Machine Learning

Conditional Generative Adversarial Networks (GAN)

Instructor: Prof. Yi Fang

yfang@nyu.edu

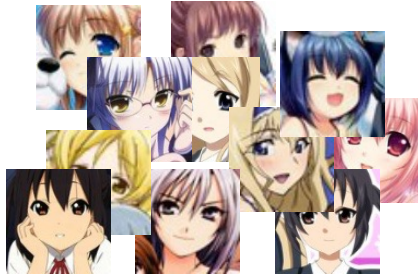# Generative Adversarial Network

GAN's Architecture
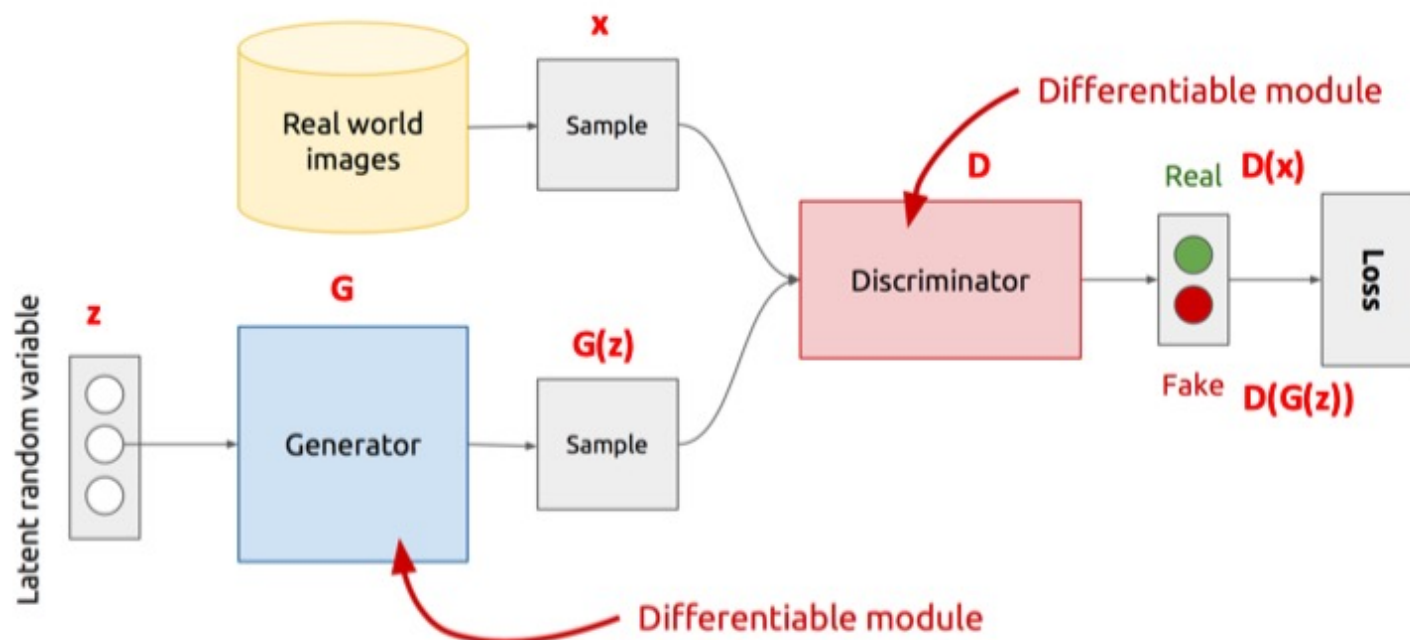
- **Z** is some random noise (Gaussian/Uniform).
- **Z** can be thought as the latent representation of the image.

Three Categories of GAN

1. *Typical GAN*

$$\begin{bmatrix} -0.3 \\ 0.1 \\ \vdots \\ 0.9 \end{bmatrix}$$ → Generator → image

random vector

2. *Conditional GAN*

blue eyes, red hair, short hair

paired data

"Girl with red hair" text → Generator → image
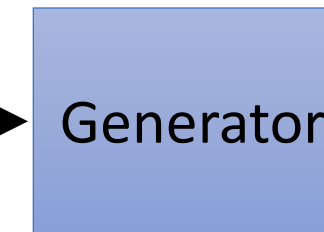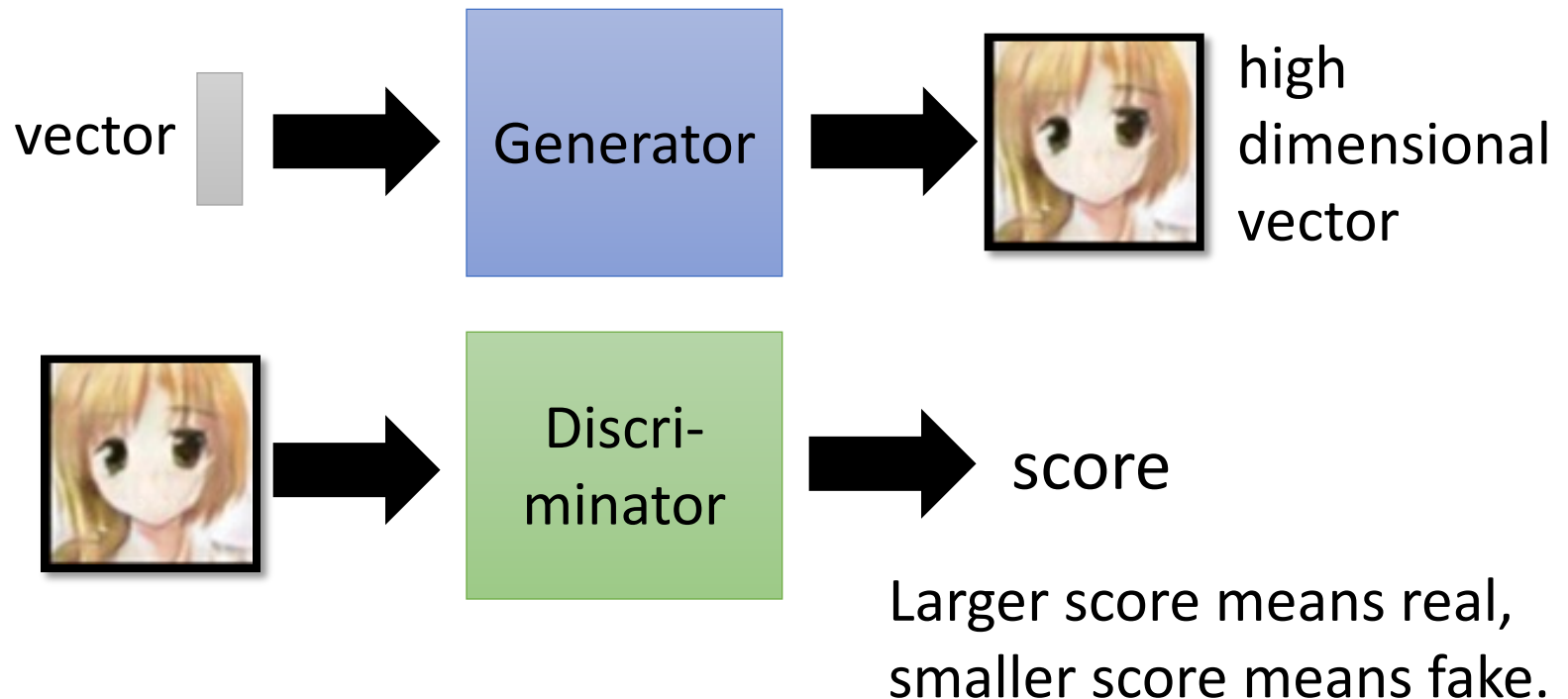
3. *Unsupervised Conditional GAN*

domain x   domain y

unpaired data

x Photo → Generator → y

Vincent van Gogh's style

Generative Adversarial Network (GAN)

