Generative Adversarial Network
and its Applications to Human Language Processing

李宏毅
Hung-yi Lee

Full version of the tutorial

臺灣大學
National Taiwan University

Outline

National Taiwan University

Part I: General Introduction of Generative Adversarial Network (GAN)

Part II: Applications to Natural Language Processing

Part III: Applications to Speech Processing

All Kinds of GAN …

https://github.com/hindupuravinash/the-gan-zoo

GAN

ACGAN

BGAN

CGAN

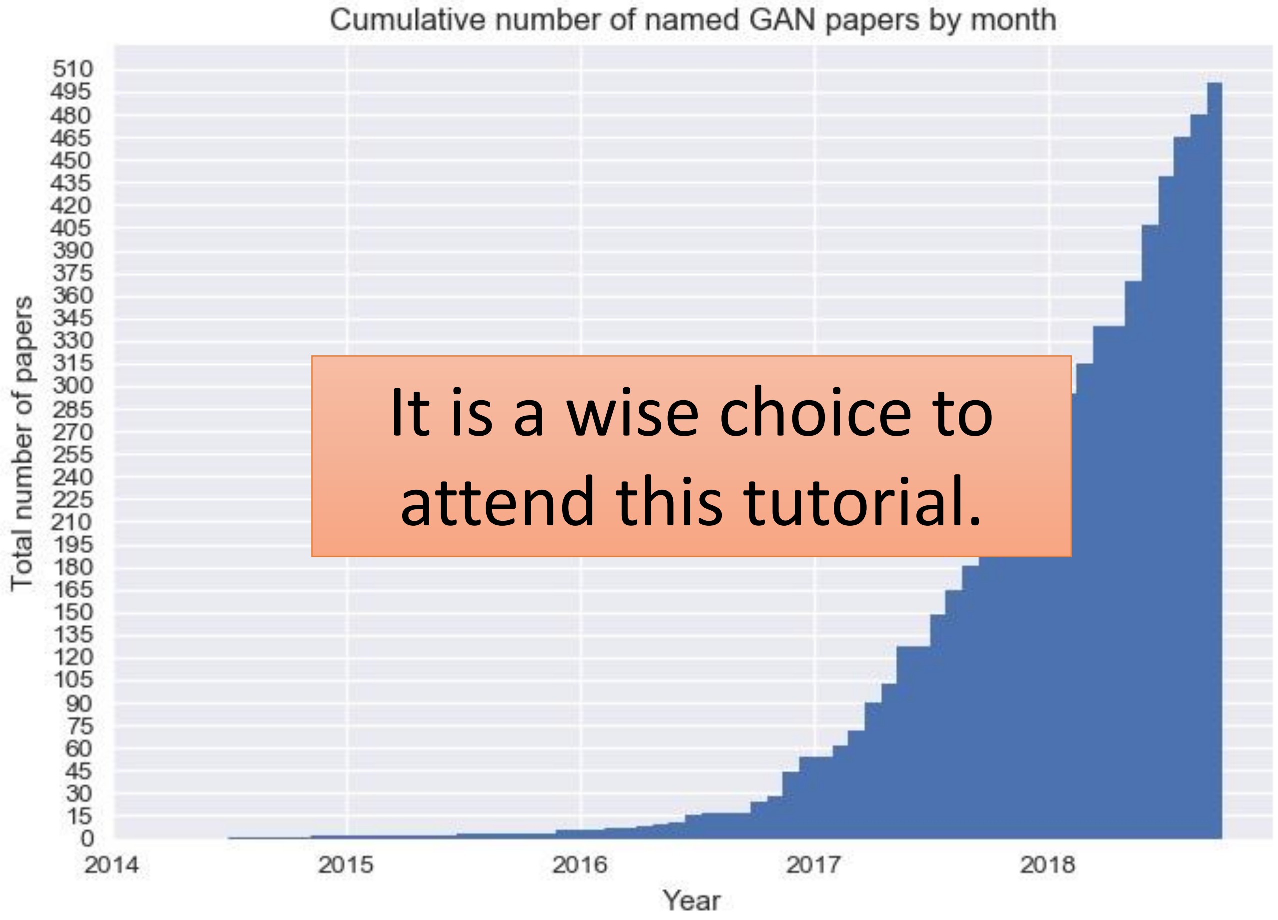
DCGAN

EBGAN

fGAN

GoGAN

⋮



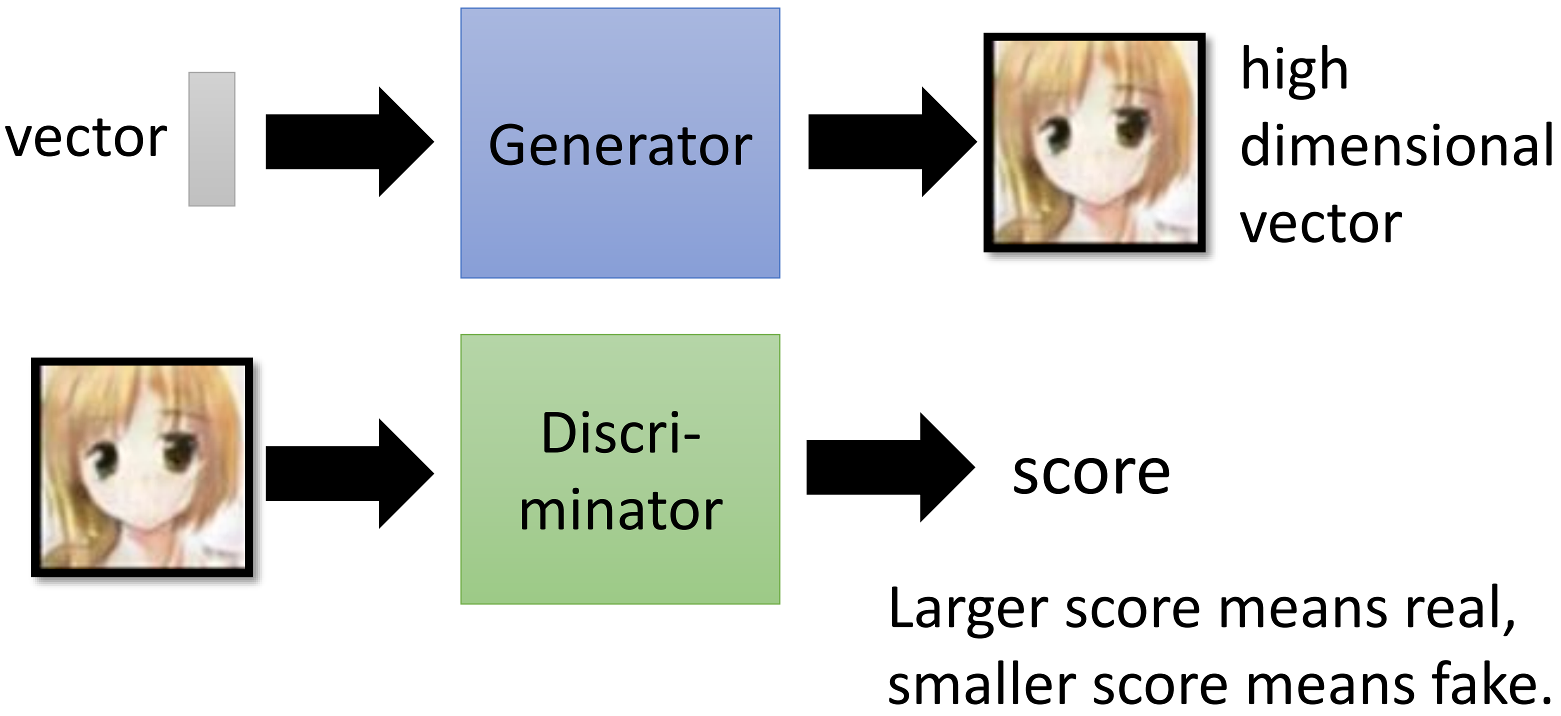
It is a wise choice to attend this tutorial.

Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, Shakir Mohamed, “Variational Approaches for Auto-Encoding Generative Adversarial Networks”, arXiv, 2017

[2]We use the Greek α prefix for α-GAN, as AEGAN and most other Latin prefixes seem to have been taken https://deephunt.in/the-gan-zoo-79597dc8c347.

Generative Adversarial Network (GAN)

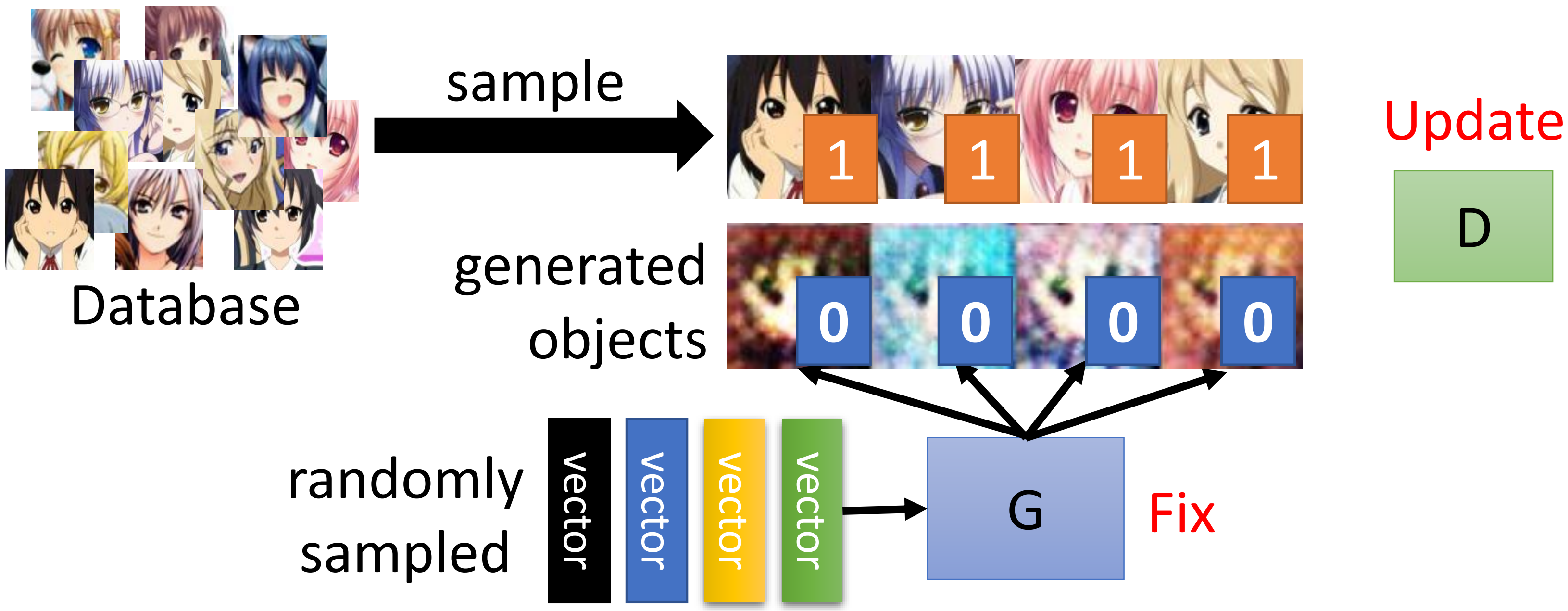
- Anime face generation as example



Algorithm

- Initialize generator and discriminator G D
- In each training iteration:

Step 1: Fix generator G, and update discriminator D



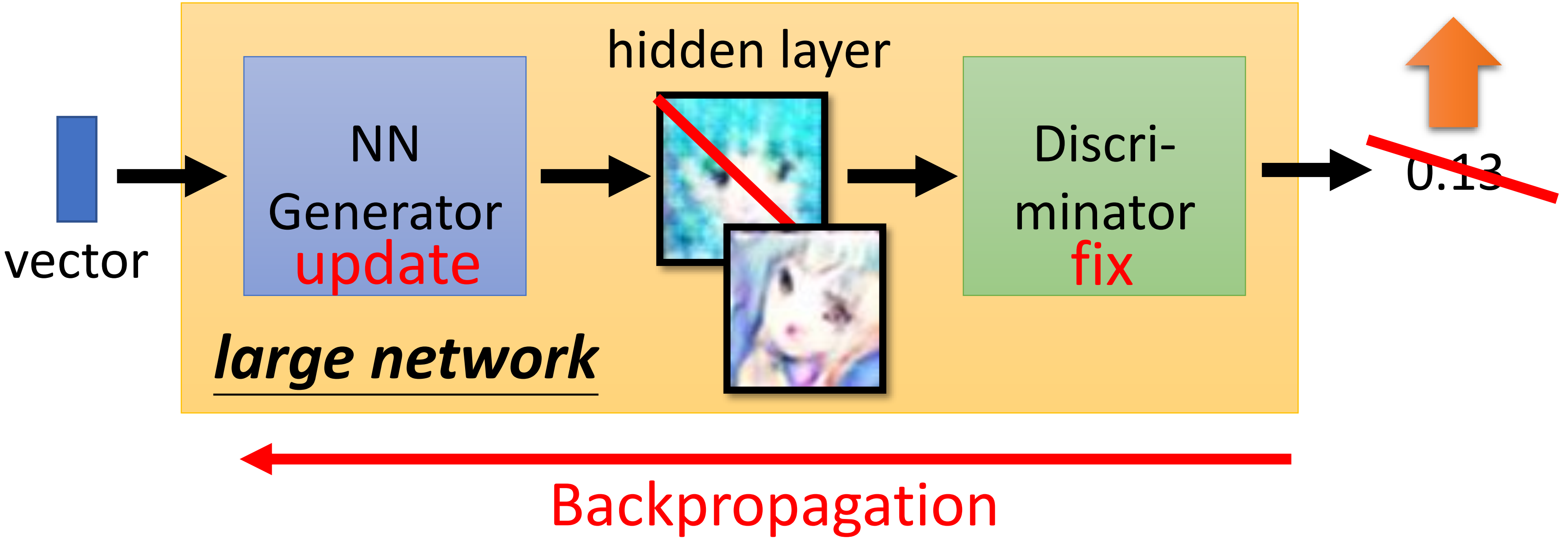
Discriminator learns to assign high scores to real objects and low scores to generated objects.

Algorithm

- Initialize generator and discriminator G D
- In each training iteration:

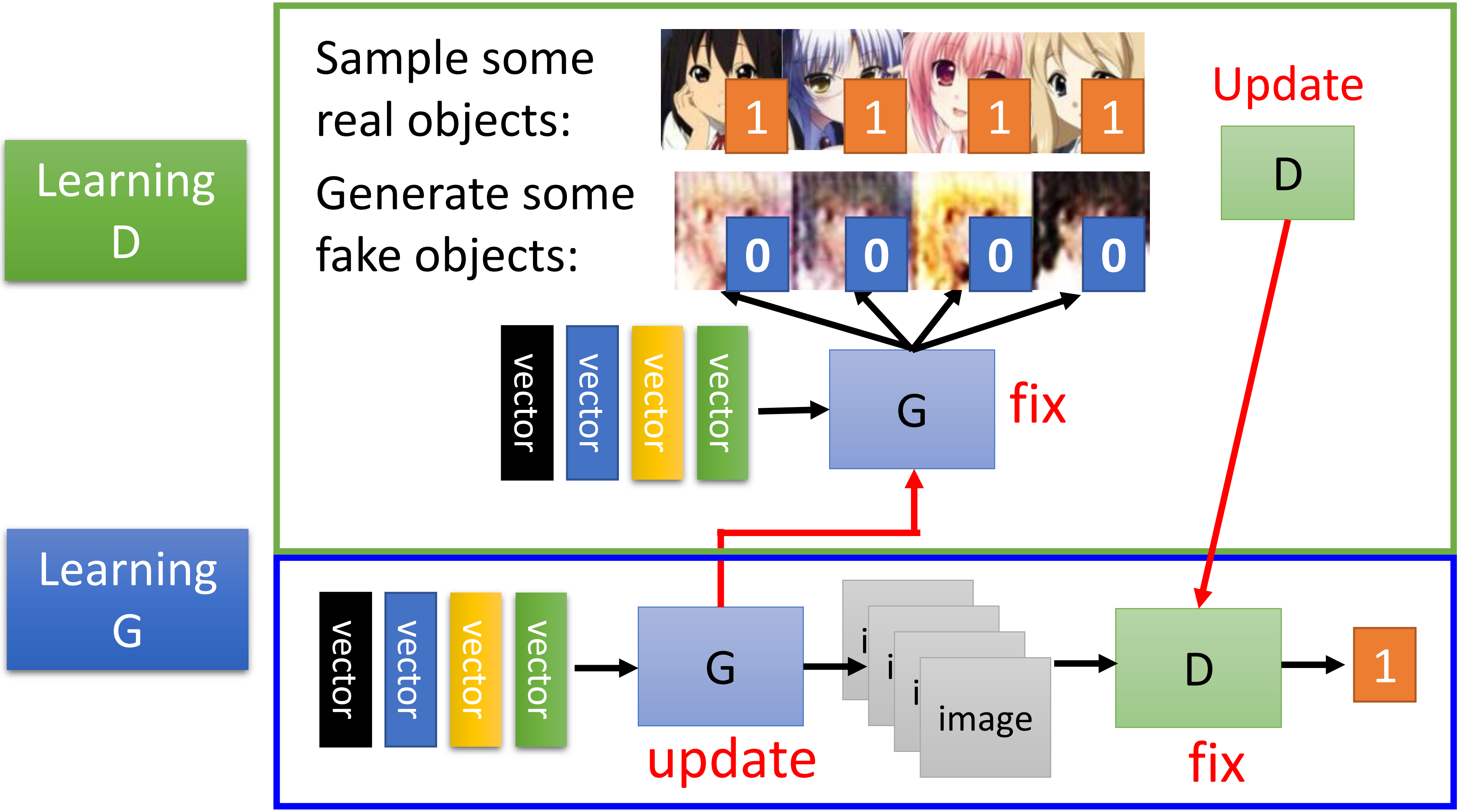
Step 2: Fix discriminator D, and update generator G

Generator learns to "*fool*" the discriminator



Algorithm

- Initialize generator and discriminator G D
- In each training iteration:



The faces generated by machine.

The images are generated by Yen-Hao Chen, Po-Chun Chien, Jun-Chen Xie, Tsung-Han Wu.

Conditional Generation

Generation

$$\begin{bmatrix} 0.3 \\ -0.1 \\ \vdots \\ -0.7 \end{bmatrix} \begin{bmatrix} 0.1 \\ -0.1 \\ \vdots \\ 0.7 \end{bmatrix} \begin{bmatrix} -0.3 \\ 0.1 \\ \vdots \\ 0.9 \end{bmatrix} \rightarrow \boxed{\text{NN Generator}} \rightarrow$$

In a specific range

Conditional Generation

"Girl with red hair and red eyes"

"Girl with yellow ribbon"

$\rightarrow \boxed{\text{NN Generator}} \rightarrow$

paired data

blue eyes
red hair
short hair

Conditional GAN

c: red hair

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: (red hair, ) 1

(blue hair , ) 0

(red hair, Image) 0

[Scott Reed, et al, ICML, 2016]

The images are generated by Yen-Hao Chen, Po-Chun Chien, Jun-Chen Xie, Tsung-Han Wu.

