
Thrs Sept 15	CLIP2StyleGAN: Unsupervised Extraction of StyleGAN Edit Directions.	Sam Ehrenstein
	DyStyle: Dynamic Neural Network for Multi-Attribute-Conditioned Style Editing.	Qiwei Zhao

590: Next Assignment will be about using StyleGAN inversion and editing on your images!

Additional Reading:

- https://towardsdatascience.com/explained-a-style-based-generator-architecture-for-gans-generating-and-tuning-realistic-6cb2be0f431
- https://jonathan-hui.medium.com/gan-stylegan-stylegan2-479bdf256299
- Next steps
- StyleGAN2
- StyleGAN3
- StyleGAN-ADA

Future research potential:

- StyleGAN allows detailed editing of faces
- Diffusion model still lacks the ability to perform fine-grained editing.
- But Diffusion models can produce images with more diversity!
- Faces in Diffusion models often look horrible!
- Can we somehow merge the best of the both?

Today's class

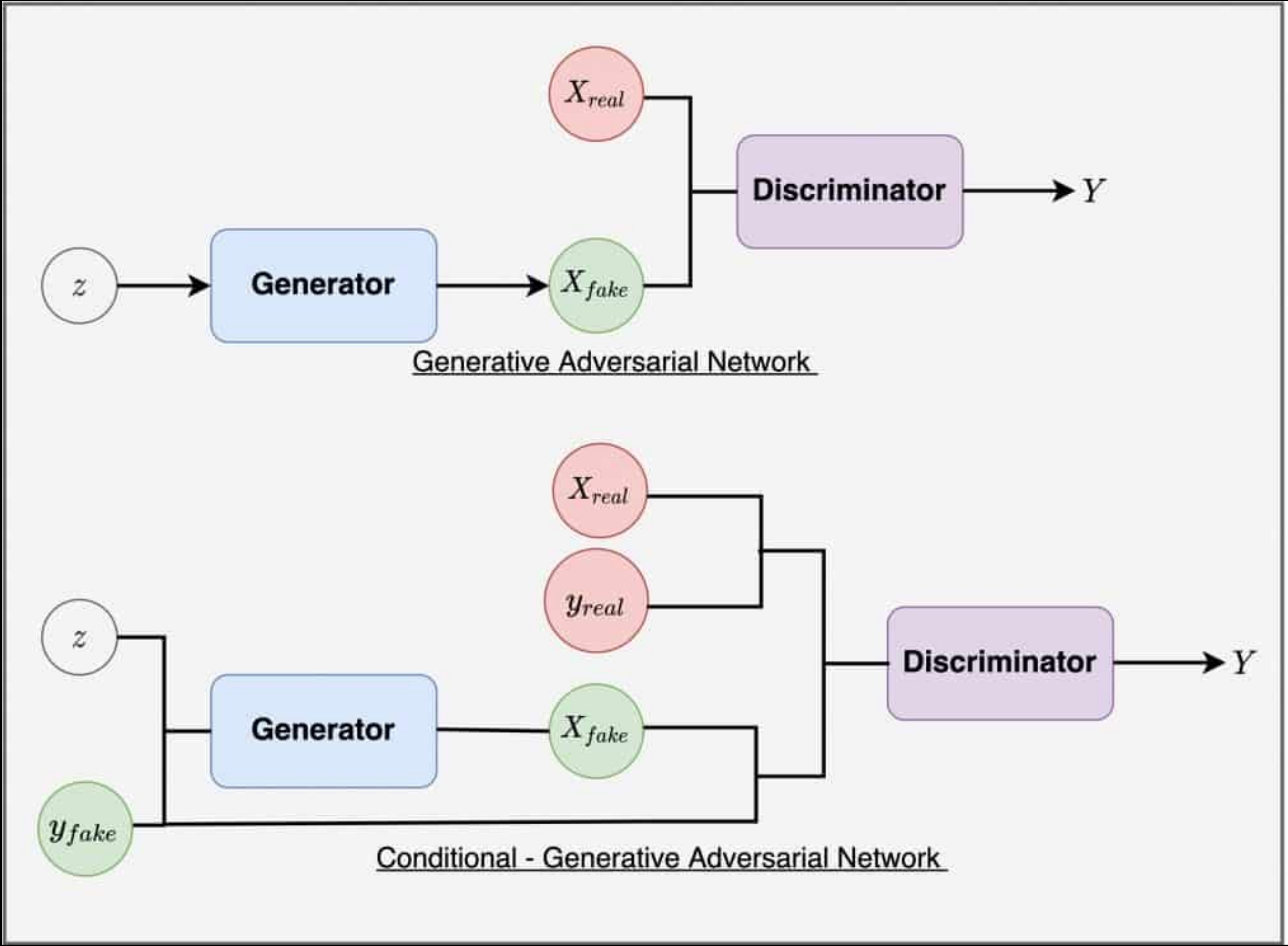
- Unconditional Image generation
  - DC-GAN
  - Wasserstein GAN
  - Progressive GAN
  - StyleGAN
- Conditional Image generation
  - Class conditional (Big GAN)
  - Paired (Pix2Pix)
  - Unpaired (CycleGAN)

Conditional GAN

Today's class

- Unconditional Image generation
  - DC-GAN
  - Wasserstein GAN
  - Progressive GAN
  - StyleGAN

- Conditional Image generation
  - Class conditional (Big GAN)
  - Paired (Pix2Pix)
  - Unpaired (CycleGAN)



X_{real}
Discriminator
Y
z
Generator
X_{fake}
Generative Adversarial Network
X_{real}
y_{real}
z
Discriminator
Y
Generator
X_{fake}
y_{fake}
Conditional - Generative Adversarial Network

Conditional GANs: Conditional Batch Normalization

Batch Normalization

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{i,j}$$

$$\sigma_j^2 = \frac{1}{N}\sum_{i=1}^{N} (x_{i,j} - \mu_j)^2$$

$$\hat{x}_{i,j} = \frac{x_{i,j} - \mu_j}{\sqrt{\sigma_j^2 + \epsilon}}$$

$$y_{i,j} = \gamma_j \hat{x}_{i,j} + \beta_j$$

Learn a separate scale and shift for each different label y

Conditional Batch Normalization

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{i,j}$$

$$\sigma_j^2 = \frac{1}{N}\sum_{i=1}^{N} (x_{i,j} - \mu_j)^2$$

$$\hat{x}_{i,j} = \frac{x_{i,j} - \mu_j}{\sqrt{\sigma_j^2 + \epsilon}}$$

$$y_{i,j} = \boldsymbol{\gamma_j^y} \hat{x}_{i,j} + \boldsymbol{\beta_j^y}$$

Dumoulin et al, "A learned representation for artistic style", ICLR 2017

Similar in idea with AdaIN
Except this is batch normalization vs instance normalization!