- Anime face generation as example

vector → Generator → high dimensional vector

Discriminator → score

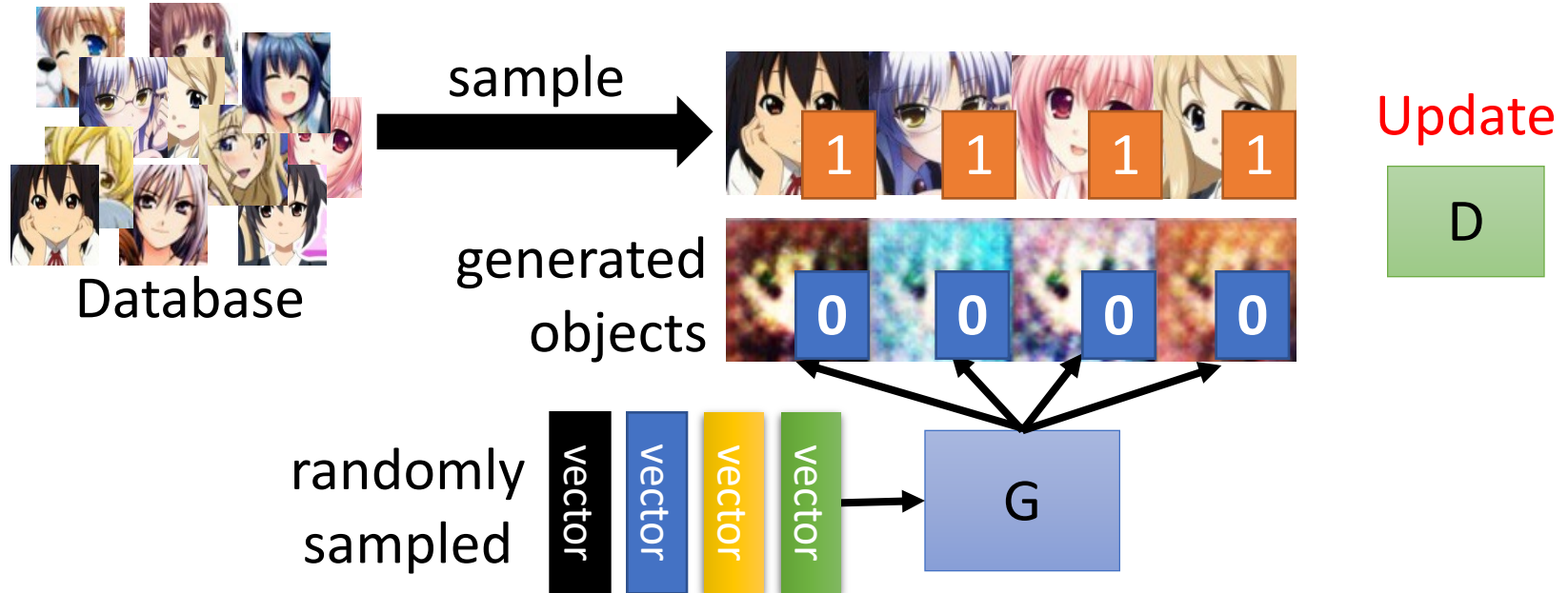Larger score means real, smaller score means fake.

Algorithm

- Initialize generator and discriminator **G** **D**

- In each training iteration:

Step 1: Fix generator G, and update discriminator D

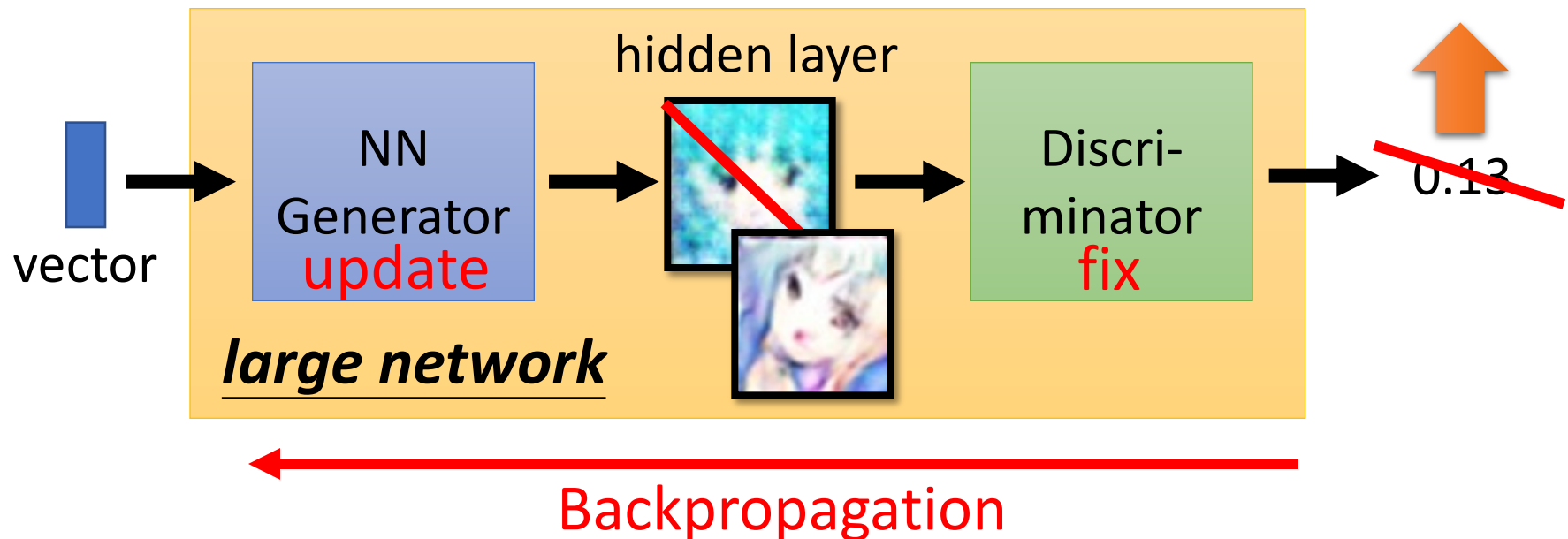Discriminator learns to assign high scores to real objects and low scores to generated objects.