Conditional GAN

[Scott Reed, et al, ICML, 2016]

paired data

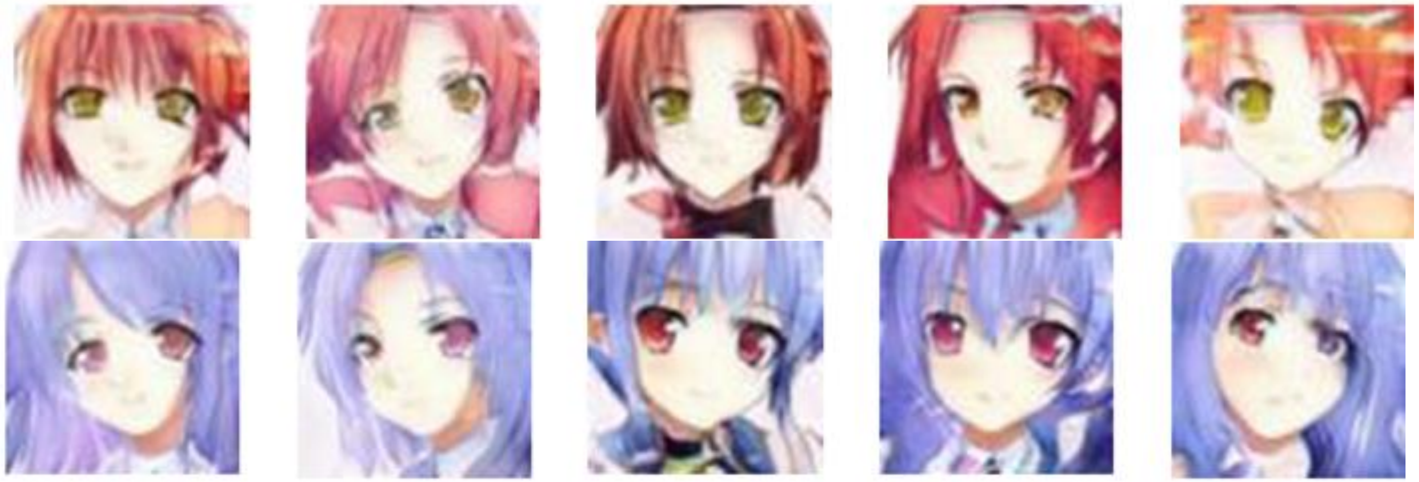
blue eyes
red hair
short hair

$x = G(c,z)$

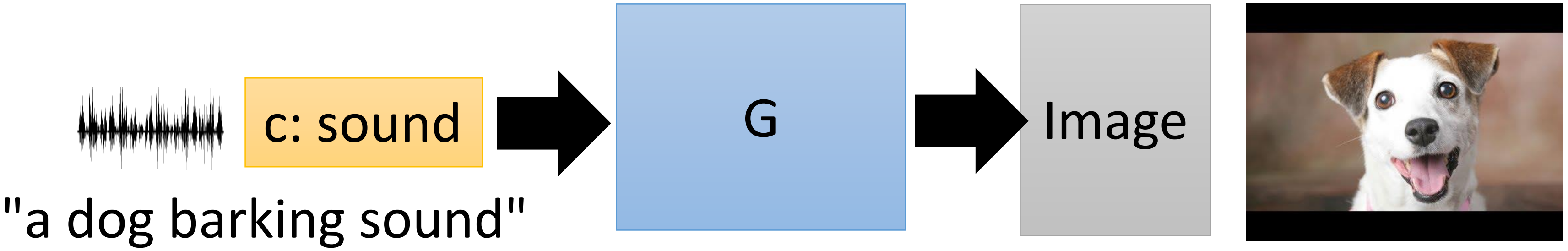
c: text → G → Image

red hair,
green eyes

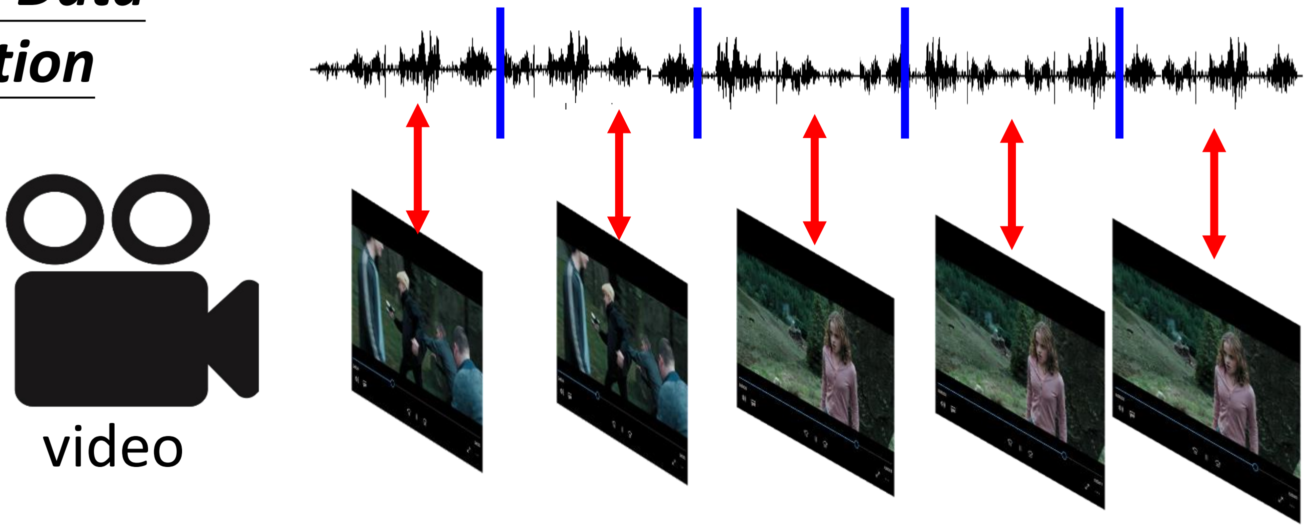
blue hair,
red eyes



Conditional GAN



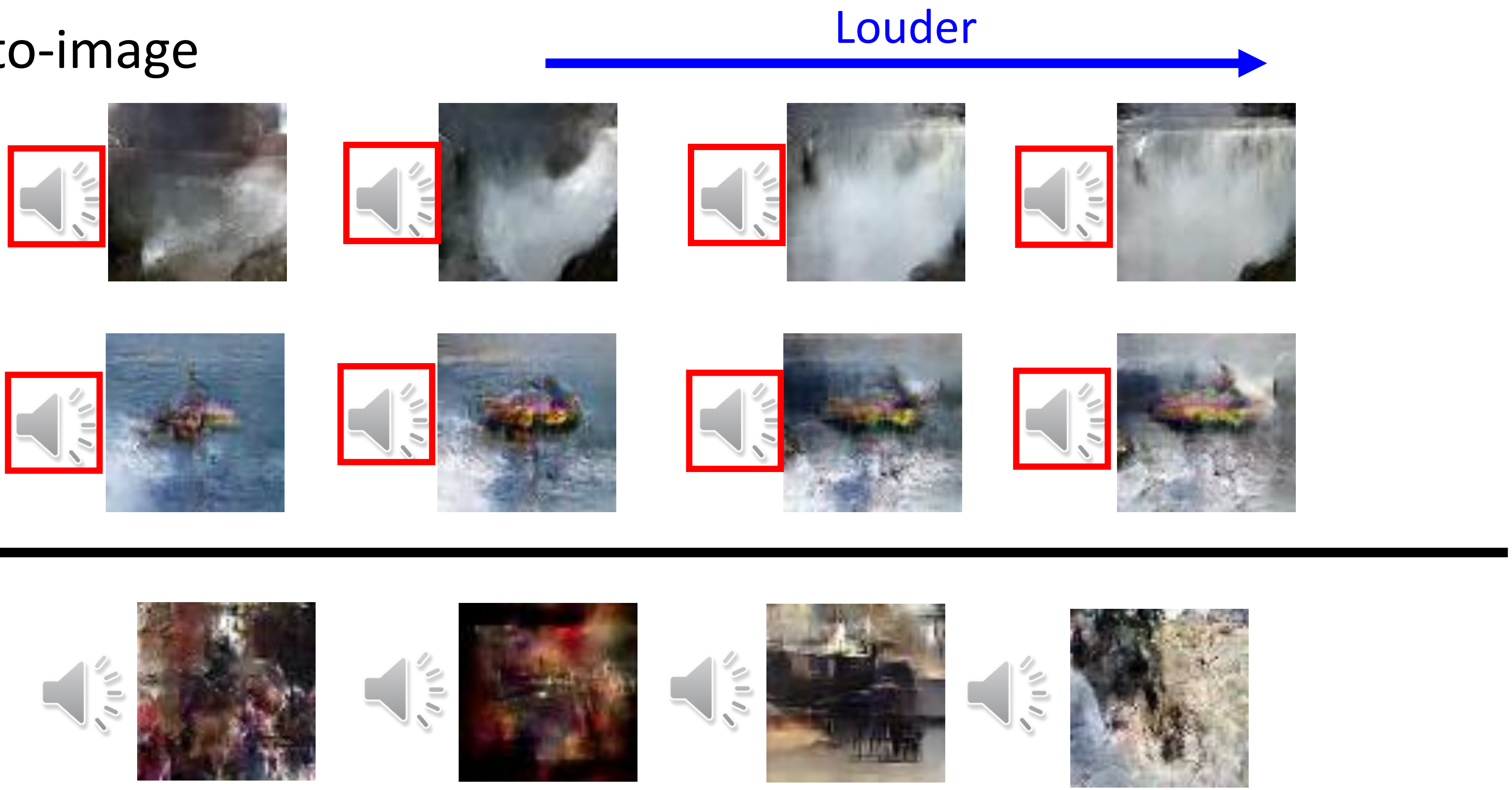
Training Data Collection



Conditional GAN

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

- Audio-to-image



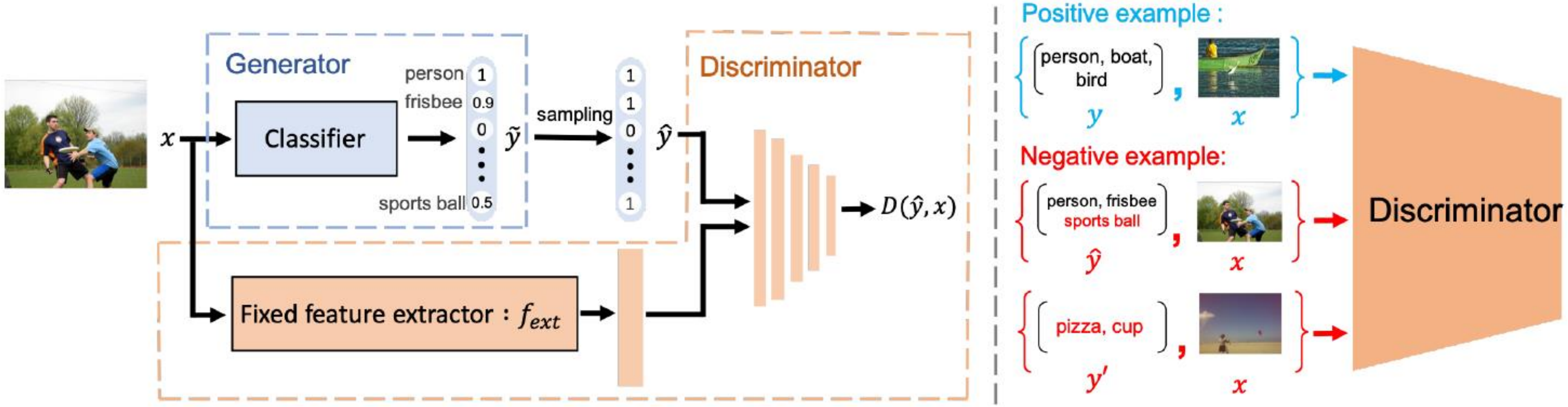
Conditional GAN - Image-to-label

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.