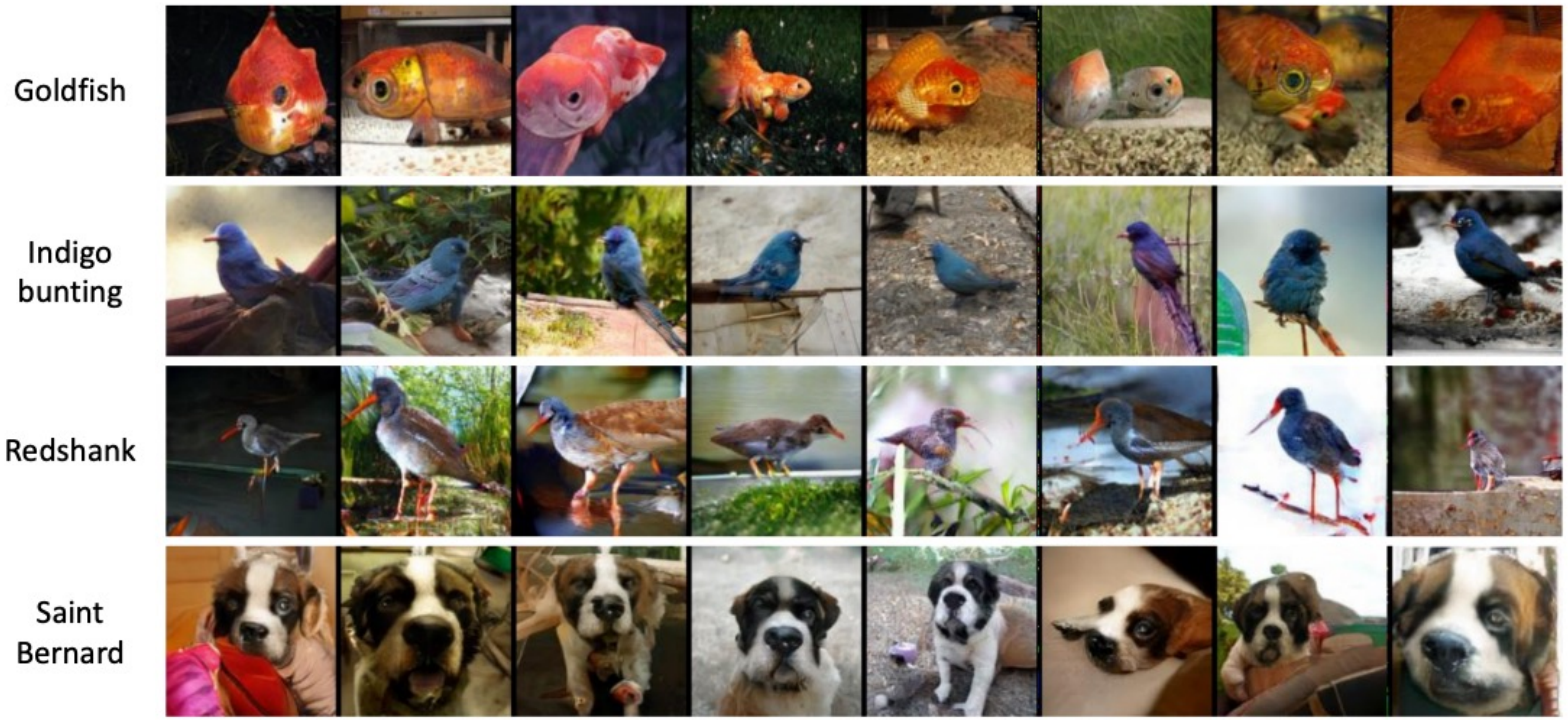
Conditional GANs: BigGAN

Brock et al, "Large Scale GAN Training for High Fidelity Natural Image Synthesis", ICLR 2019

512x512 images on ImageNet

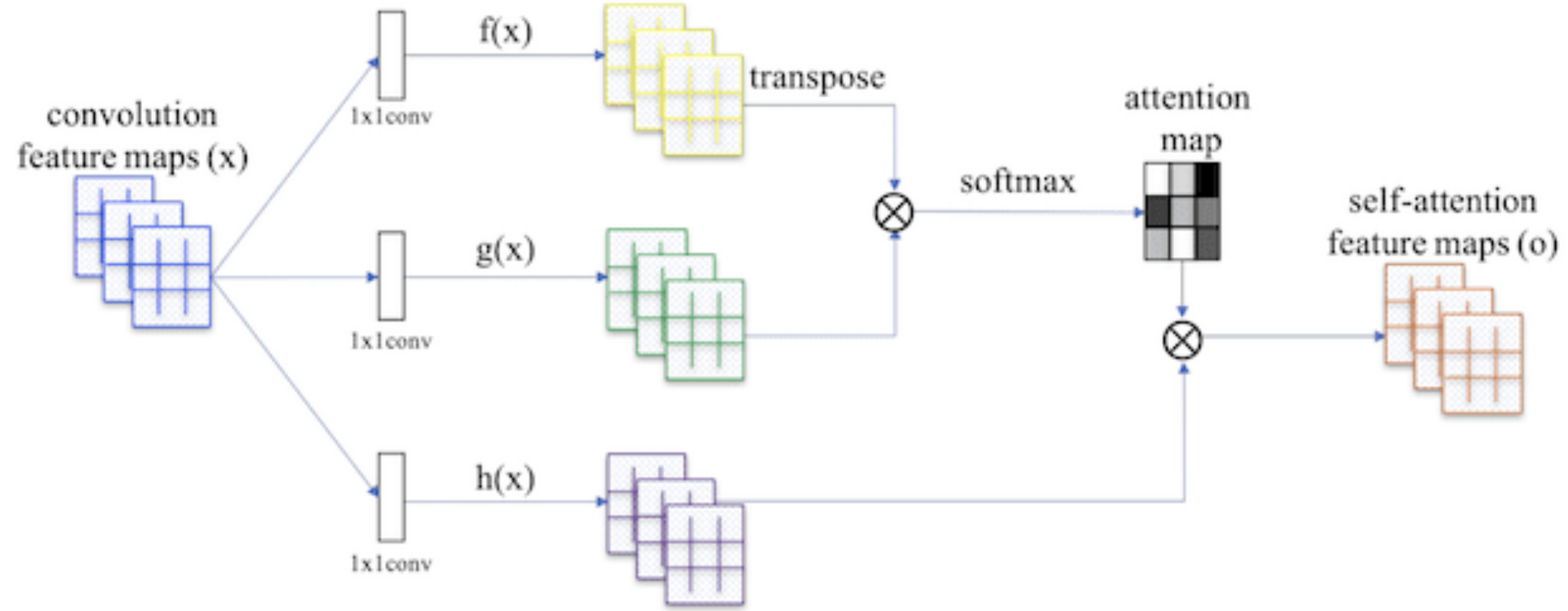
- Image generation is conditioned on input class from ImageNet
- Includes Self-attention module
- Many engineering changes:
  - Update discriminator more than generator
  - Larger batch size and more model parameters

Conditional GANs: Self-Attention



Zhang et al, "Self-Attention Generative Adversarial Networks", ICML 2019

128x128 images on ImageNet



Conditioning on more than labels! Text to Image

This bird is red and brown in color, with a stubby beak

The bird is short and stubby with yellow on its body

A bird with a medium orange bill white body gray wings and webbed feet

This small black bird has a short, slightly curved bill and long legs

A picture of a very clean living room

A group of people on skis stand in the snow

Eggs fruit candy nuts and meat served on white dish

A street sign on a stoplight pole in the middle of a day

Zhang et al, "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks.", TPAMI 2018
Zhang et al, "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks.", ICCV 2017
Reed et al, "Generative Adversarial Text-to-Image Synthesis", ICML 2016

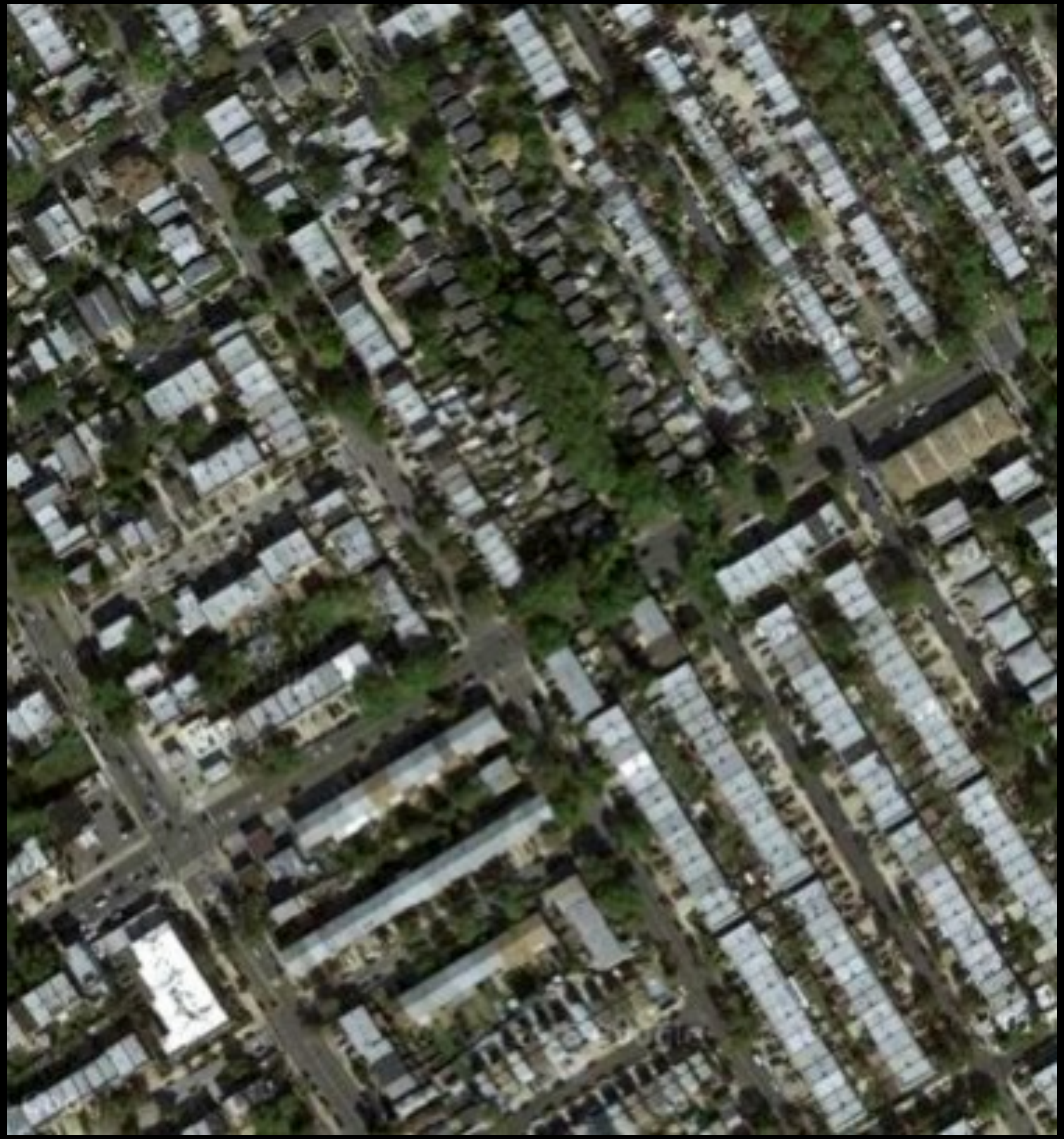
Today’s class

- Unconditional Image generation
  - DC-GAN
  - Wasserstein GAN
  - Progressive GAN
  - StyleGAN

