Generative Adversarial Network

李宏毅
Hung-yi Lee

Three Categories of GAN

1. *Typical GAN*

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

random vector → Generator → 

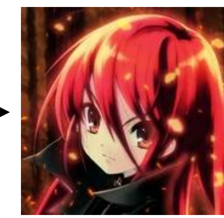
image

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

unpaired data

x

Photo

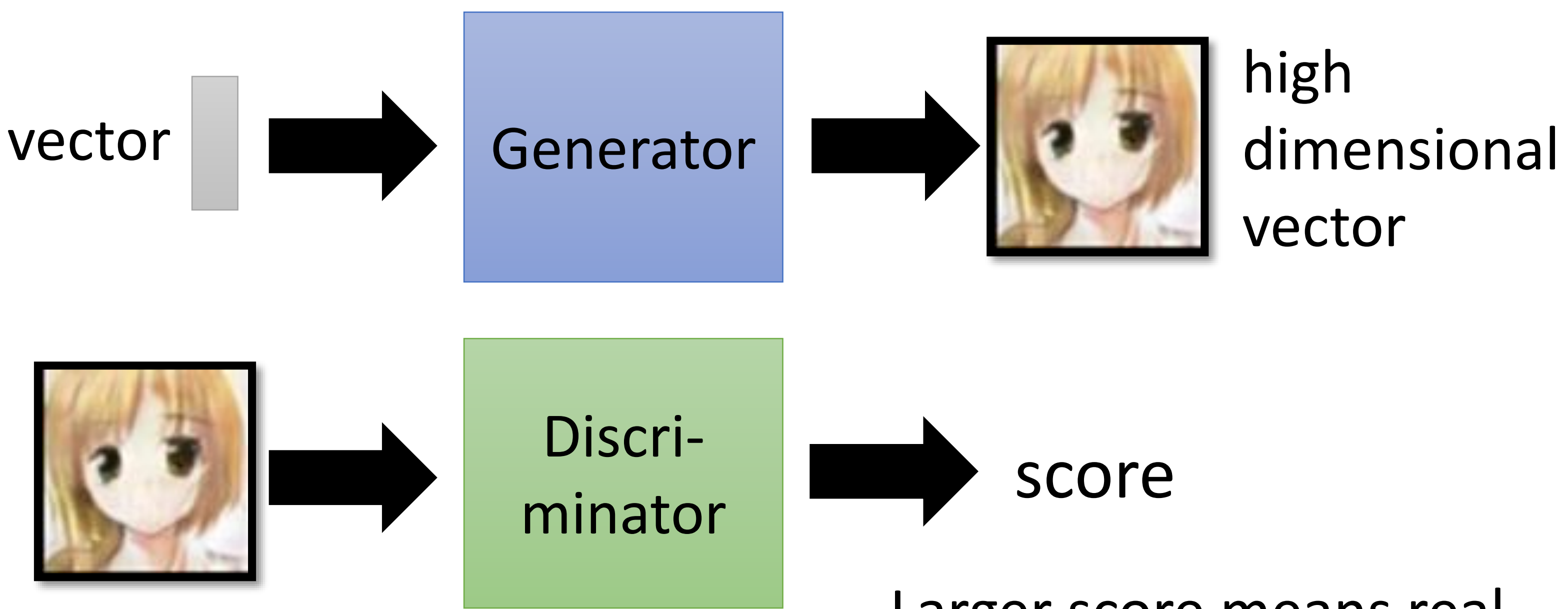
→ Generator →

y

Vincent van Gogh's style

Generative Adversarial Network (GAN)

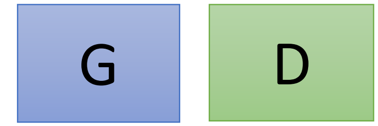
- Anime face generation as example



Larger score means real,
smaller score means fake.

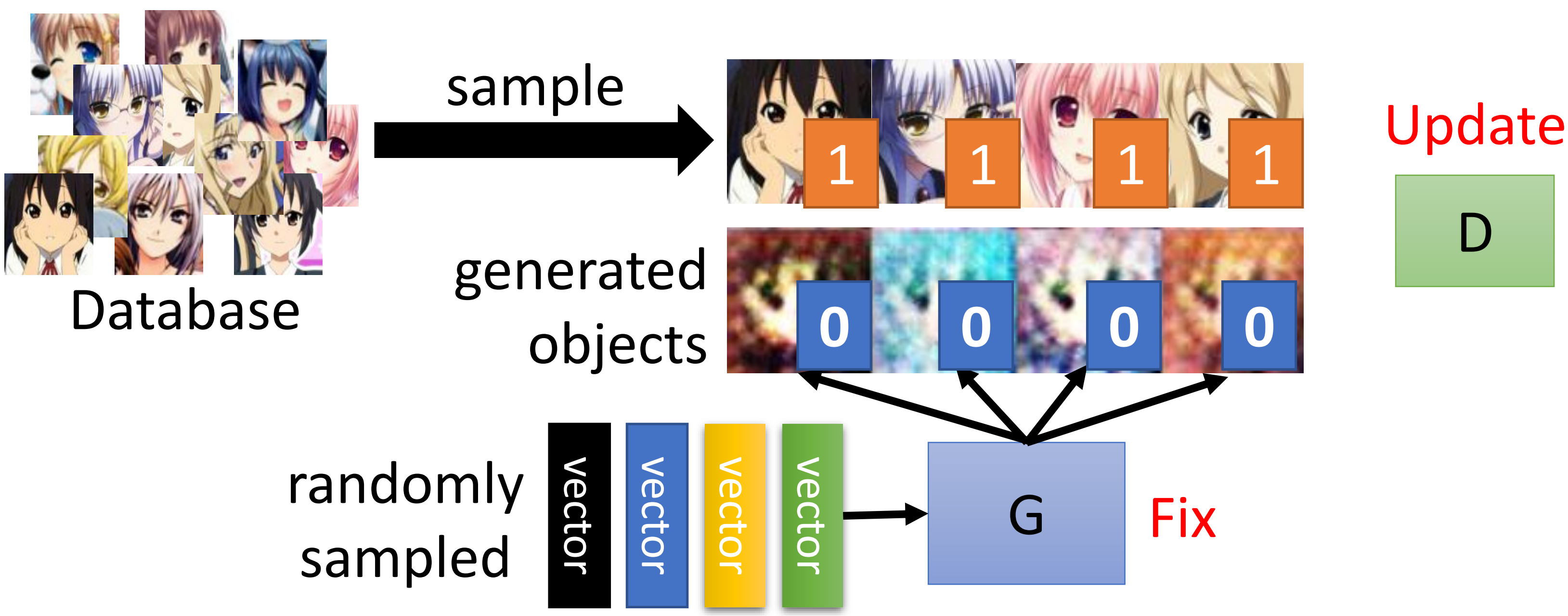
Algorithm

- Initialize generator and discriminator



- In each training iteration:

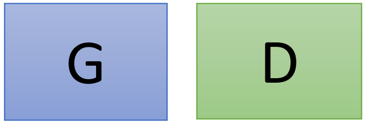
Step 1: Fix generator G, and update discriminator D



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

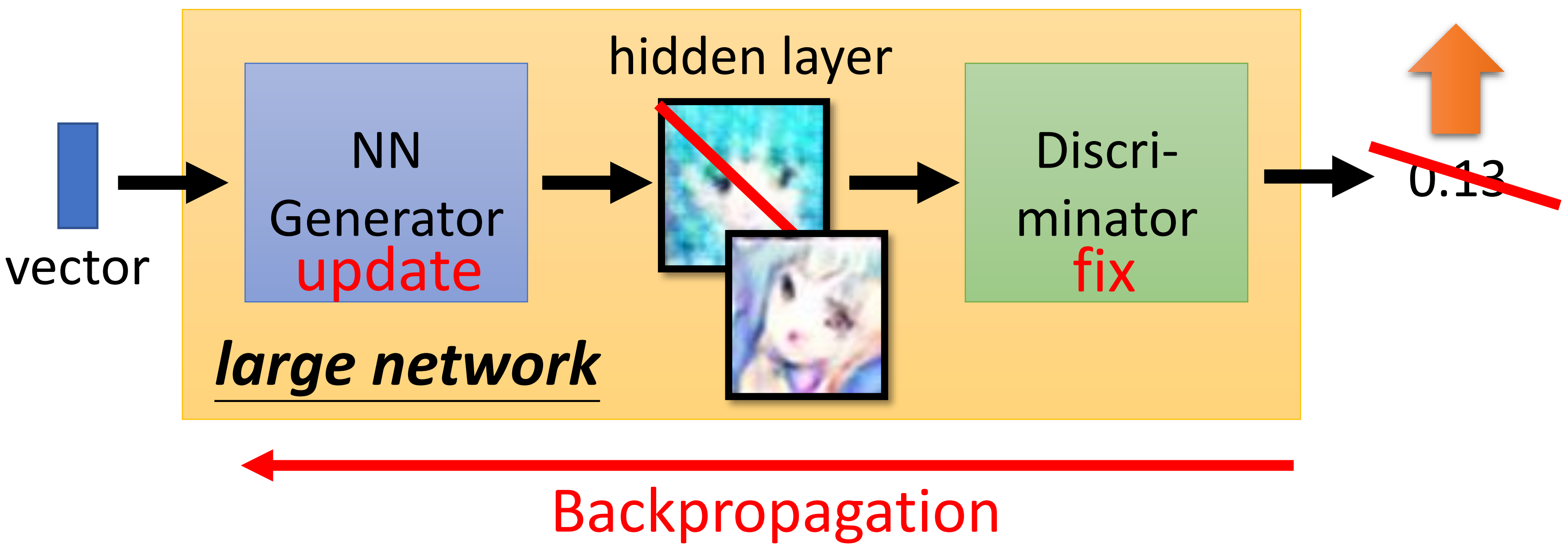
Algorithm

- Initialize generator and discriminator
- In each training iteration:



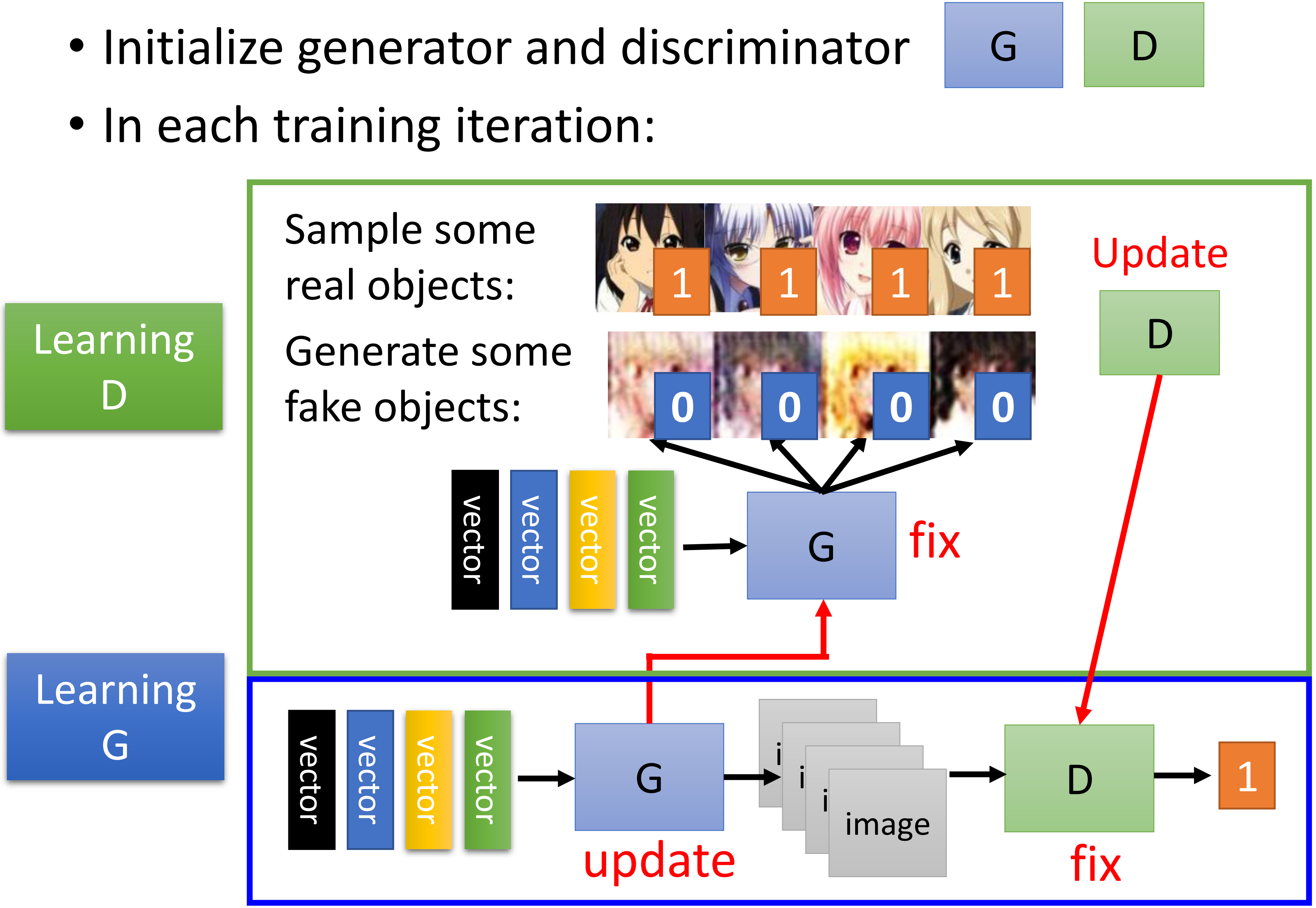
Step 2: Fix discriminator D, and update generator G

Generator learns to "*fool*" the discriminator



Algorithm

- Initialize generator and discriminator
- In each training iteration:



https://crypko.ai/#/

GAN is hard to train ……

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

Three Categories of GAN

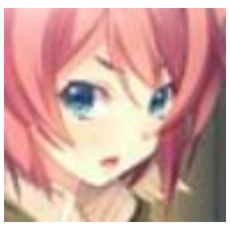
1. Typical GAN

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

random vector → Generator →

image

2. Conditional GAN



blue eyes,
red hair,
short hair

paired data

“Girl with red hair”

text

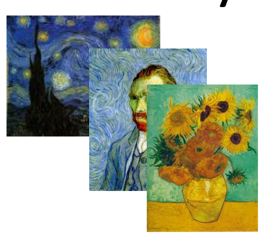
→ Generator →

image

3. Unsupervised Conditional GAN

domain x

domain y



unpaired data

x

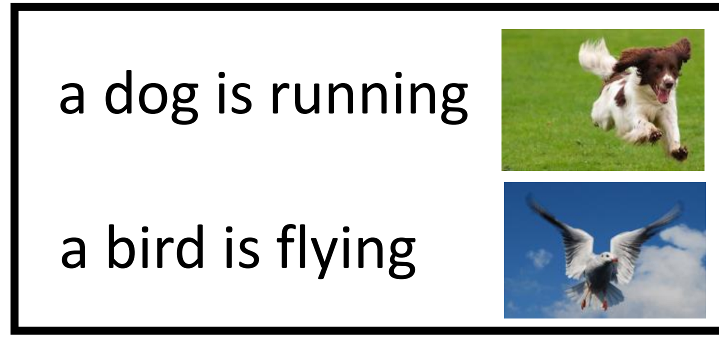
Photo

→ Generator →

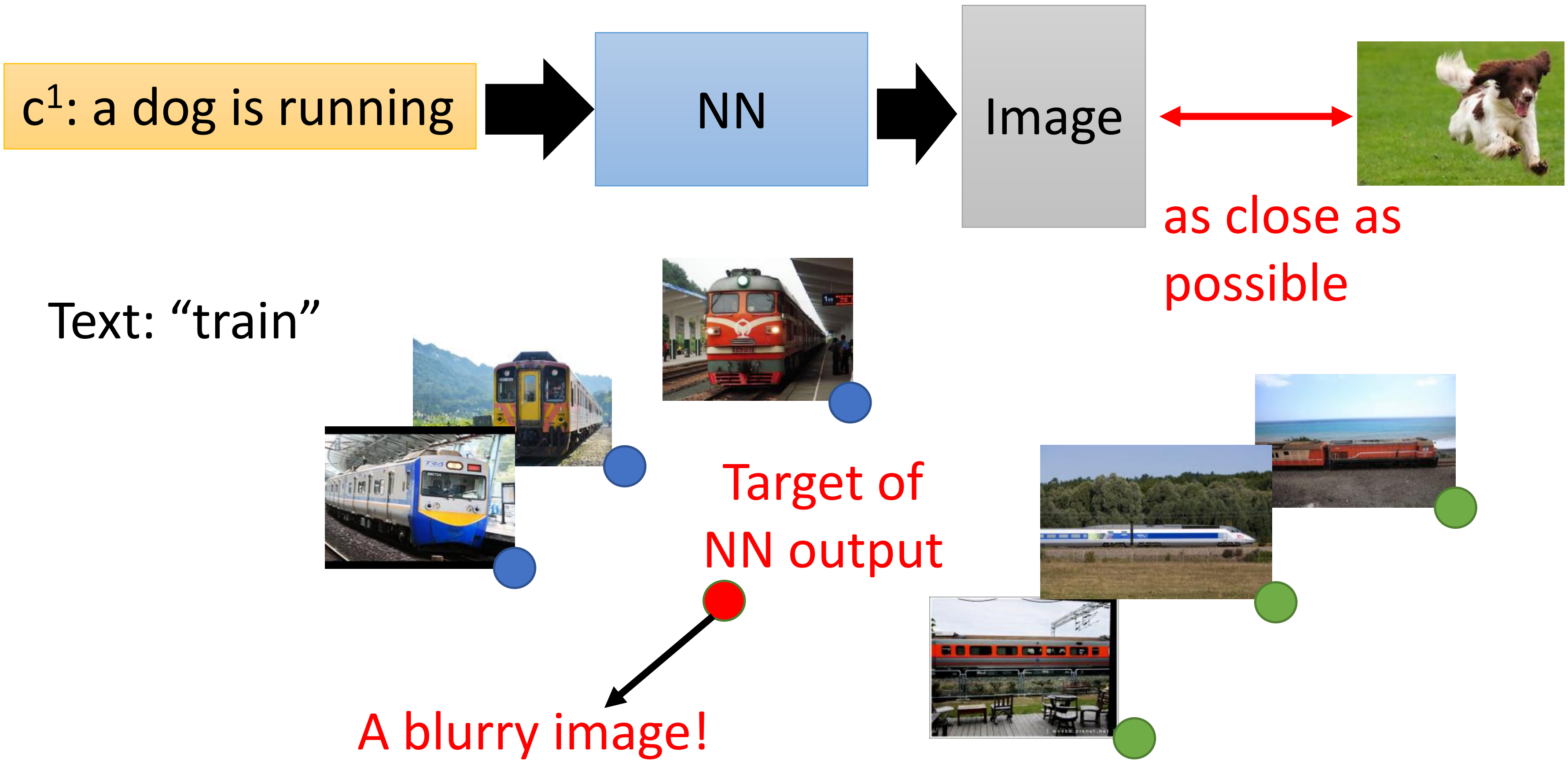
y

Vincent van Gogh’s style

Text-to-Image

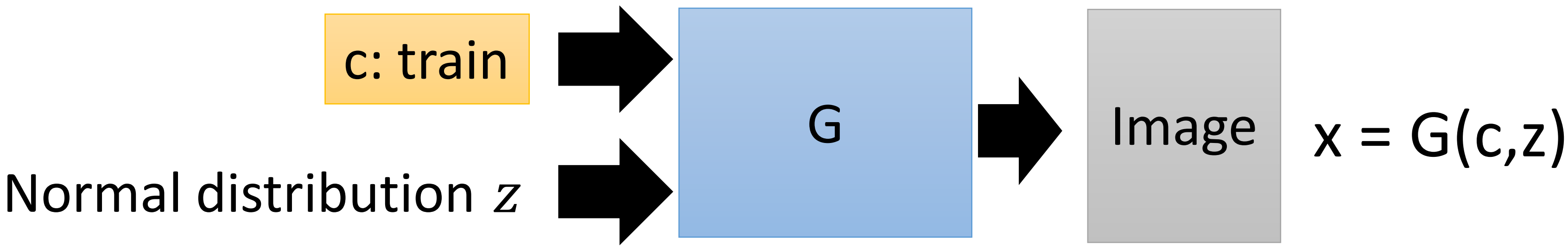


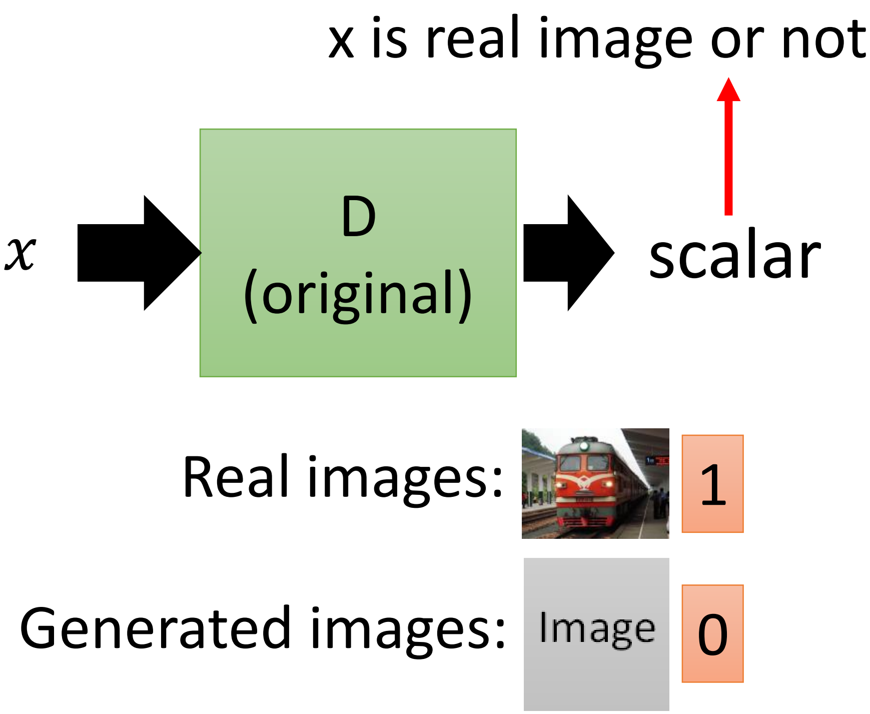
- **Traditional supervised approach**



[Scott Reed, et al, ICML, 2016]

Conditional GAN





Generator will learn to generate realistic images ....

But completely ignore the input conditions.

