Lecture 8
Generative Adversarial Networks (GANs)

M2 Data Science and AI

Yaohui Wang

http://www-sop.inria.fr/members/Yaohui.Wang/

informatics mathematics

- Generative Adversarial Networks: *Yaohui Wang*
- DeepFake Detection: *Dr. Antitza Dantcheva*
- Labs (TP): *David Anghelone*

Question: VAE ?

Ian Goodfellow

Generative Adversarial Networks [NIPS 2014]

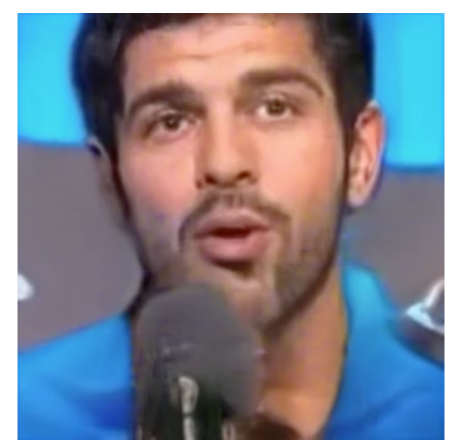

"GANs are the most interesting idea in the last 10 years in ML"
- Yann LeCun

Image Generation

Style Transfer

Video Generation

Video Generation

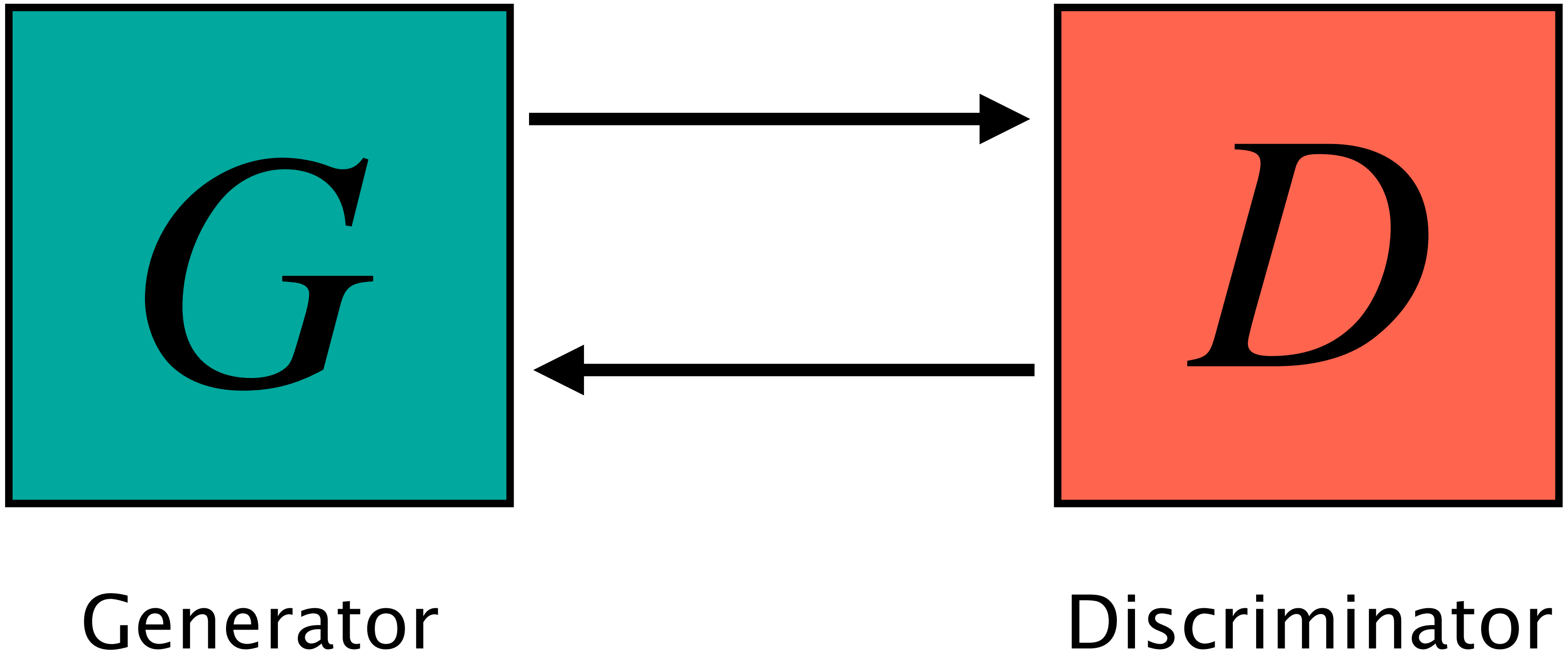


Outline

- Basic Idea of GAN
- Image Generation
  - Conditional GAN (CGAN, ACGAN)
  - Modern GANs (StyleGAN, BigGAN)
  - Image-to-image translation (Pix2Pix, CycleGAN)
- Video-to-video translation
- GANs Evaluation
- Video Generation
- Lab (DCGAN for manga face generation)

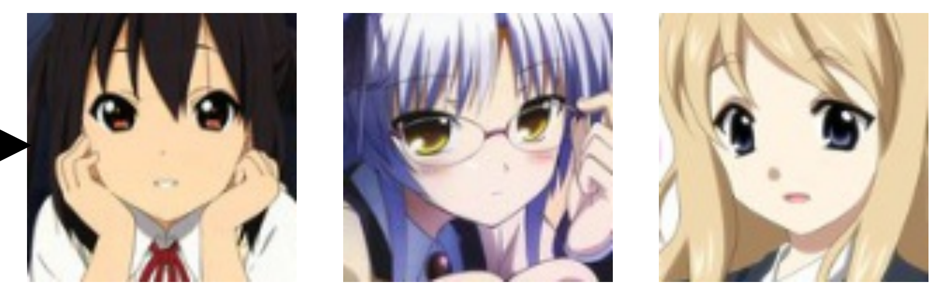
Basic idea of GAN

Basic idea of GAN



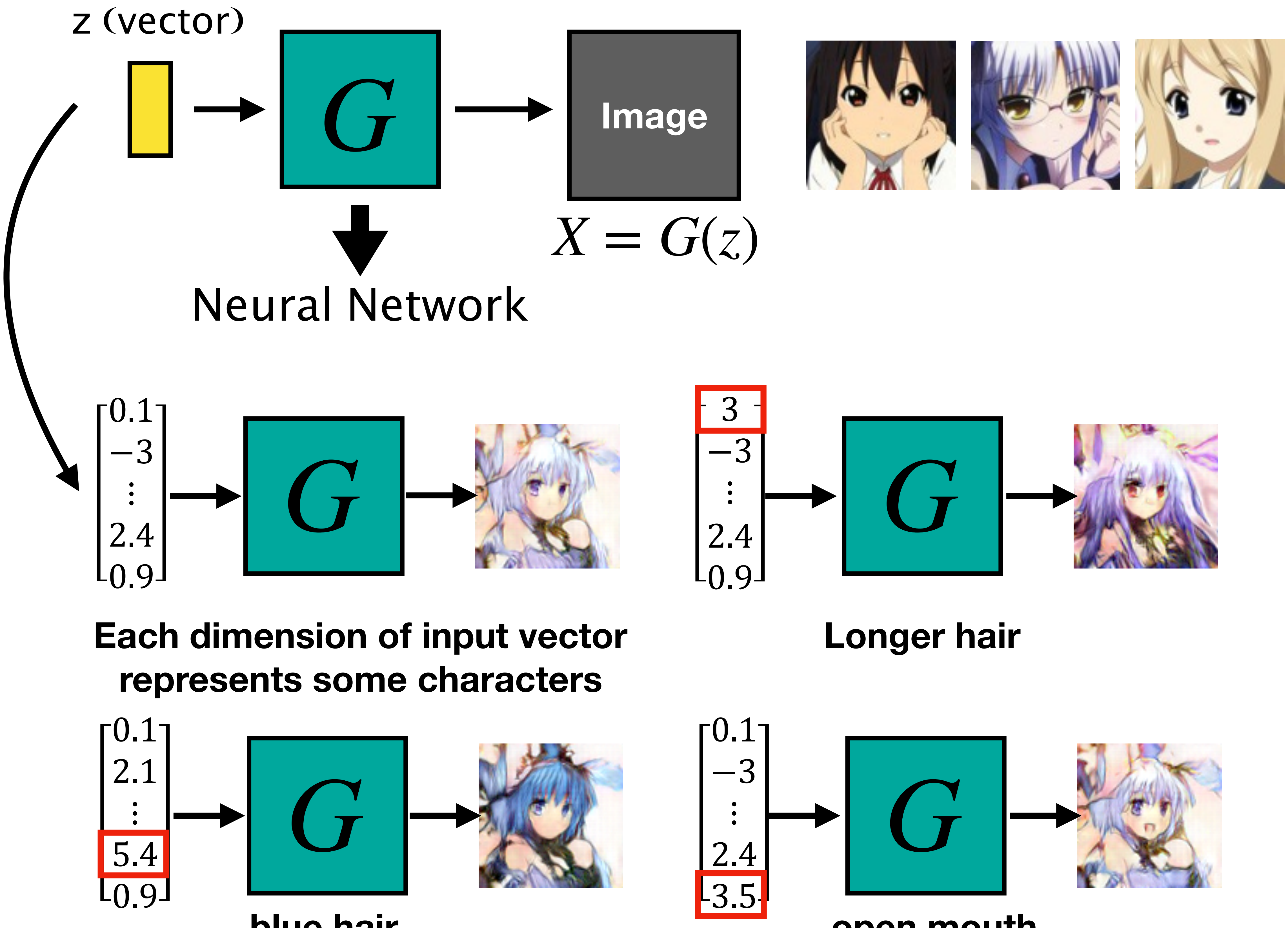
Basic idea of GAN

$$\begin{bmatrix} 0.3 \\ -0.1 \\ \vdots \\ -0.7 \end{bmatrix} \begin{bmatrix} 0.1 \\ -0.1 \\ \vdots \\ 0.7 \end{bmatrix} \begin{bmatrix} -0.3 \\ 0.1 \\ \vdots \\ 0.9 \end{bmatrix}$$

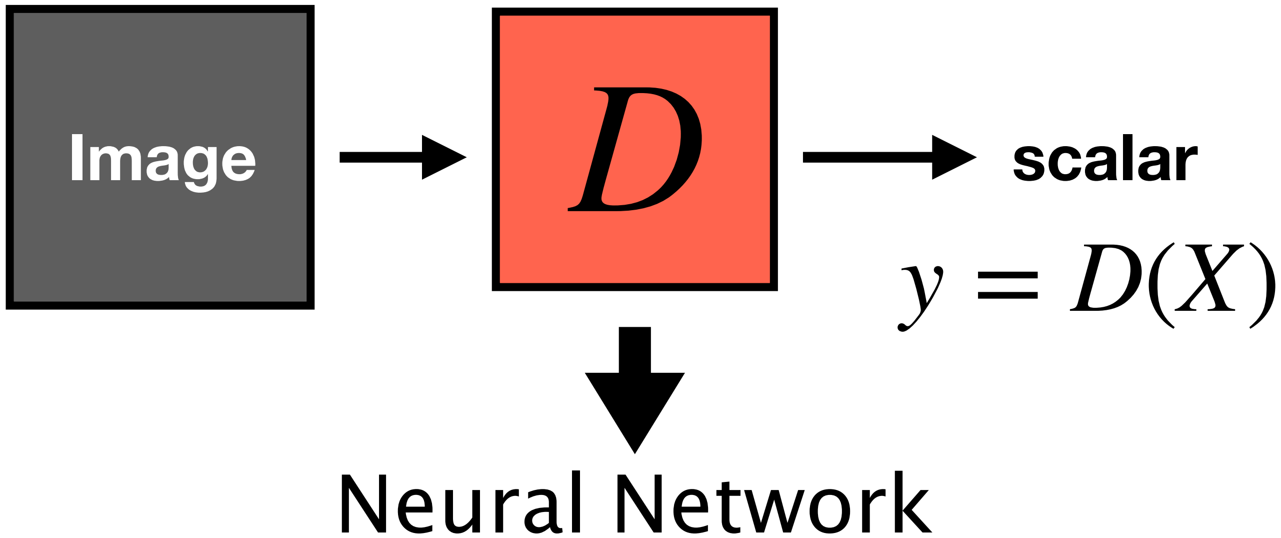


in a specific range (Gaussian, ...)

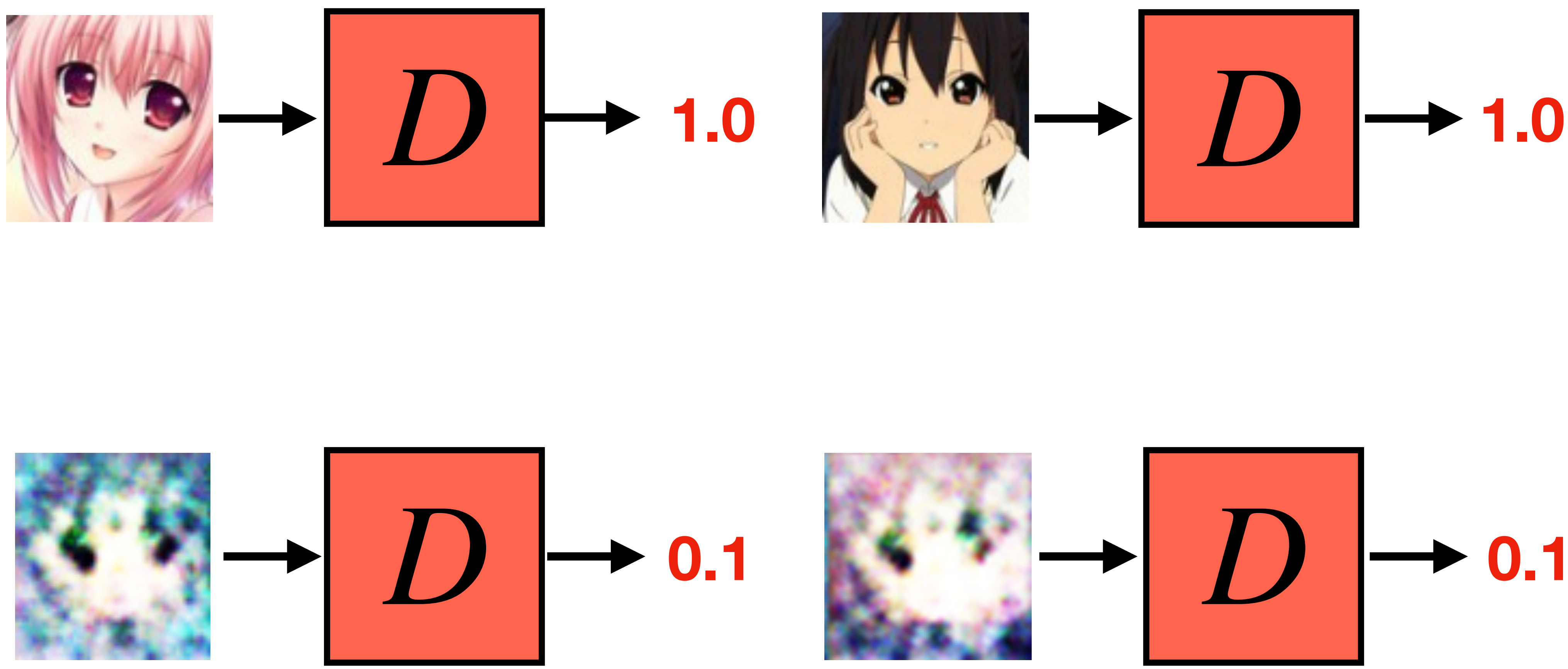
Basic idea of GAN



Basic idea of GAN

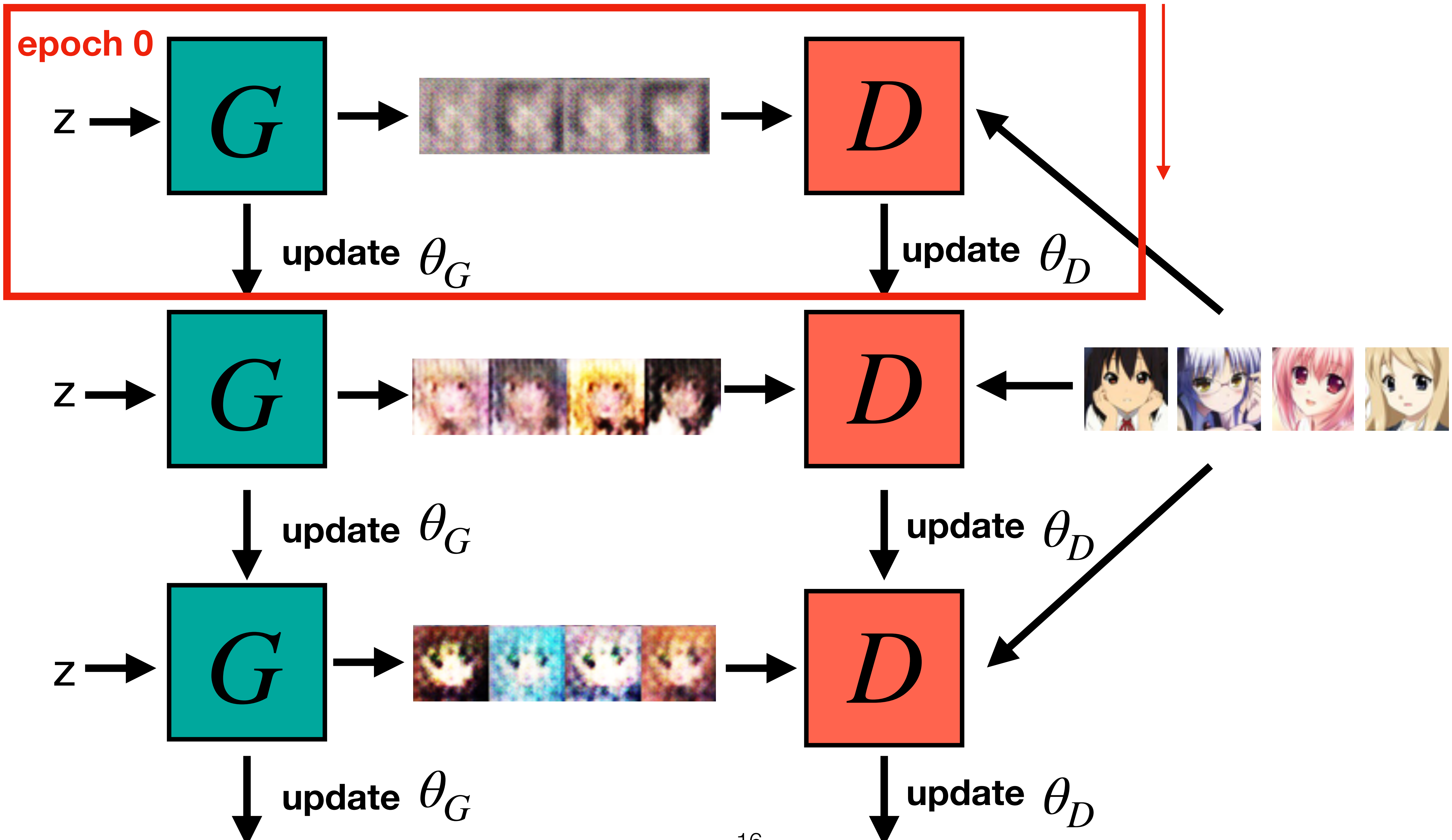


higher value: more realistic
lower value: less realistic



Basic idea of GAN

Adversarial Training (Generative Adversarial Networks)