Algorithm

- Initialize generator and discriminator $\boxed{G}$ $\boxed{D}$

- In each training iteration:

Step 2: Fix discriminator D, and update generator G

Generator learns to "*fool*" the discriminator

Algorithm
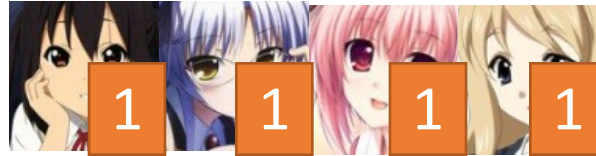
- Initialize generator and discriminator G D

- In each training iteration:

Learning D

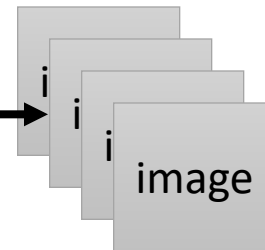Sample some real objects:

Generate some fake objects:

Update

D

fix

vector vector vector vector → G

Learning G

vector vector vector vector → G → image → D → 1

update

fix

GAN is hard to train ……
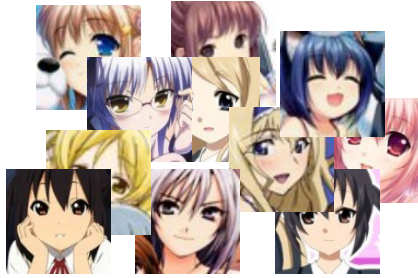

(I found this joke from 陳柏文's facebook.)

Three Categories of GAN

1. Typical GAN

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

random vector → Generator → image

2. Conditional GAN

blue eyes, red hair, short hair

paired data

"Girl with red hair"
text → Generator → image

3. Unsupervised Conditional GAN

domain x   domain y   x

unpaired data

Photo → Generator → y

Vincent van Gogh's style

Text-to-Image
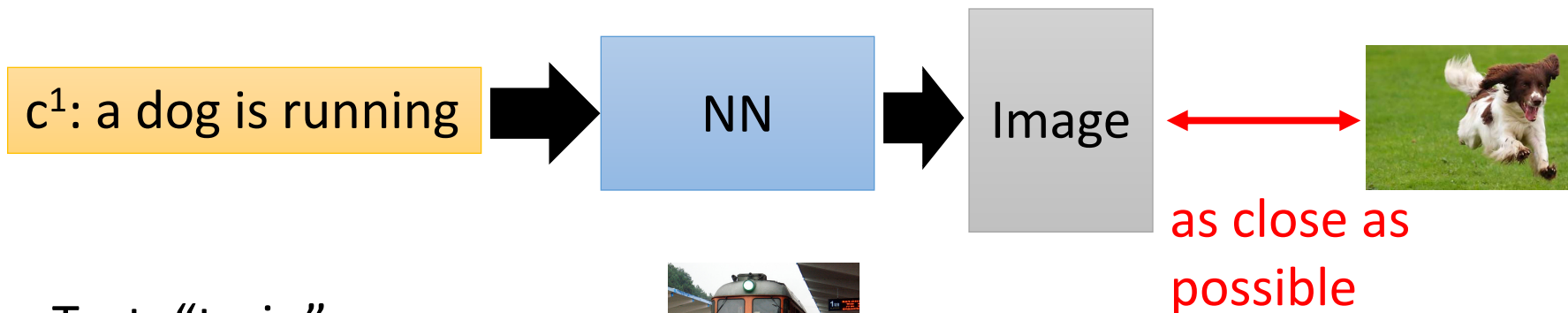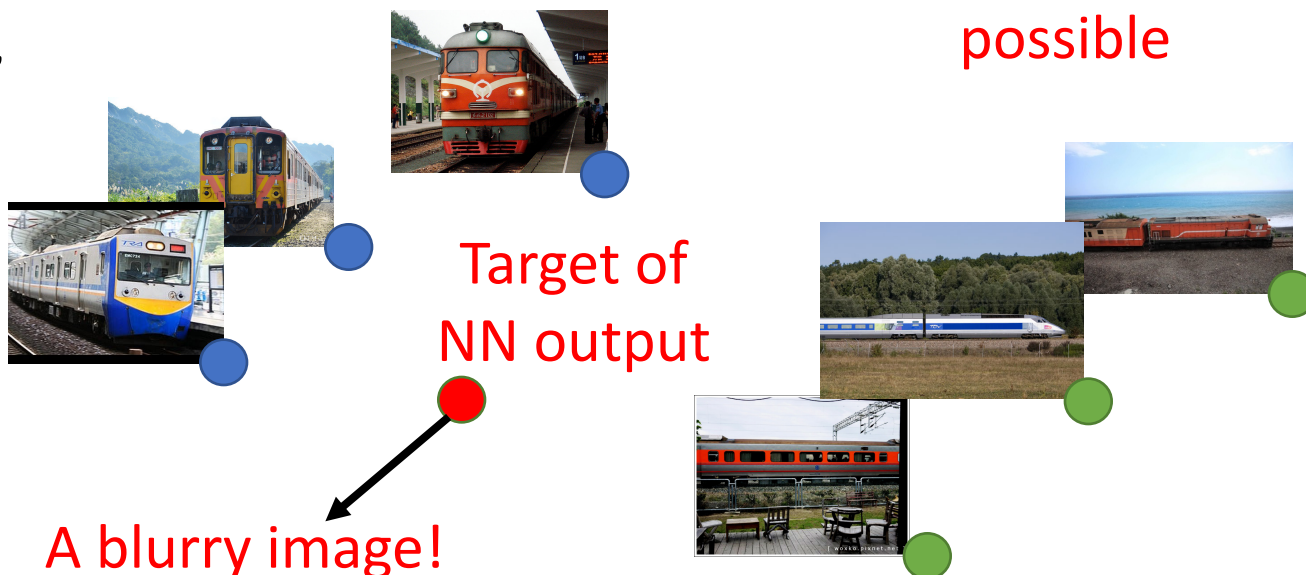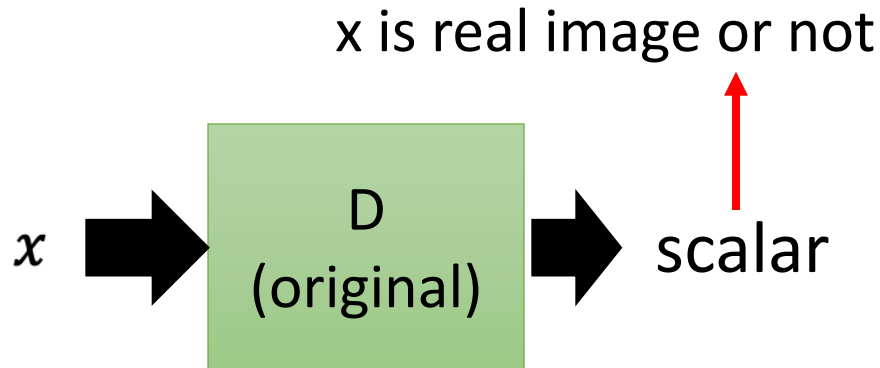
a dog is running