- Conditional Image generation
  - Class conditional (Big GAN)
  - Paired (Pix2Pix)
  - Unpaired (CycleGAN)

input: edges

output: image



input: satellite view

output: map

Colorful Image Colorization

Richard Zhang, Phillip Isola and Alexei Efros

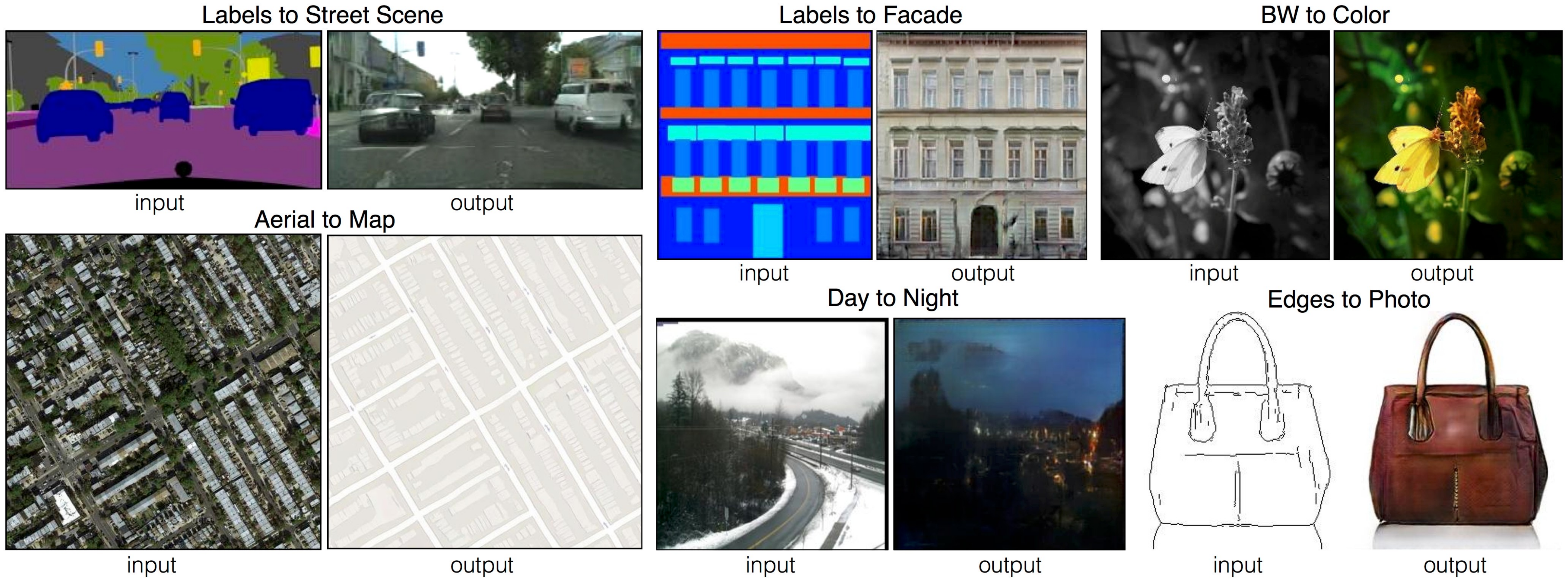

colorize-it.com

Image-to-Image Translation with Conditional Adversarial Networks

Phillip Isola Jun-Yan Zhu Tinghui Zhou Alexei A. Efros

Berkeley AI Research (BAIR) Laboratory, UC Berkeley