Basic idea of GAN

Adversarial Training (Generative Adversarial Networks)

Algorithm Initialize θ_d for D and θ_g for G

Learning D

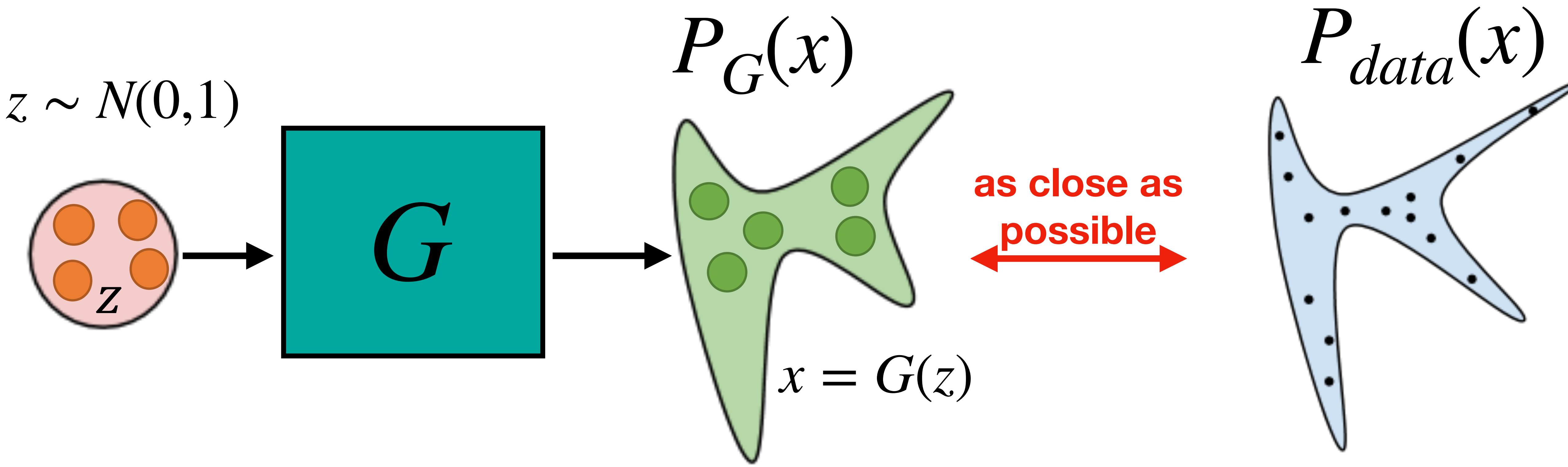
- In each training iteration:
  - Sample m examples $\{x^1, x^2, \dots, x^m\}$ from database
  - Sample m noise samples $\{z^1, z^2, \dots, z^m\}$ from a distribution
  - Obtaining generated data $\{\tilde{x}^1, \tilde{x}^2, \dots, \tilde{x}^m\}$, $\tilde{x}^i = G(z^i)$
  - Update discriminator parameters θ_d to maximize
    - $\tilde{V} = \frac{1}{m}\sum_{i=1}^{m} logD(x^i) + \frac{1}{m}\sum_{i=1}^{m} log\left(1 - D(\tilde{x}^i)\right)$
    - $\theta_d \leftarrow \theta_d + \eta\nabla\tilde{V}(\theta_d)$

Learning G

  - Sample m noise samples $\{z^1, z^2, \dots, z^m\}$ from a distribution
  - Update generator parameters θ_g to maximize
    - $\tilde{V} = \frac{1}{m}\sum_{i=1}^{m} log\left(D\left(G(z^i)\right)\right)$
    - $\theta_g \leftarrow \theta_g - \eta\nabla\tilde{V}(\theta_g)$

Basic idea of GAN

Generator: G is a network. It defines a probability distribution P_G



$$G^* = \arg\min_G \boxed{Div(P_G, P_{data})}$$

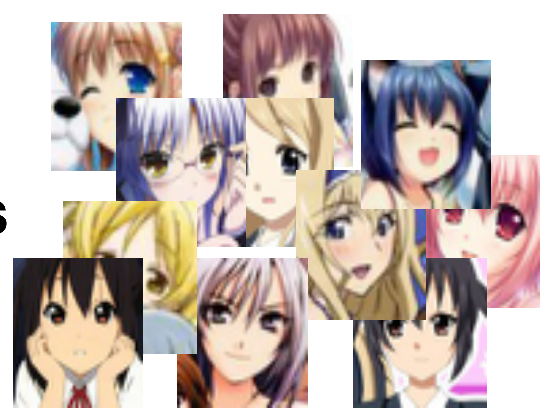
how to compute the divergence between two distributions ?

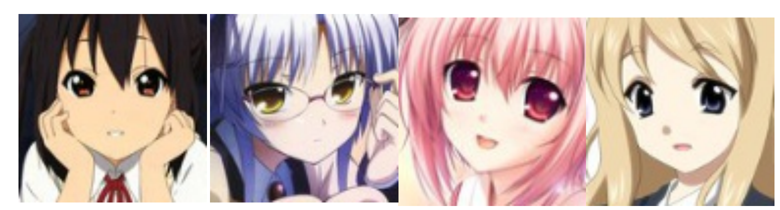
Basic idea of GAN

Discriminator $$G^* = \arg\min_G Div(P_G, P_{data})$$

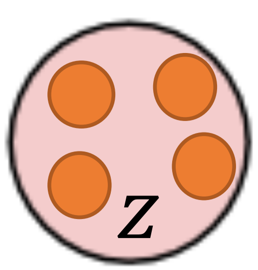
Although we do not know the distributions of $P_G(x)$ and $P_{data}(x)$, we can still sample from them

real images





sampling from $P_{data}(x)$



sample from normal distribution

sampling from $P_G(x)$

Basic idea of GAN

Discriminator $$G^* = \arg\min_G Div(P_G, P_{data})$$

Objective function for D

$$V(G,D) = E_{x\sim P_{data}}[\log D(x)] + E_{x\sim P_G}[\log(1 - D(x))]$$

(G is fixed)

$$D^* = arg\boxed{\max_D V(G,D)}$$ **= binary classification**

JS Divergence

Basic idea of GAN

Discriminator

$$G^* = \underset{G}{argmin}\ Div(P_G, P_{data})$$

Objective function for G

$$G^* = \underset{G}{argmin}(\cancel{E_{x\sim P_{data}}[log D(x)]} + \underline{E_{x\sim P_G}[log(1 - D(G(z)))]})$$

(D is fixed)

$$E_{x\sim P_G}[-log(D(G(z)))])$$

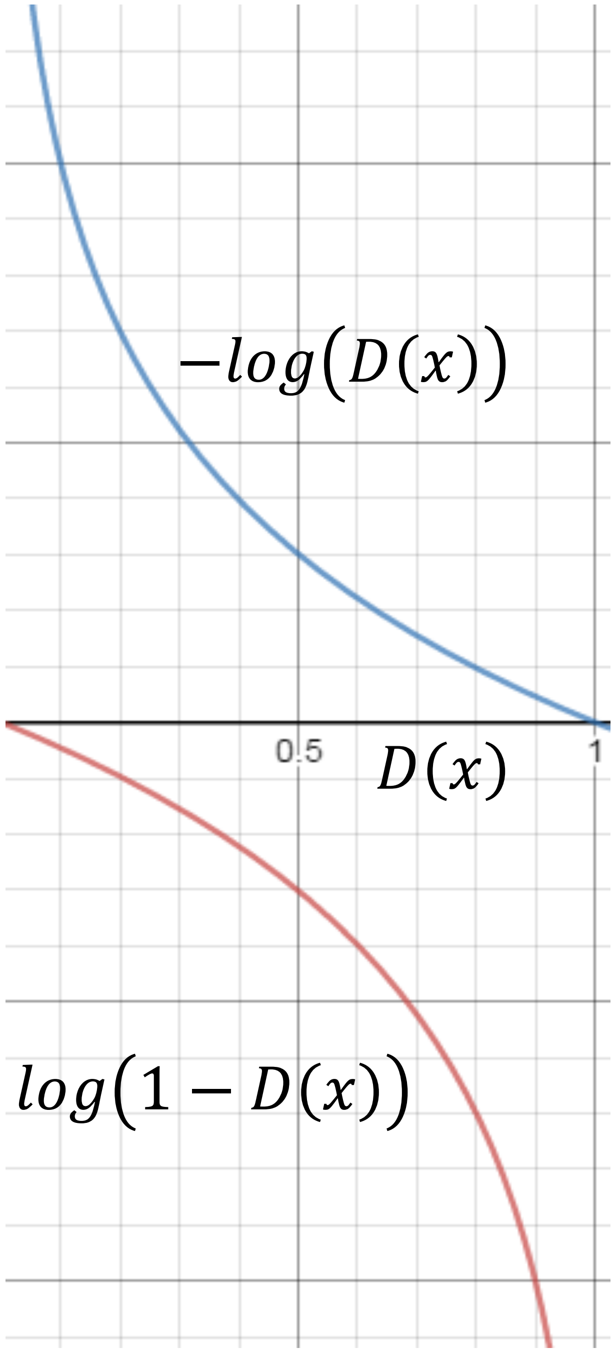
Basic idea of GAN

$$E_{x \sim P_G}[log(1 - D(G(z)))])$$

slow at the beginning

$$E_{x \sim P_G}[-log(D(G(z)))])$$

real implementation



Basic idea of GAN

Different GANs

- Wasserstein GAN
- Wasserstein GAN-GP (gradient penalty)
- LSGAN
- …

Basic idea of GAN

$$V(G, D) = E_{x \sim P_{data}}[logD(x)] + E_{z \sim P_z}[log(1 - D(G(z)))]$$

$$G* = \underset{G}{argmin}\ \underset{D}{max} V(G, D)$$

Training Steps:

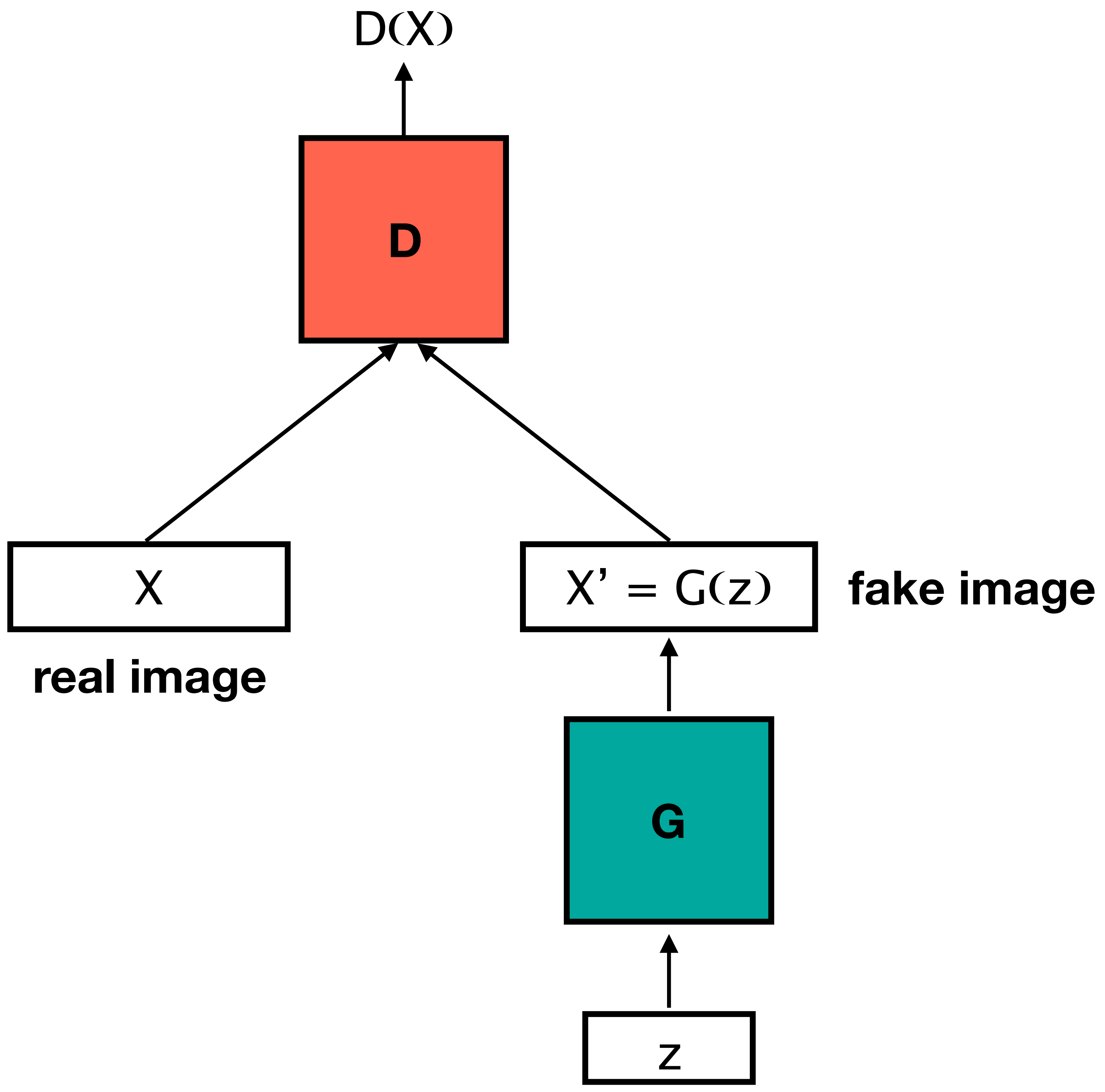
- Initialize Generator and Discriminator
- In each training iteration:

  Step 1: Fix Generator G, and update Discriminator D

  Step 2: Fix Discriminator D, and update Generator G

Vanilla GAN (unconditional)

Vanilla GAN [Ian Goodfellow, et al, NIPS 2014]

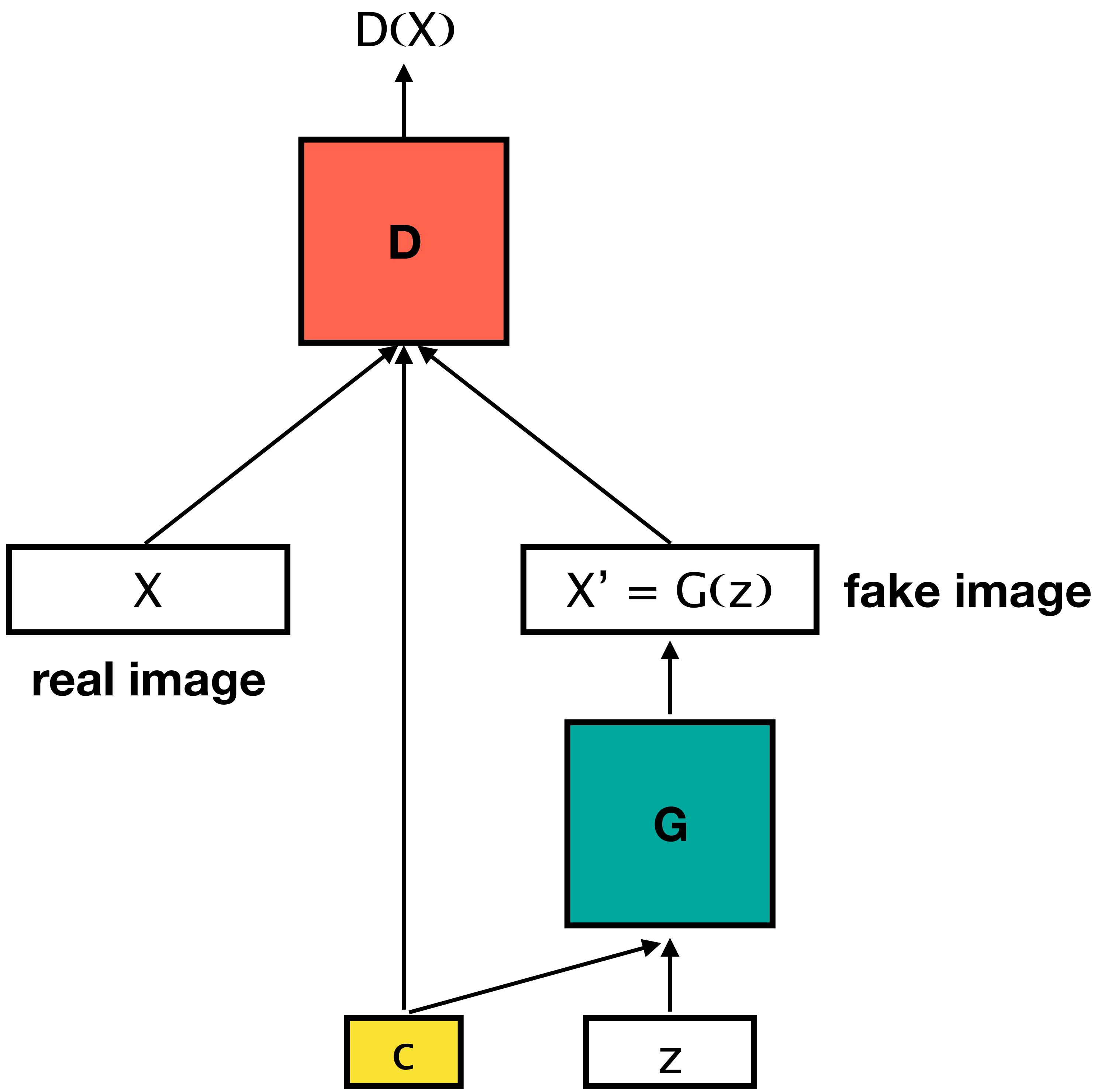


Conditional GAN

Conditional GAN

[M Mirza, et al, arXiv 2014]