[Scott Reed, et al, ICML, 2016]

Conditional GAN

c: train

Normal distribution z

G

Image

x = G(c,z)

c

x

D
(better)

scalar

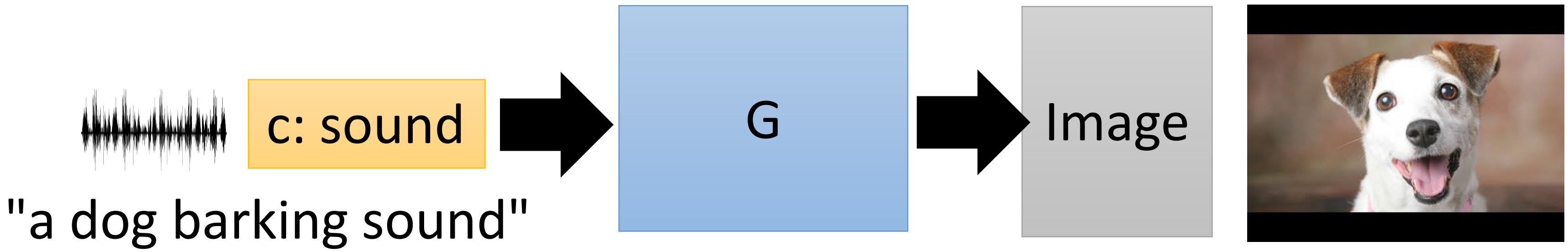
x is realistic or not +
c and x are matched or not

True text-image pairs: (train , ) 1

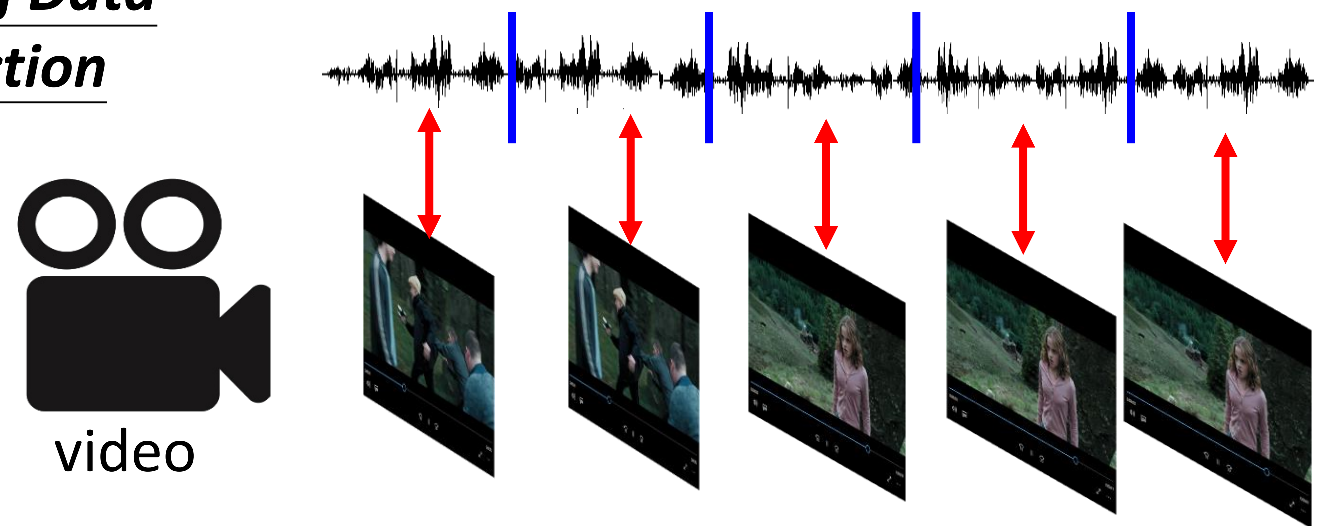
(cat , ) 0

(train , Image) 0

Conditional GAN
- Sound-to-image



Training Data Collection

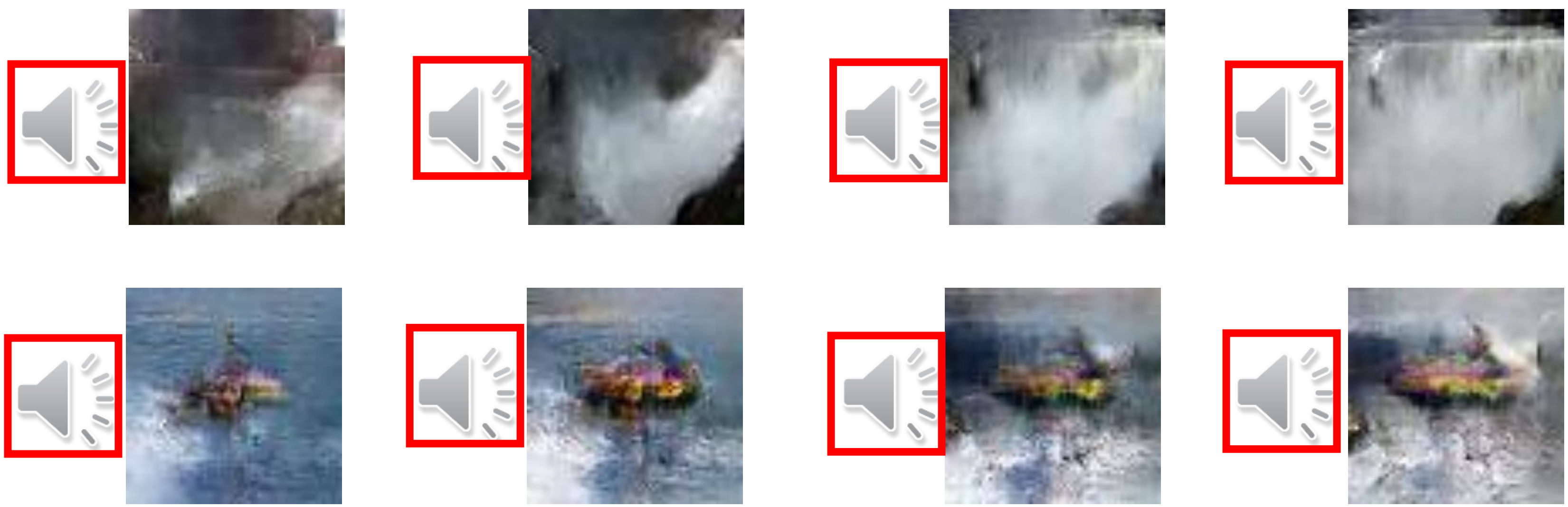


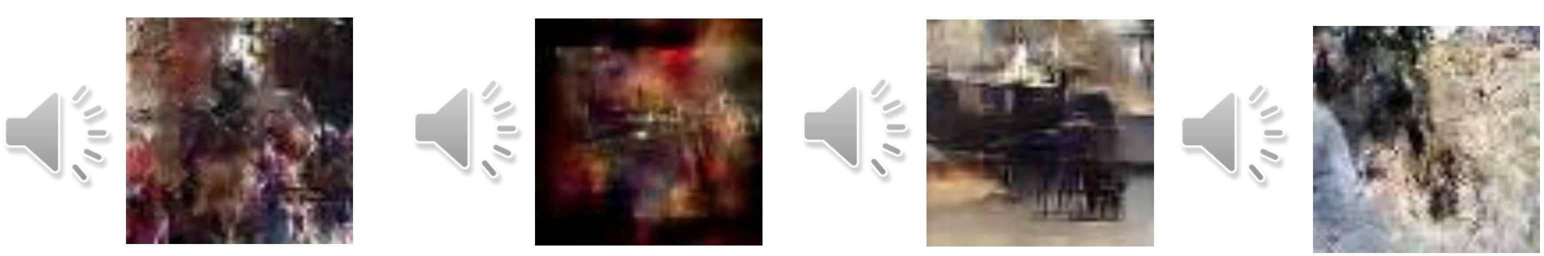
Conditional GAN - Sound-to-image

The images are generated by Chia-Hung Wan and Shun-Po Chuang.
https://wjohn1483.github.io/audio_to_scene/index.html

- Audio-to-image

Louder





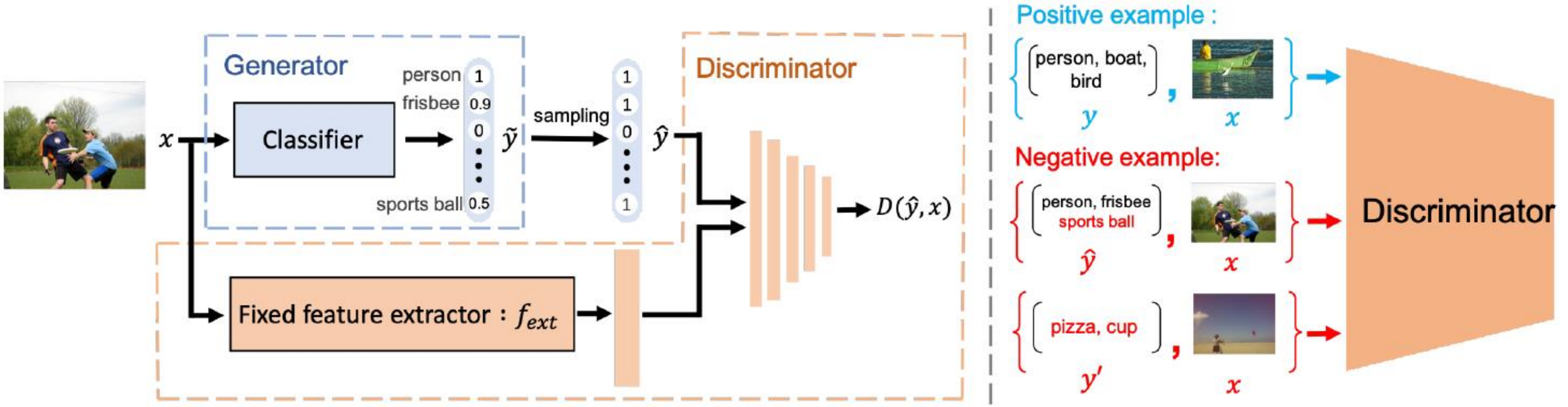
Conditional GAN - Image-to-label

Multi-label Image Classifier

person, sports ball, baseball bat, baseball glove

Input condition

Generated output



Conditional GAN - Image-to-label

The classifiers can have different architectures.

The classifiers are trained as conditional GAN.