{isola,junyanz,tinghuiz,efros}@eecs.berkeley.edu



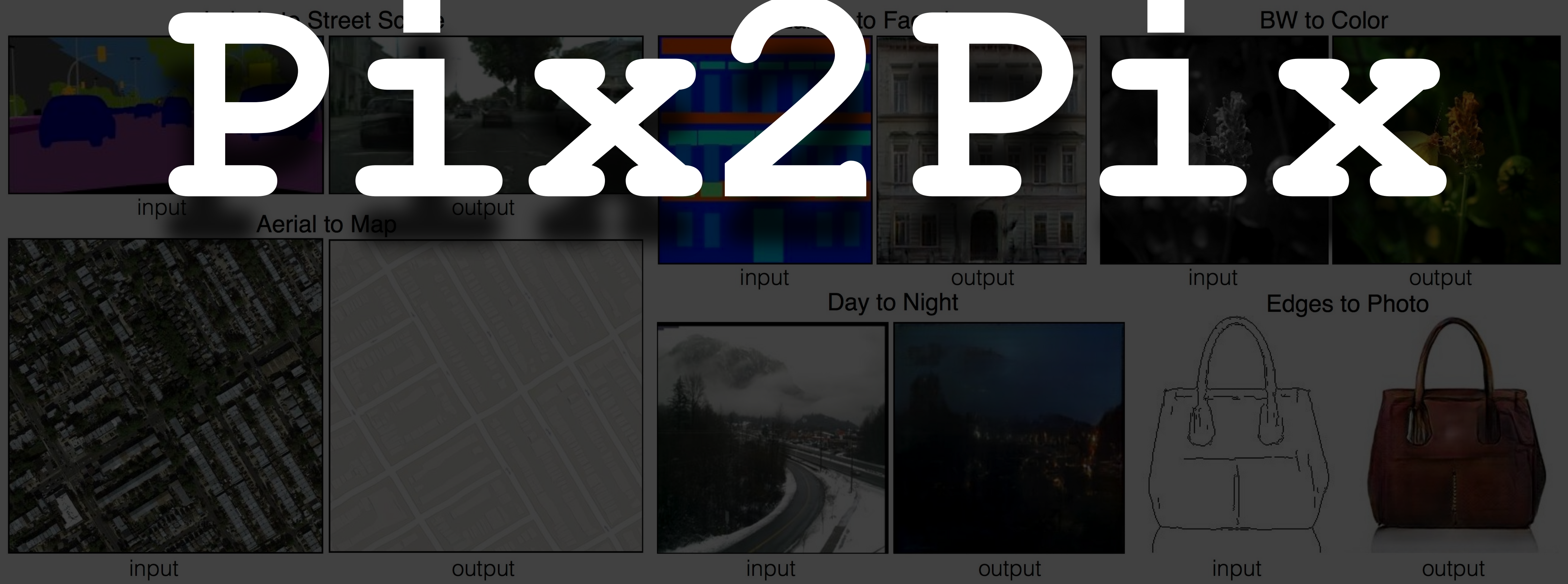
Pix2Pix

Image-to-Image Translation with Conditional Adversarial Networks

Phillip Isola Jun-Yan Zhu Tinghui Zhou Alexei A. Efros

Berkeley AI Research (BAIR) Laboratory, UC Berkeley

{isola,junyanz,tinghuiz,efros}@eecs.berkeley.edu



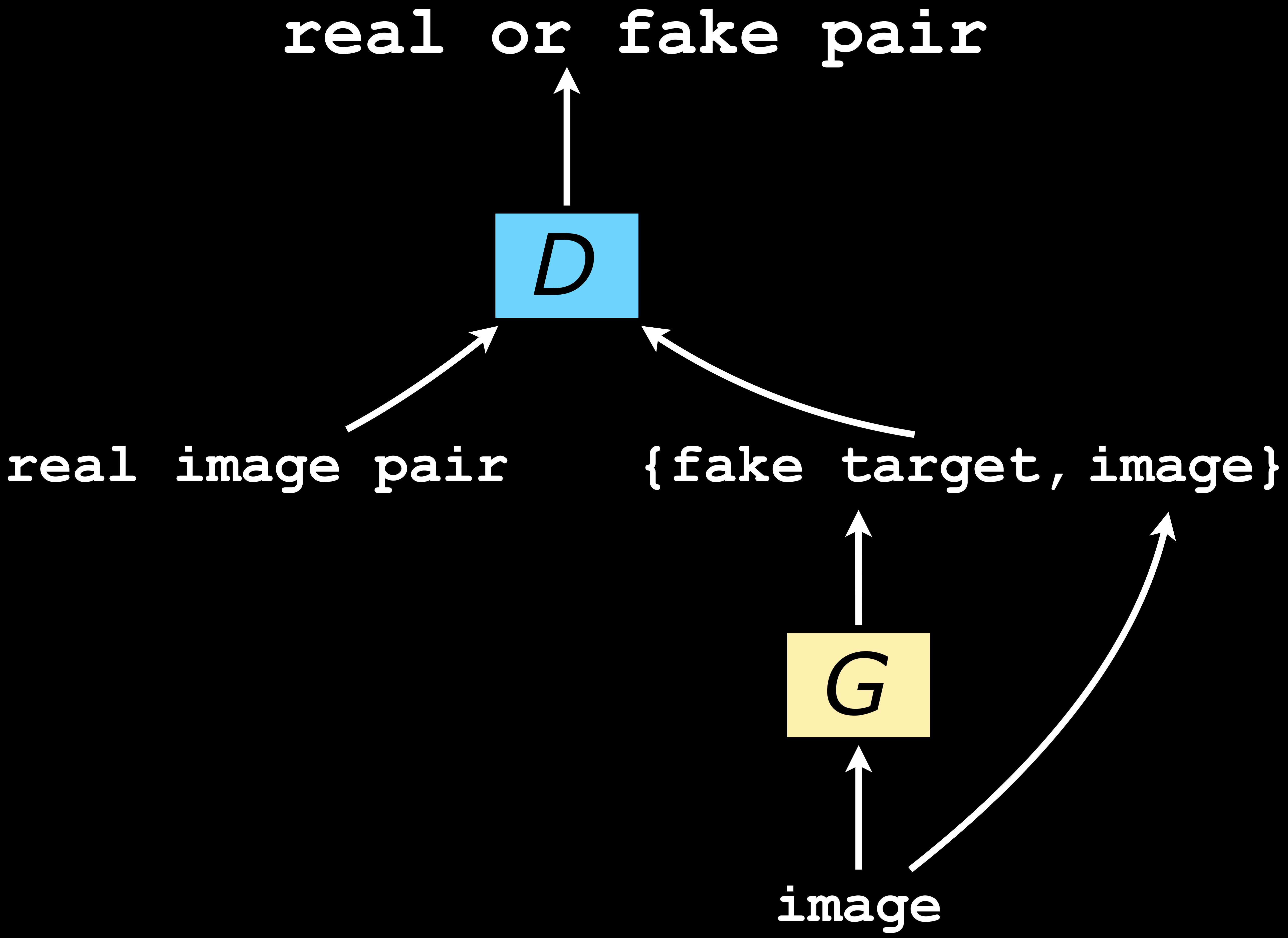
assumption

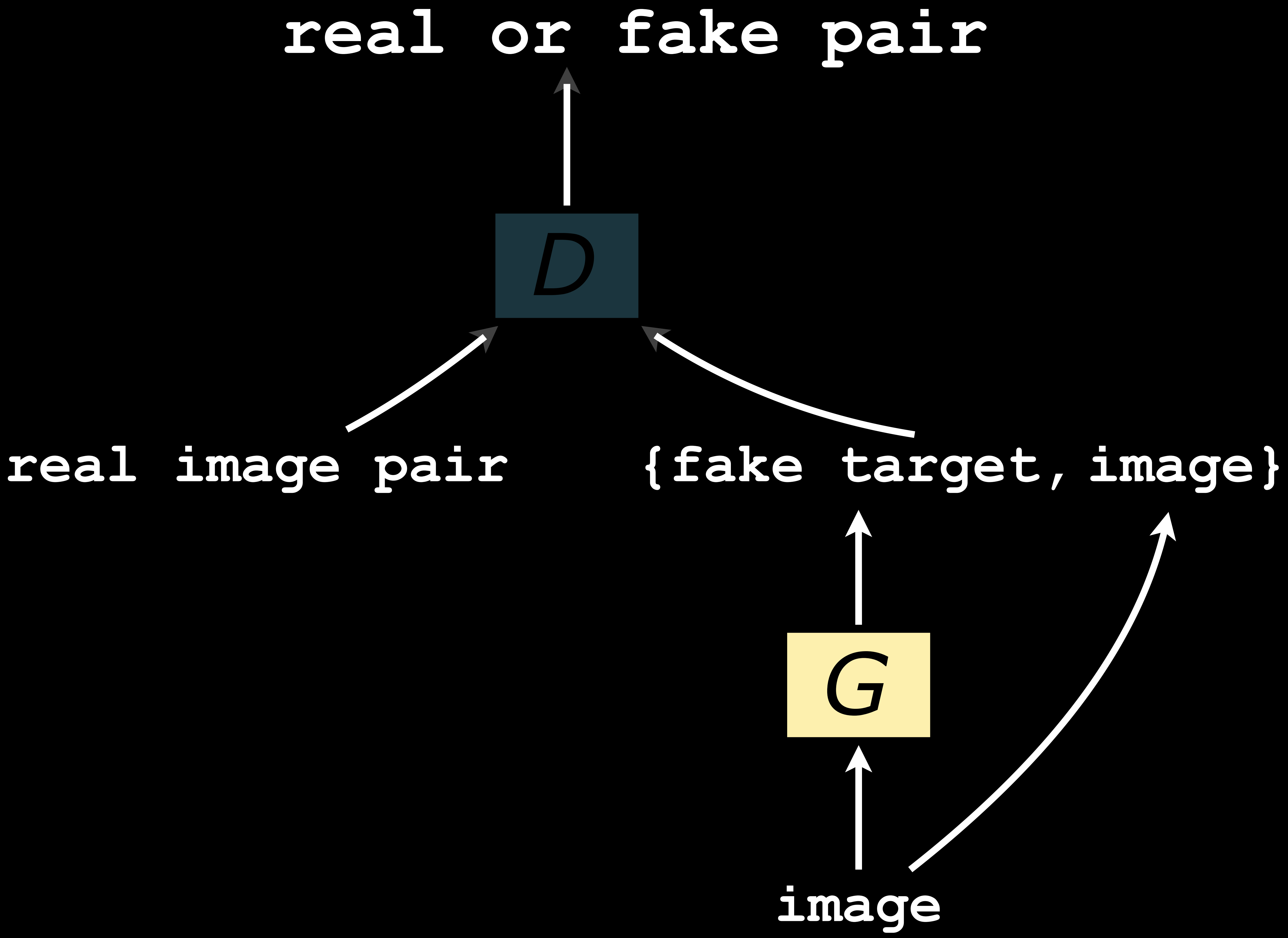
input: edges