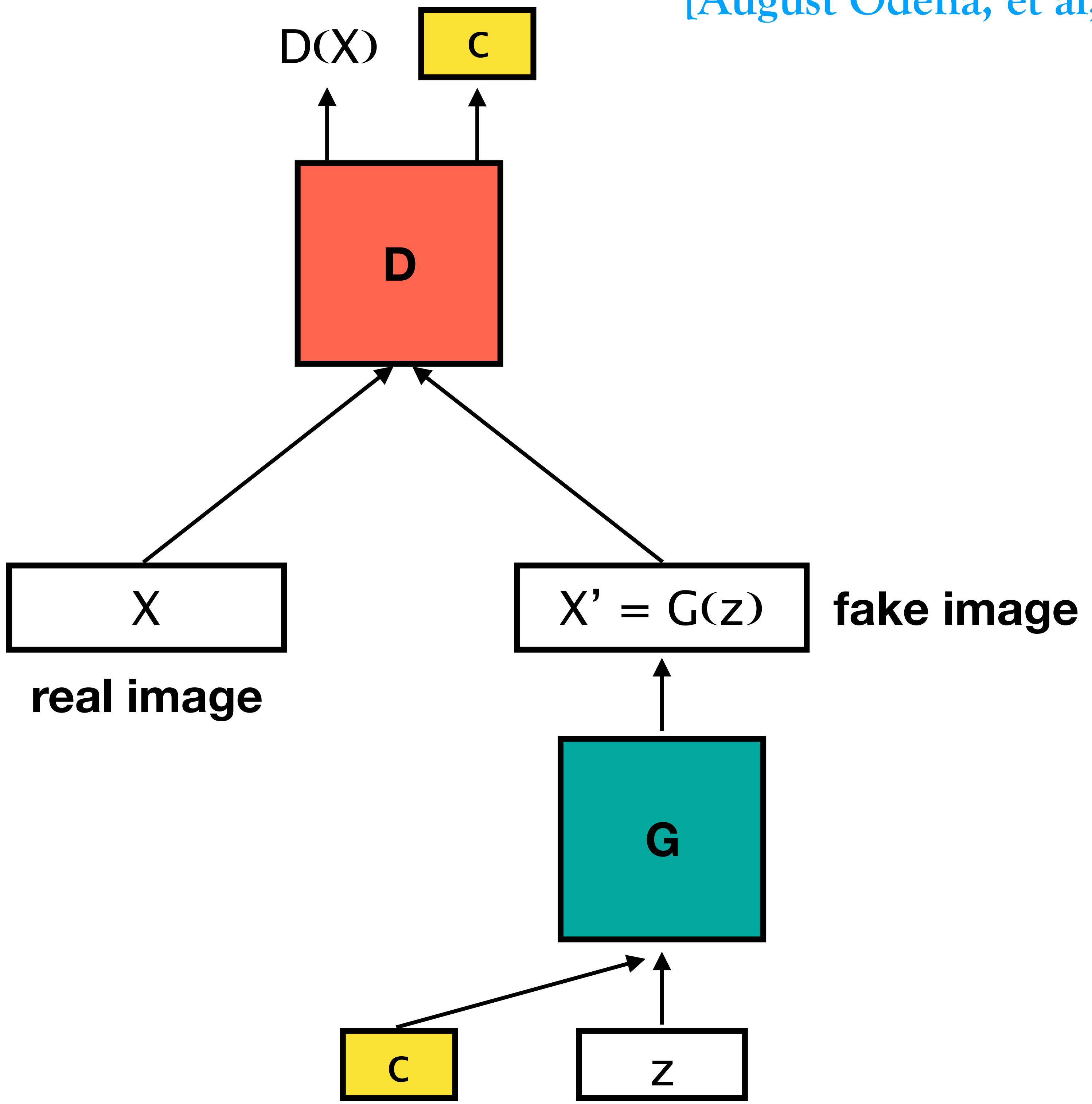
CGAN



Conditional GAN

[August Odena, et al, ICML 2016]

ACGAN



Conditional GAN

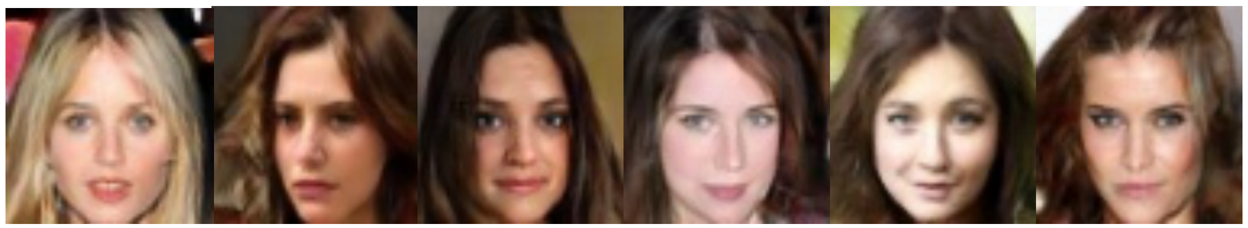
male, with glasses

female, with glasses

male, without glasses

female, without glasses

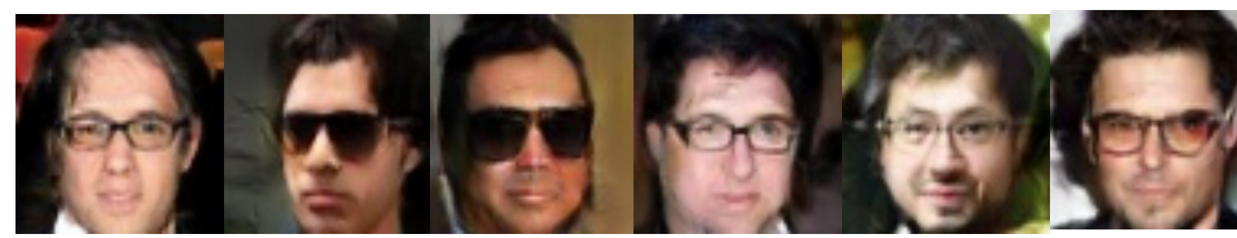
Conditional GAN



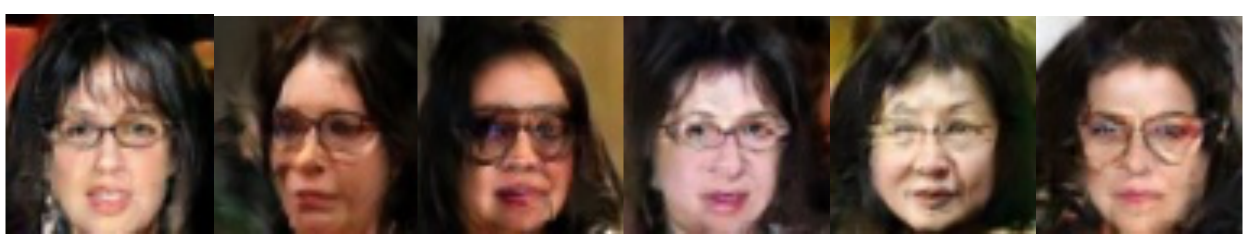
without glasses, female, no black hair, no smiling, young

without glasses, male, no black hair, smiling, young

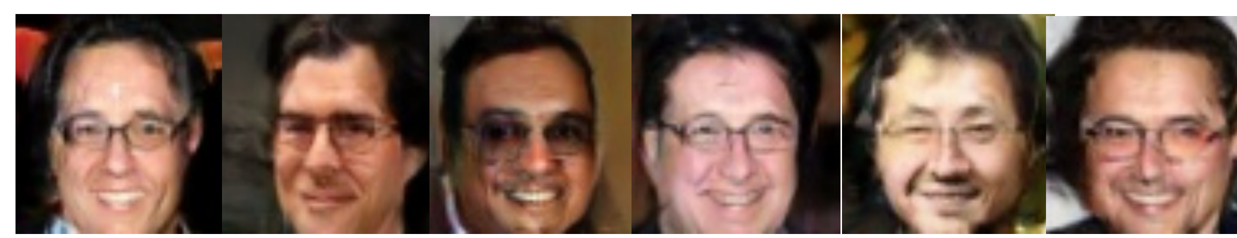
without glasses, female, black hair, smiling, young



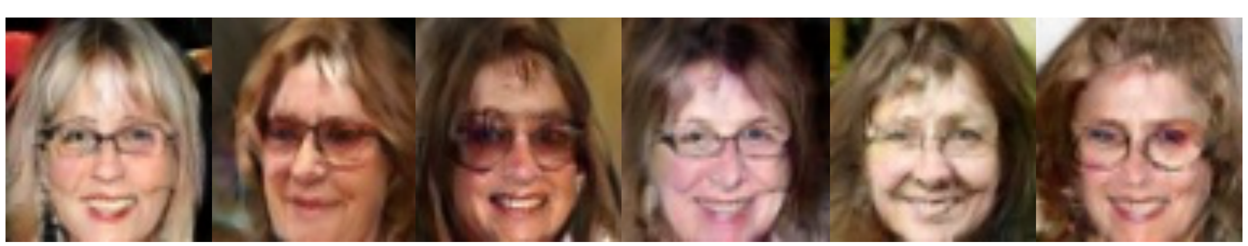
with glasses, male, black hair, no smiling, young



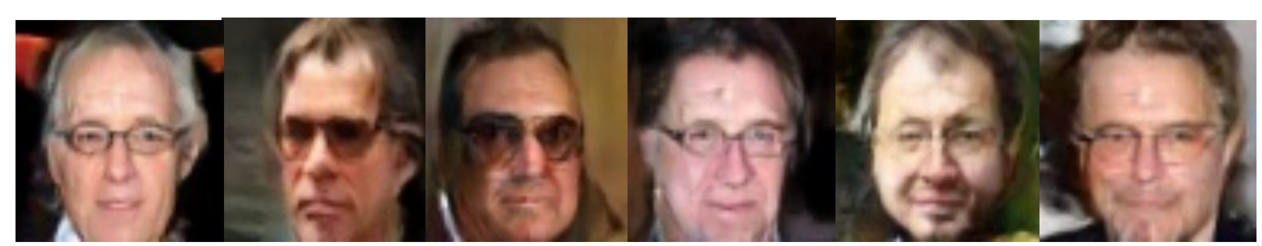
with glasses, female, black hair, no smiling, old



with glasses, male, black hair, smiling, old



with glasses, female, no black hair, smiling, old



without glasses, male, no black hair, no smiling, old

Conditional GAN

[Scott Reed, et al, ICML 2016]

Text-to-image Generation

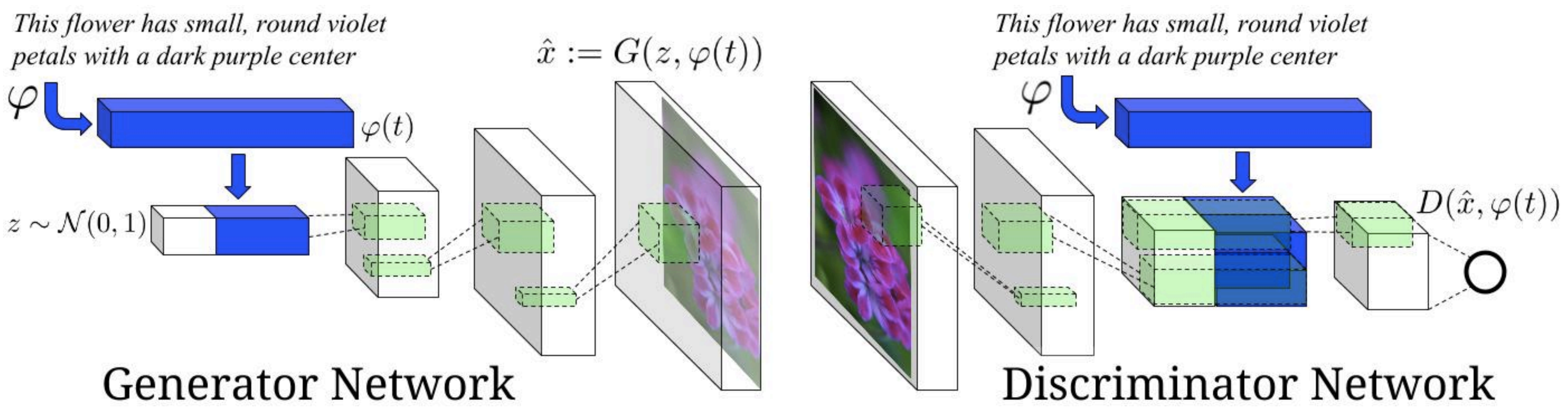
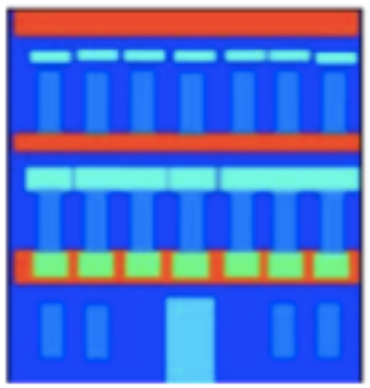


Image-to-image translation

Image-to-image translation

- Traditional method



NN

Image

as close as possible

L1 / L2 loss

Testing:

It is blurry,
what is the problem here ?

Image-to-image translation

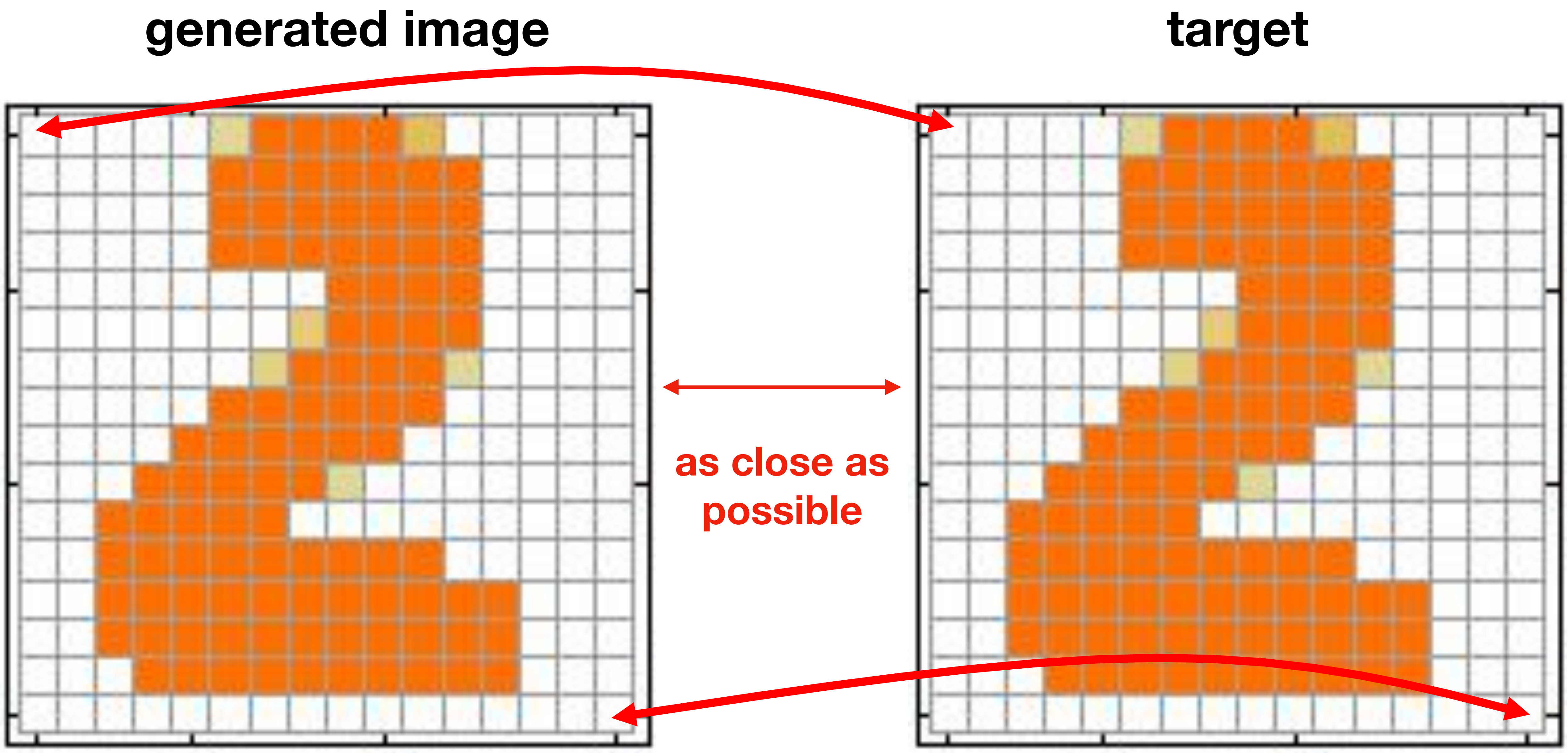
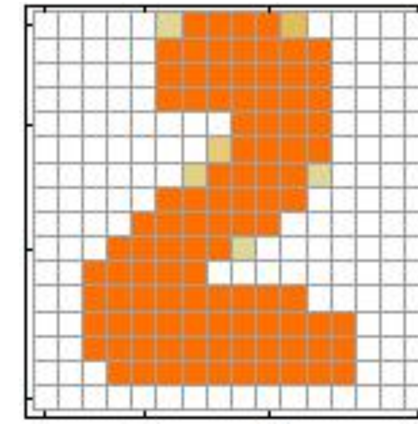
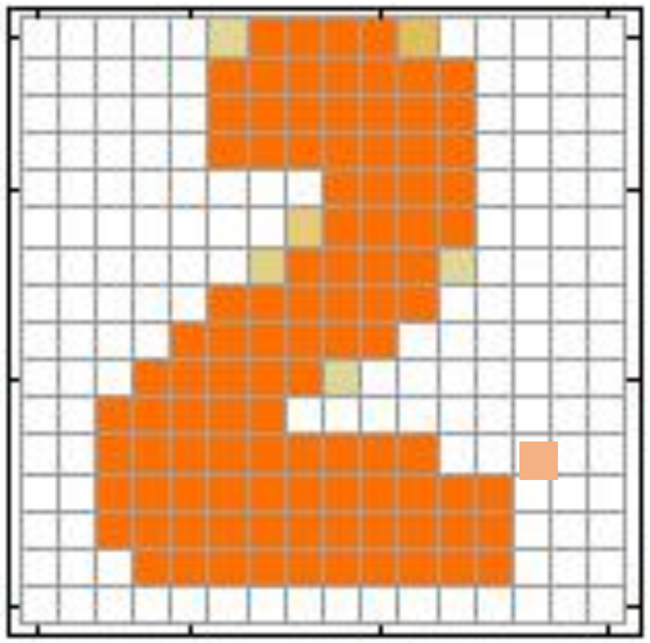