F1	MS-COCO	NUS-WIDE
VGG-16	56.0	33.9
+ GAN	60.4	41.2
Inception	62.4	53.5
+GAN	63.8	55.8
Resnet-101	62.8	53.1
+GAN	64.0	55.4
Resnet-152	63.3	52.1
+GAN	63.9	54.1
Att-RNN	62.1	54.7
RLSD	62.0	46.9

[Tsai, et al., submitted to ICASSP 2019]

Conditional GAN - Image-to-label

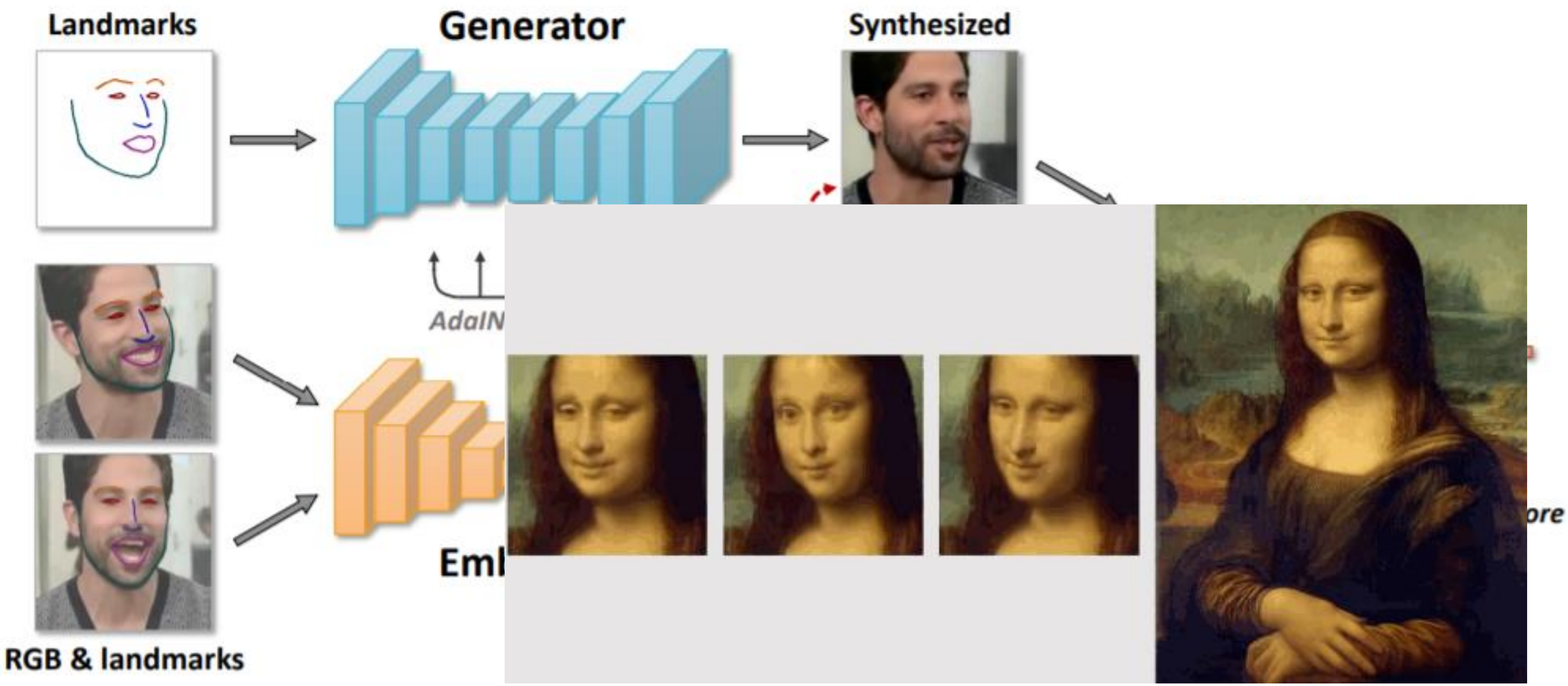
The classifiers can have different architectures.

The classifiers are trained as conditional GAN.

Conditional GAN outperforms other models designed for multi-label.

F1	MS-COCO	NUS-WIDE
VGG-16	56.0	33.9
+ GAN	60.4	41.2
Inception	62.4	53.5
+GAN	63.8	55.8
Resnet-101	62.8	53.1
+GAN	64.0	55.4
Resnet-152	63.3	52.1
+GAN	63.9	54.1
Att-RNN	62.1	54.7
RLSD	62.0	46.9

Talking Head



https://arxiv.org/abs/1905.08233

Three Categories of GAN

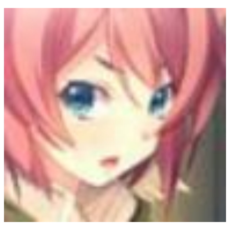
1. Typical GAN

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

random vector → Generator → 

image

2. Conditional GAN



blue eyes,
red hair,
short hair

paired data

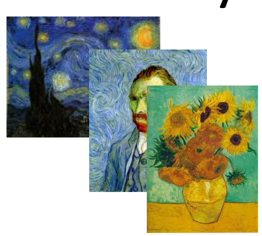
"Girl with red hair"
text → Generator → 

image

3. Unsupervised Conditional GAN

domain x

domain y



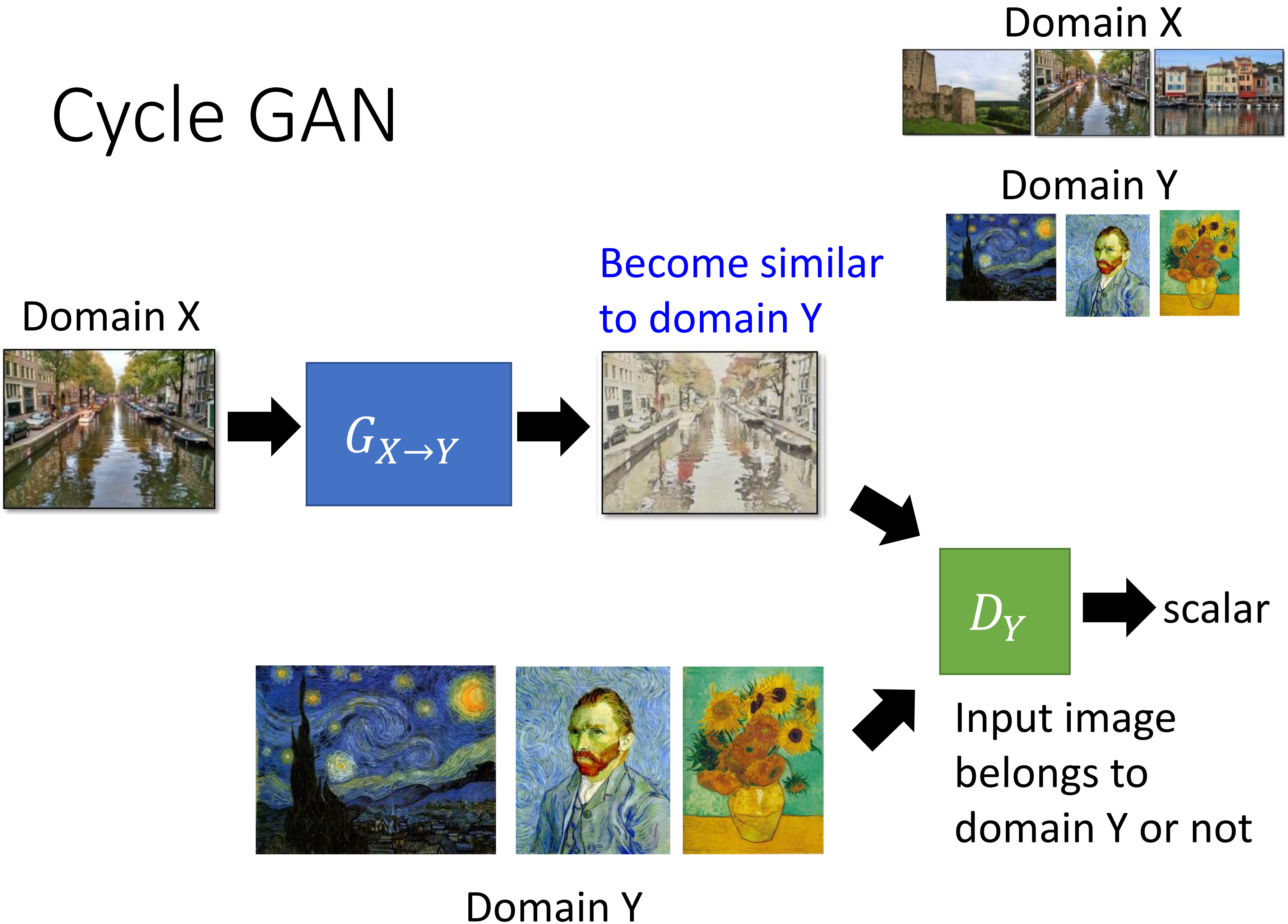
unpaired data

x

Photo → Generator → y

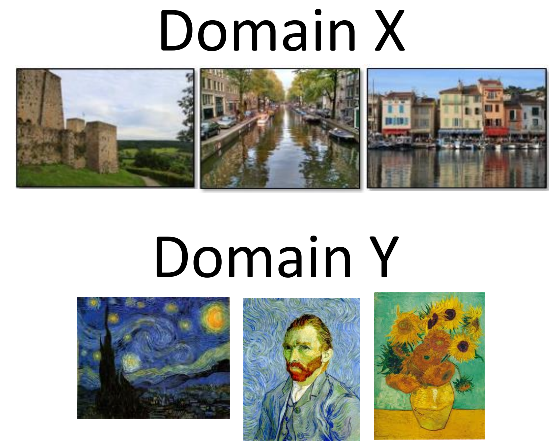
Vincent van Gogh's style

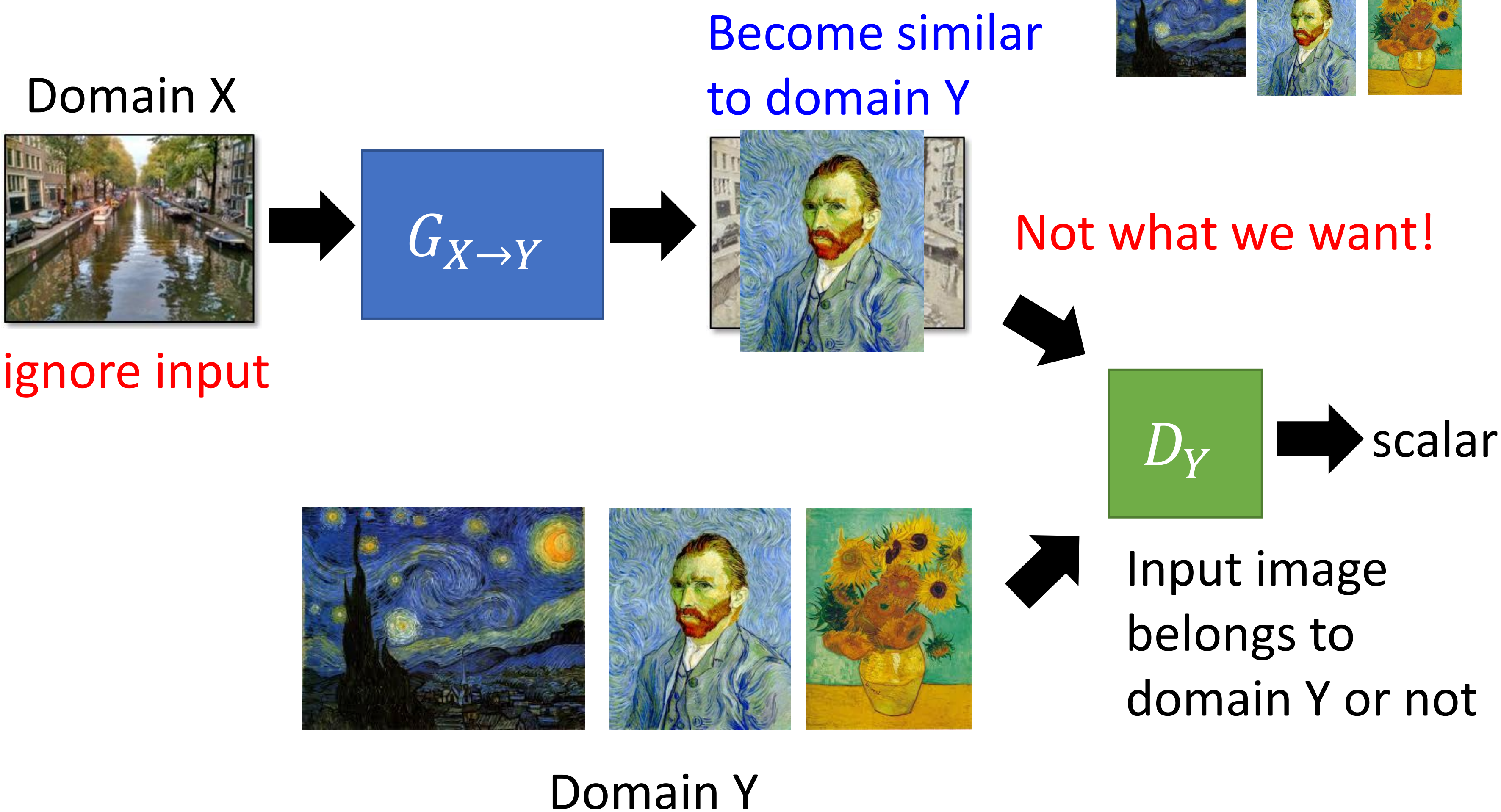
Cycle GAN



Domain X
Domain Y
Become similar to domain Y
Domain X
$G_{X \to Y}$
D_Y
scalar
Input image belongs to domain Y or not
Domain Y