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

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

Conditional GAN - Image-to-label

The classifiers can have different architectures.

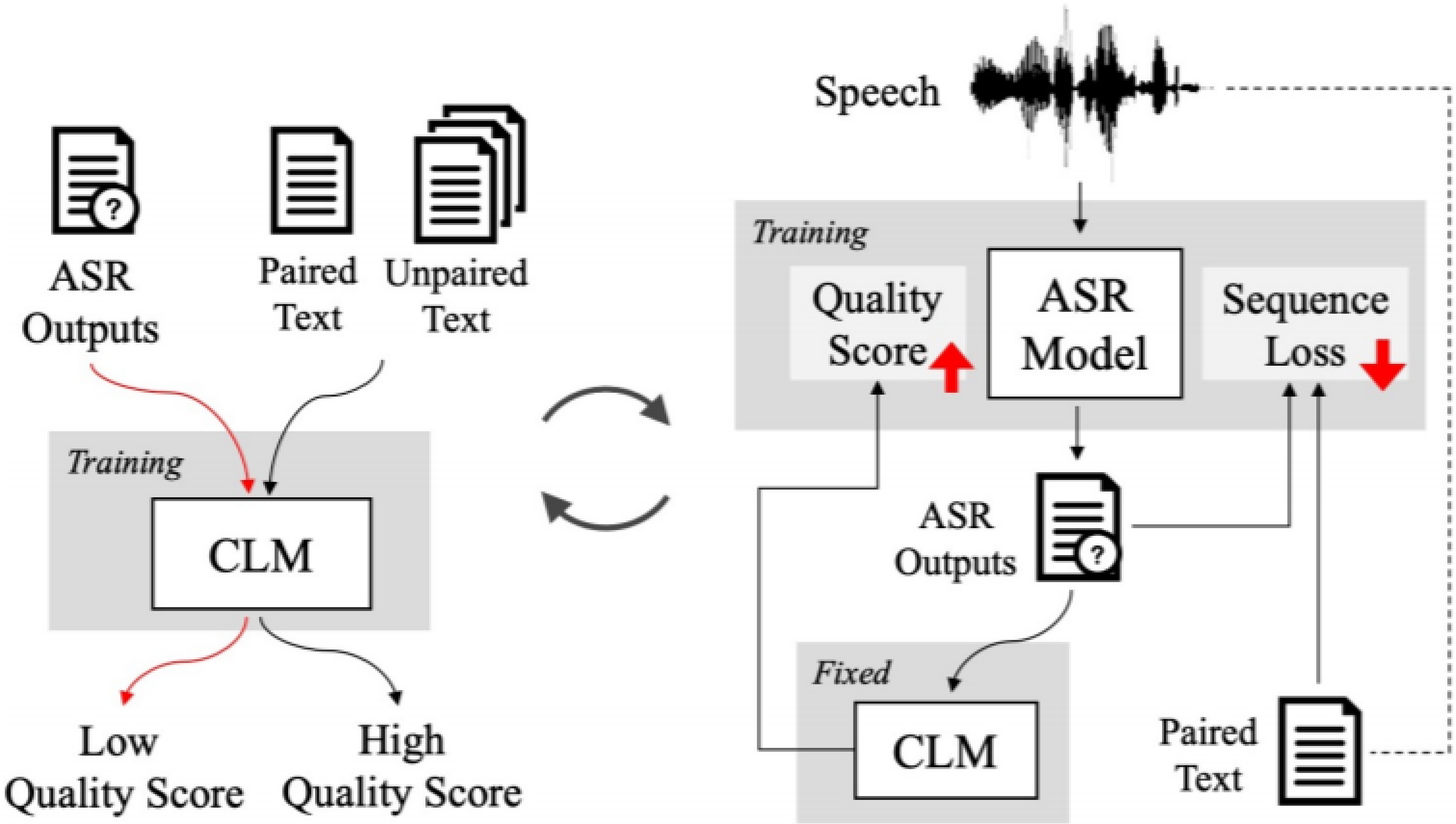
The classifiers are trained as conditional GAN.

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	$\boxed{55.8}$
Resnet-101	62.8	53.1
+GAN	$\boxed{64.0}$	55.4
Resnet-152	63.3	52.1
+GAN	63.9	54.1
Att-RNN	62.1	54.7
RLSD	62.0	46.9

Conditional GAN – Speech Recognition

Adversarial Training of End-to-end Speech Recognition Using a Criticizing Language Model, https://arxiv.org/abs/1811.00787



(a) CLM learning step

(b) ASR model learning step

Table 1. Speech recognition performance. "+LM" refers to shallow fusion decoding jointly with RNN-LM [13], "+AT" refers to the adversarial training proposed here, "+Both" indicates training with AT and joint decoding with RNN-LM, and BT is the prior work of back-translation [21].

Data	Method	CER/WER (%) Dev	CER/WER (%) Test	WER Δ† Test
(A) w/o unpair text	(a) Baseline	10.5 / 21.6	10.5 / 21.7	-
	(b) +LM	10.9 / 20.0	11.1 / 20.3	6.5%
	(c) +AT	**9.5 / 19.9**	**9.6 / 20.1**	**7.4%**
	(d) +Both	9.4 / 17.9	9.7 / 18.3	15.7%
(B) w/ 360hrs text	(e) +LM	10.5 / 19.6	10.6 / 19.6	9.7%
	(f) +AT	**9.1 / 19.1**	**9.5 / 19.2**	$\boxed{\textbf{11.5\%}}$
	(g) +Both	9.0 / 17.1	9.1 / 17.3	20.3%
	(h) BT‡	10.3 / 23.5	10.3 / 23.6	6.3%
	(i) BT+LM‡	9.8 / 21.6	10.0 / 22.0	12.7%
(C) w/ 860hrs text	(j) +LM	9.9 / 18.6	10.2 / 18.8	13.4%
	(k) +AT	**8.6 / 18.5**	**8.8 / 18.7**	$\boxed{\textbf{13.8\%}}$
	(l) +Both	7.9 / 15.3	8.2 / 15.8	27.2%

† Relative improvement with respect to the baseline.
‡ Prior work [21], baseline WER 25.2% on test set reported.

Unsupervised Conditional GAN

Condition
Object in Domain X → **G** → Generated Object
Object in Domain Y

Transform an object from one domain to another ***without paired data*** (e.g. style transfer)