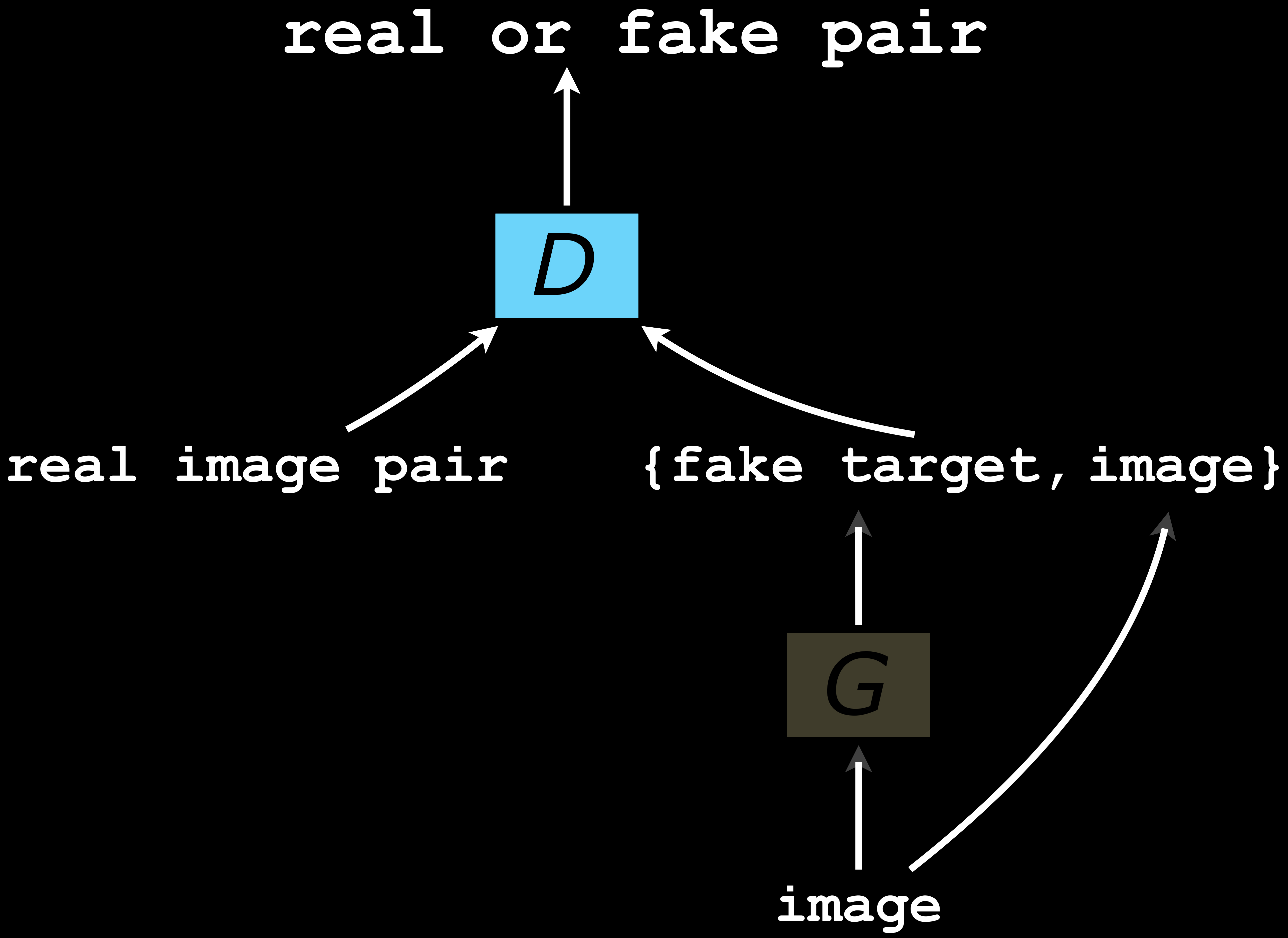
output: image

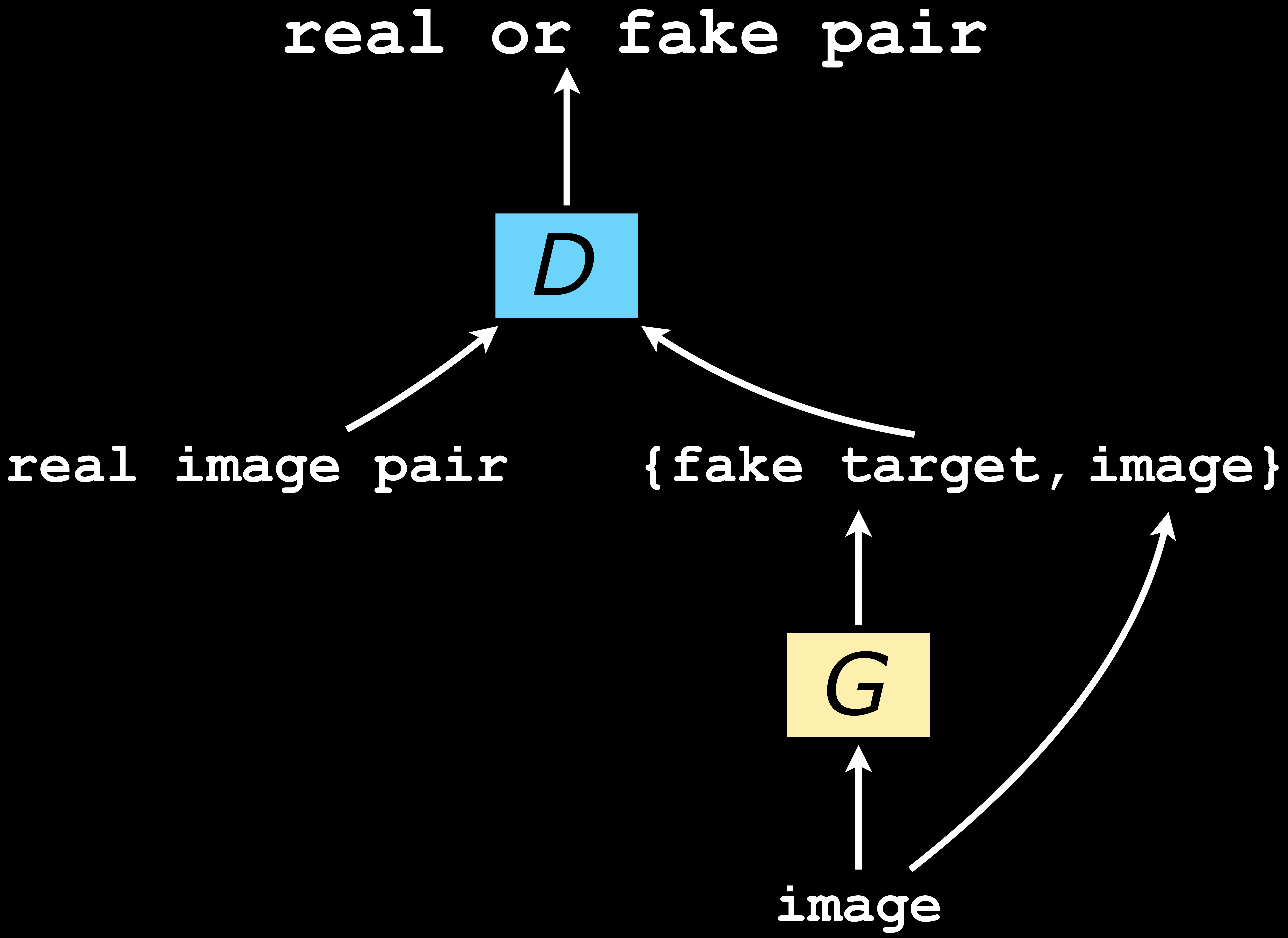
Training data consists of such pairs.

assumption


real or fake pair
D
real image pair
{fake target, image}
G
image


real or fake pair
D
real image pair
{fake target, image}
G
image


real or fake pair
D
real image pair
{fake target, image}
G
image



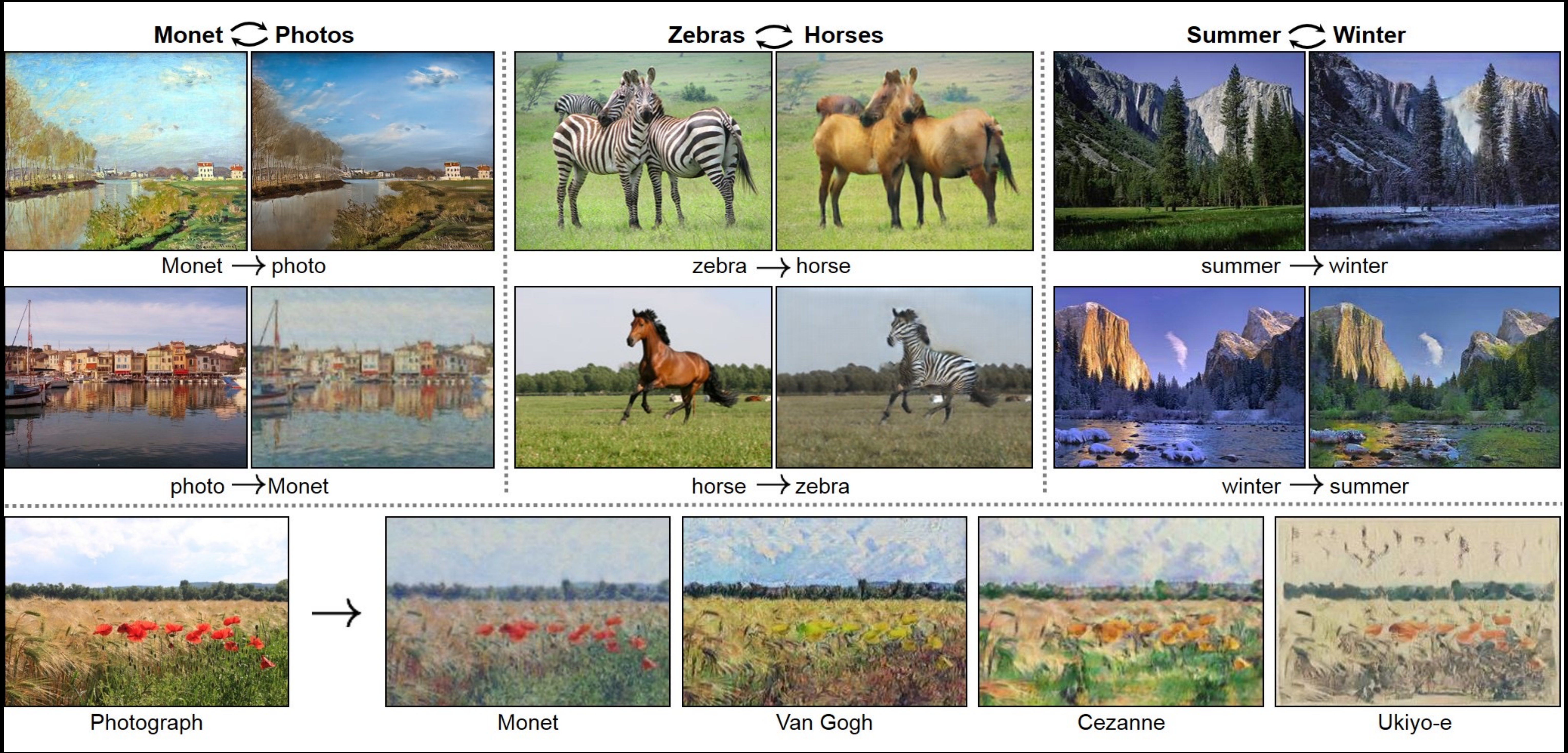
real or fake pair
D
real image pair
{fake target, image}
G
image