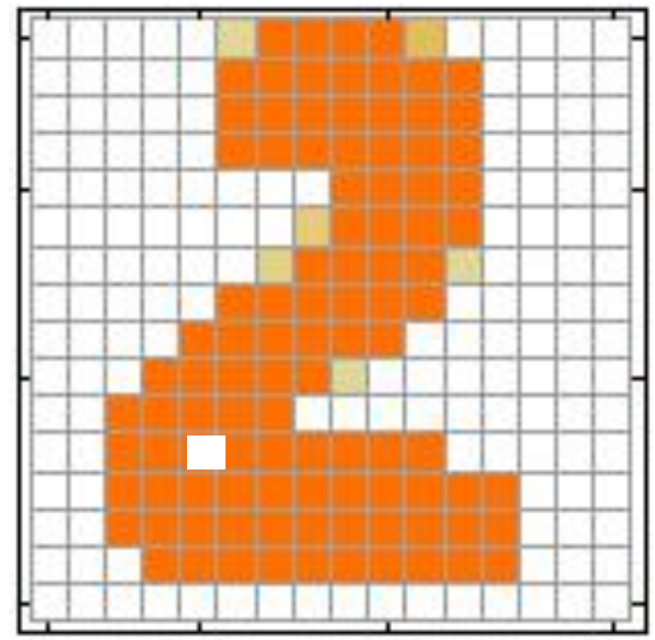
Image-to-image translation



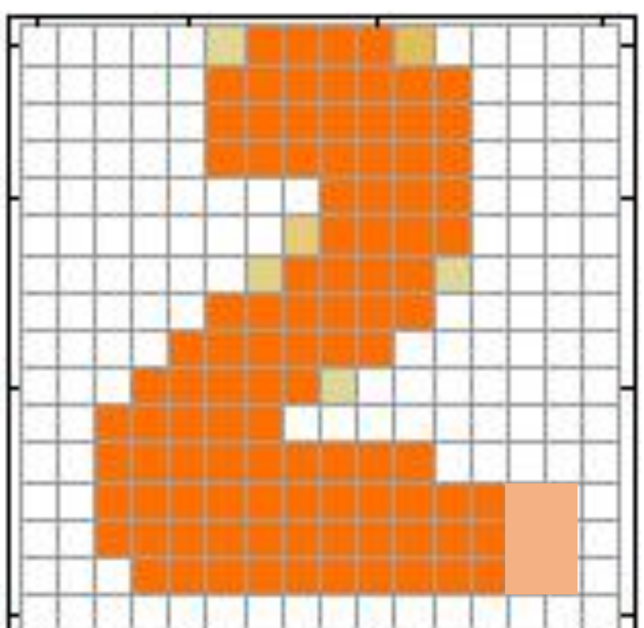
target



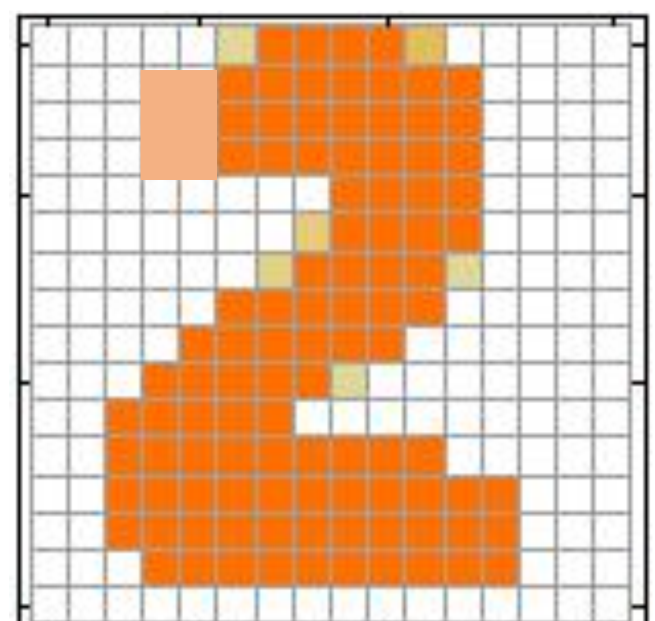
1 pixel error
not realistic



1 pixel error
not realistic



6 pixel error
realistic



6 pixel error
realistic

Image-to-image translation

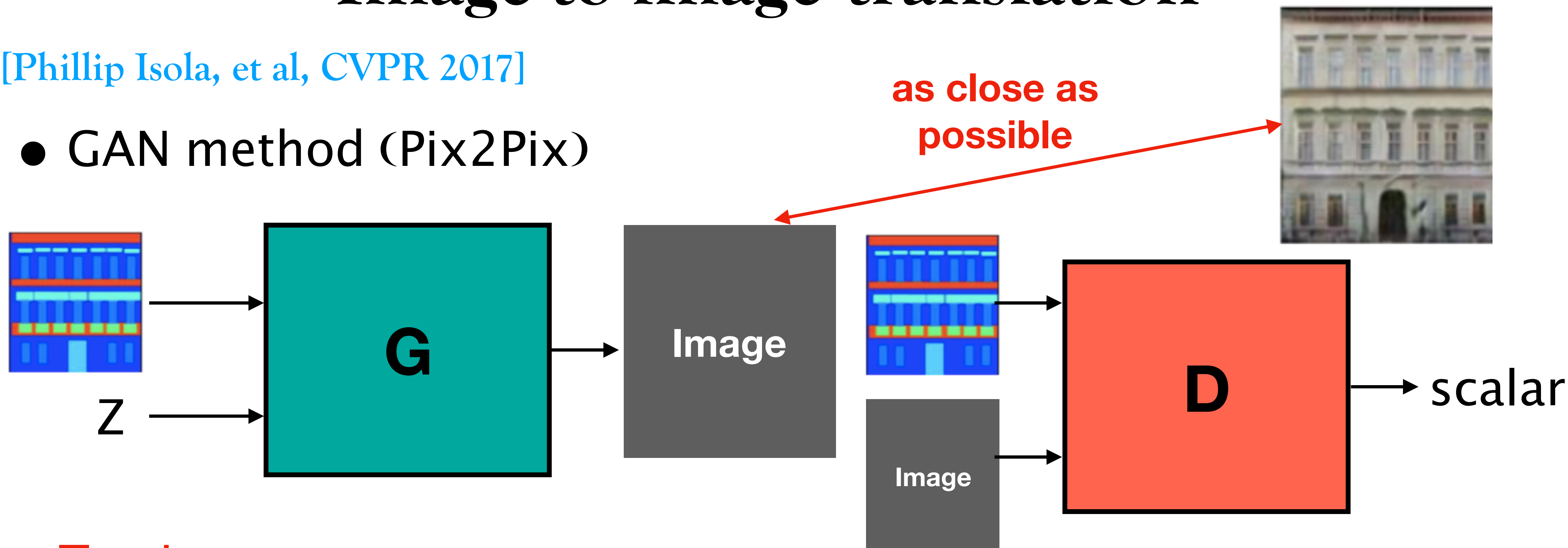
Reconstruction loss can not provide a sharp generation, what should be the solution ?

Since we can not find a good metric, we can use GAN to learn the metric !

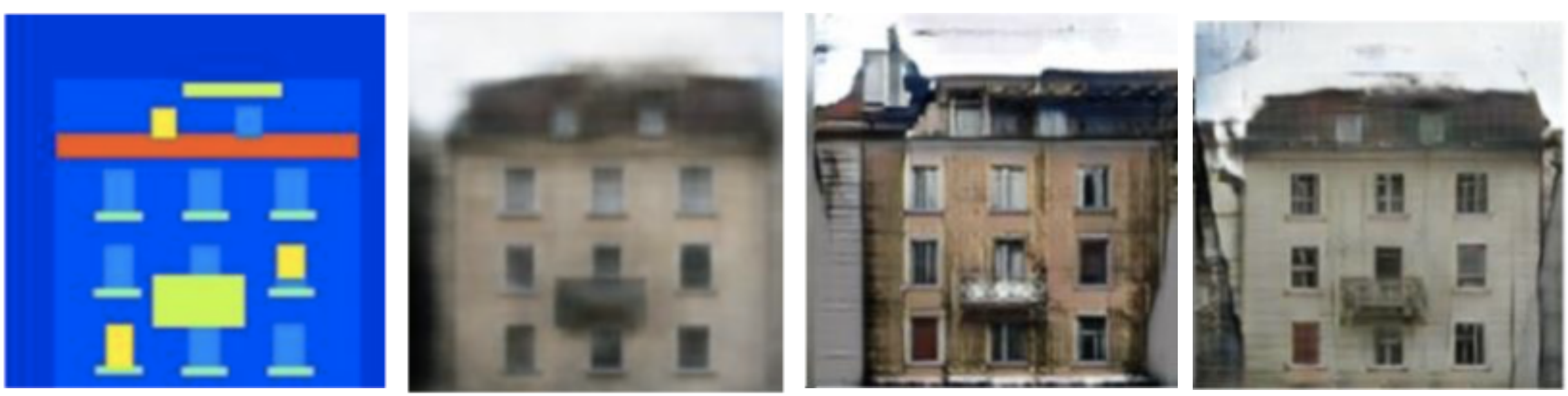
Image-to-image translation

[Phillip Isola, et al, CVPR 2017]

- GAN method (Pix2Pix)



Testing:



Input Reconstruct GAN GAN + Reconstruct

Image-to-image translation

- What about unpaired data (no ground truth of target image) ?

X: zebra

Y: horse

X: summer

Y: winter

Image-to-image translation

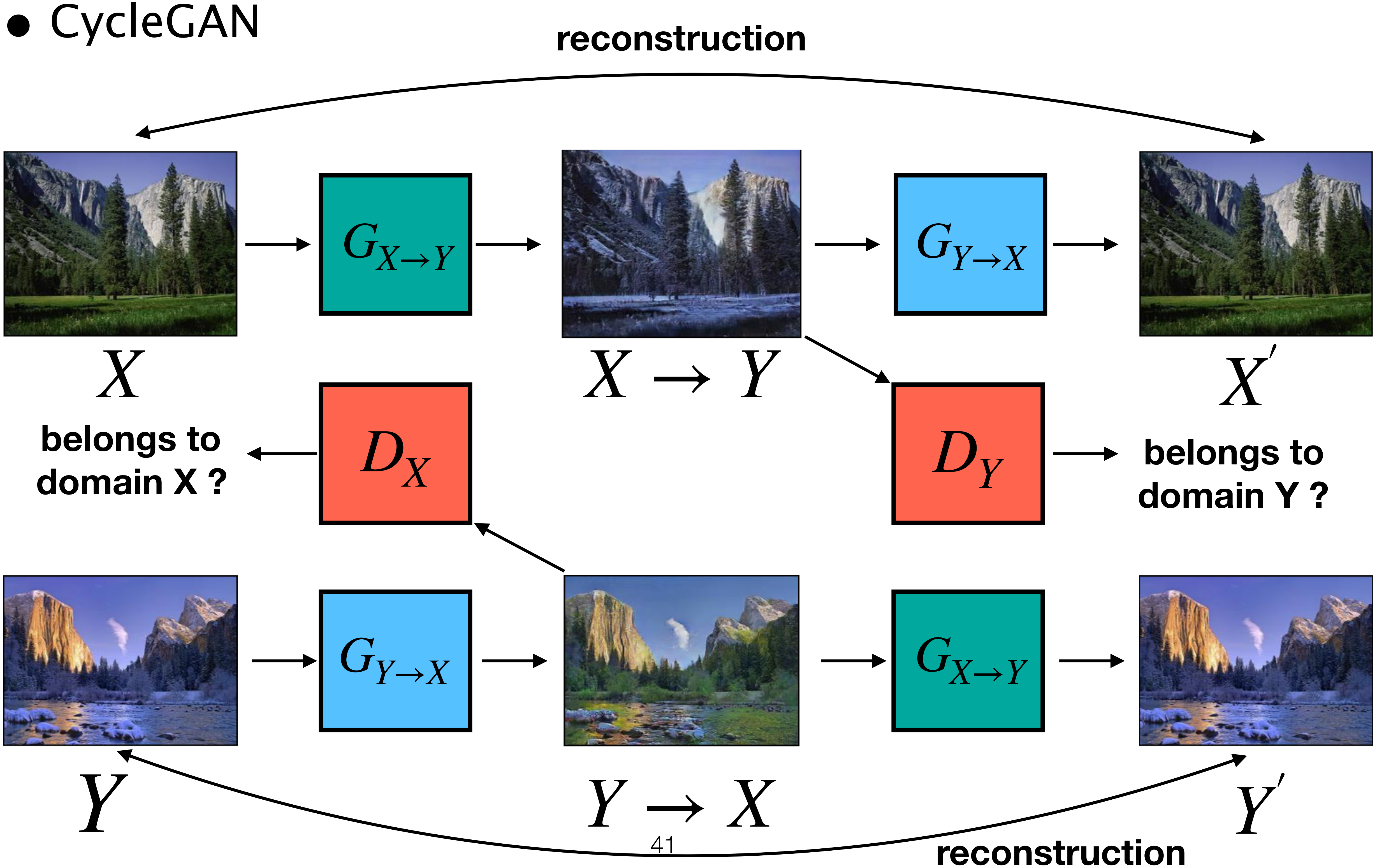
[Jun-yan Zhu, et al, ICCV 2017]

- CycleGAN

Image-to-image translation

[Jun-yan Zhu, et al, ICCV 2017]

- CycleGAN



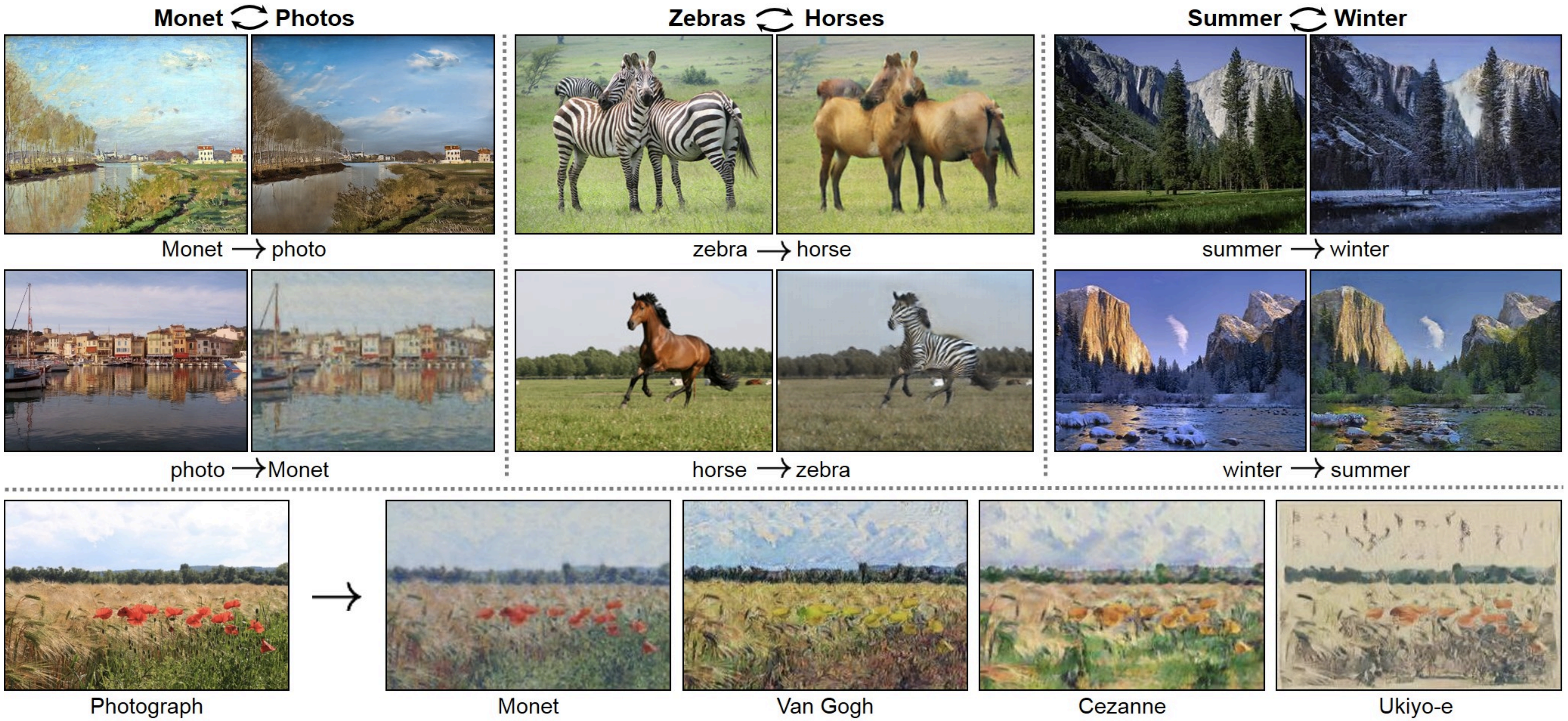

Monet ⇄ Photos
Zebras ⇄ Horses
Summer ⇄ Winter
Monet → photo
zebra → horse
summer → winter
photo → Monet
horse → zebra
winter → summer
Photograph
Monet
Van Gogh
Cezanne
Ukiyo-e

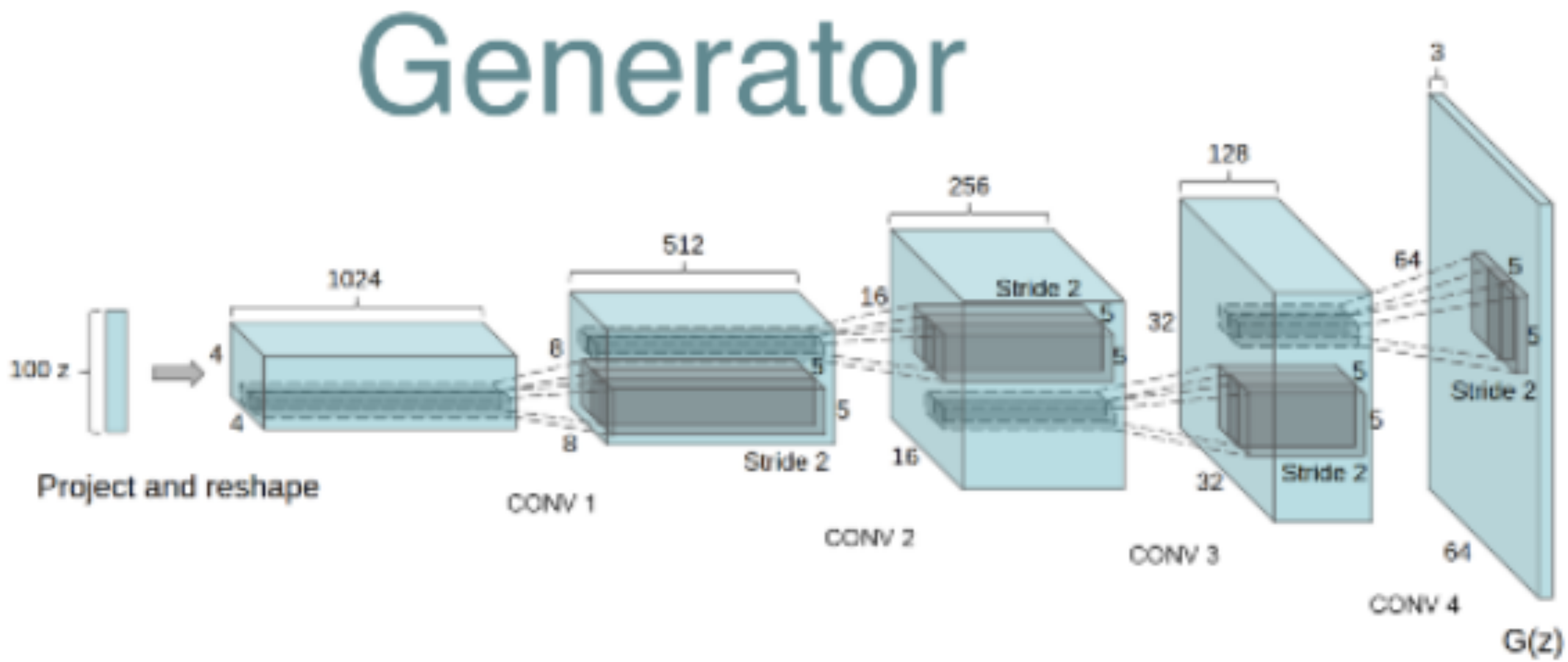
Image-to-image translation

- UNIT
- MUNIT
- …

Modern GAN Architectures

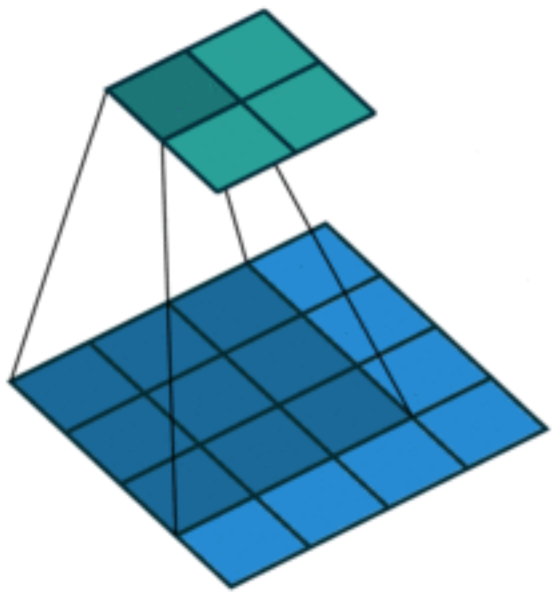
Modern Architectures

DCGAN

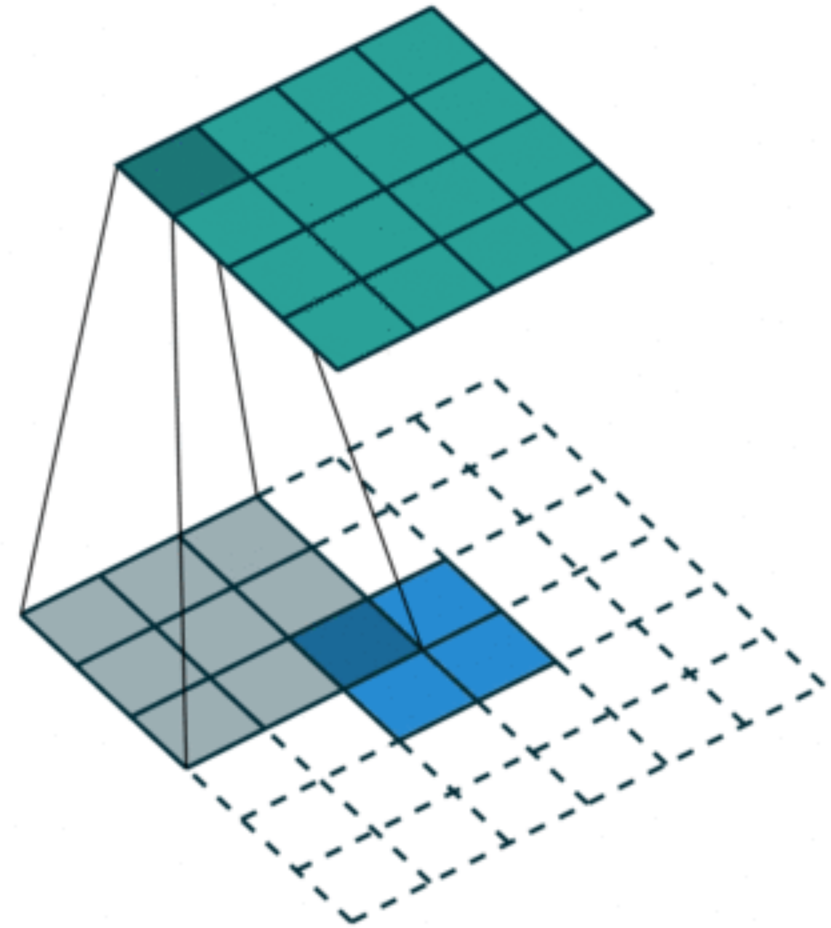


https://github.com/vdumoulin/conv_arithmetic

[A Radford, et al, arXiv 2015]