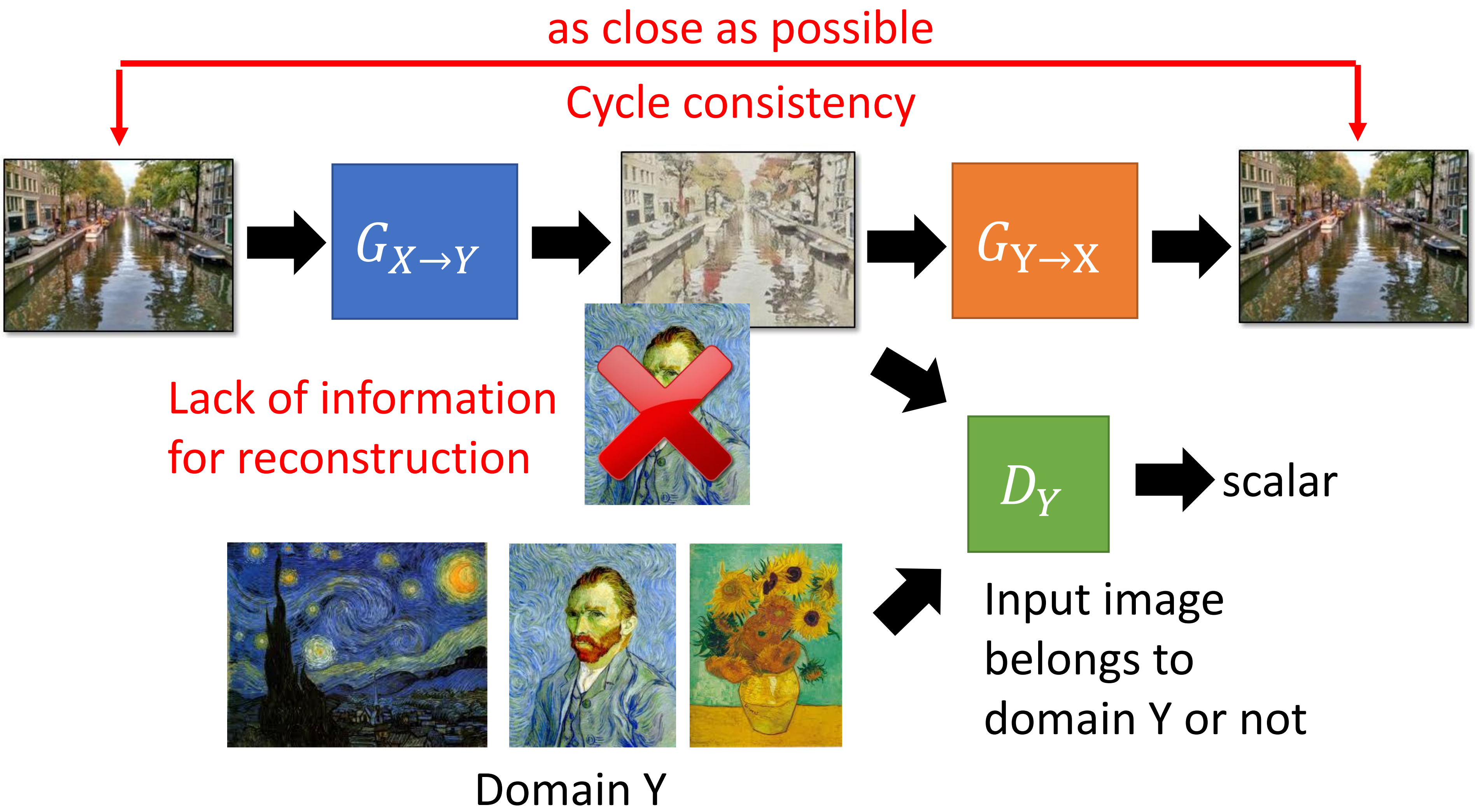
Cycle GAN


Domain X
Domain Y

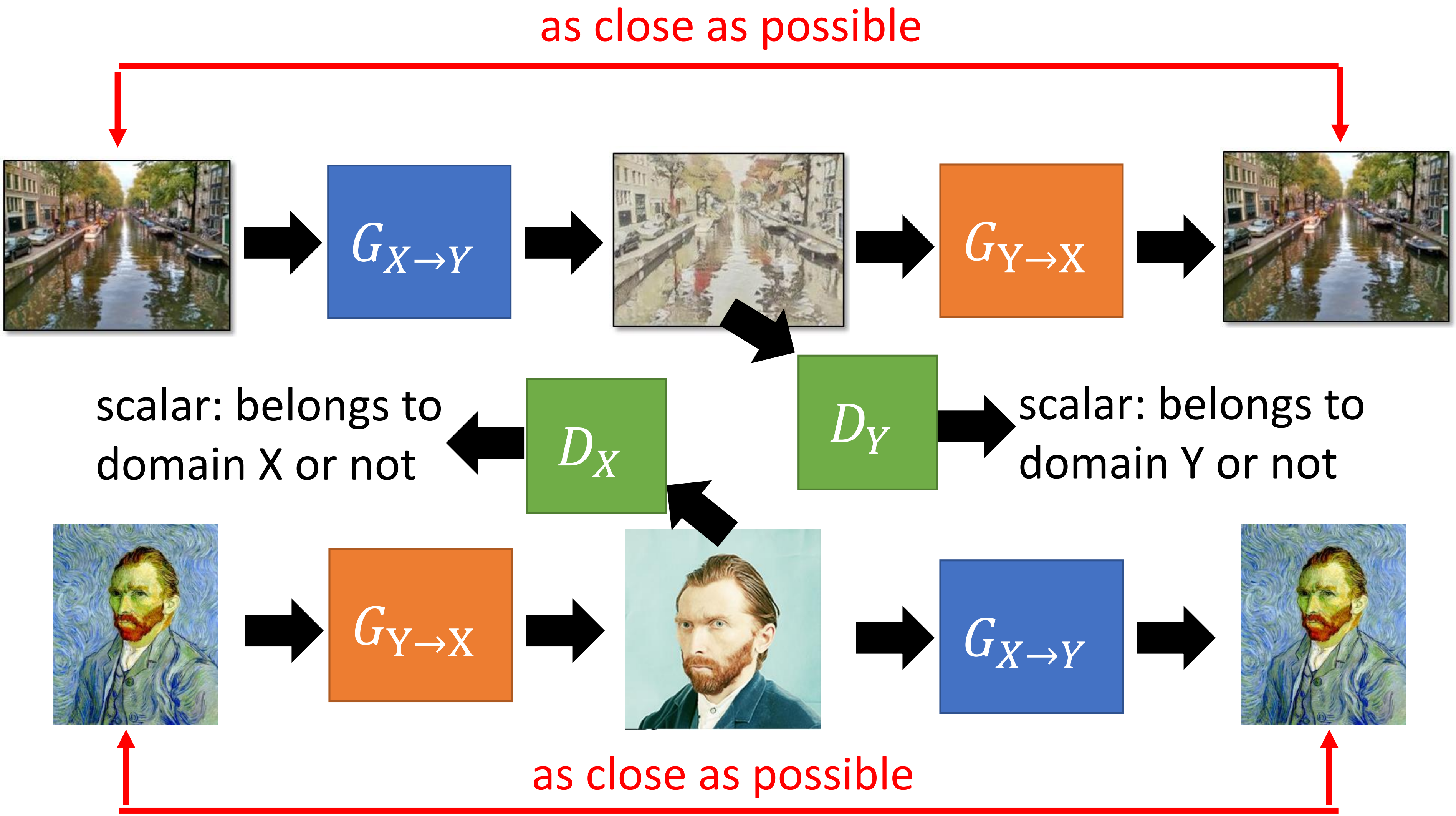

Domain X
$G_{X \to Y}$
Become similar
to domain Y
Not what we want!
ignore input
D_Y
scalar
Input image
belongs to
domain Y or not
Domain Y

[Jun-Yan Zhu, et al., ICCV, 2017]