Today's class

- Unconditional Image generation
  - DC-GAN
  - Wasserstein GAN
  - Progressive GAN
  - StyleGAN

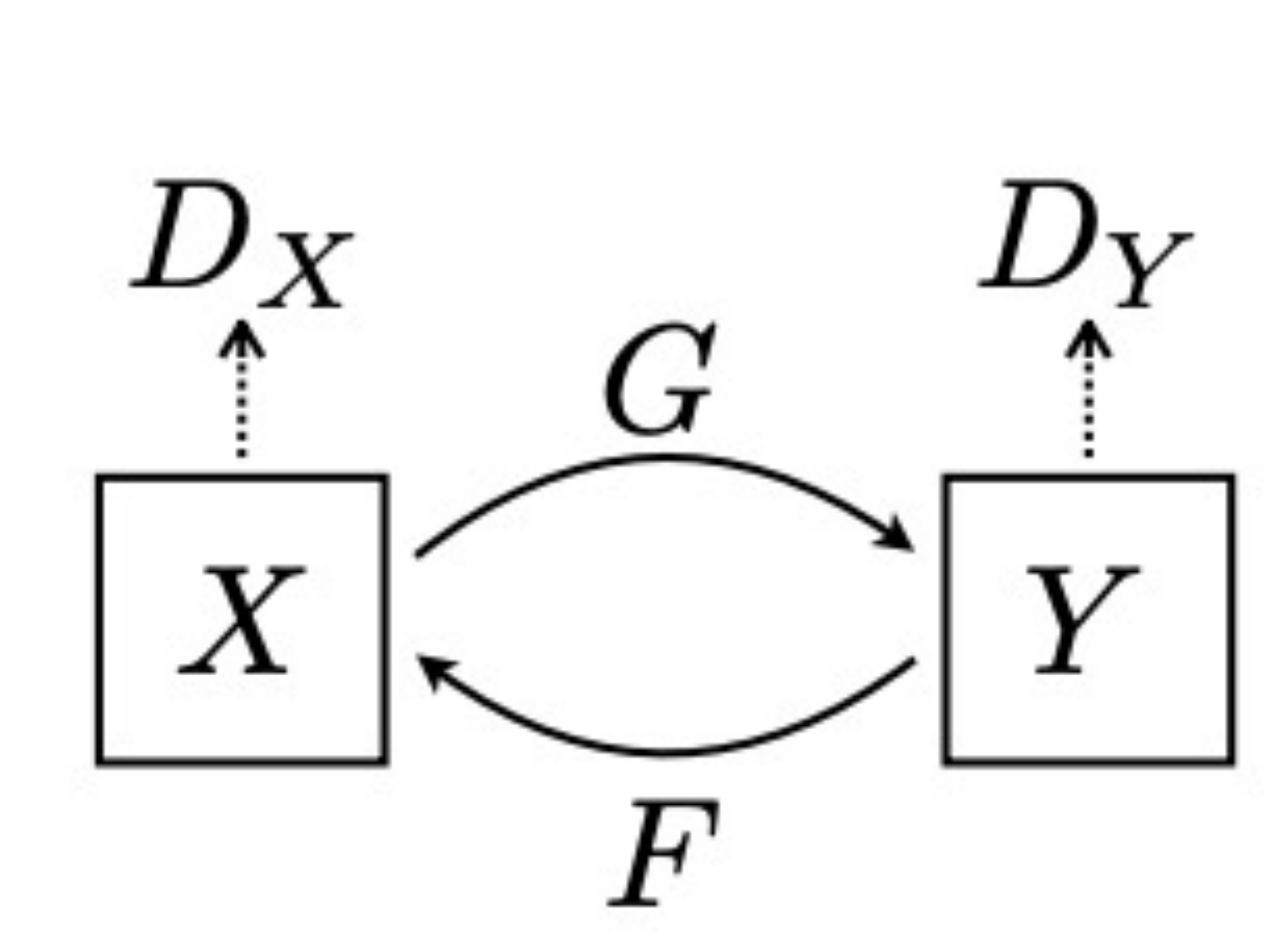
- Conditional Image generation
  - Class conditional (Big GAN)
  - Paired (Pix2Pix)
  - Unpaired (CycleGAN)

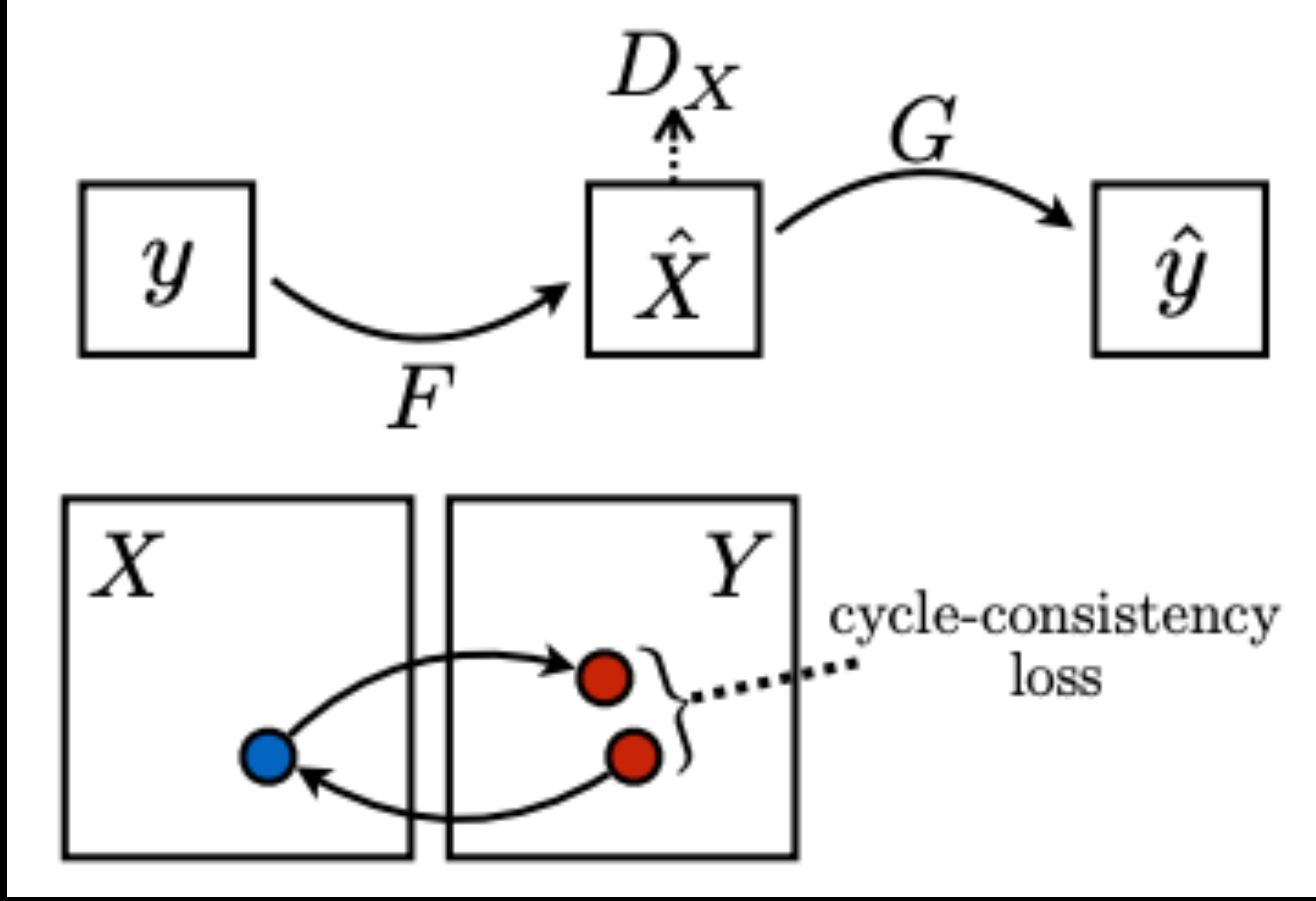
CycleGAN: Unpaired Image to Image Translation