a bird is flying

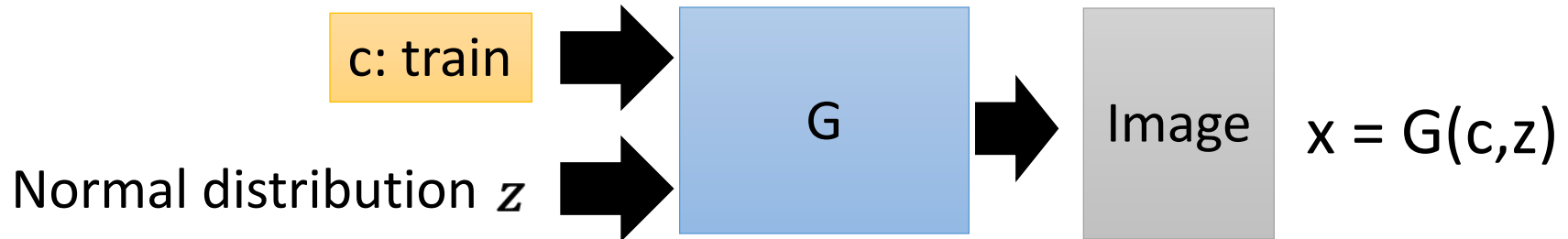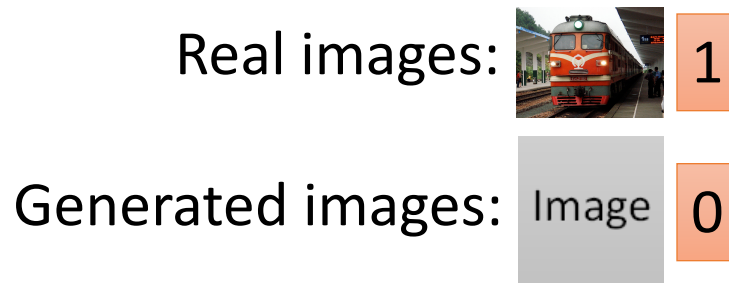- **Traditional supervised approach**

c^1: a dog is running → NN → Image ←→

as close as possible

Text: "train"

Target of NN output

A blurry image!

Conditional GAN

c: train → G → Image $x = G(c,z)$

Normal distribution z →
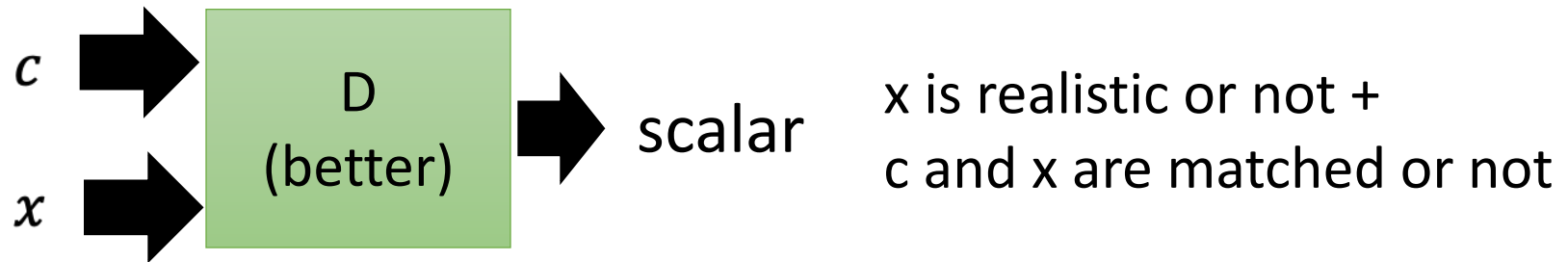
x is real image or not

x → D (original) → scalar

Generator will learn to generate realistic images ....

But completely ignore the input conditions.

Real images: 1

Generated images: Image 0

Conditional GAN

c: train →

Normal distribution z →

G → Image $x = G(c,z)$

c →

x →

D (better) → scalar

x is realistic or not +
c and x are matched or not
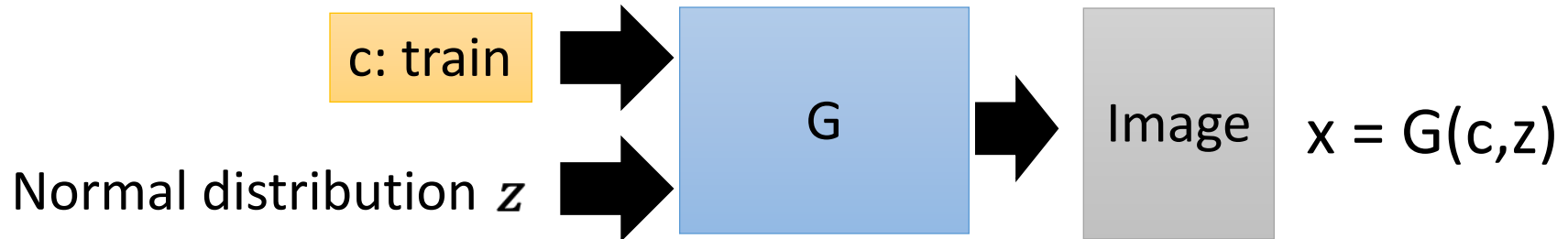
True text-image pairs: (train ,   ) 1