convolution



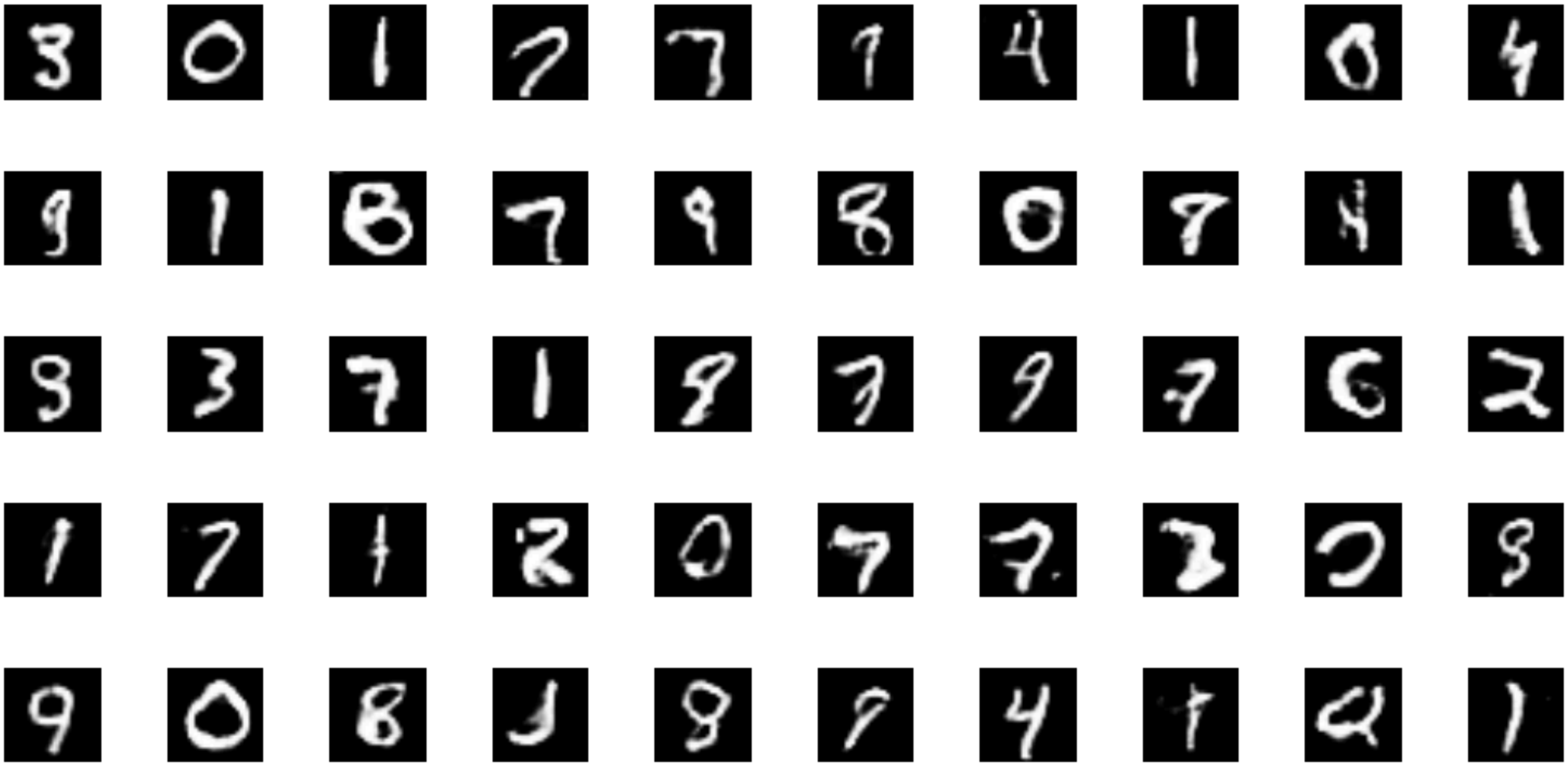
transposed convolution

Modern Architectures

https://github.com/vdumoulin/conv_arithmetic

[A Radford, et al, arXiv 2015]

Results - MNIST



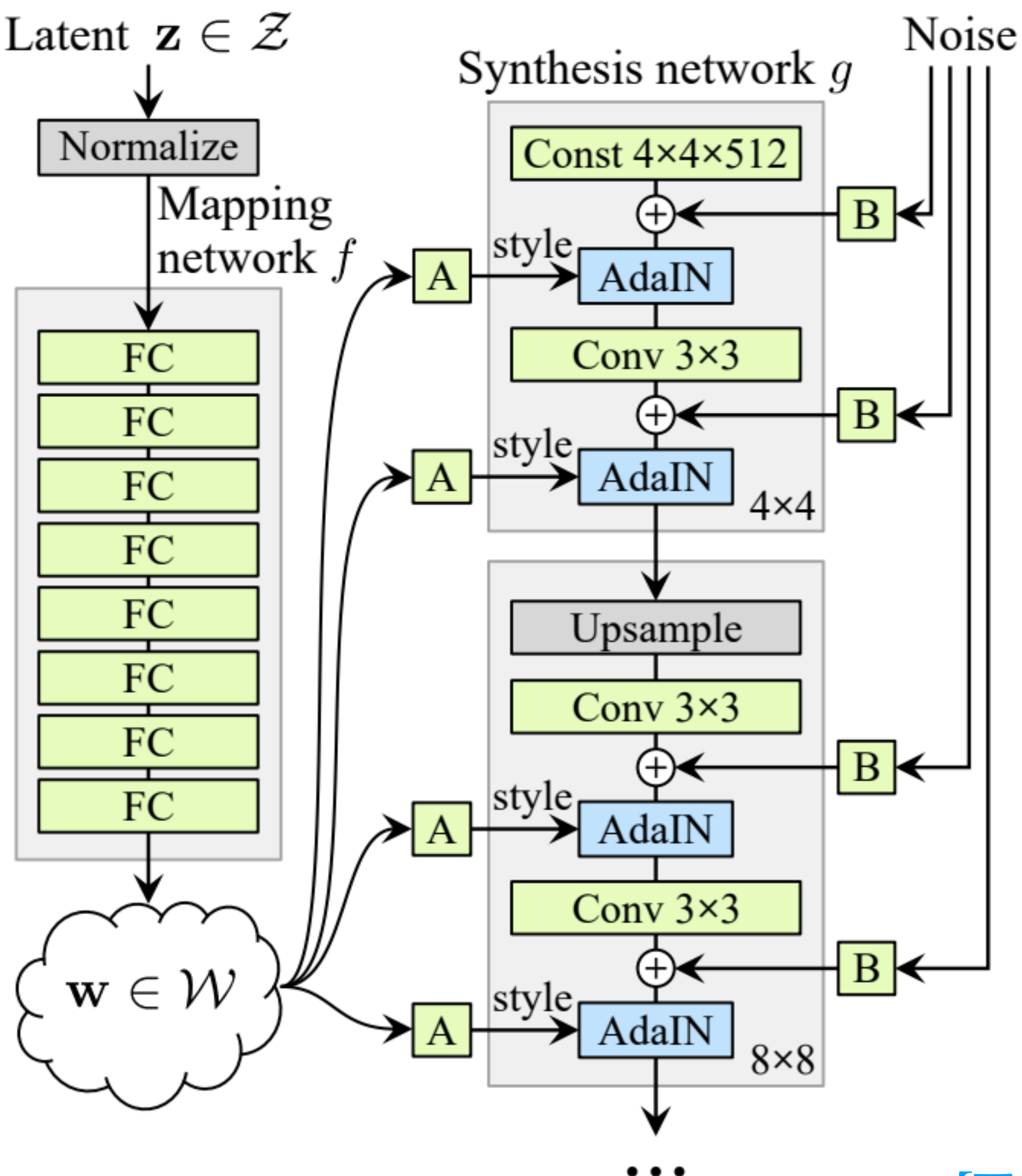
Results - CelebA (faces)

Results - LSUN (bedrooms)

Modern Architectures

StyleGAN (NVIDIA)

https://github.com/NVlabs/stylegan



[T Karras, et al, CVPR 2019]

Modern Architectures

StyleGAN

https://www.youtube.com/watch?v=kSLJriaOumA

Karras et al, A Style-Based Generator Architecture for Generative Adversarial Networks, CVPR 2019

Modern Architectures

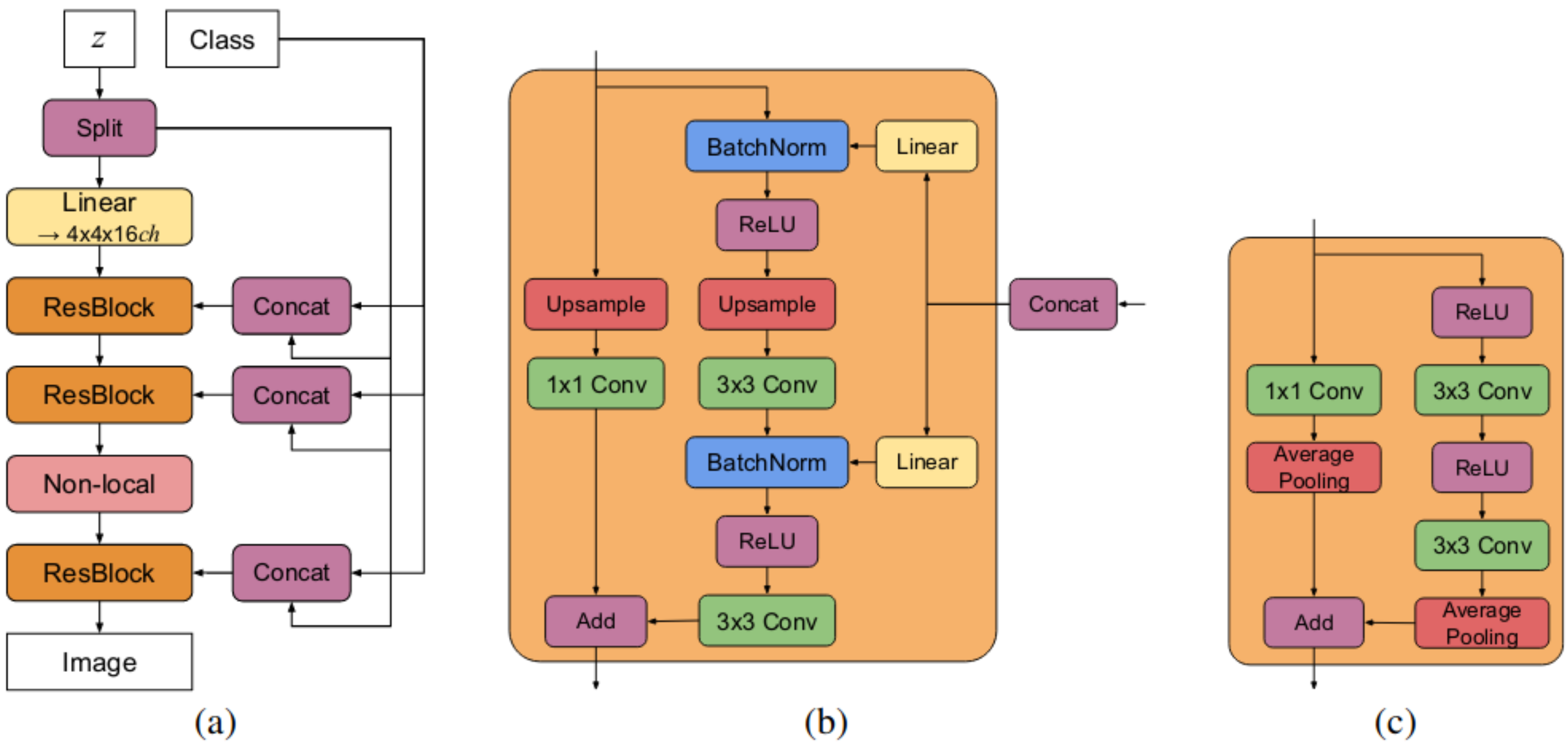
StyleGAN

GPUs	1024×1024	512×512	256×256
1	41 days 4 hours	24 days 21 hours	14 days 22 hours
2	21 days 22 hours	13 days 7 hours	9 days 5 hours
4	11 days 8 hours	7 days 0 hours	4 days 21 hours
8	6 days 14 hours	4 days 10 hours	3 days 8 hours

Modern Architectures

BigGAN (DeepMind)

https://github.com/ajbrock/BigGAN-PyTorch



[A Brock, et al, ICLR 2019]

Modern Architectures

BigGAN

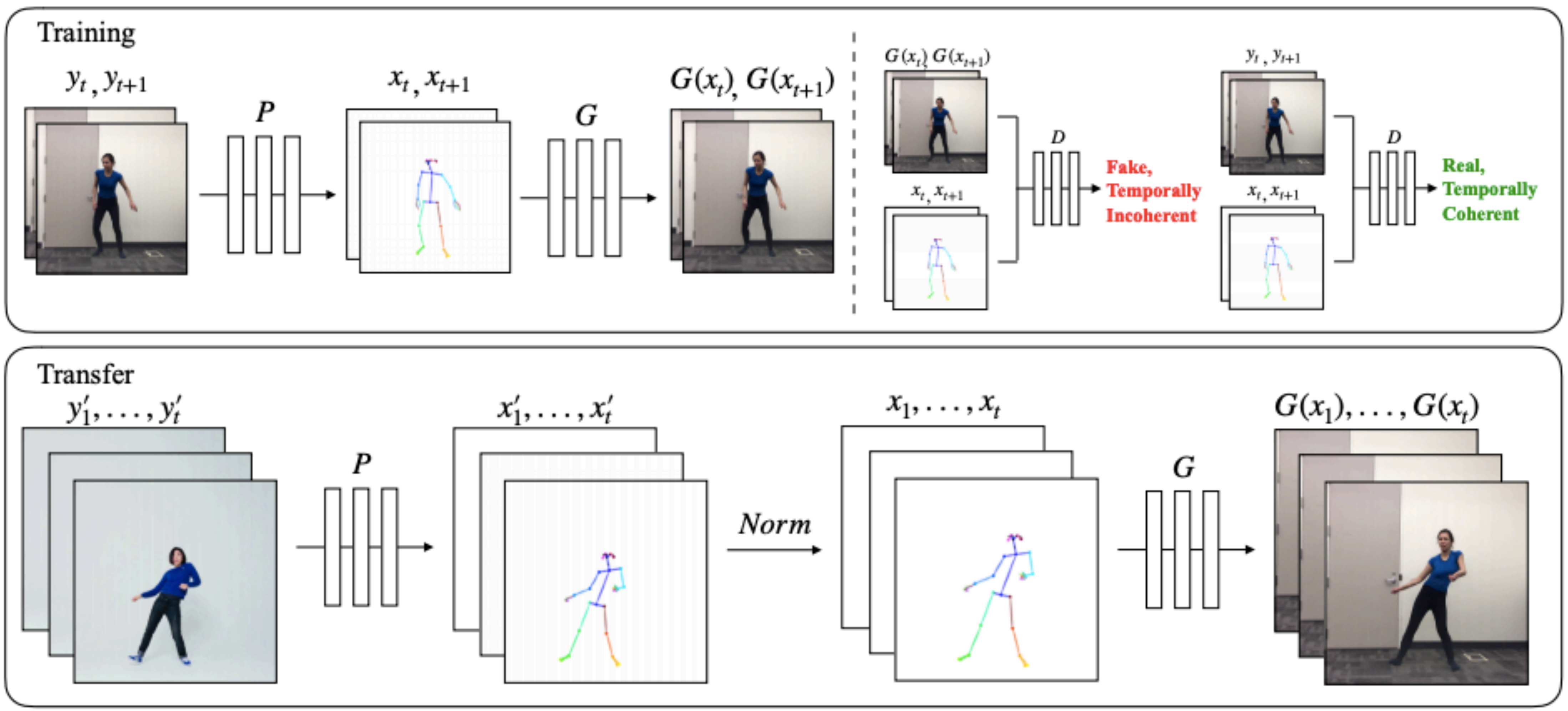
On 8xV100 with full-precision training (no Tensor cores), this script takes 15 days to train to 150k iterations.