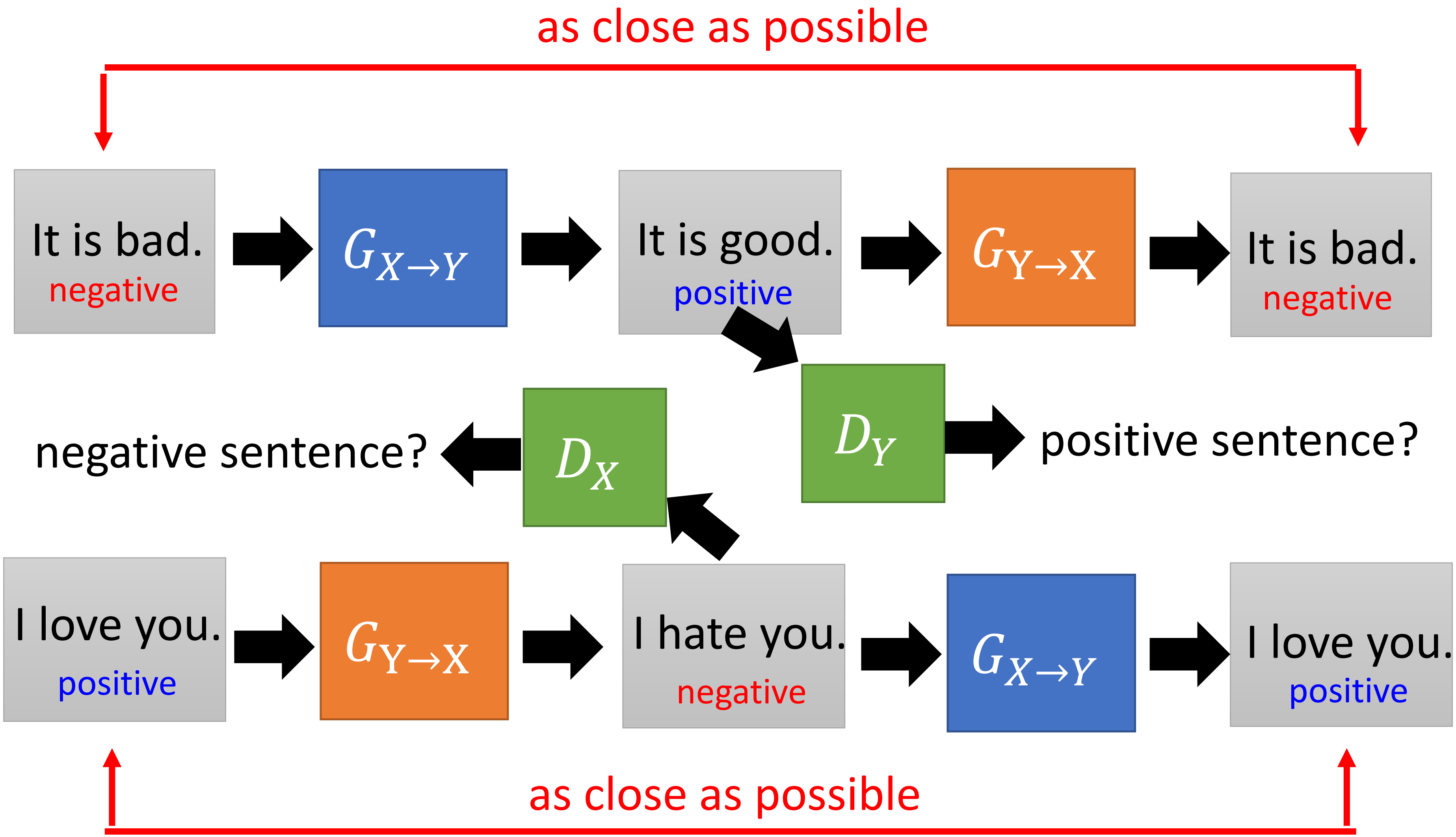
Cycle GAN



Cycle GAN

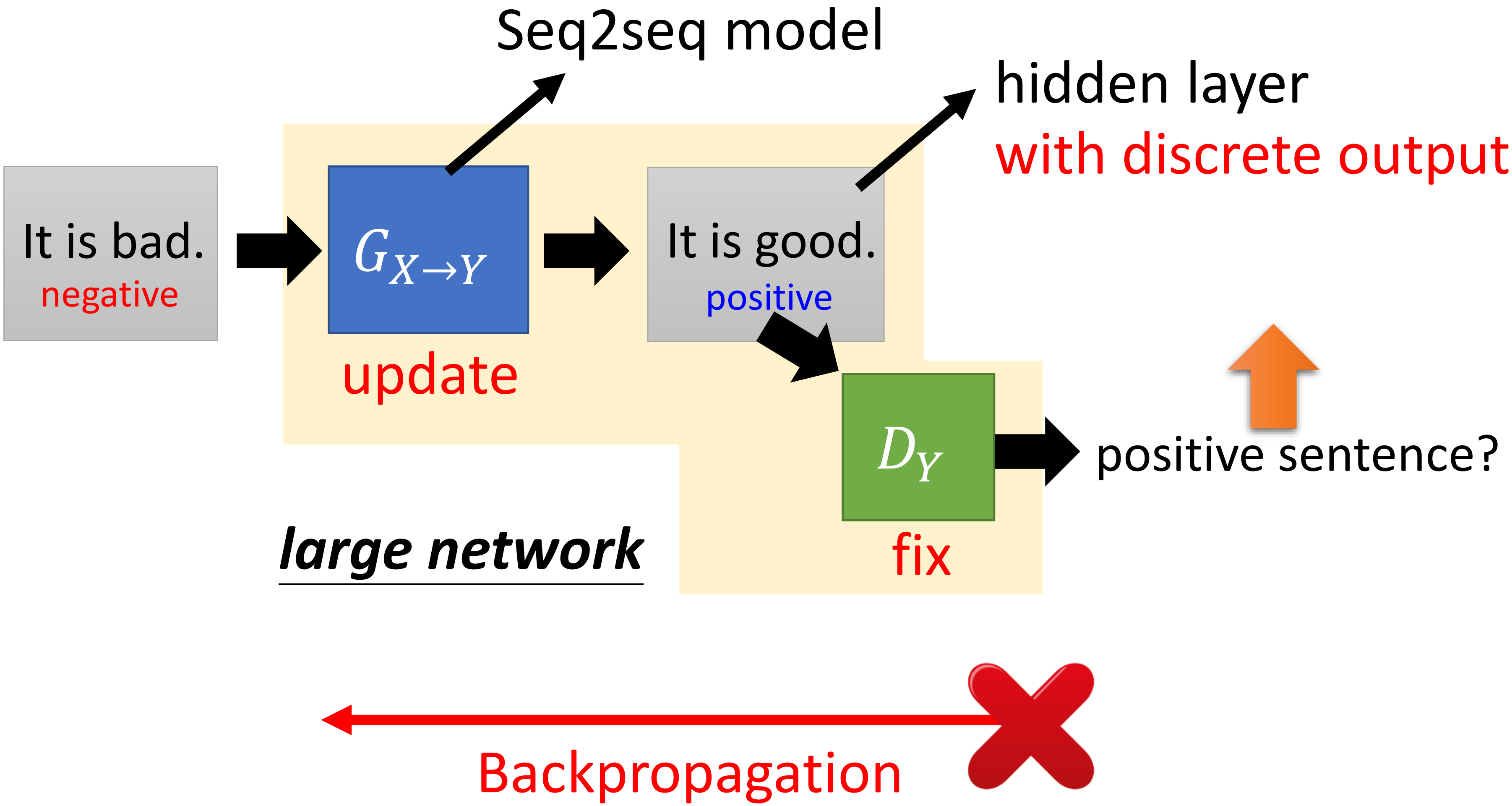

as close as possible
$G_{X\to Y}$
$G_{Y\to X}$
D_Y
scalar: belongs to domain Y or not
D_X
scalar: belongs to domain X or not
$G_{Y\to X}$
$G_{X\to Y}$
as close as possible

Cycle GAN



as close as possible
It is bad.
negative
$G_{X\to Y}$
It is good.
positive
$G_{Y\to X}$
It is bad.
negative
D_Y
positive sentence?
D_X
negative sentence?
I love you.
positive
$G_{Y\to X}$
I hate you.
negative
$G_{X\to Y}$
I love you.
positive
as close as possible

Discrete Issue



Three Categories of Solutions

Gumbel-softmax

- [Matt J. Kusner, et al, arXiv, 2016]

Continuous Input for Discriminator

- [Sai Rajeswar, et al., arXiv, 2017][Ofir Press, et al., ICML workshop, 2017][Zhen Xu, et al., EMNLP, 2017][Alex Lamb, et al., NIPS, 2016][Yizhe Zhang, et al., ICML, 2017]

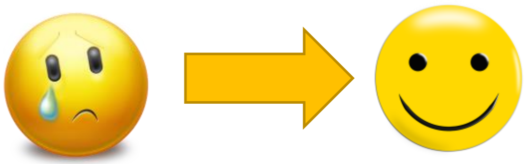
“Reinforcement Learning”

- [Yu, et al., AAAI, 2017][Li, et al., EMNLP, 2017][Tong Che, et al, arXiv, 2017][Jiaxian Guo, et al., AAAI, 2018][Kevin Lin, et al, NIPS, 2017][William Fedus, et al., ICLR, 2018]

文句改寫

感謝 王耀賢 同學提供實驗結果

Negative sentence to positive sentence:
it's a crappy day -> it's a great day
i wish you could be here -> you could be here
it's not a good idea -> it's good idea
i miss you -> i love you
i don't love you -> i love you
i can't do that -> i can do that
i feel so sad -> i happy
it's a bad day -> it's a good day
it's a dummy day -> it's a great day
sorry for doing such a horrible thing -> thanks for doing a great thing
my doggy is sick -> my doggy is my doggy
my little doggy is sick -> my little doggy is my little doggy



文句改寫

感謝 張瓊之 同學提供實驗結果

Negative sentence to positive sentence:

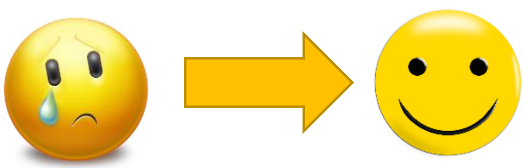
胃疼,沒睡醒,各種不舒服 -> 生日快樂,睡醒,超級舒服

我都想去上班了,真夠賤的! -> 我都想去睡了,真帥的！

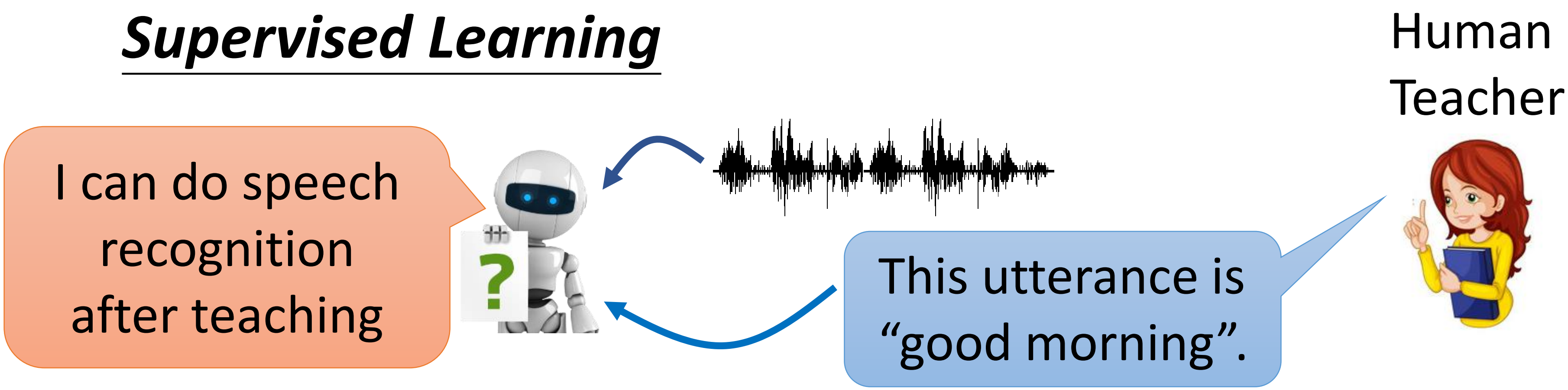
暈死了,吃燒烤、竟然遇到個變態狂 -> 哈哈好~,吃燒烤~竟然遇到帥狂

我肚子痛的厲害 -> 我生日快樂厲害

感冒了,難受的說不出話來了！-> 感冒了,開心的說不出話來！



Speech Recognition



- Supervised learning needs lots of annotated speech.
- However, most of the languages are low resourced.

Speech Recognition

Supervised Learning

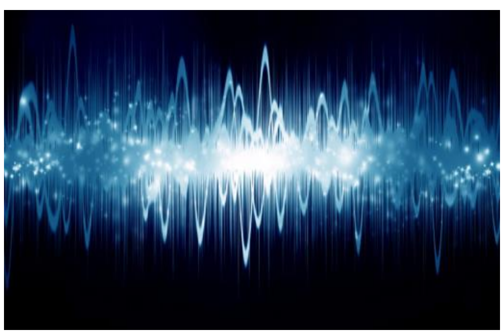
Human Teacher

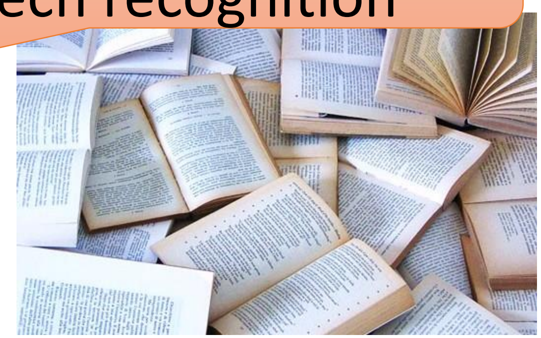
I can do speech recognition after teaching

This utterance is “good morning”.