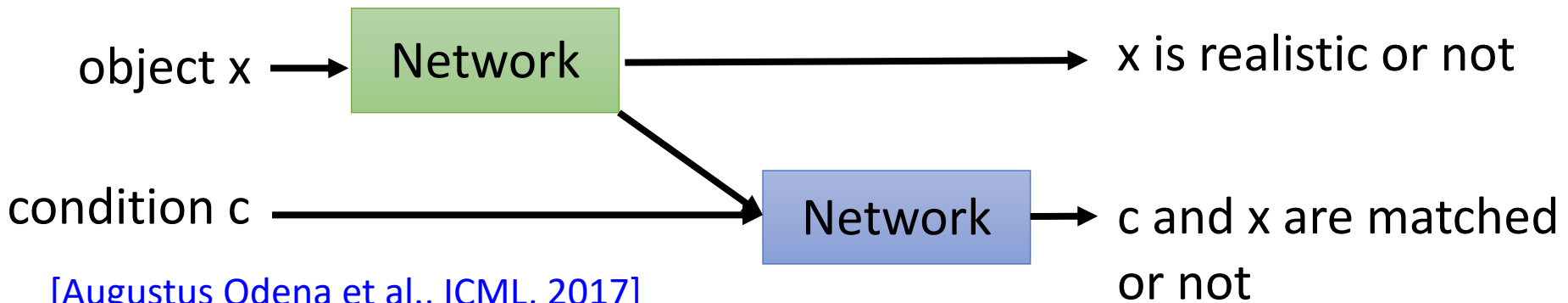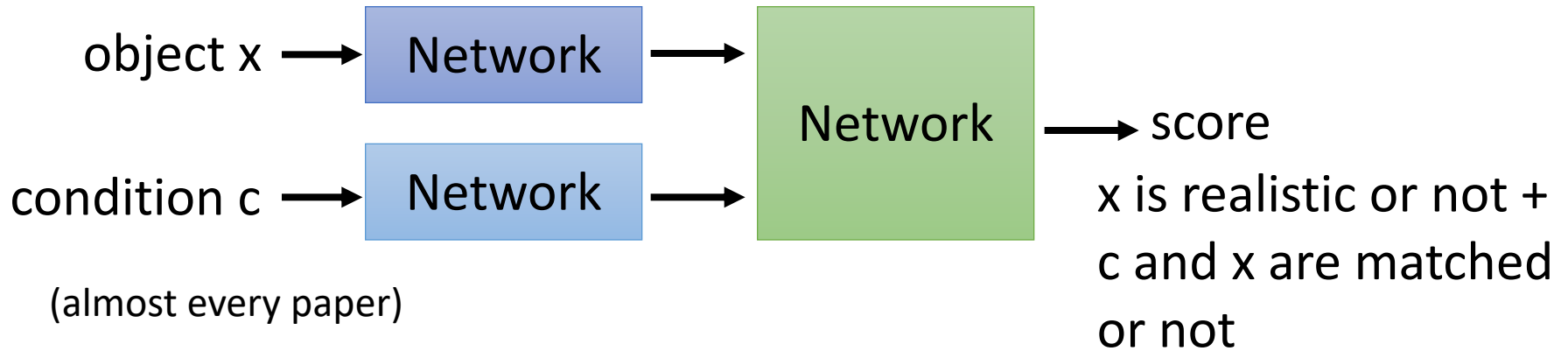(cat ,   ) 0   (train , Image) 0

- In each training iteration:

Learning D

- Sample m positive examples $\{(c^1, x^1), (c^2, x^2), \ldots, (c^m, x^m)\}$ from database
- Sample m noise samples $\{z^1, z^2, \ldots, z^m\}$ from a distribution
- Obtaining generated data $\{\hat{x}^1, \tilde{x}^2, \ldots, \tilde{x}^m\}$, $\tilde{x}^i = G(c^i, z^i)$
- Sample m objects $\{\hat{x}^1, \hat{x}^2, \ldots, \hat{x}^m\}$ from database
- Update discriminator parameters θ_d to maximize
  - $\tilde{V} = \frac{1}{m} \sum_{i=1}^{m} log D(c^i, x^i)$
    $$+ \frac{1}{m} \sum_{i=1}^{m} log\left(1 - D(c^i, \tilde{x}^i)\right) + \frac{1}{m} \sum_{i=1}^{m} log\left(1 - D(c^i, \hat{x}^i)\right)$$
  - $\theta_d \leftarrow \theta_d + \eta \nabla \tilde{V}(\theta_d)$

Learning G

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

Conditional GAN - Discriminator

object x → Network →

condition c → Network →

→ Network → score

x is realistic or not +

c and x are matched or not

(almost every paper)

object x → Network → x is realistic or not

condition c → Network → c and x are matched or not

[Augustus Odena et al., ICML, 2017]
[Takeru Miyato, et al., ICLR, 2018]
[Han Zhang, et al., arXiv, 2017]

Conditional GAN

paired data
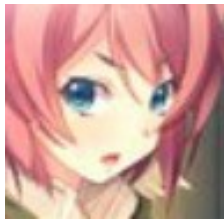
blue eyes
red hair
short hair