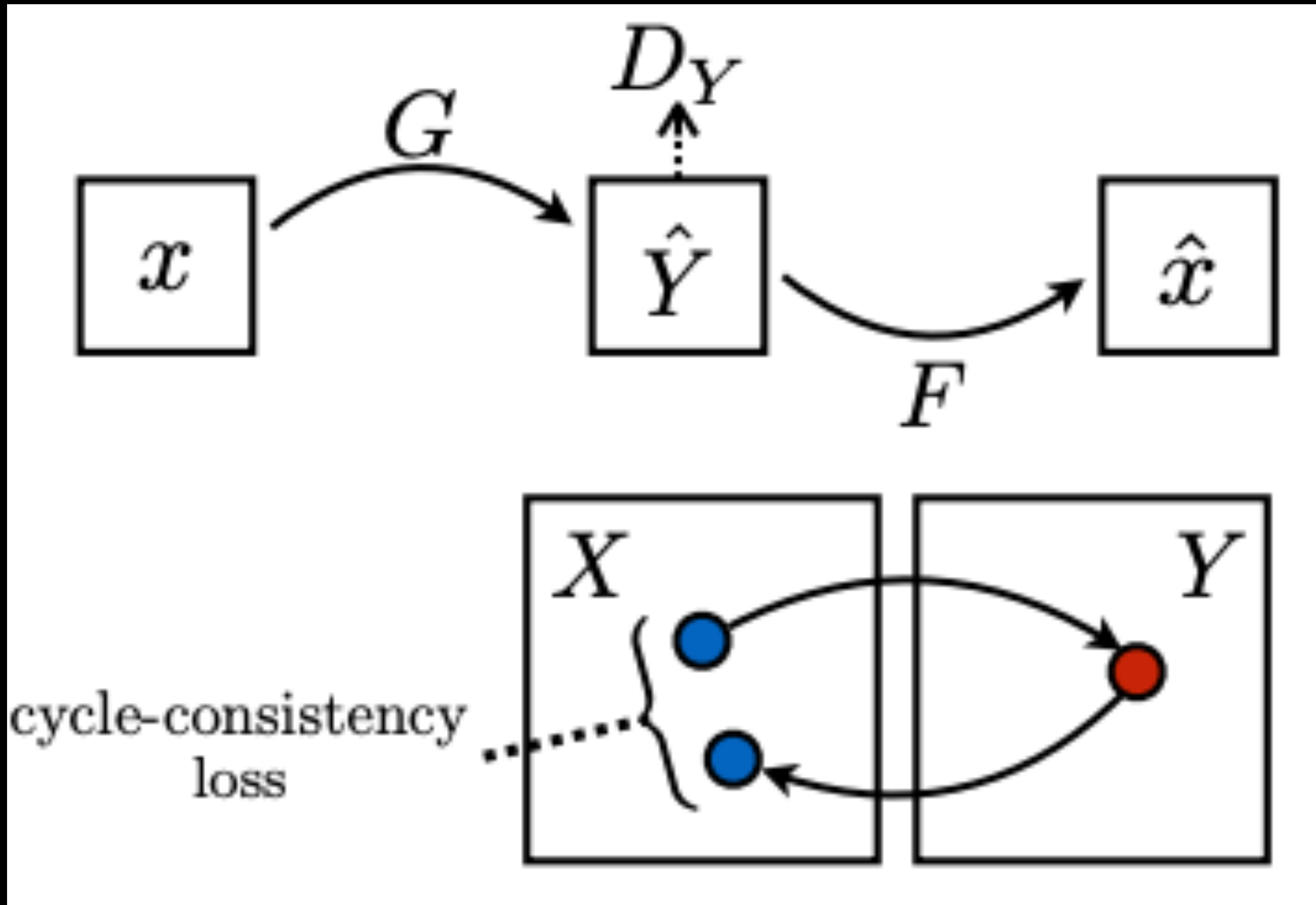
Training data: A set of images of style X + A set of images of style Y

Test: Given an image of style X, generate the same image in style Y

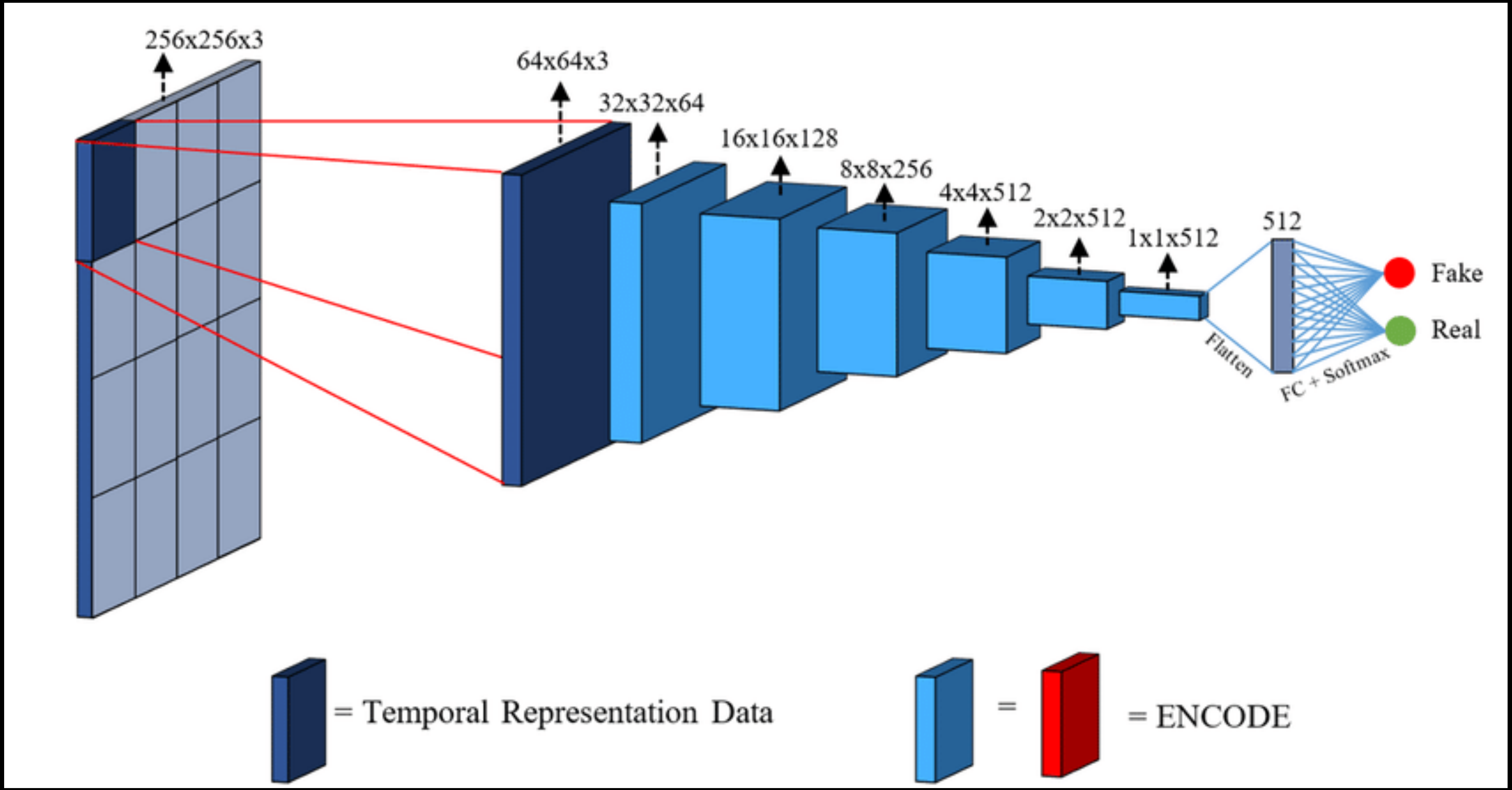




In addition to regular GAN loss on domain X and Y respectively, also add cycle-consistency loss for domain X and Y.