Domain X

photos

Not Paired

Domain Y

Vincent van Gogh's paintings

Unsupervised Conditional Generation

- Approach 1: Direct Transformation

Domain X

$\rightarrow G_{X\rightarrow Y} \rightarrow$

Domain Y

For texture or color change

- Approach 2: Projection to Common Space

Domain X

$\rightarrow EN_X \rightarrow$ Face Attribute $\rightarrow DE_Y \rightarrow$

Encoder of domain X

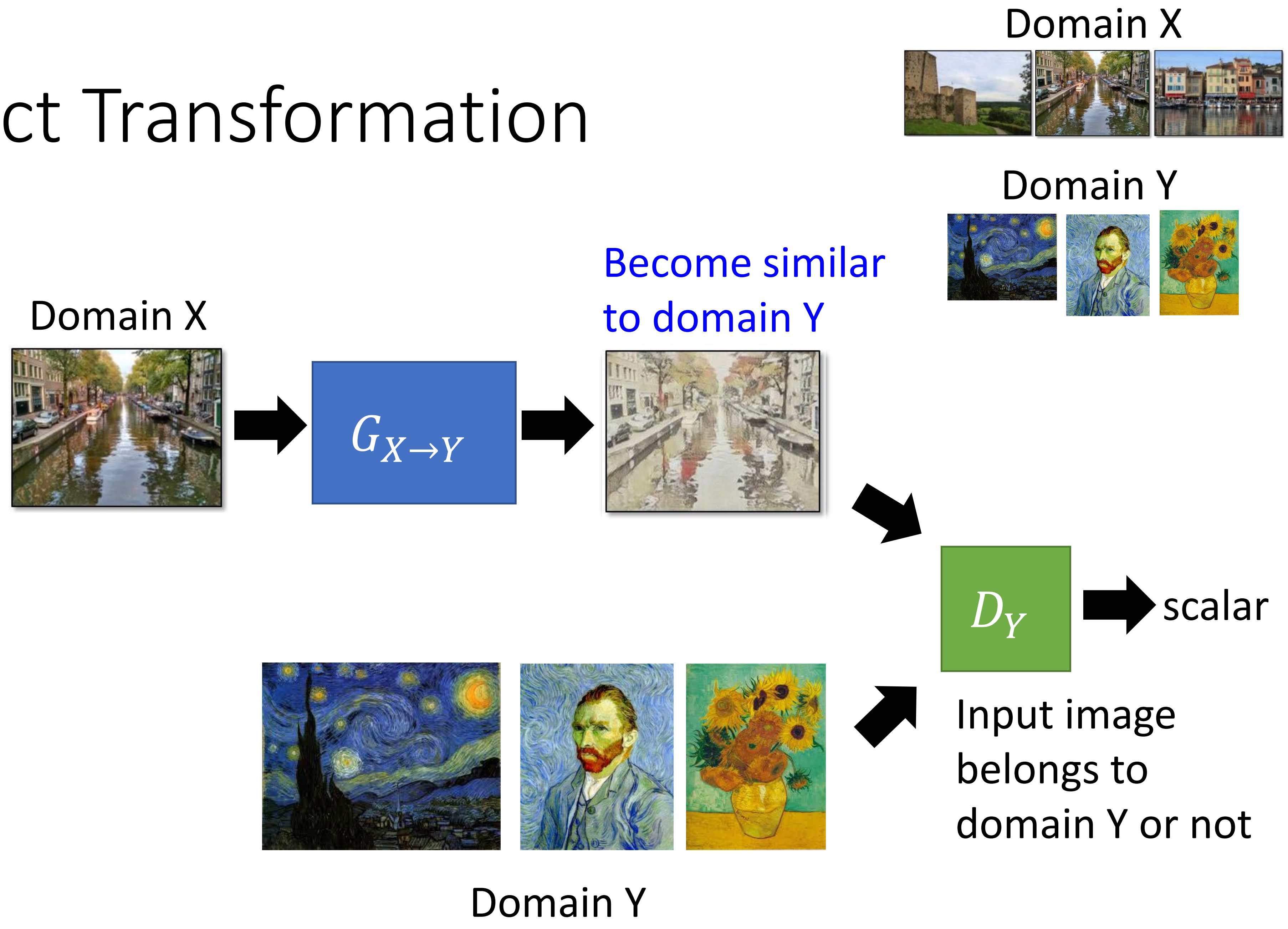
Face Attribute

Decoder of domain Y

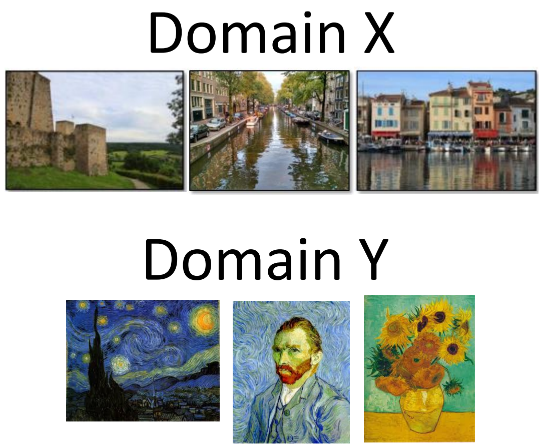
Domain Y

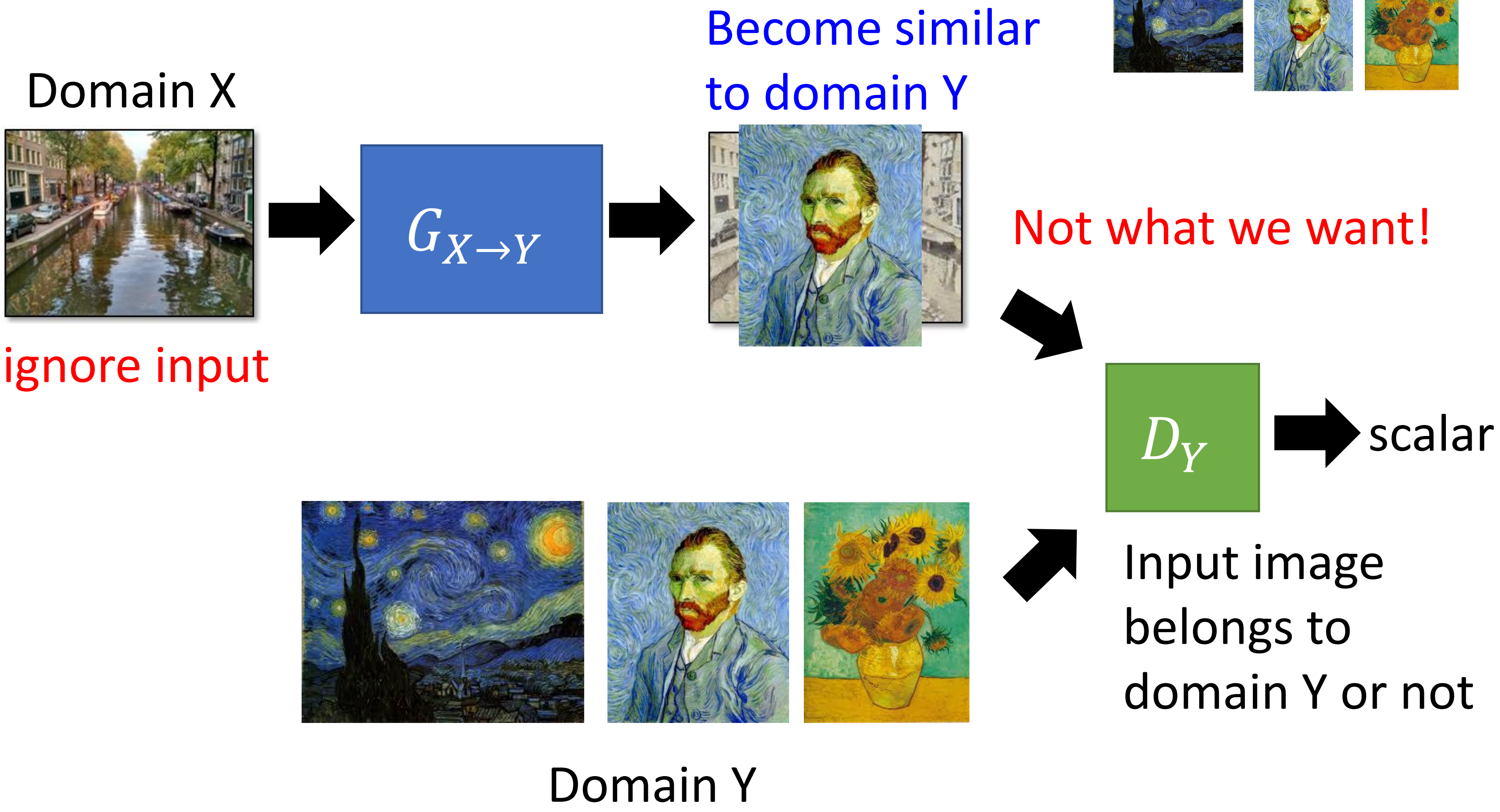