Collecting anime faces and the description of its characteristics

red hair,
green eyes

blue hair,
red eyes

Stack GAN

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, Dimitris Metaxas, "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks". ICCV. 2017
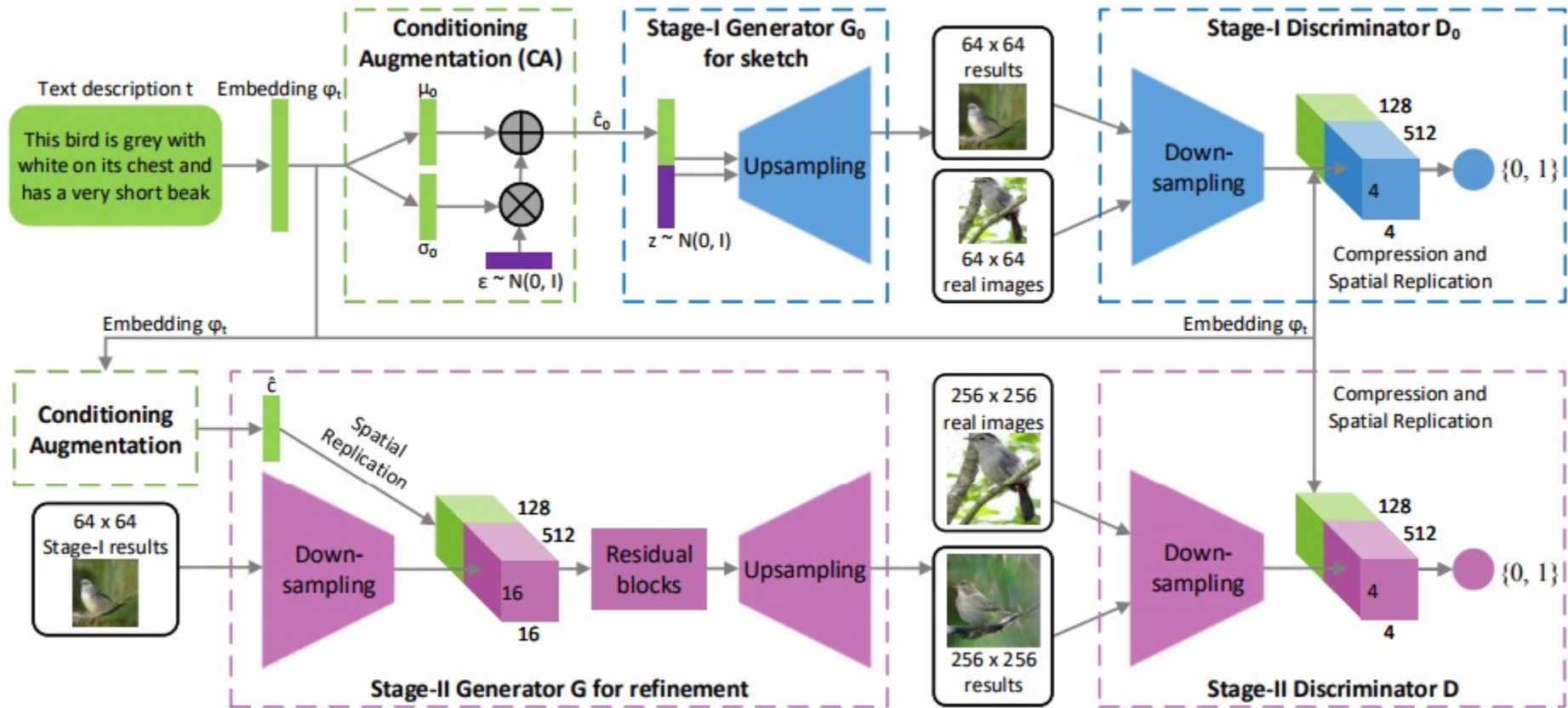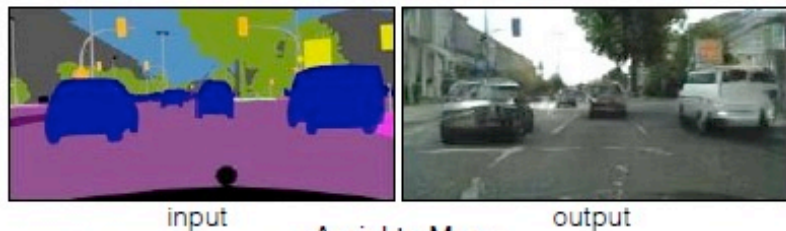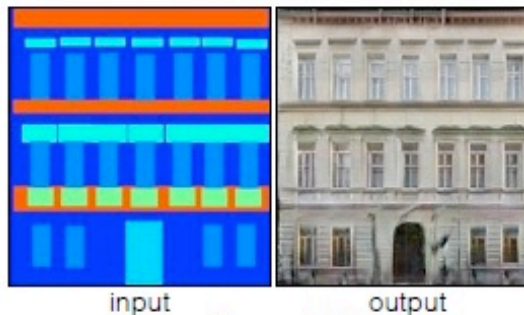
Image-to-image
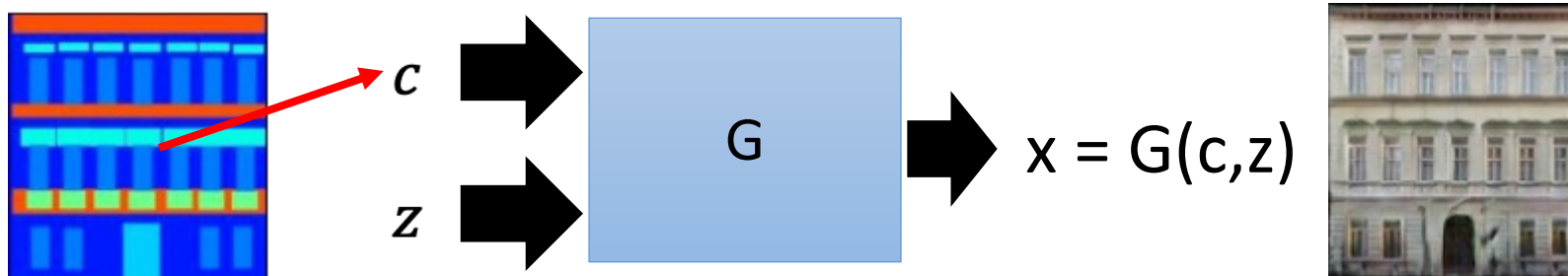
c → G → $x = G(c,z)$

z →

Labels to Street Scene
input   output

Aerial to Map
input   output

Labels to Facade
input   output

Day to Night
input   output

BW to Color
input   output

Edges to Photo
input   output

https://arxiv.org/pdf/1611.07004
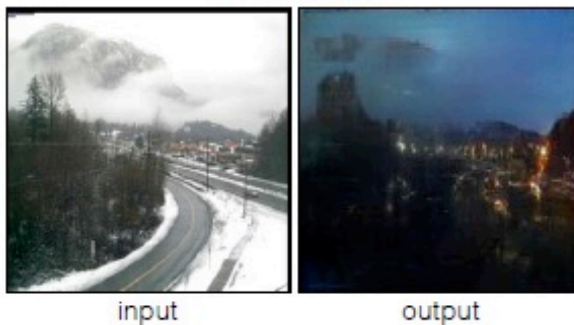
Image-to-image

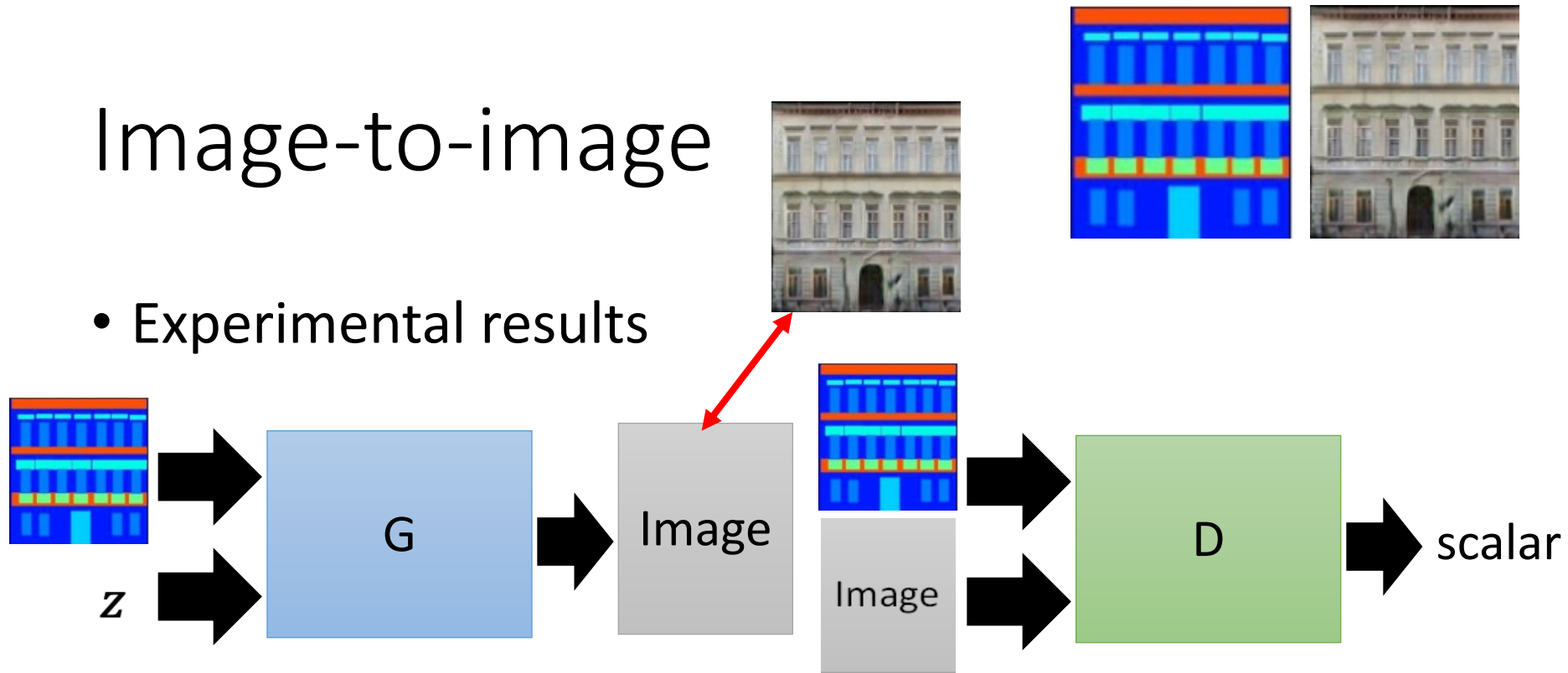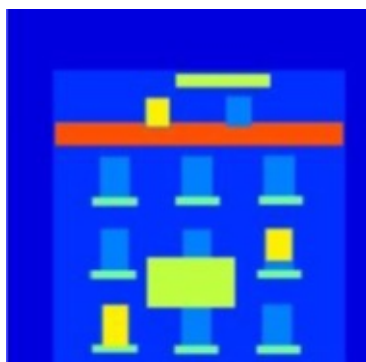- Traditional supervised approach