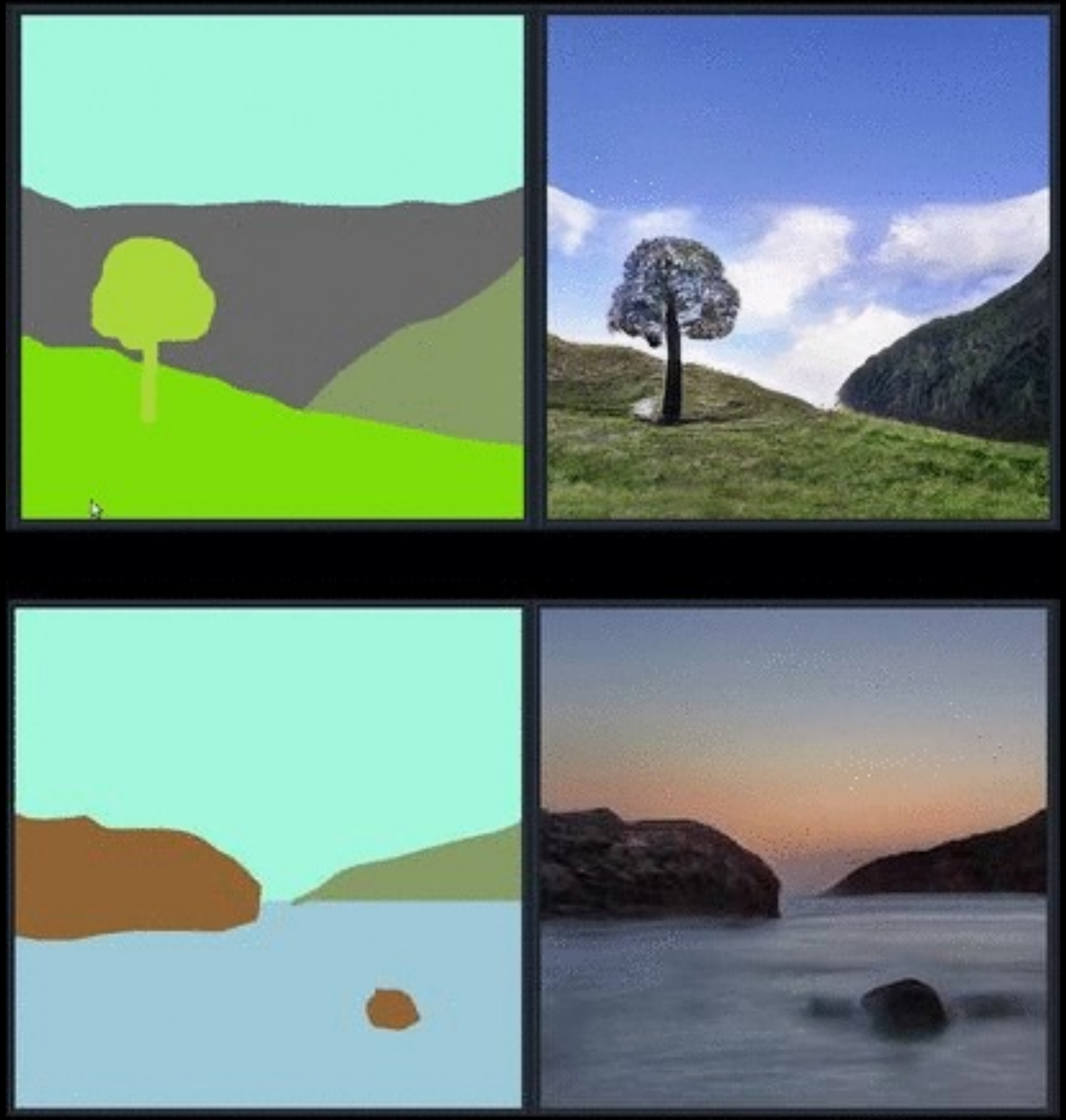
Patch Discriminator

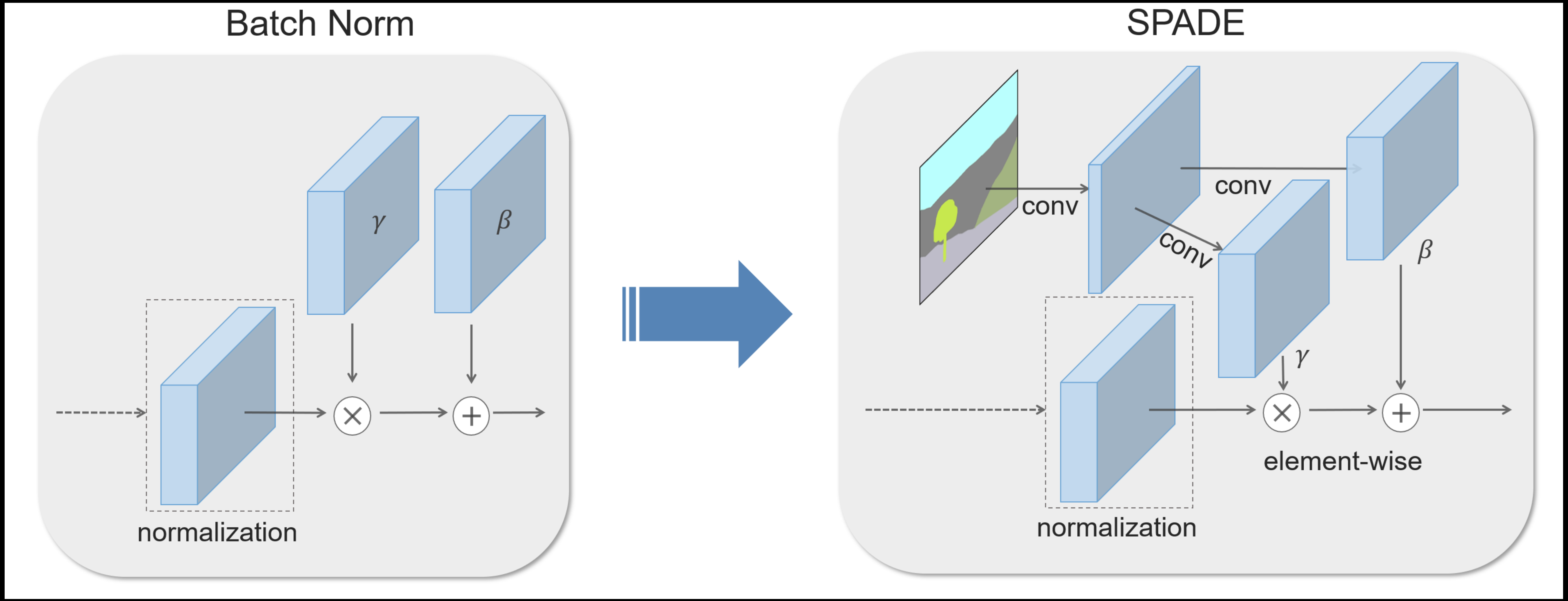


In practice Patch discriminator is applied at multiple resolutions (e.g. 64x64, 128x128, 256x256), and often the patches are overlapping.

Figure courtesy: Unsupervised Anomaly Detection and Localization Based on Deep Spatiotemporal Translation Network

GauGAN: Semantic Image Synthesis with Spatially-Adaptive Normalization





- Activation parameters (gamma and beta) are learned from the input segmentation map.
- In Conditional BatchNorm or AdaIN, activation parameters are spatially invariant (and only depends on feature size, batchsize for AdaIN).
- In SPADE activation parameters are spatially varying.

Paper Presentation: Thursday, Sept 8

Multimodal Conditional Image Synthesis with Product-of-Experts GANs

Xun Huang Arun Mallya Ting-Chun Wang Ming-Yu Liu

NVIDIA Corporation

ECCV 2022

Research project behind the AI tool GauGAN2 and GauGAN360

Paper (arxiv) | Demo (GauGAN2) | Demo (GauGAN 360) | Code (coming soon)

TL;DR: PoE-GAN can synthesize images conditioned on an arbitrary combination of multiple modalities.

Day	Papers	Presenter	Reviewer #1	Reviewer #2
Thrs Sept 8	Multimodal Conditional Image Synthesis with Product-of-Experts GANs.	Akshay Paruchuri	Max Christman	Andrew Buchanan
	Local Relighting of Real Scenes.	Max Lennon	Sofia Wong	Yufan Liu
Tue Sept 13	Pivotal Tuning for Latent-based Editing of Real Images.	William Stanford	Xiaolong Huang	William Zhao
	Third Time's the Charm? Image and Video Editing with StyleGAN3.	Nurislam Tursynbek	Yulu Pan	Savitha Patil
Thrs Sept 15	CLIP2StyleGAN: Unsupervised Extraction of StyleGAN Edit Directions.	Sam Ehrenstein	Maddison Khire	Mariana Rodriguez
	DyStyle: Dynamic Neural Network for Multi-Attribute-Conditioned Style Editing.	Qiwei Zhao	Longtian Ye	Andrea Dunn Beltran
Thrs Sept 22	Dynamic View Synthesis from Dynamic Monocular Video.	Jade kandel	Ziheng Wang	Aidan Carter Scott
	Neural Light Transport for Relighting and View Synthesis	JUN MYEONG CHOI	Zenan Wang	Sofia Wong

Important Deadlines

- 590: Assignment 2 announced, due Sept 8.
- 590/790: Paper presentation/review schedule announced
- 790: Deadline to register your project group, Sept 2 (Friday - TOMORROW)!
- 1 points deducted per late day!
- 790: Project Proposal presentation is due Sept 20!

Slide Credits

EECS 6322 Deep Learning for Computer Vision, Kosta Derpanis (York University)
Many amazing research papers!