Larger change, only keep the semantics

Direct Transformation



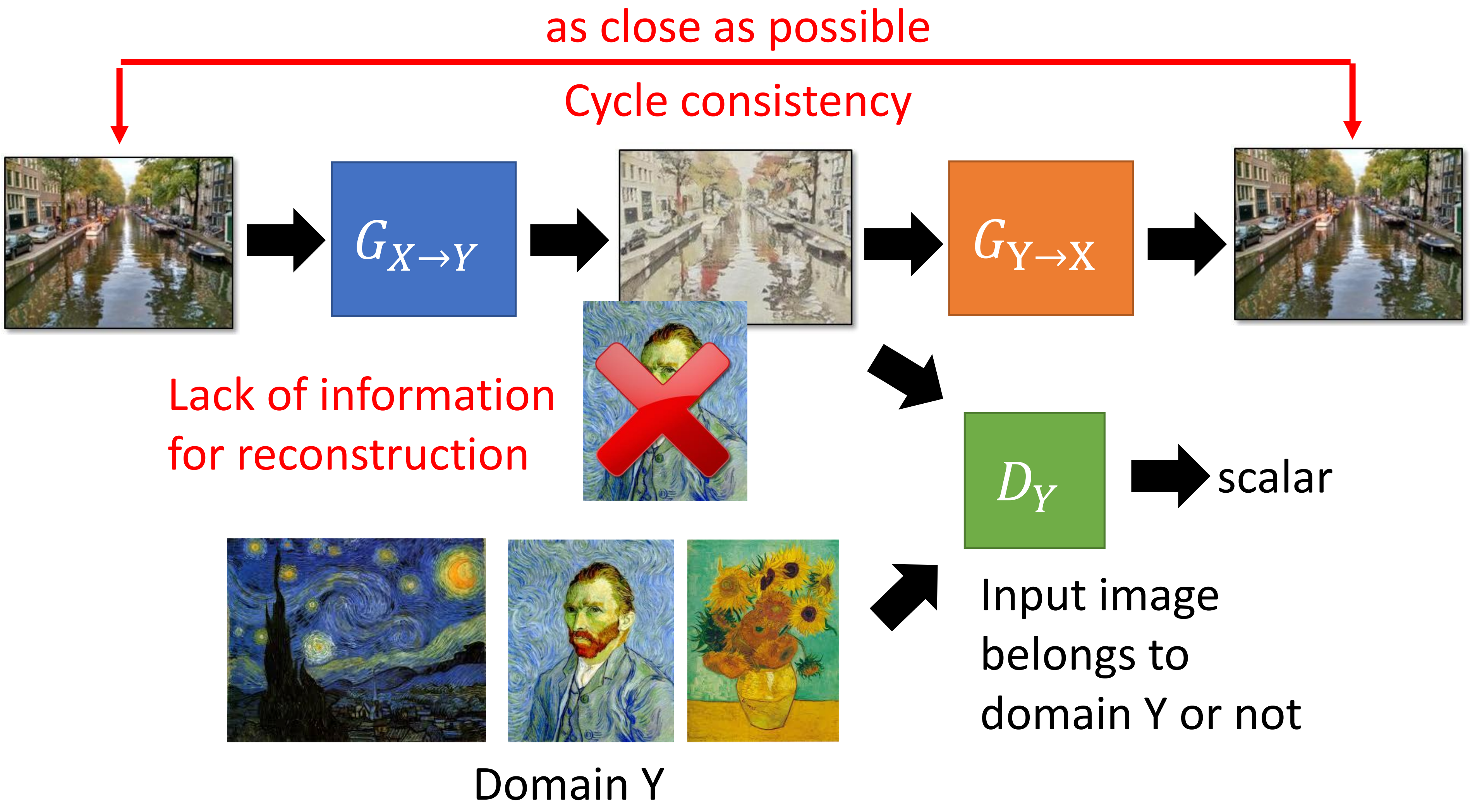
Direct Transformation



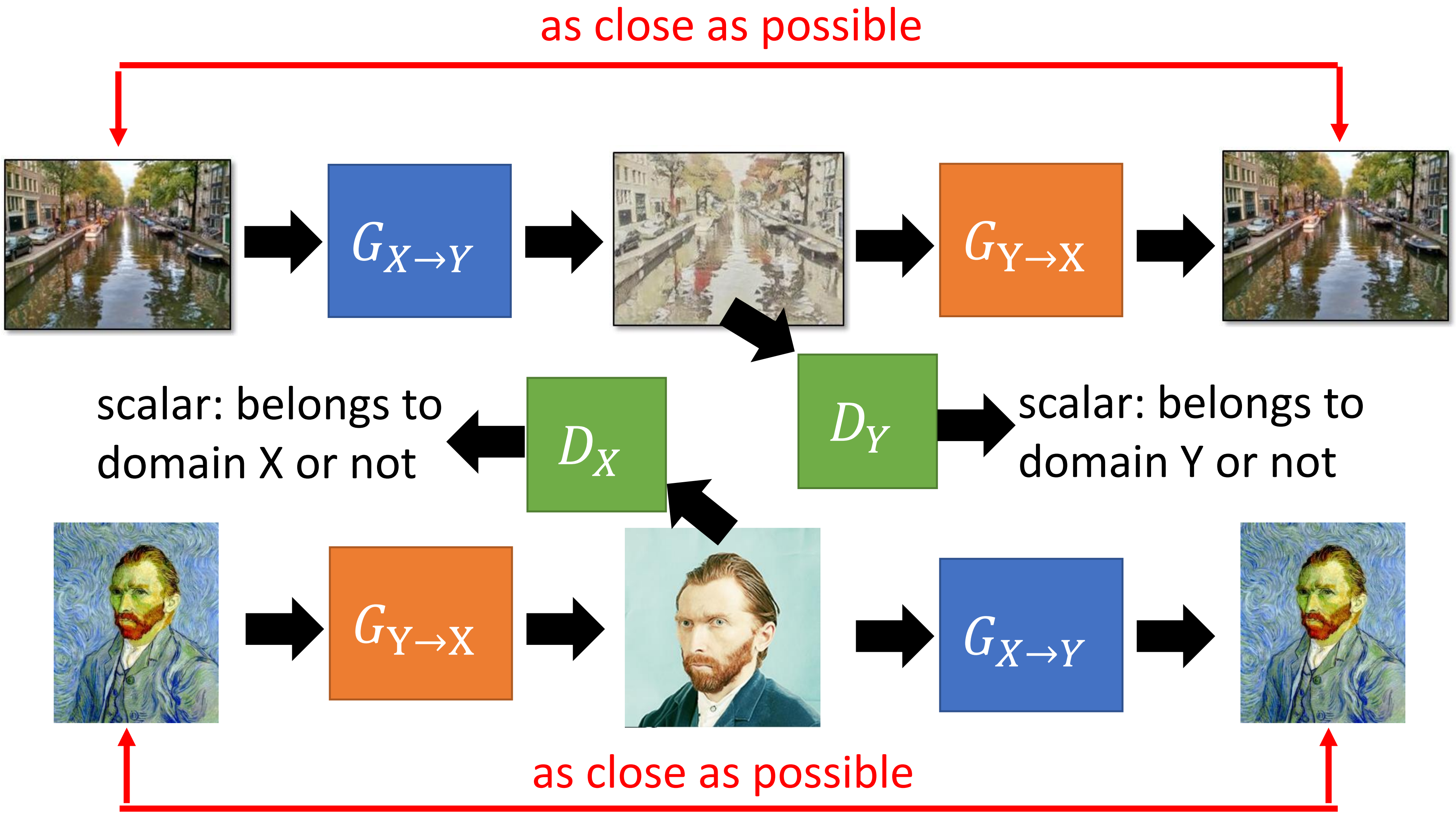


[Jun-Yan Zhu, et al., ICCV, 2017]

Direct Transformation

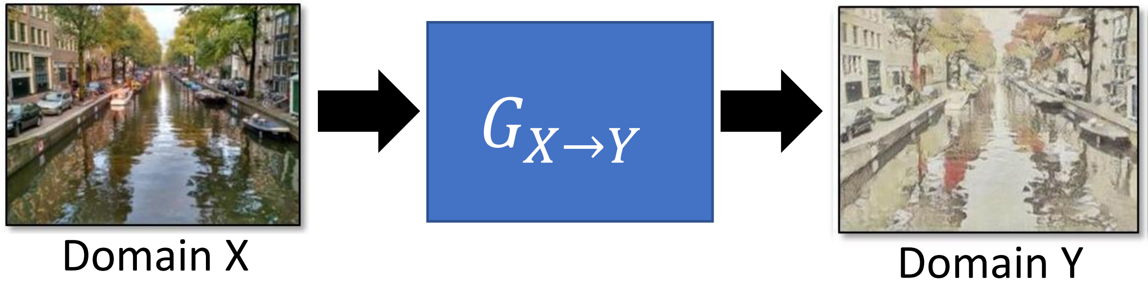


Cycle GAN


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

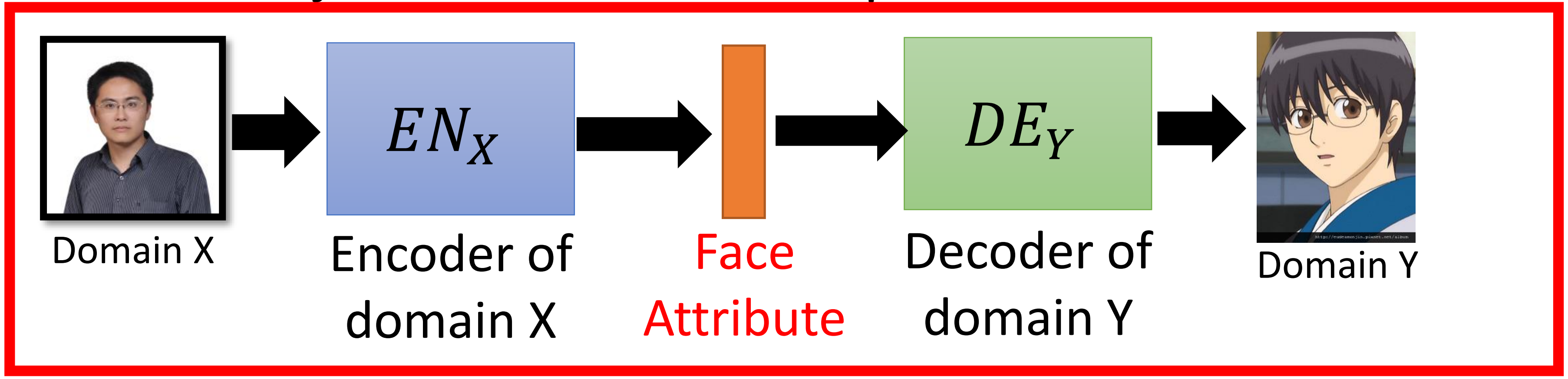
Unsupervised Conditional Generation

- Approach 1: Direct Transformation



For texture or color change

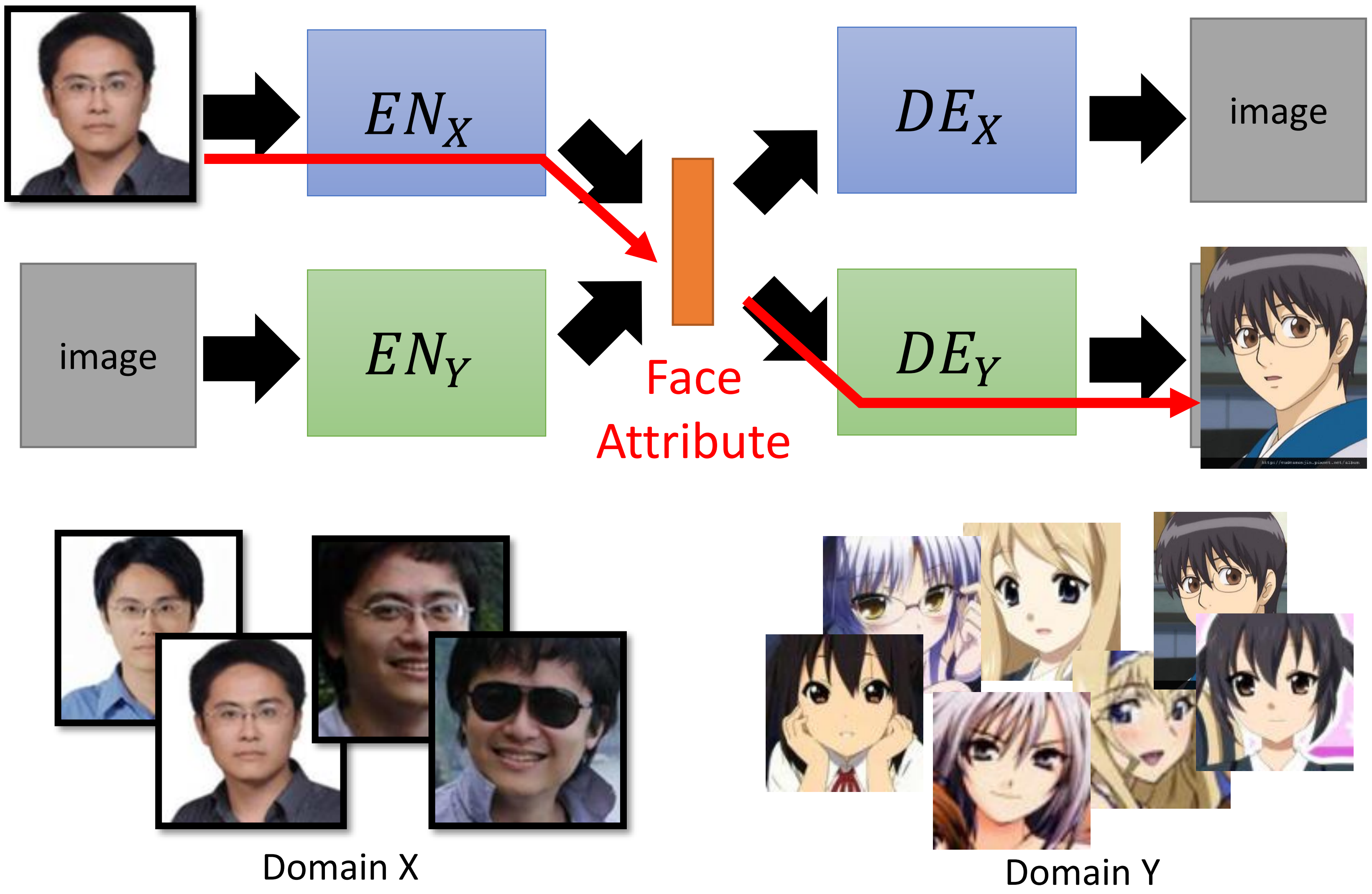
- Approach 2: Projection to Common Space