Vid-to-vid translation

Vid-to-vid translation

[Carolin Chan, et al, ICCV 2019]

- Everybody dance now



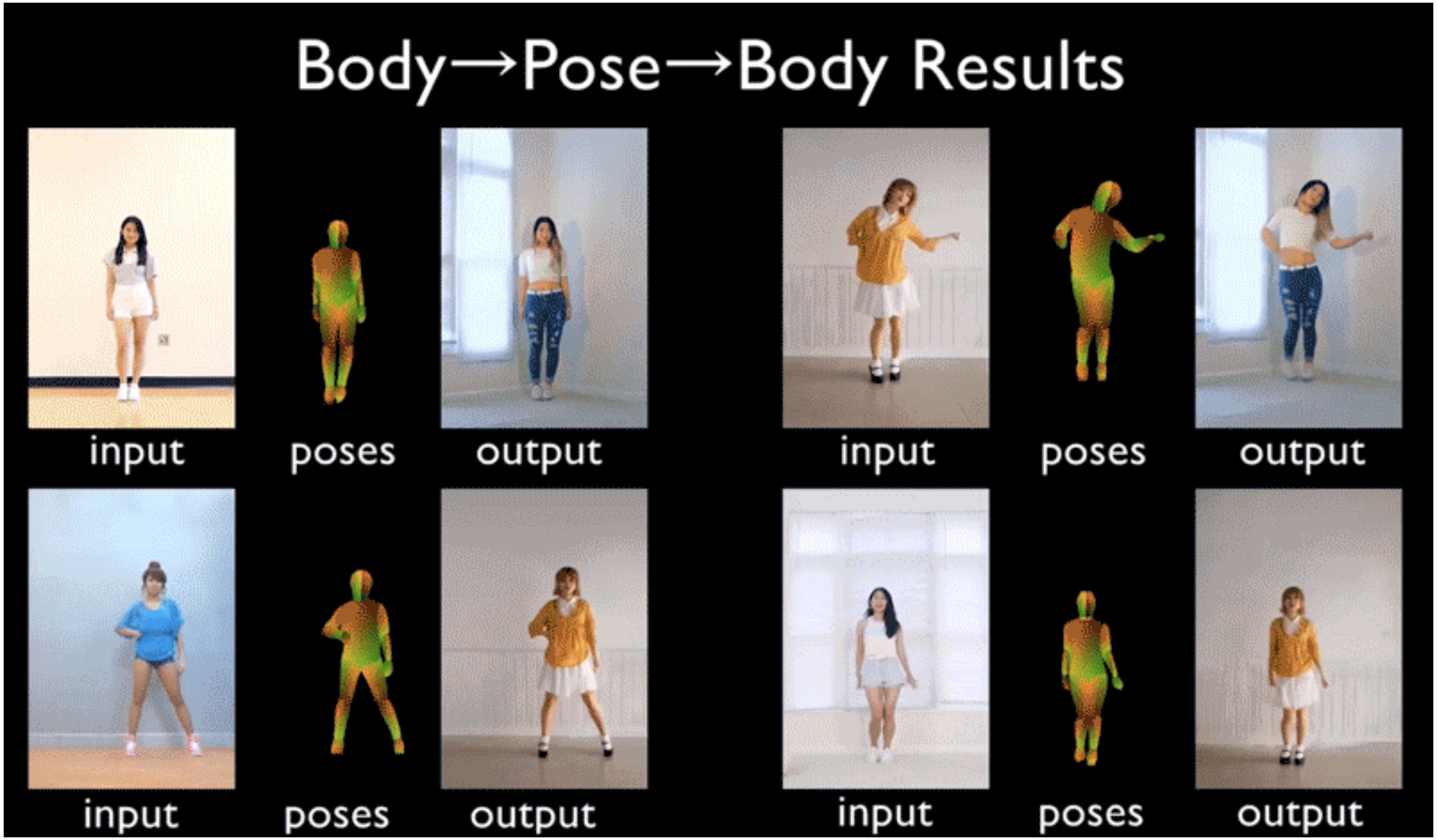
https://www.youtube.com/watch?v=PCBTZh41Ris

Vid-to-vid translation

- Video-to-video synthesis

[Ting-chun Wang, et al, NIPS 2018]

https://github.com/NVIDIA/vid2vid



Outline

- Basic Idea of GAN
- Image Generation
  - Conditional GAN (CGAN, ACGAN)
  - Modern GANs (StyleGAN, BigGAN)
  - Image-to-image translation (Pix2Pix, CycleGAN)
- Video-to-video translation
- GANs Evaluation
- Video Generation
- Lab (DCGAN for manga face generation)

GANs Evaluation

GANs Evaluation

Two Metrics:

- Inception Score (IS) ↑
- Fréchet Inception Distance (FID) ↓

GANs Evaluation-IS

Requirements:

- High-quality (clear contents, sharp images)
- Diversity (different contents)

GANs Evaluation - IS

Conditional generation

Definition： $IS(G) = exp(\mathbb{E}_{x \sim p_g} D_{KL}(p(y|x) \ || \ p(y)))$

x **generated sample**

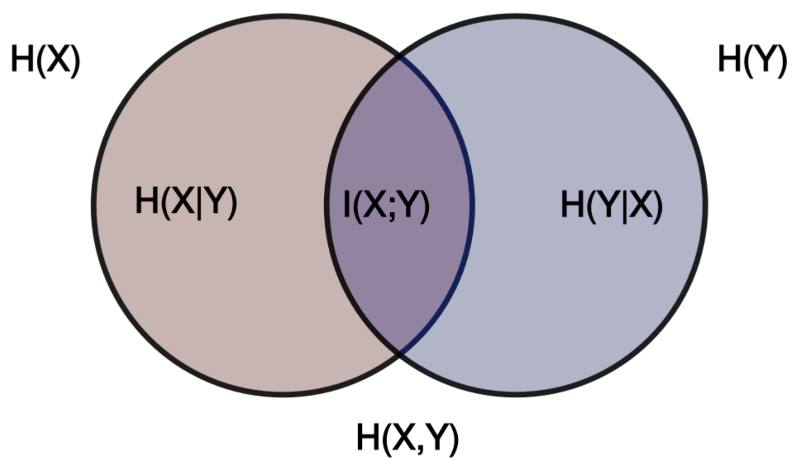
$p(y|x)$ **Conditional distribution**

$p(y) = \int_x p(y|x)p_g(x)$ **Marginal distribution**

GANs Evaluation - IS

Definition： $IS(G) = exp(\mathbb{E}_{x \sim p_g} D_{KL}(p(y|x) \,||\, p(y)))$

$$
\begin{aligned}
ln(IS(G)) &= E_{x\sim p_g} D_{KL}(p(y|x)||p(y)) \\
&= \sum_x p(x) D_{KL}(p(y|x)||p(y)) \\
&= \sum_x p(x) \sum_i p(y=i|x) ln(\frac{p(y=i|x)}{p(y=i)}) \\
&= \sum_x \sum_i p(x, y=i) ln \frac{p(x, y=i)}{p(x)p(y=i)} \\
&= I(y;x) \quad \textbf{Mutual Information} \\
&= H(y) - H(y|x) \quad \textbf{Entropy difference}
\end{aligned}
$$



GANs Evaluation - IS

- The generative algorithm should output a high diversity of images from all the different classes in ImageNet —> $H(y)$ should be high

- The images generated should contain clear objects (i.e. the images are sharp rather than blurry) —> $H(y|x)$ should be low

GANs Evaluation - IS

$$IS(G) = exp(\mathbb{E}_{x\sim p_g} D_{KL}(p(y|x) \; || \; p(y)))$$

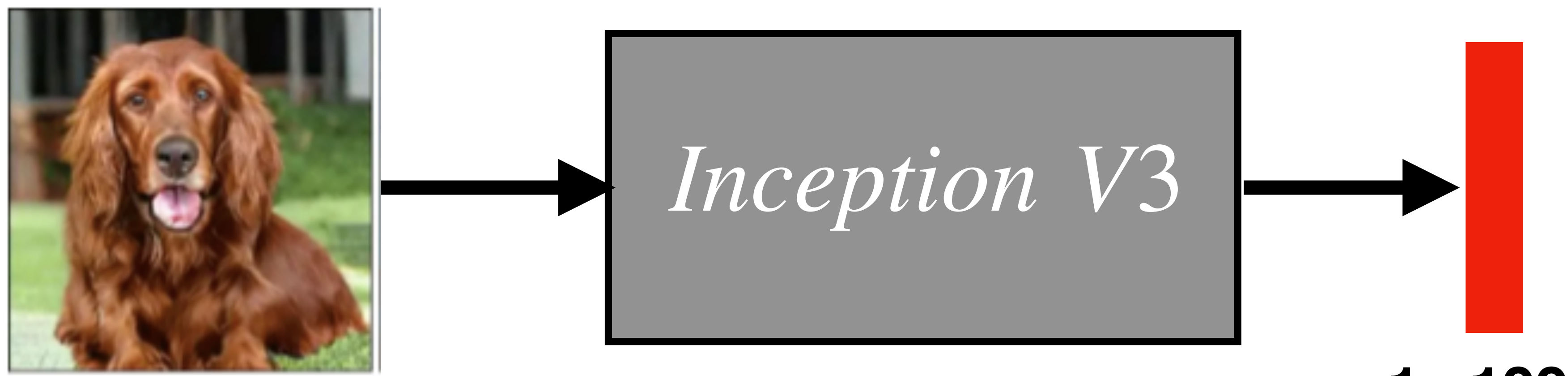
$$\Downarrow$$

$$IS(G) \approx exp(\mathbb{E}_{x\sim p_g} D_{KL}(p(y|x^{(i)}) \; || \; \hat{p}(y)))$$

$$\hat{p}(y) = \frac{1}{N} \sum_{i=1}^{N} p(y|x^{(i)})$$

GANs Evaluation - IS

$$p(y \mid x^{(i)})$$



Pre-trained on ImageNet

1x 1000

Sampled 5000 images from GANs

Attention: training and evaluation must use the same dataset

GANs Evaluation - IS

Problem of IS ?

Not considering the distribution of training dataset

GANs Evaluation - FID

$$FID = |\mu - \mu_w|^2 + tr(\Sigma + \Sigma_w - 2(\Sigma\Sigma_w)^{1/2})$$

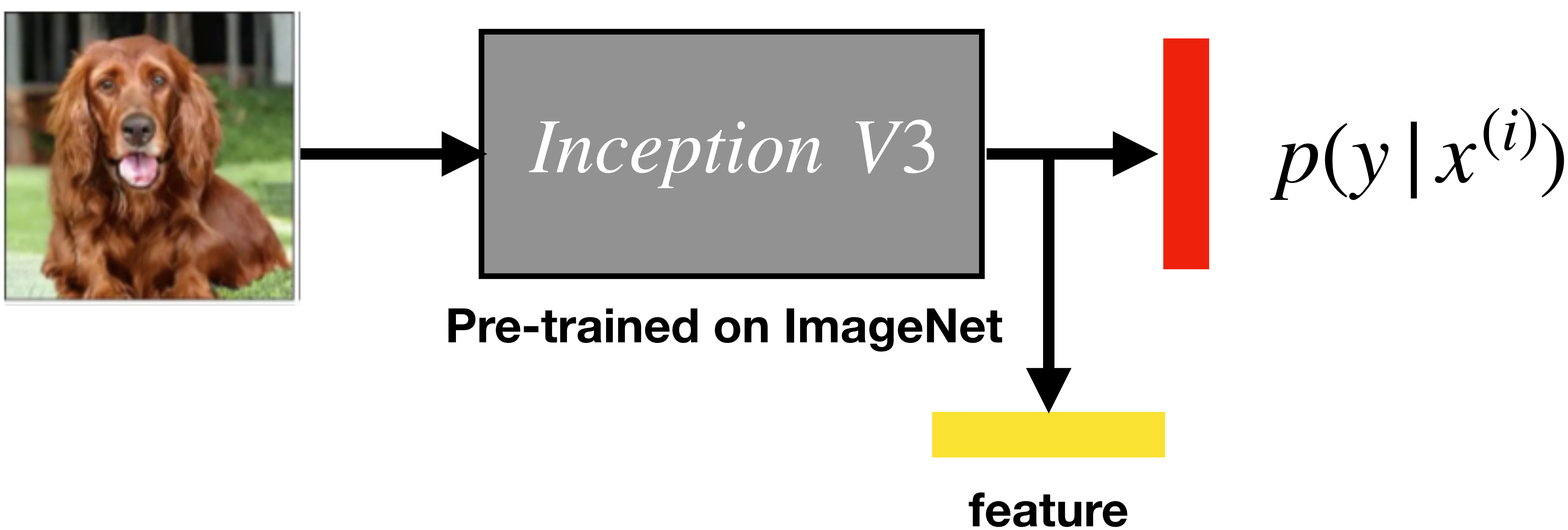
$N(\mu, \Sigma)$ distribution of generated set

$N(\mu_w, \Sigma_w)$ distribution of training set

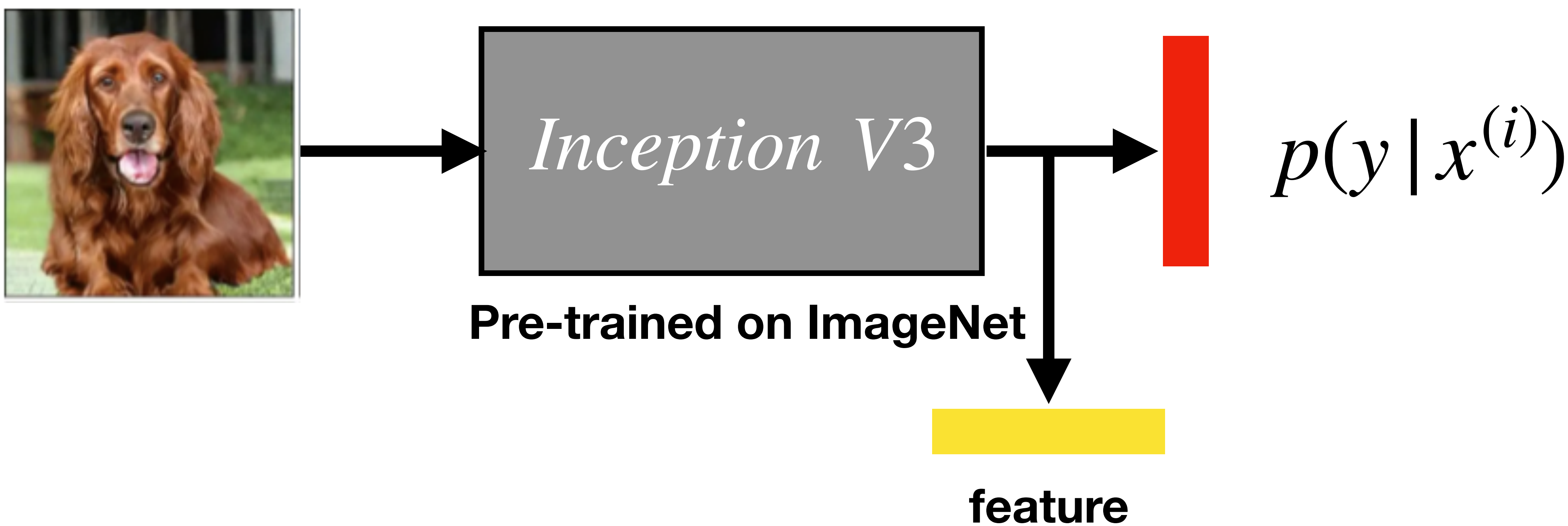
Computer Fréchet distance
between two distributions

GANs Evaluation - FID

$$FID = |\mu - \mu_w|^2 + tr(\Sigma + \Sigma_w - 2(\Sigma\Sigma_w)^{1/2})$$



GANs Evaluation - FID



Sampled approximately same number images from GANs as original training set, ideally 10000 images for ImageNet

GANs Evaluation

- IS and FID are two metrics for GANs evaluation. FID is more widely used than IS. However, both methods require large-scale generated samples.

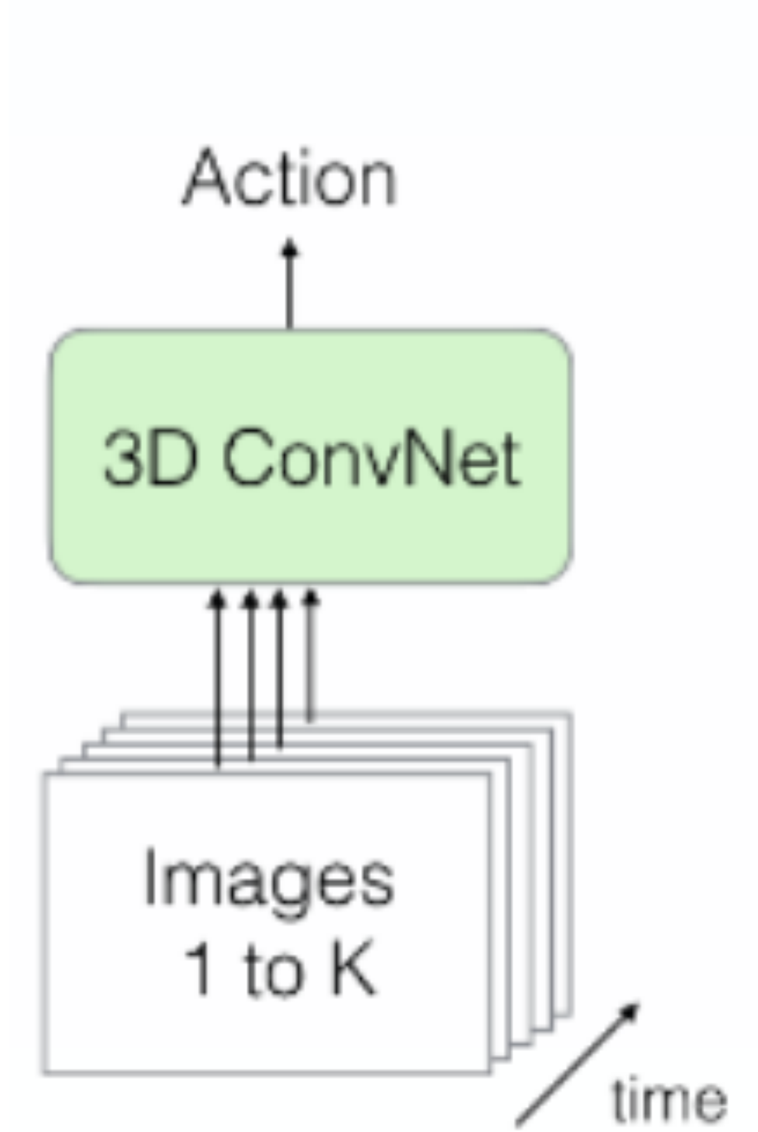
- New metric needs to be proposed to guide training process.