as close as possible
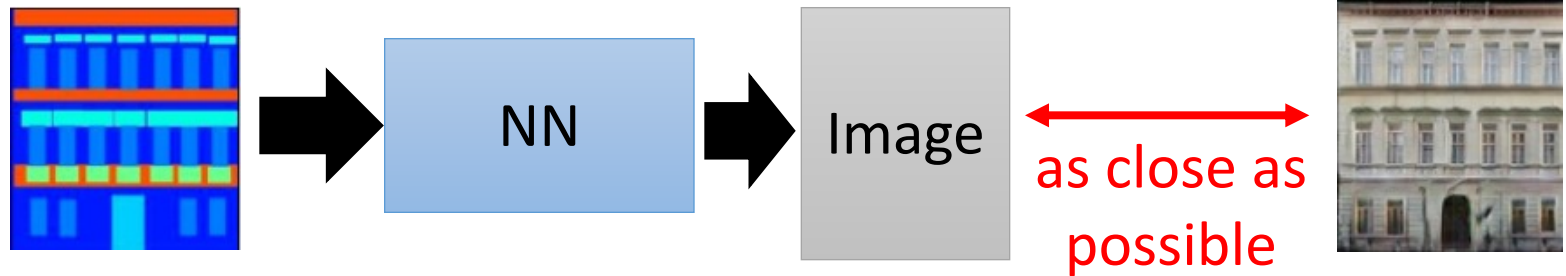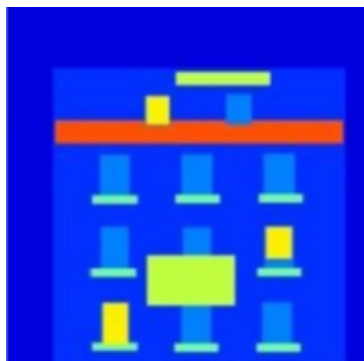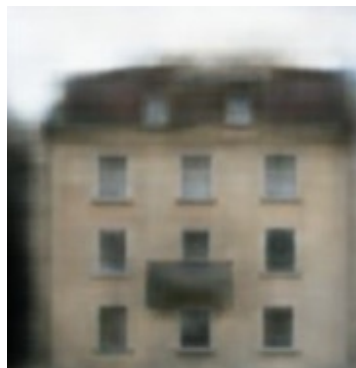
Testing:

input

close

It is blurry because it is the average of several images.
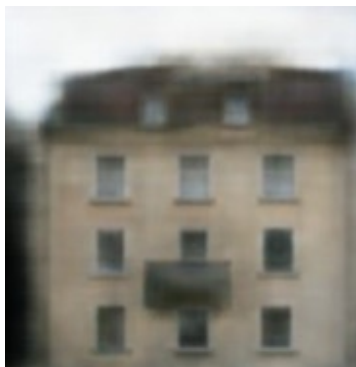
Image-to-image

- Experimental results

Testing:

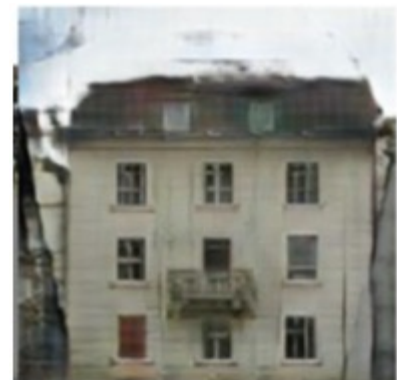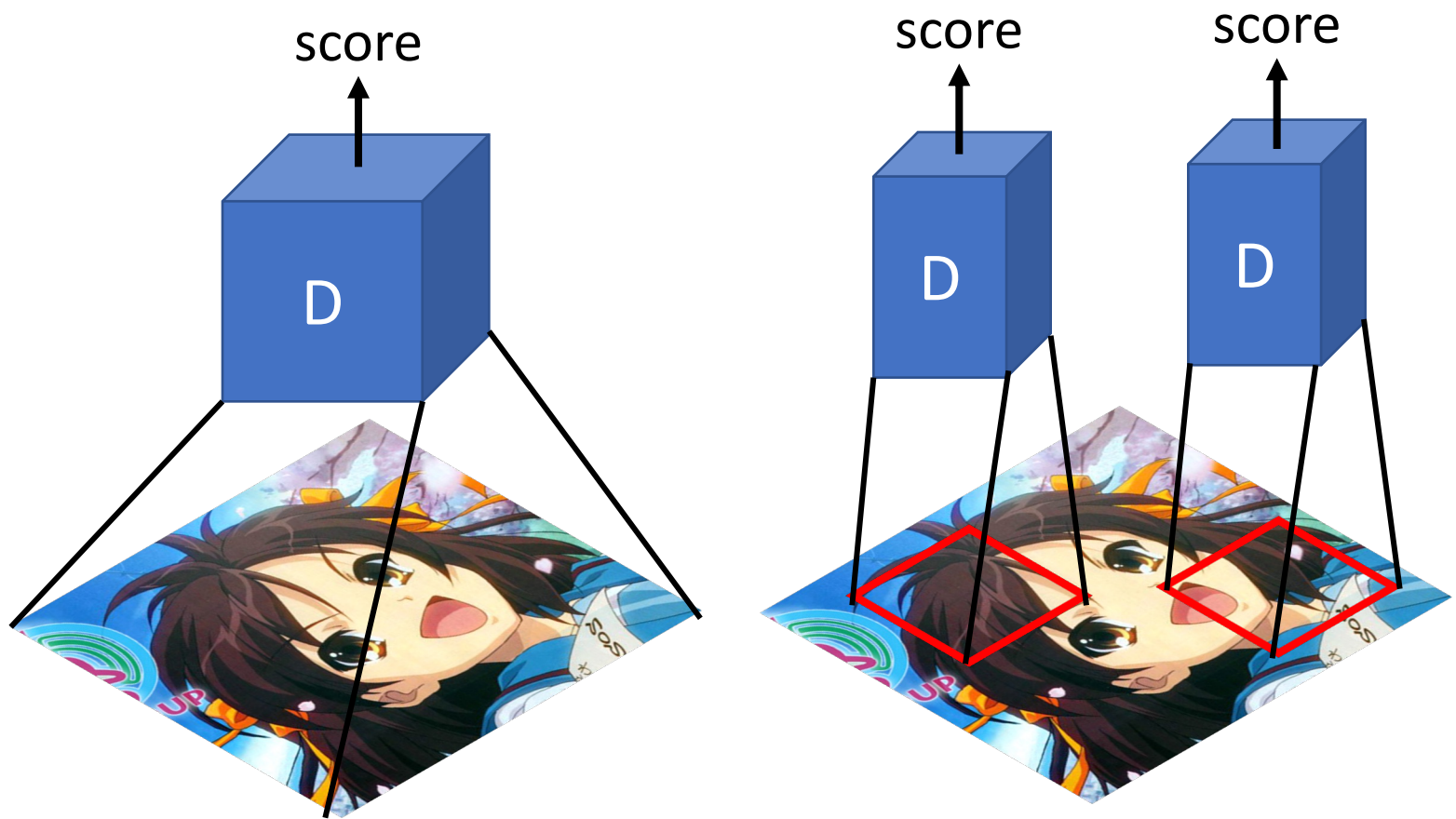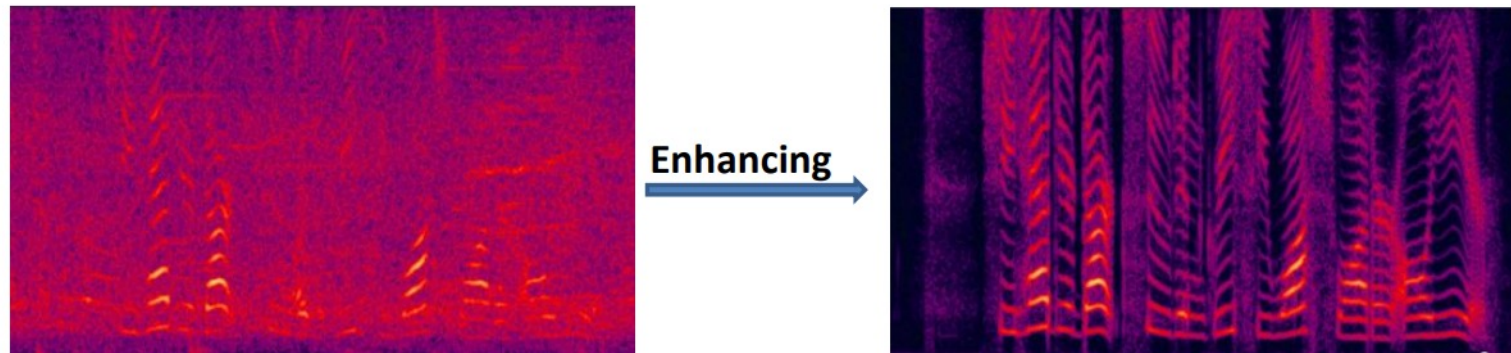input   close   GAN   GAN + close
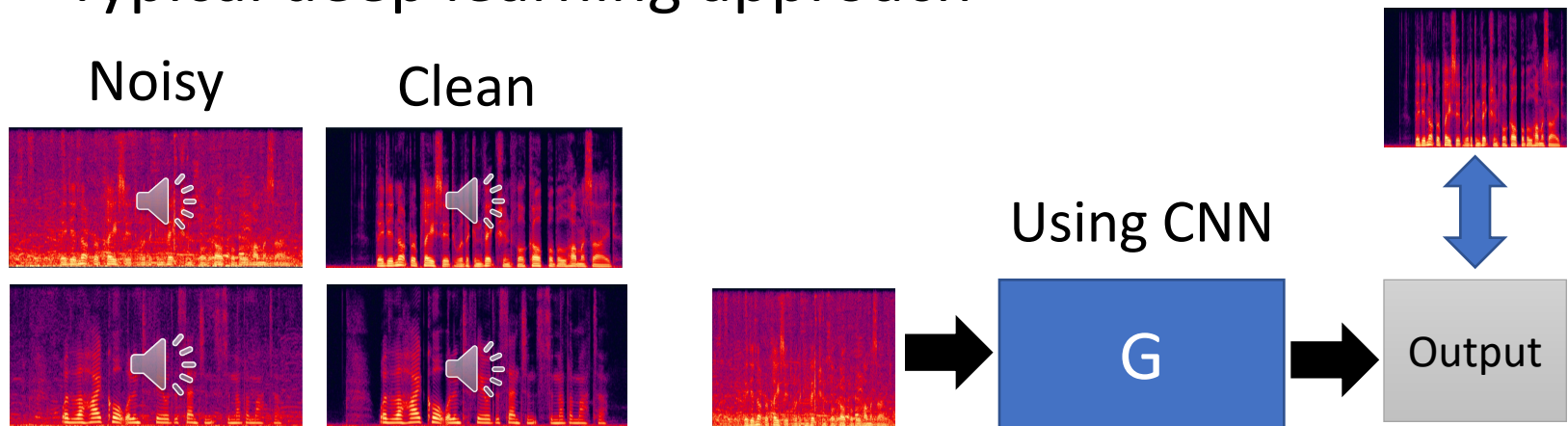
Patch GAN

https://arxiv.org/pdf/1611.07004.pdf

Speech Enhancement

Enhancing

- Typical deep learning approach

Noisy   Clean

Using CNN

G   Output

Speech Enhancement
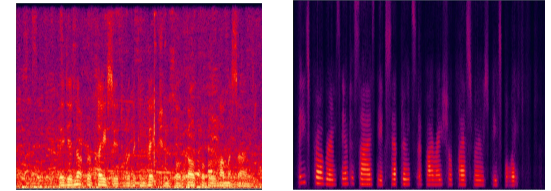
noisy   clean

- Conditional GAN

noisy → G → output ↔ clean

output → D → scalar (fake pair or not)

noisy → D

Video Generation

Deep Fake

References

- http://slazebni.cs.illinois.edu/spring17/
- https://cs.uwaterloo.ca/~mli/Deep-Learning-2017-Lecture7GAN.ppt