Larger change, only keep the semantics

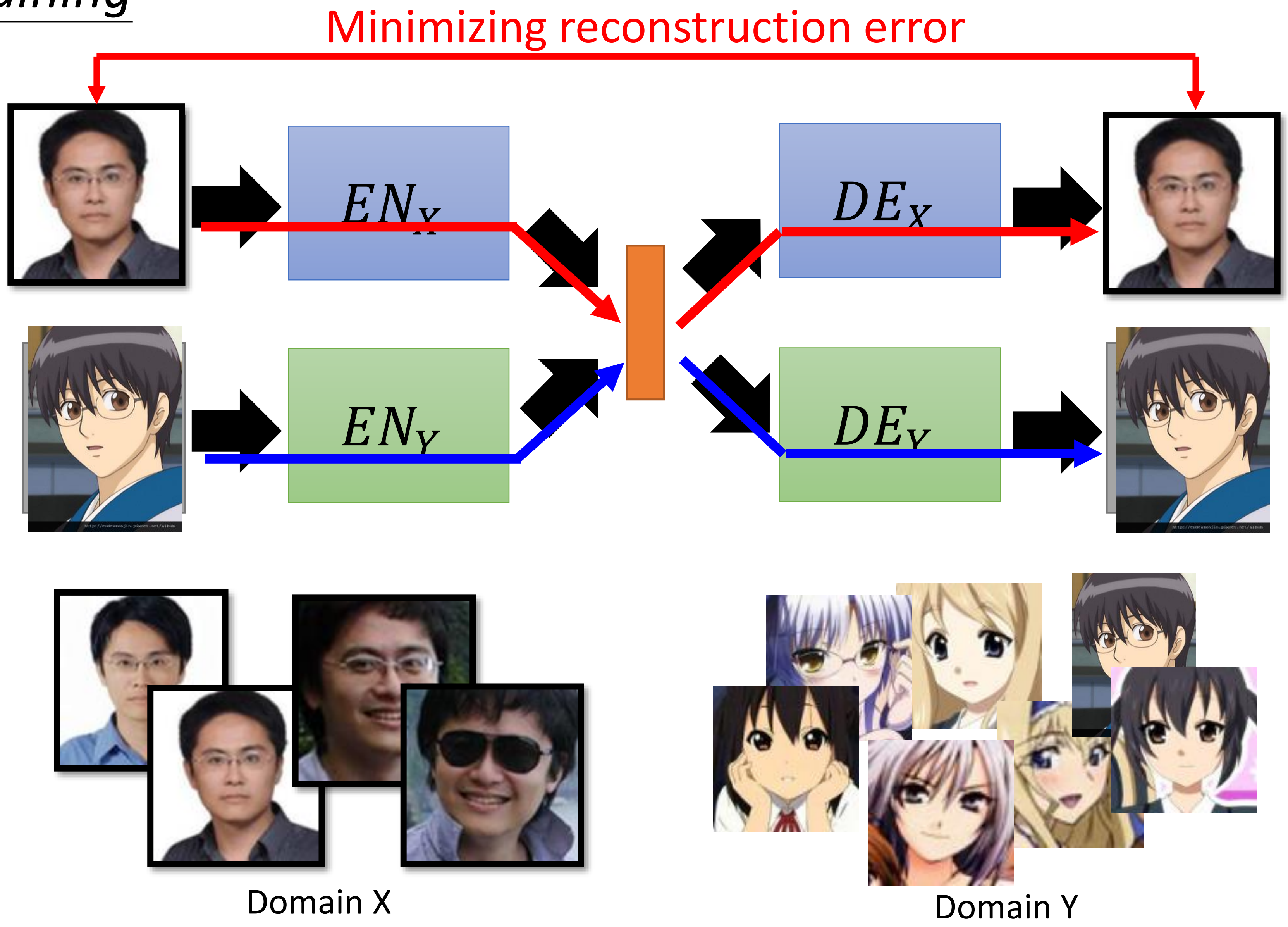
Projection to Common Space

Target



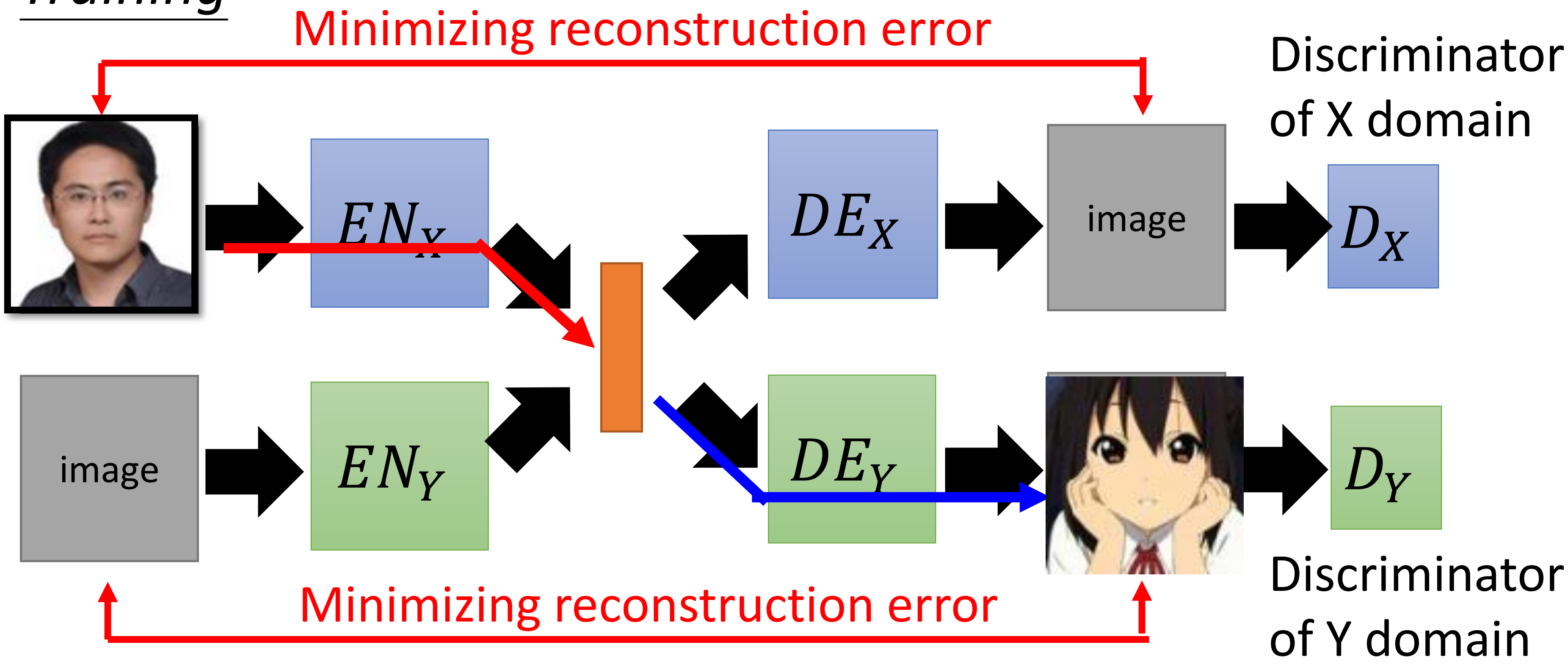
Projection to Common Space

Training



Projection to Common Space

Training

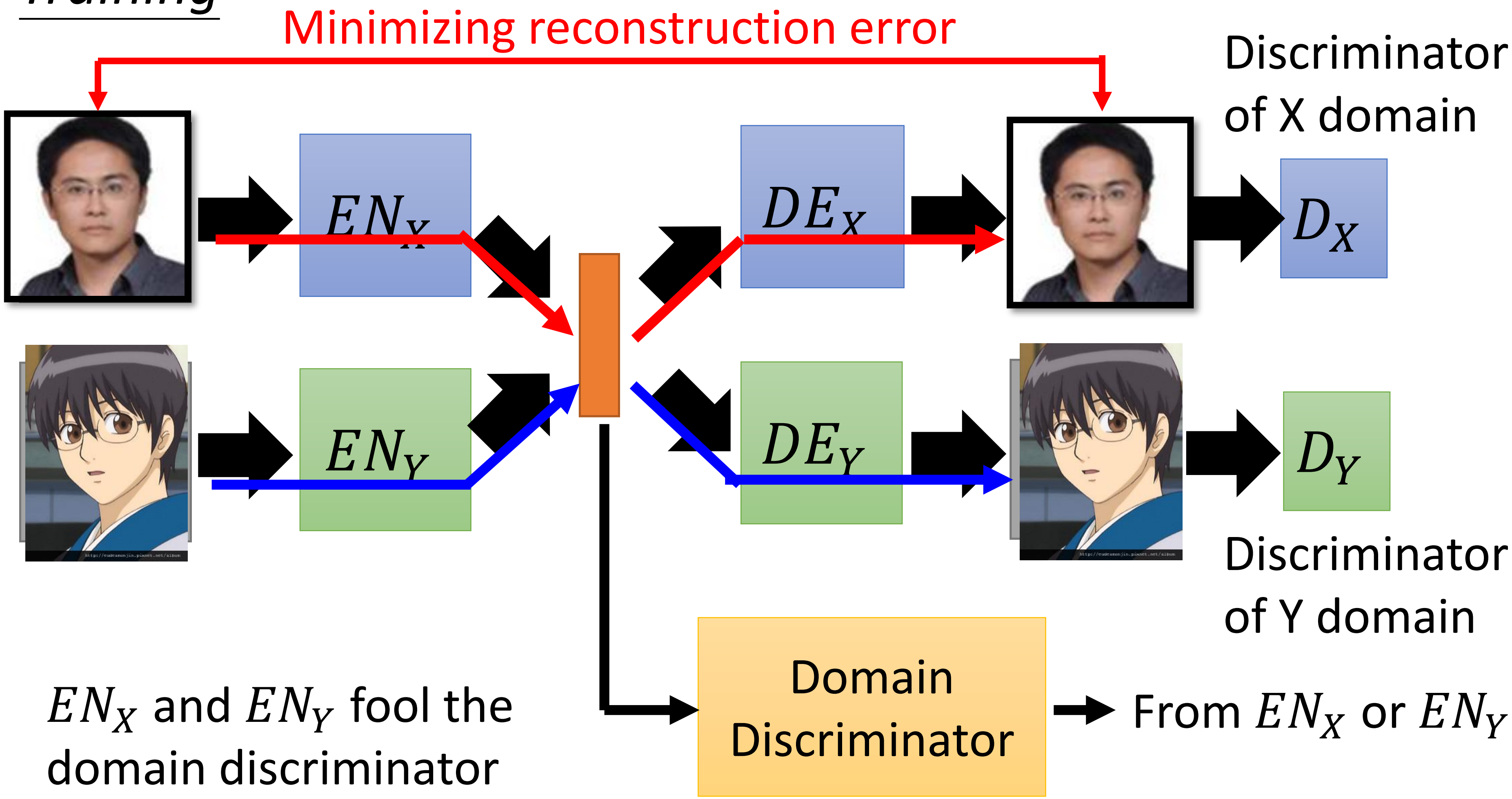


Because we train two auto-encoders separately …

The images with the same attribute may not project to the same position in the latent space.

Projection to Common Space