Video Generation

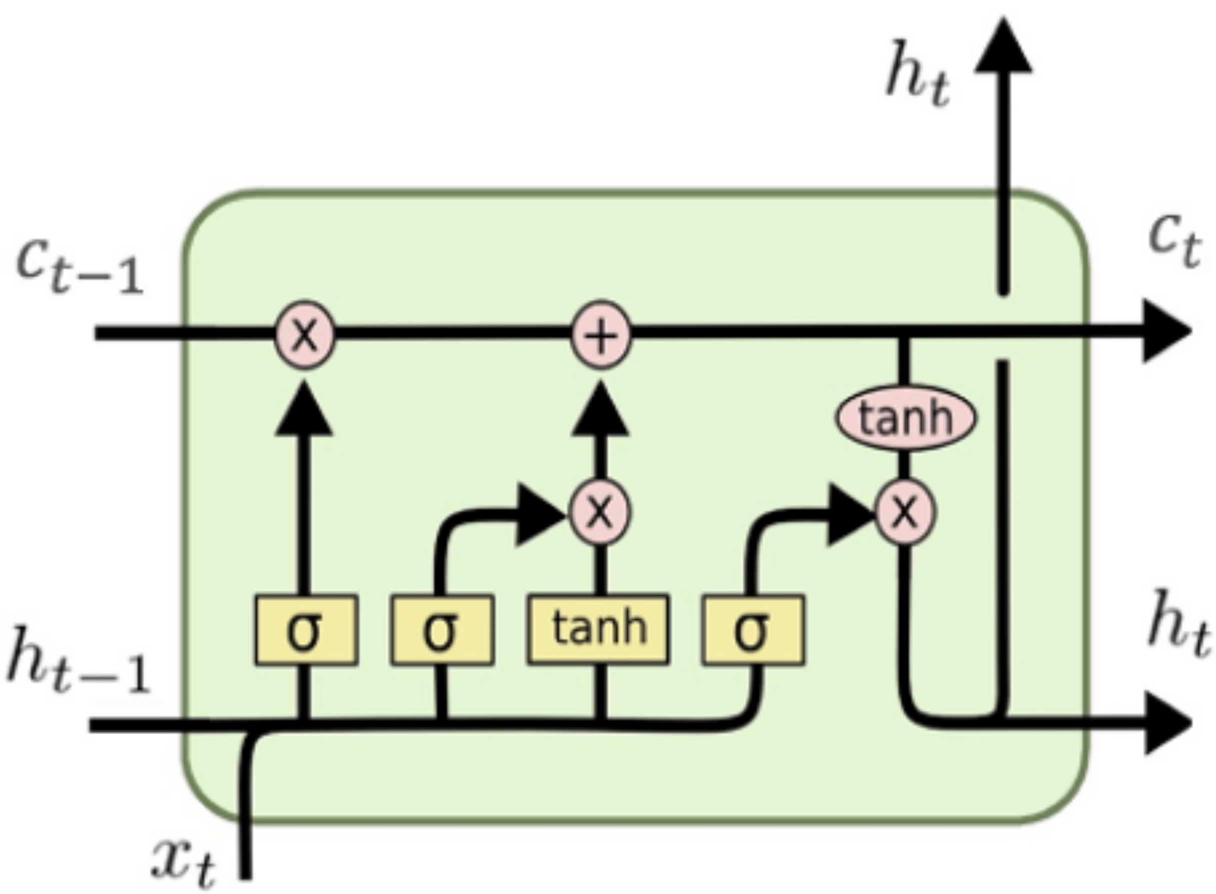
Recall: Spatio-temporal Modeling

- *Spatiotemporal CNN (3D CNN)*
- *LSTM and GRU*

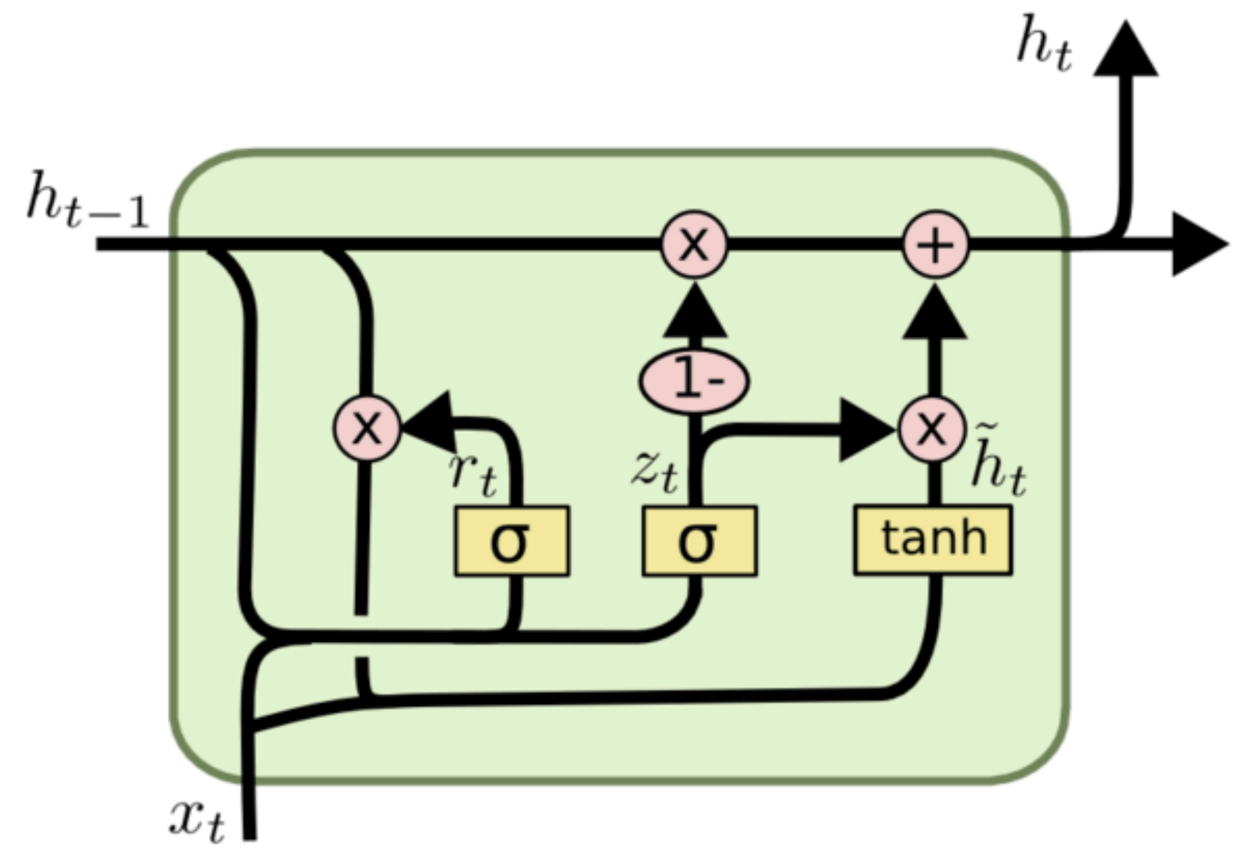
Spatiotemporal CNN (3D CNN)



LSTM and GRU

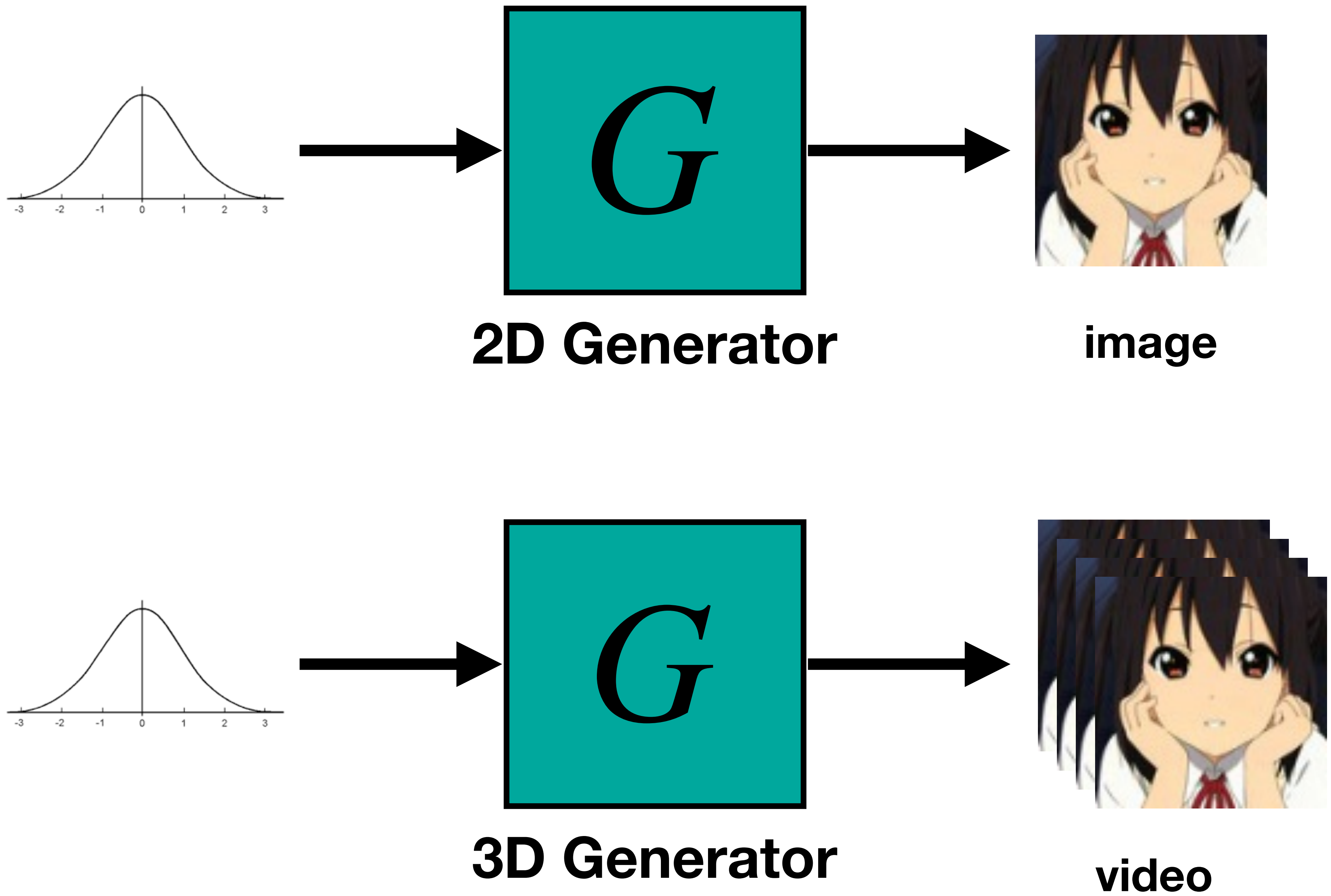


LSTM

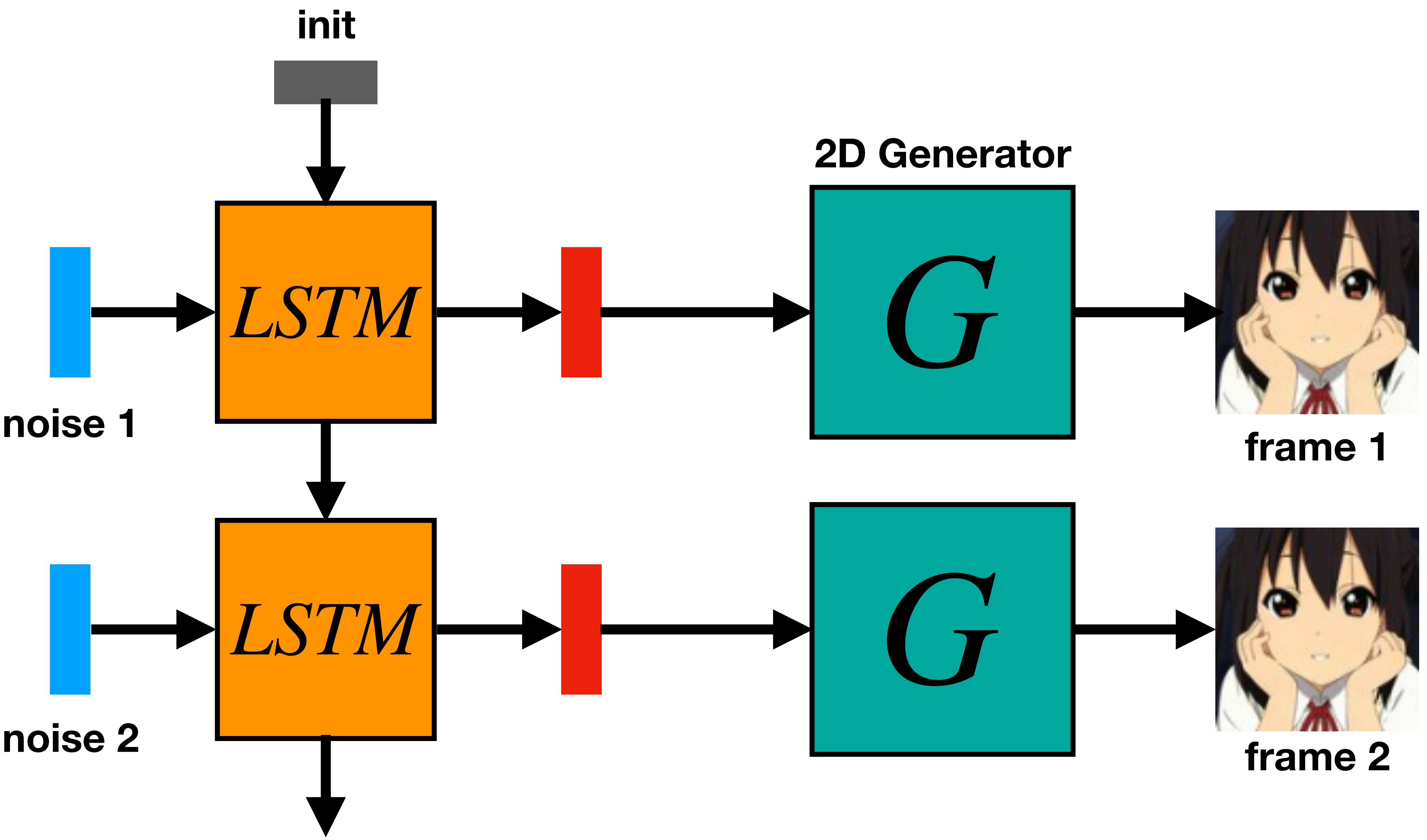


GRU

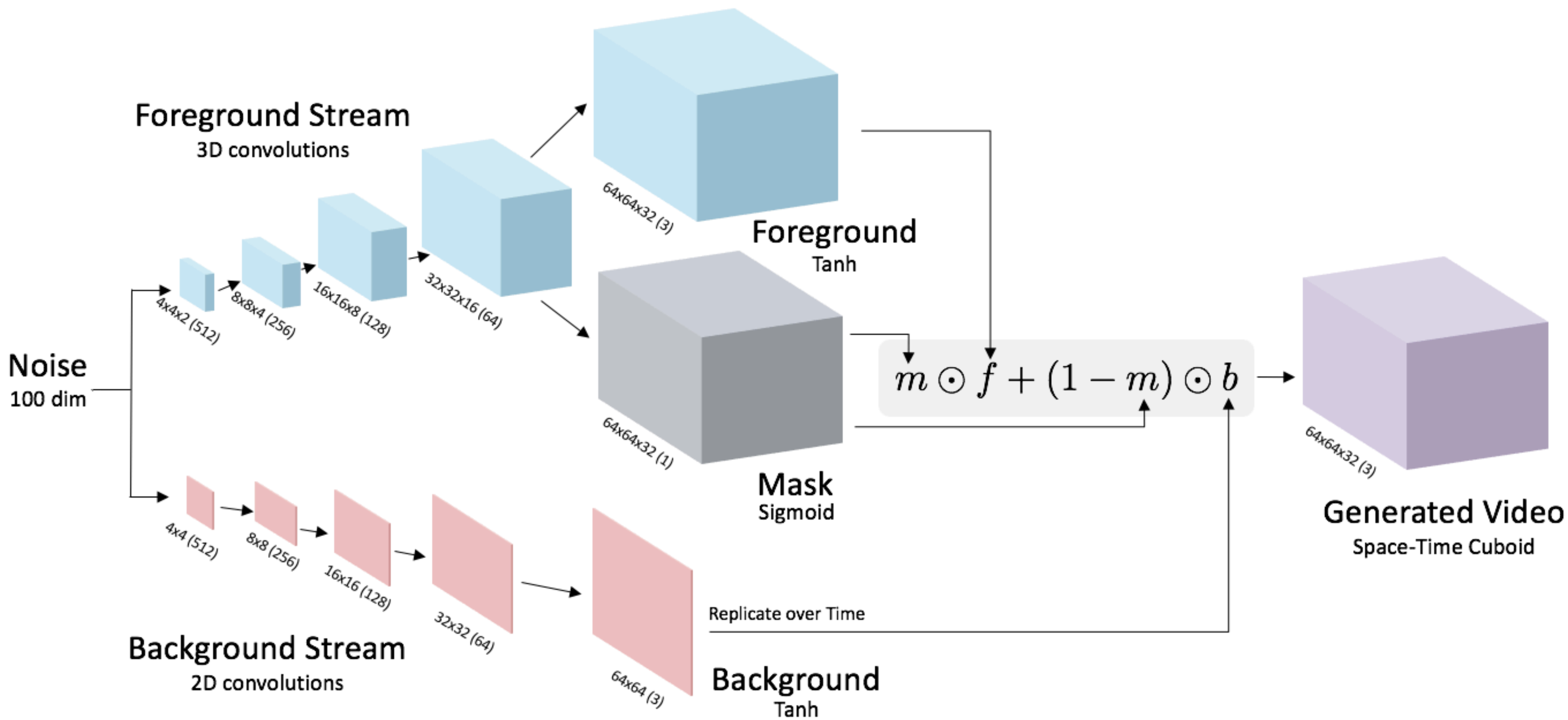
Spatiotemporal CNN (3D CNN)



LSTM and GRU

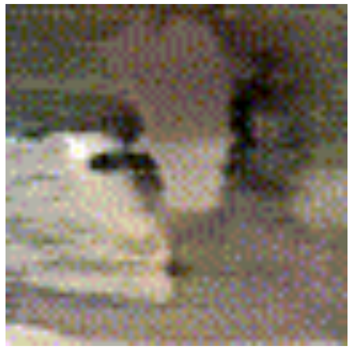
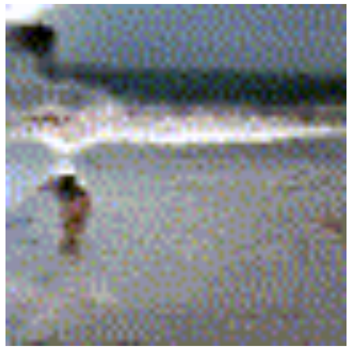