Unsupervised Learning

I can automatically learn speech recognition

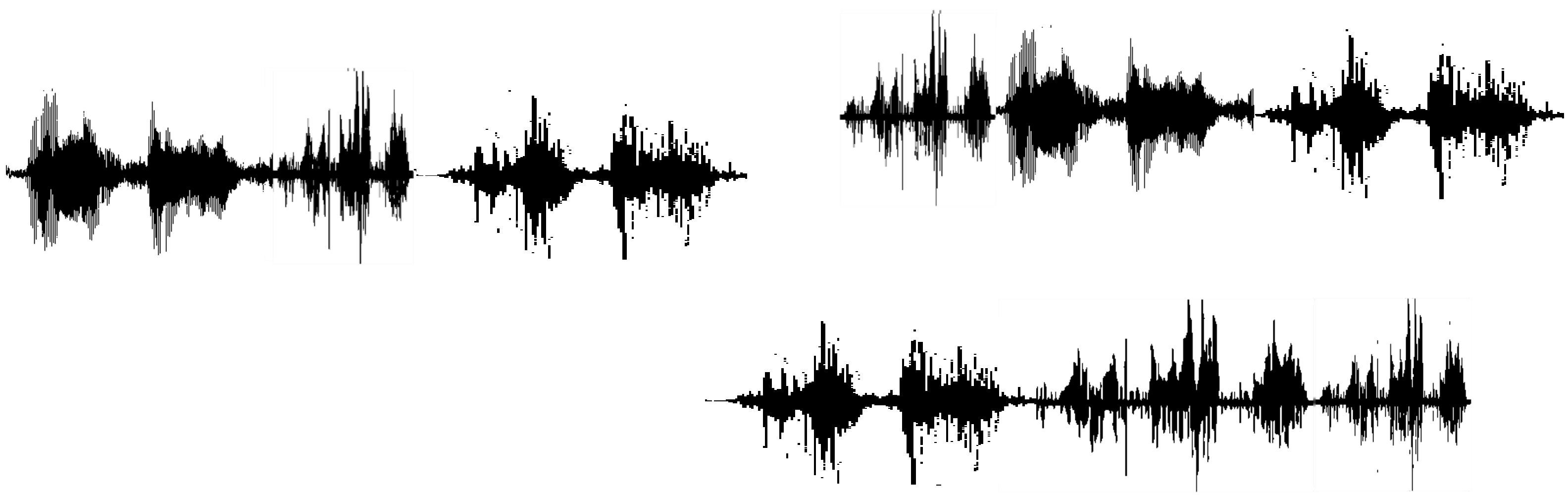




Listening to human talking

Reading text on the Internet

Acoustic Token Discovery

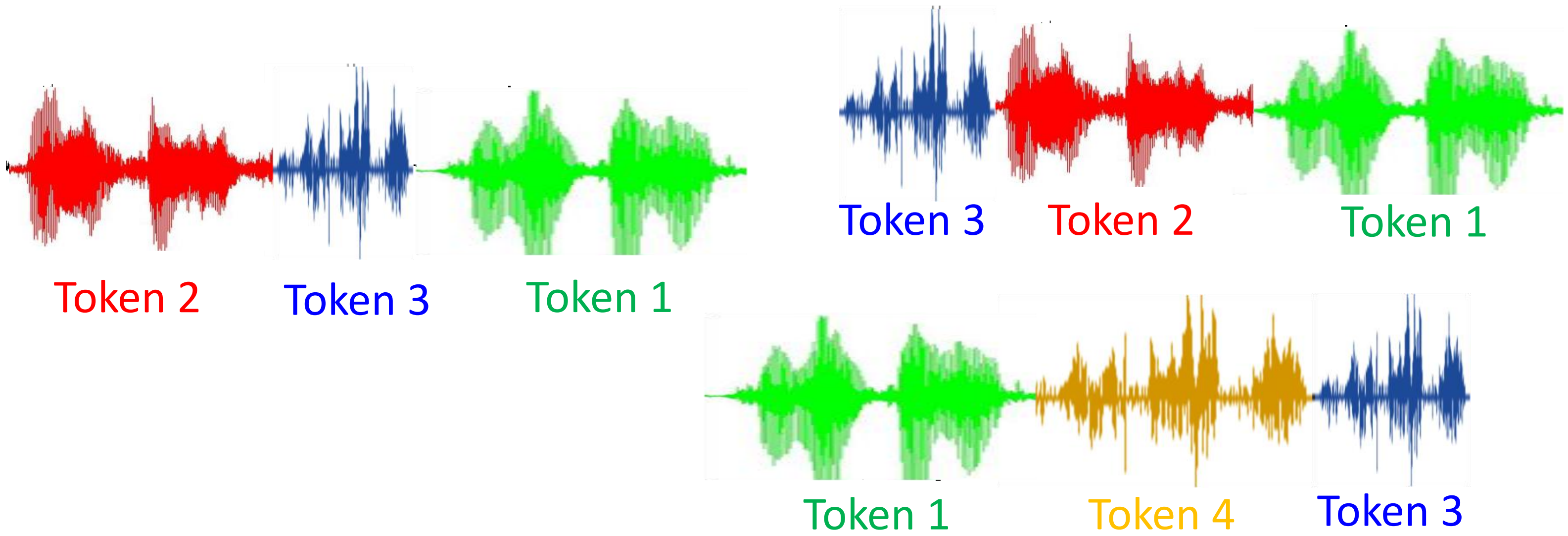


Acoustic tokens can be discovered from audio collection without text annotation.

Acoustic tokens: chunks of acoustically similar audio segments with token IDs

[Zhang & Glass, ASRU 09]
[Huijbregts, ICASSP 11]
[Chan & Lee, Interspeech 11]

Acoustic Token Discovery



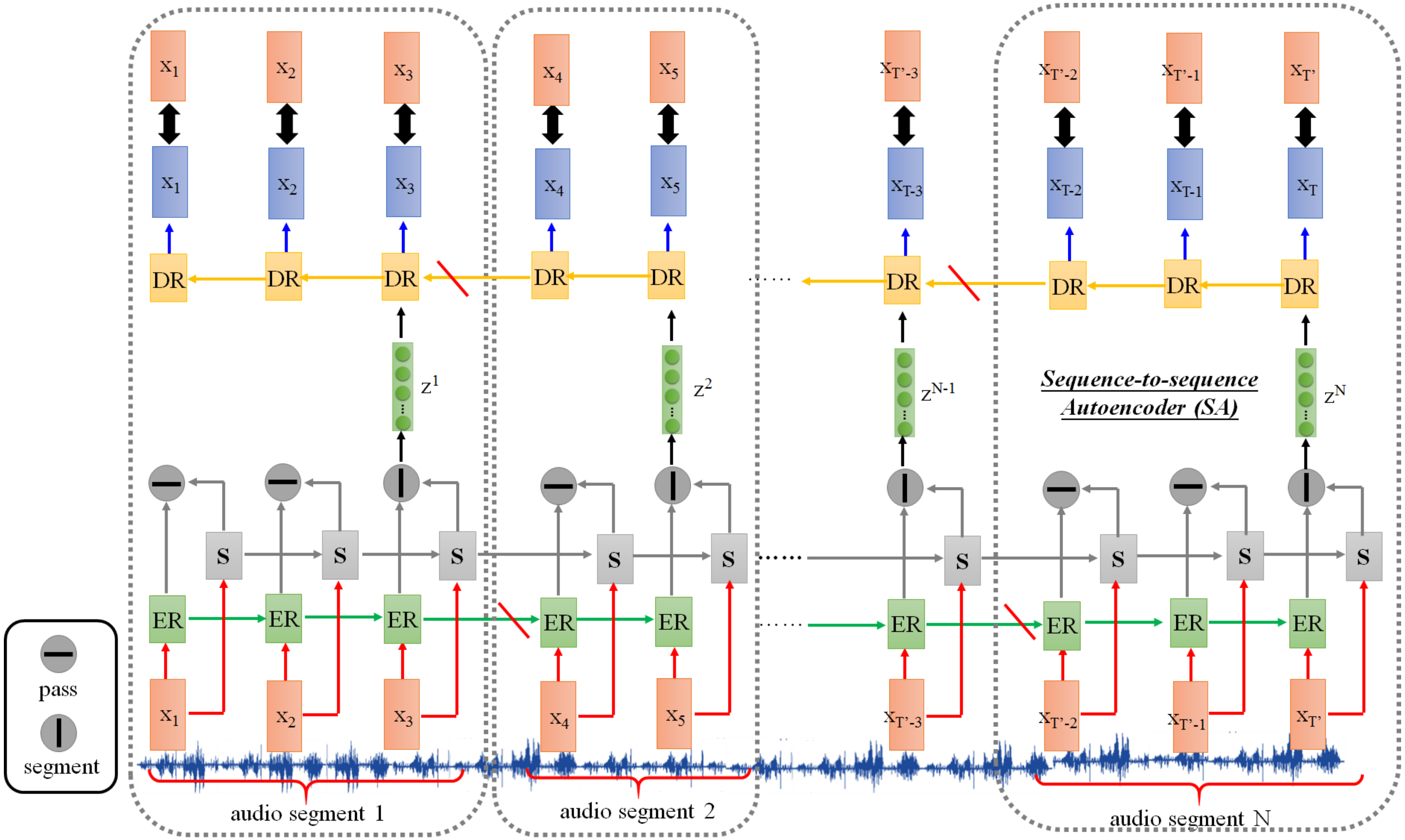
Acoustic tokens can be discovered from audio collection without text annotation.

Acoustic tokens: chunks of acoustically similar audio segments with token IDs

[Zhang & Glass, ASRU 09]
[Huijbregts, ICASSP 11]
[Chan & Lee, Interspeech 11]

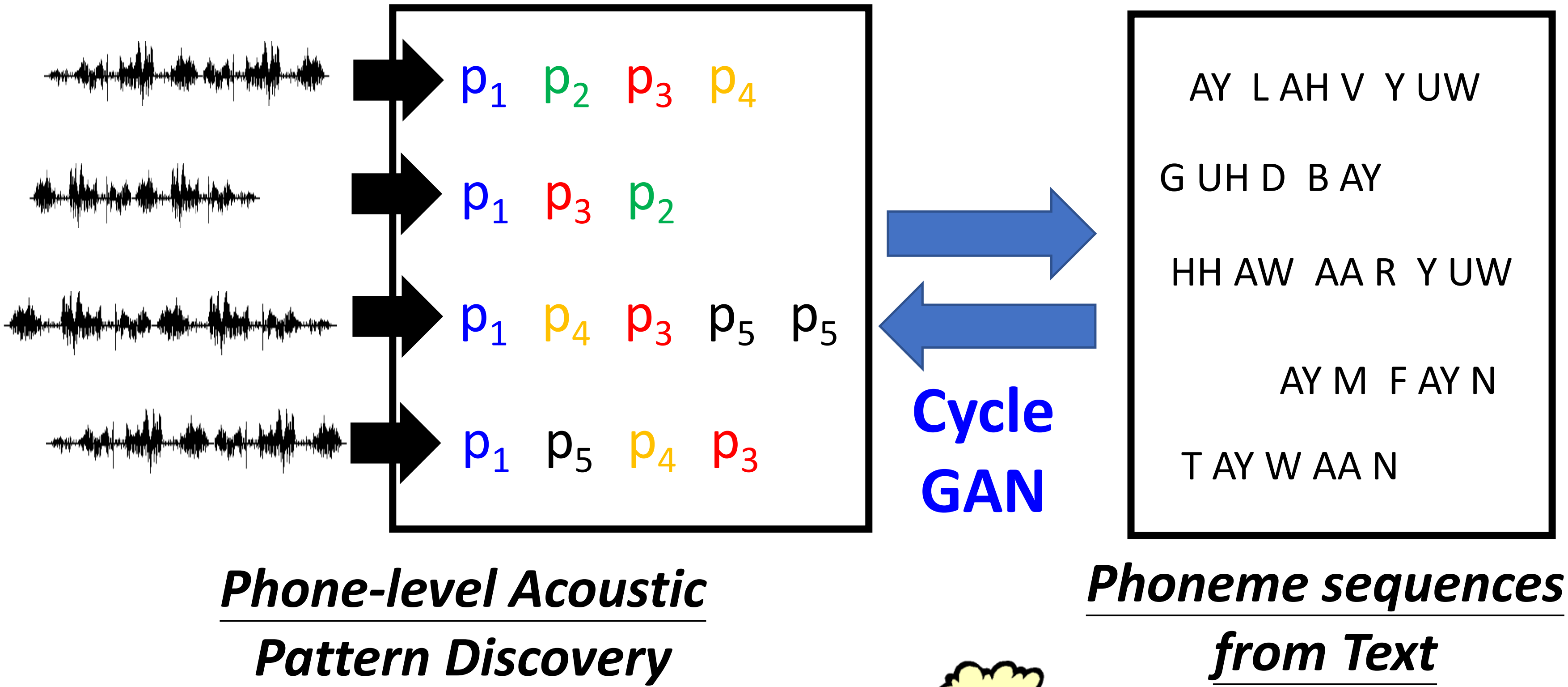
[Wang, et al., ICASSP, 2018]