VGAN



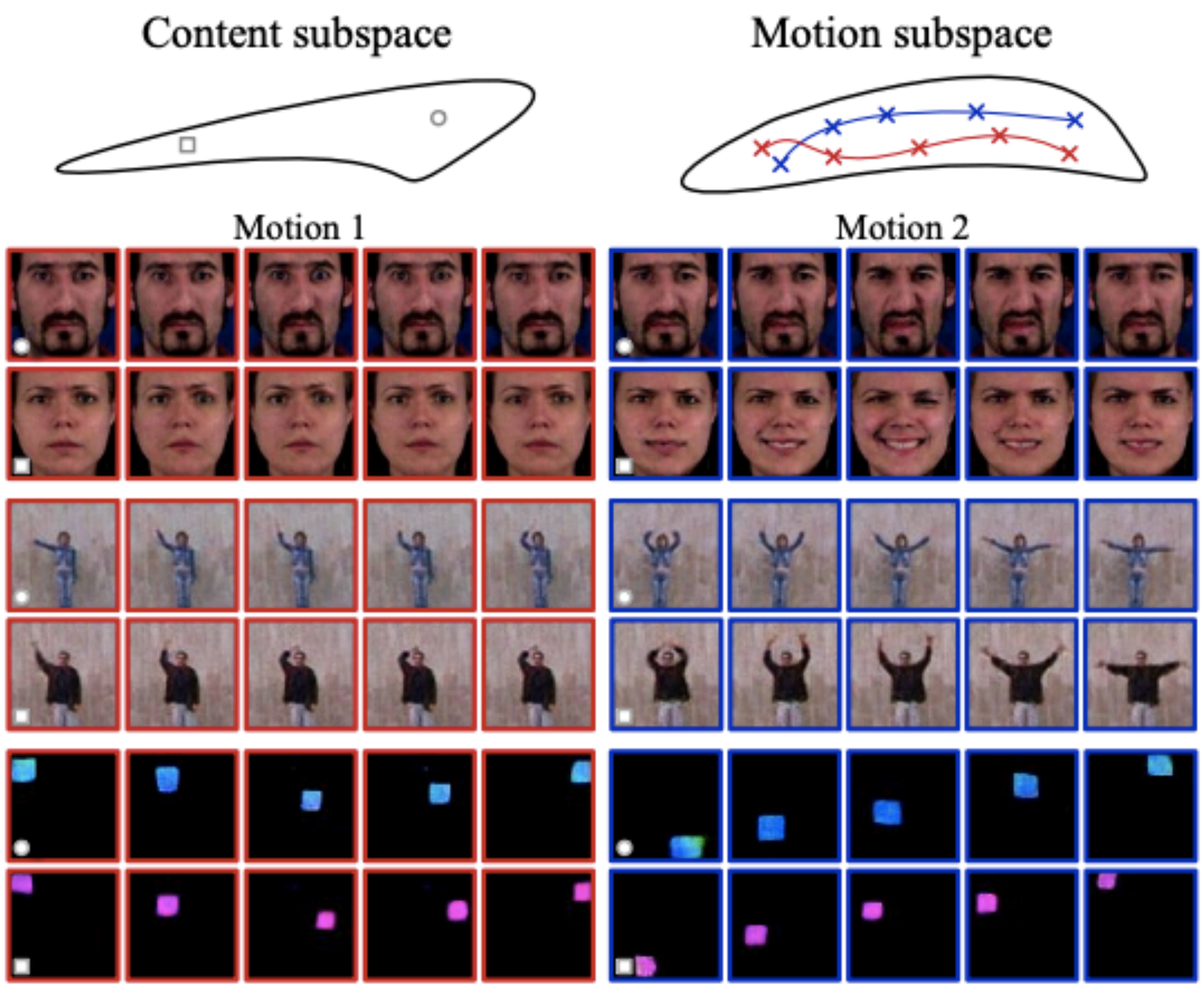
VGAN [NeurIPS’16]

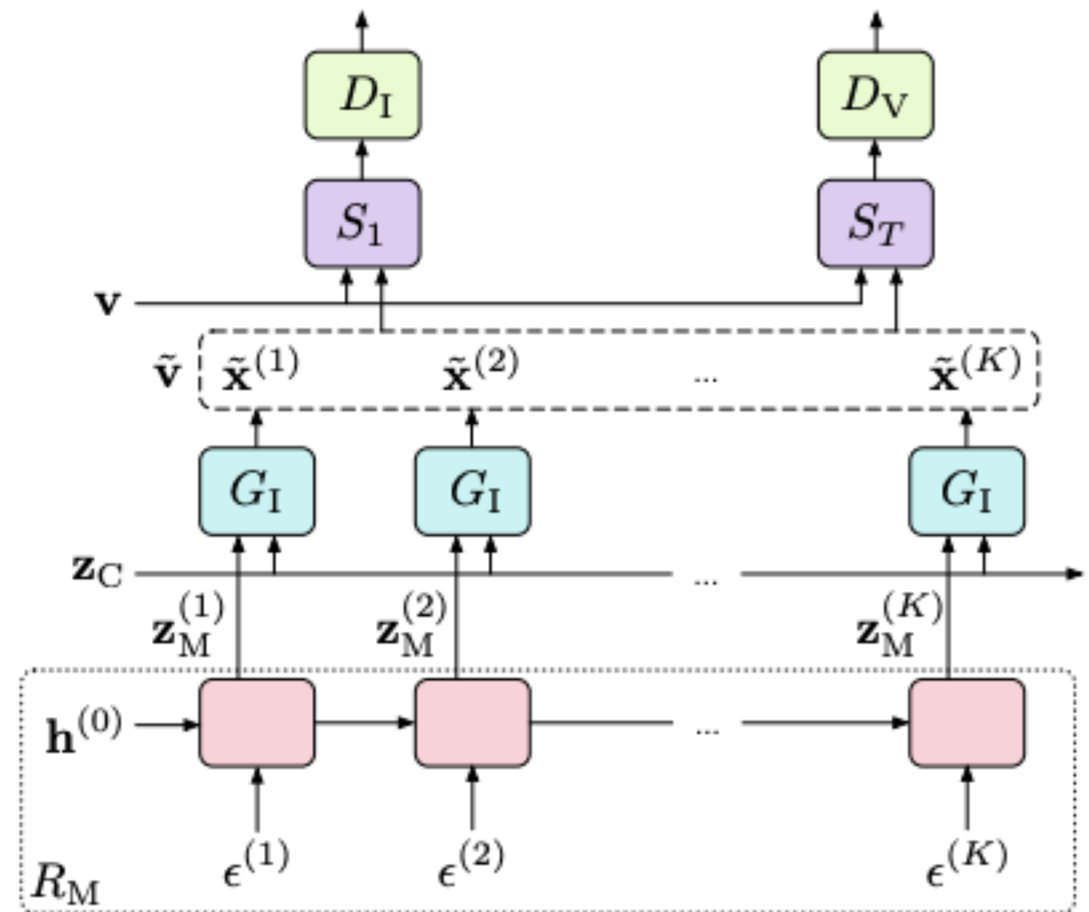
VGAN

http://www.cs.columbia.edu/~vondrick/tinyvideo/

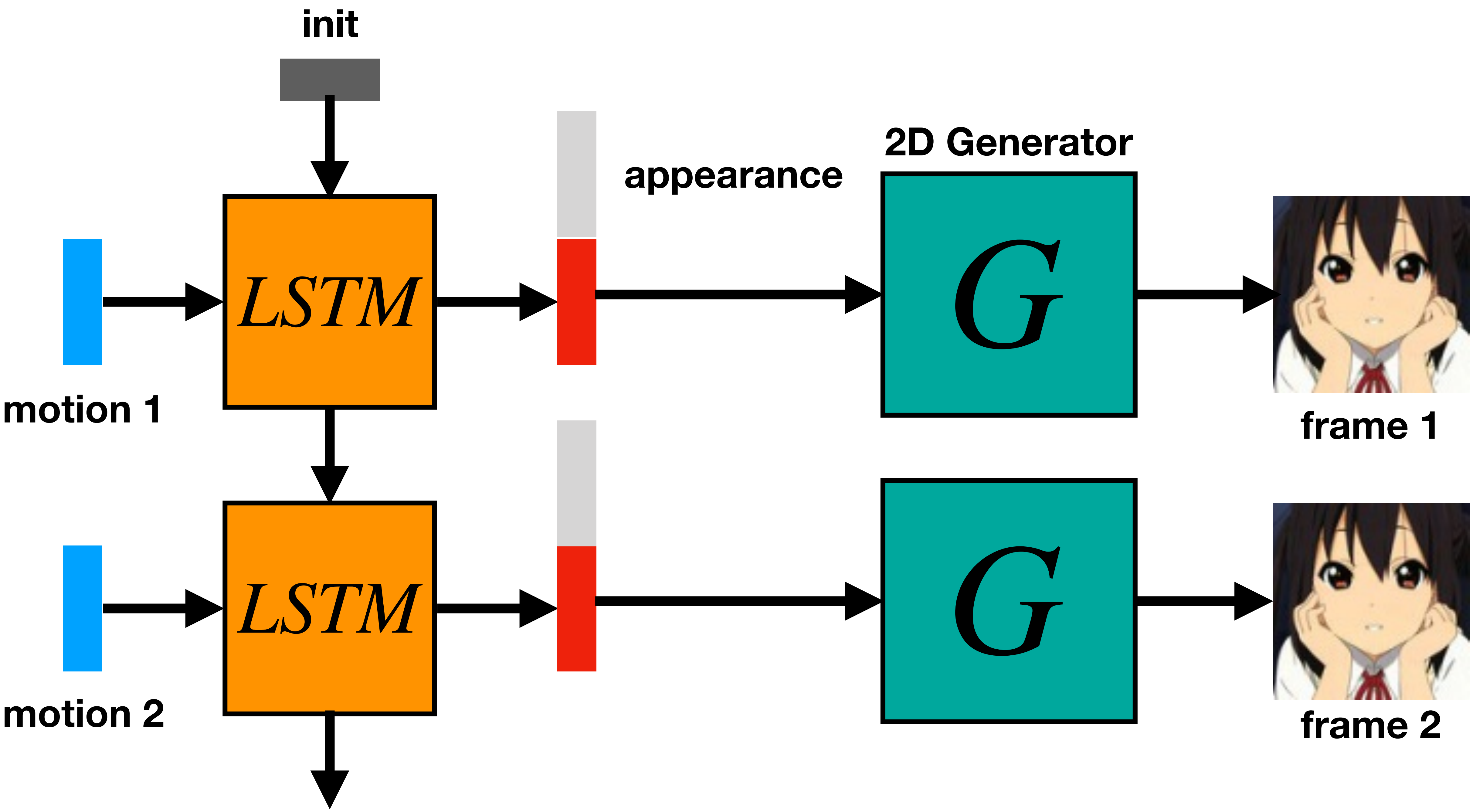
MoCoGAN



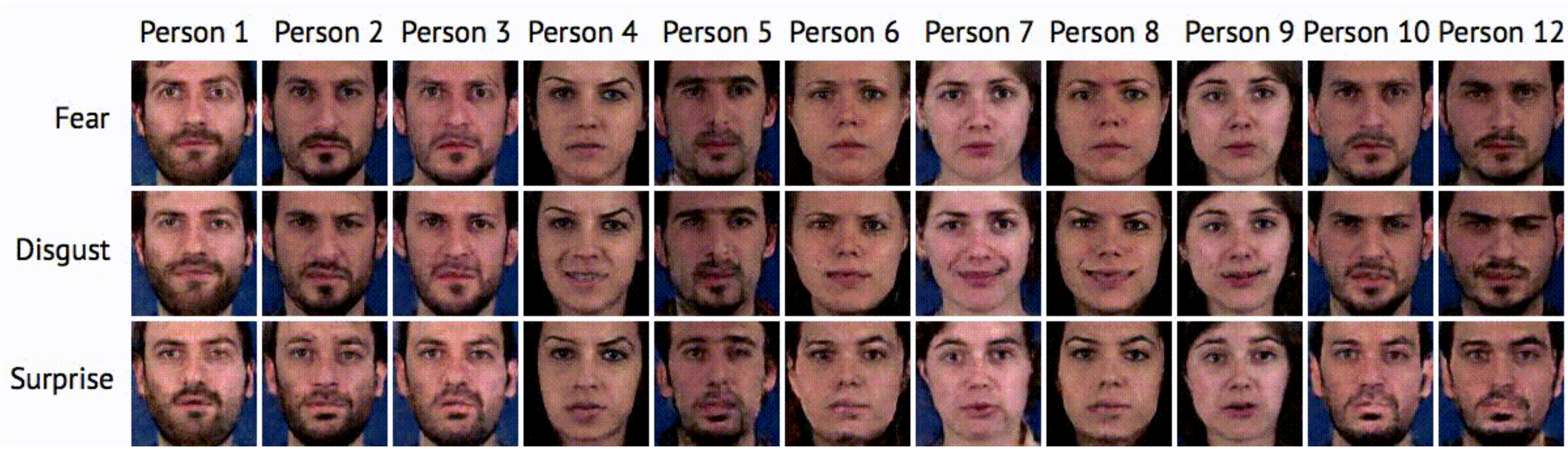


MoCoGAN [CVPR'19]

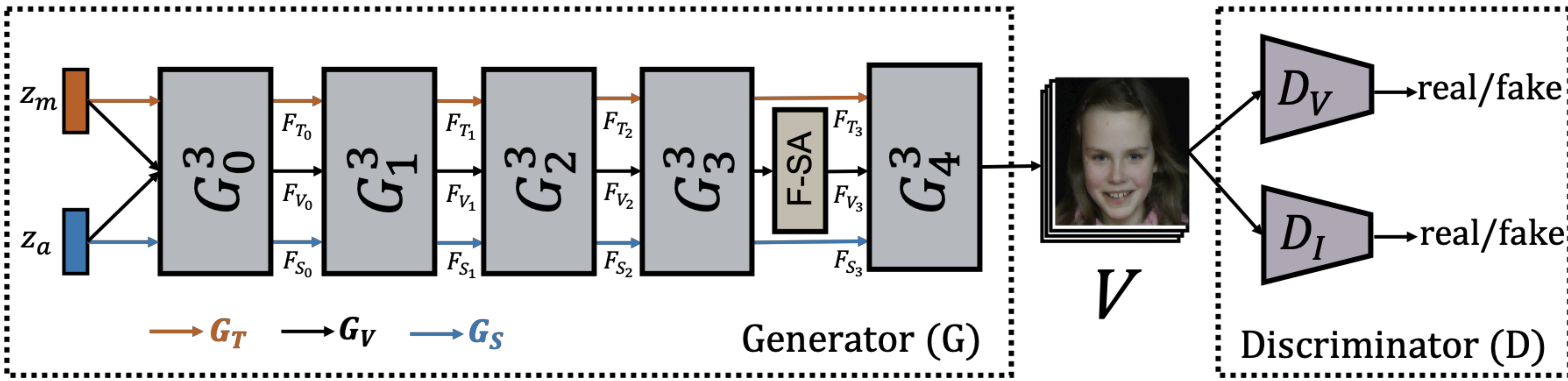
MoCoGAN

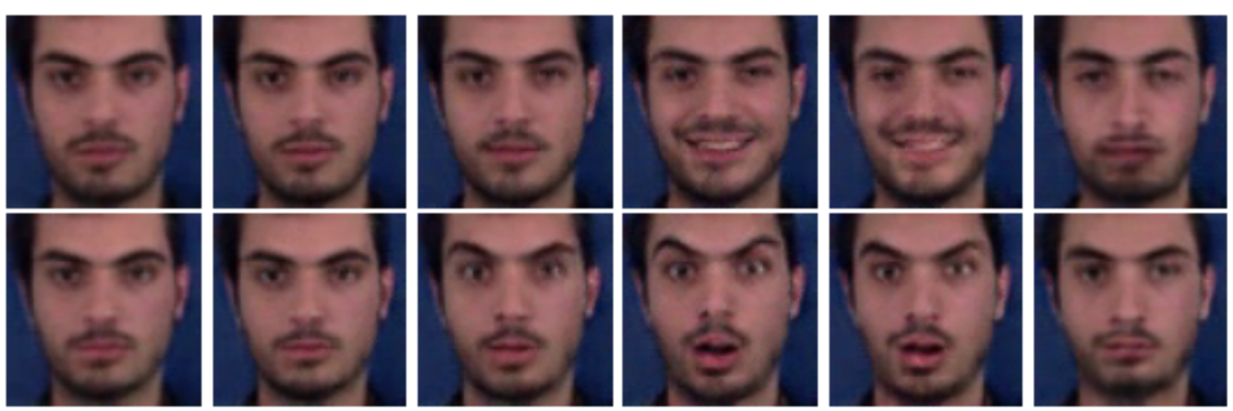

init
appearance
2D Generator
LSTM
LSTM
G
G
motion 1
motion 2
frame 1
frame 2

MoCoGAN


Person 1
Person 2
Person 3
Person 4
Person 5
Person 6
Person 7
Person 8
Person 9
Person 10
Person 12
Fear
Disgust
Surprise

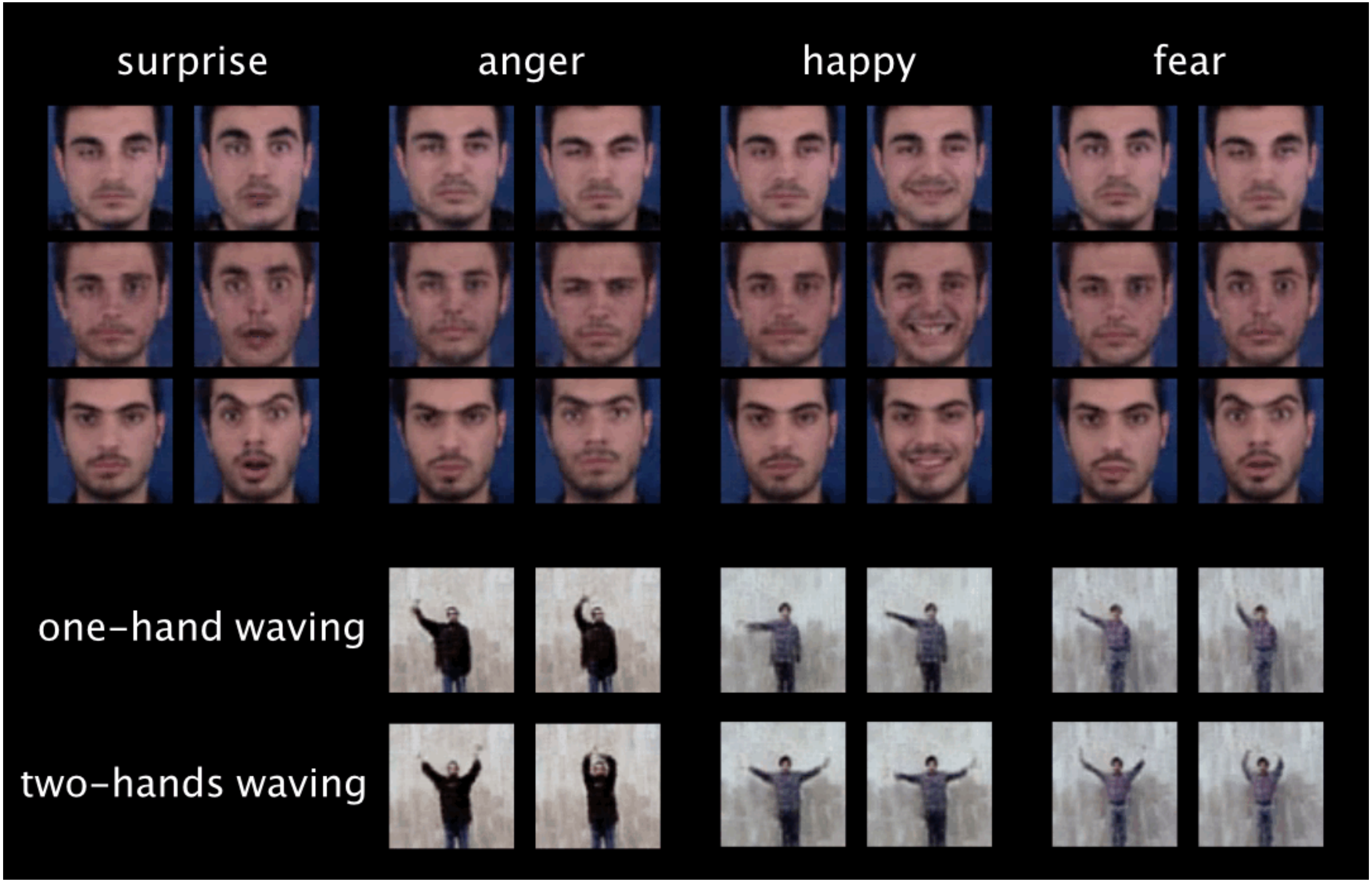
G^{3}AN





G3AN [CVPR'20]

G³AN



What I can not create, I do not understand

- R. Feynman

Thank You !