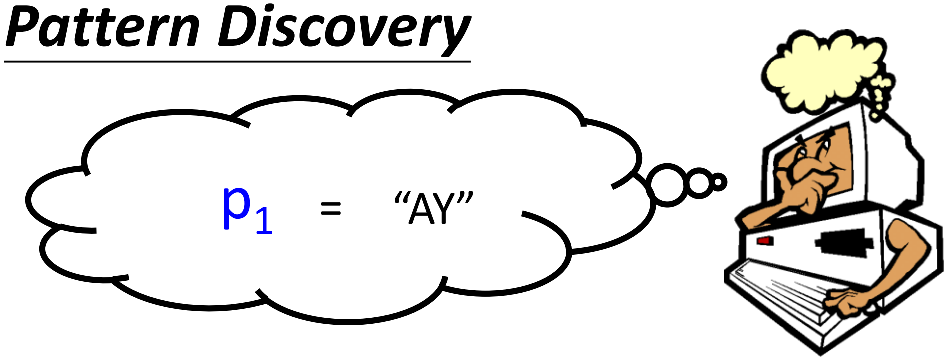
Acoustic Token Discovery



Phonetic-level acoustic tokens are obtained by segmental sequence-to-sequence autoencoder.

Unsupervised Speech Recognition

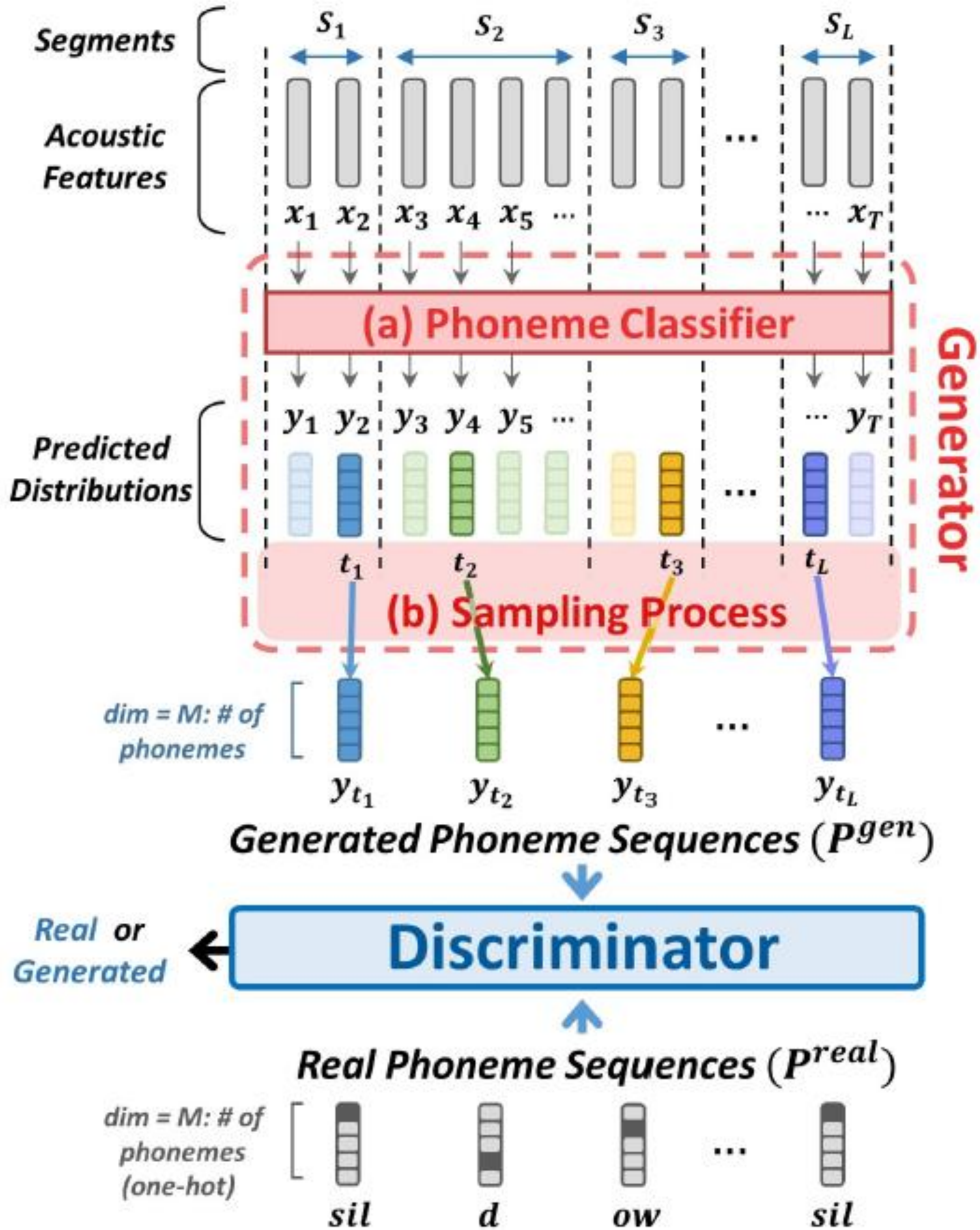




[Liu, et al., INTERSPEECH, 2018]

[Chen, et al., arXiv, 2018]

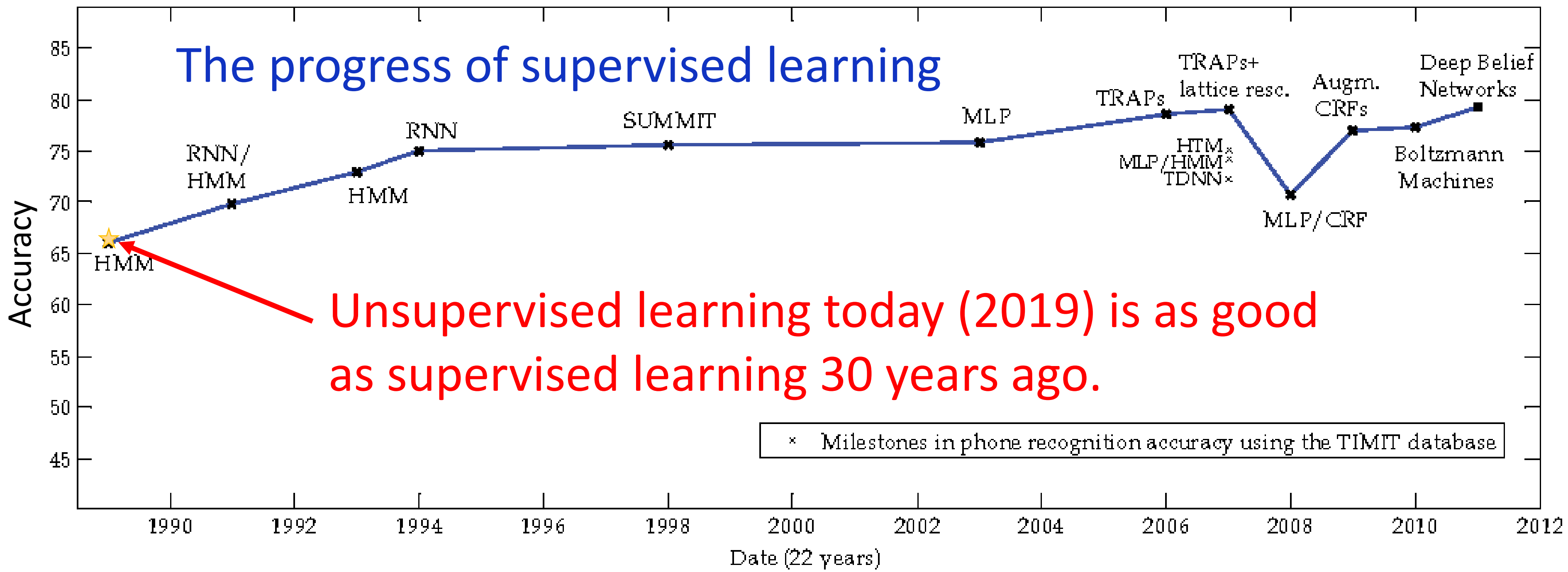
Model



Segments
S_1 S_2 S_3 S_L
Acoustic Features
x_1 x_2 x_3 x_4 x_5 ... ... x_T
(a) Phoneme Classifier
Generator
y_1 y_2 y_3 y_4 y_5 ... ... y_T
Predicted Distributions
t_1 t_2 t_3 t_L
(b) Sampling Process
dim = M: # of phonemes
y_{t_1} y_{t_2} y_{t_3} y_{t_L}
Generated Phoneme Sequences (P^{gen})
Real or Generated
Discriminator
Real Phoneme Sequences (P^{real})
dim = M: # of phonemes (one-hot)
sil d ow sil

Experimental Results

Approaches		Matched (all 4000)		Nonmatched (3000/1000)	
		FER	PER	FER	PER
(I) Supervised (labeled)					
(a) RNN Transducer [23]		-	17.7	-	-
(b) standard HMMs		-	21.5	-	-
(c) Phoneme classifier		27.0	28.9	-	-
(II) Unsupervised (with oracle boundaries)					
(d) Relationship mapping GAN [22]		40.5	40.2	43.6	43.4
(e) Segmental Emperical-ODM [23]		33.3	32.5	40.0	40.1
(f) Proposed: GAN		27.6	28.5	32.7	34.3
(III) Completely unsupervised (no label at all)					
(g) Segmental Emperical-ODM [23]		-	36.5	-	41.6
Proposed iteration 1	(h) GAN	48.3	48.6	50.3	50.0
	(i) GAN/HMM	-	30.7	-	39.5
Proposed iteration 2	(j) GAN	41.0	41.0	44.3	44.3
	(k) GAN/HMM	-	27.0	-	35.5
Proposed iteration 3	(l) GAN	39.7	38.4	45.0	44.2
	(m) GAN/HMM	-	26.1	-	33.1



The image is modified from: Phone recognition on the TIMIT database Lopes, C. and Perdigão, F., 2011. Speech Technologies, Vol 1, pp. 285--302.

Three Categories of GAN

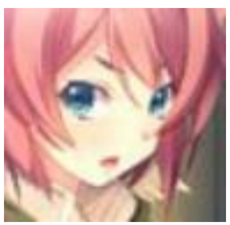
1. Typical GAN

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

random vector → Generator →

image

2. Conditional GAN



blue eyes,
red hair,
short hair

paired data

"Girl with red hair"
text

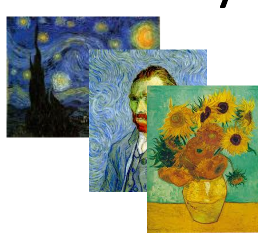
→ Generator →

image

3. Unsupervised Conditional GAN

domain x

domain y



unpaired data

x

Photo

→ Generator →

y

Vincent van Gogh's style

To Learn More …

You can learn more from the YouTube Channel