Training

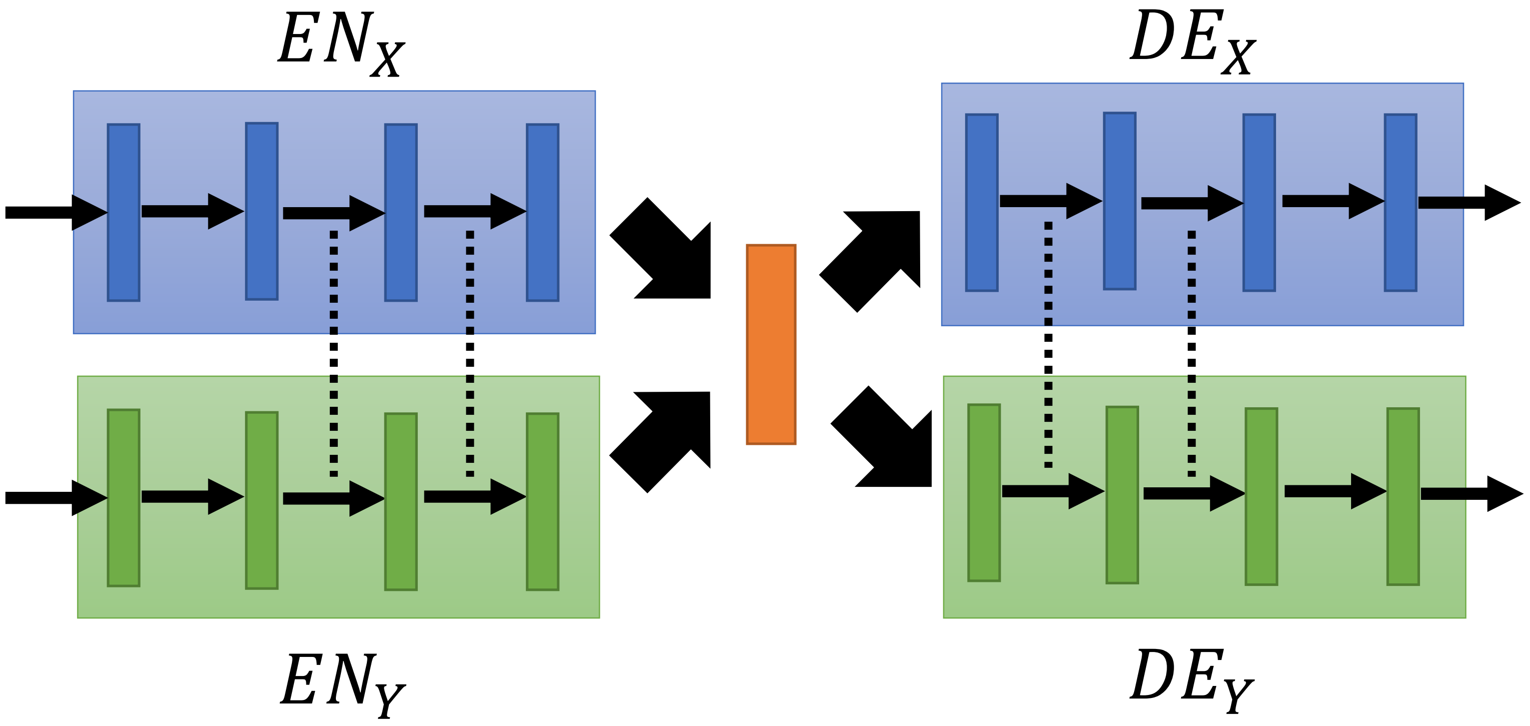


The domain discriminator forces the output of EN_X and EN_Y have the same distribution.

[Guillaume Lample, et al., NIPS, 2017]

Projection to Common Space

Training



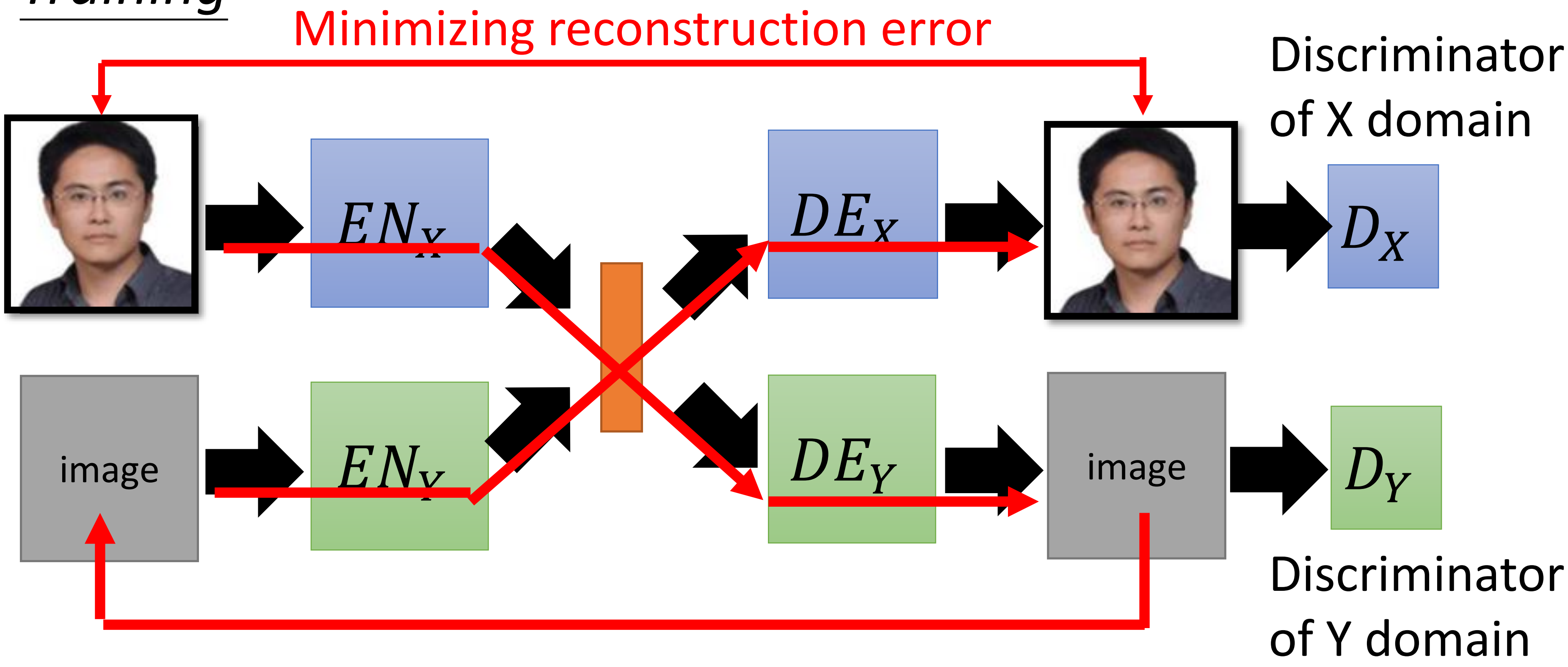
Sharing the parameters of encoders and decoders

Couple GAN[Ming-Yu Liu, et al., NIPS, 2016]

UNIT[Ming-Yu Liu, et al., NIPS, 2017]

Projection to Common Space

Training

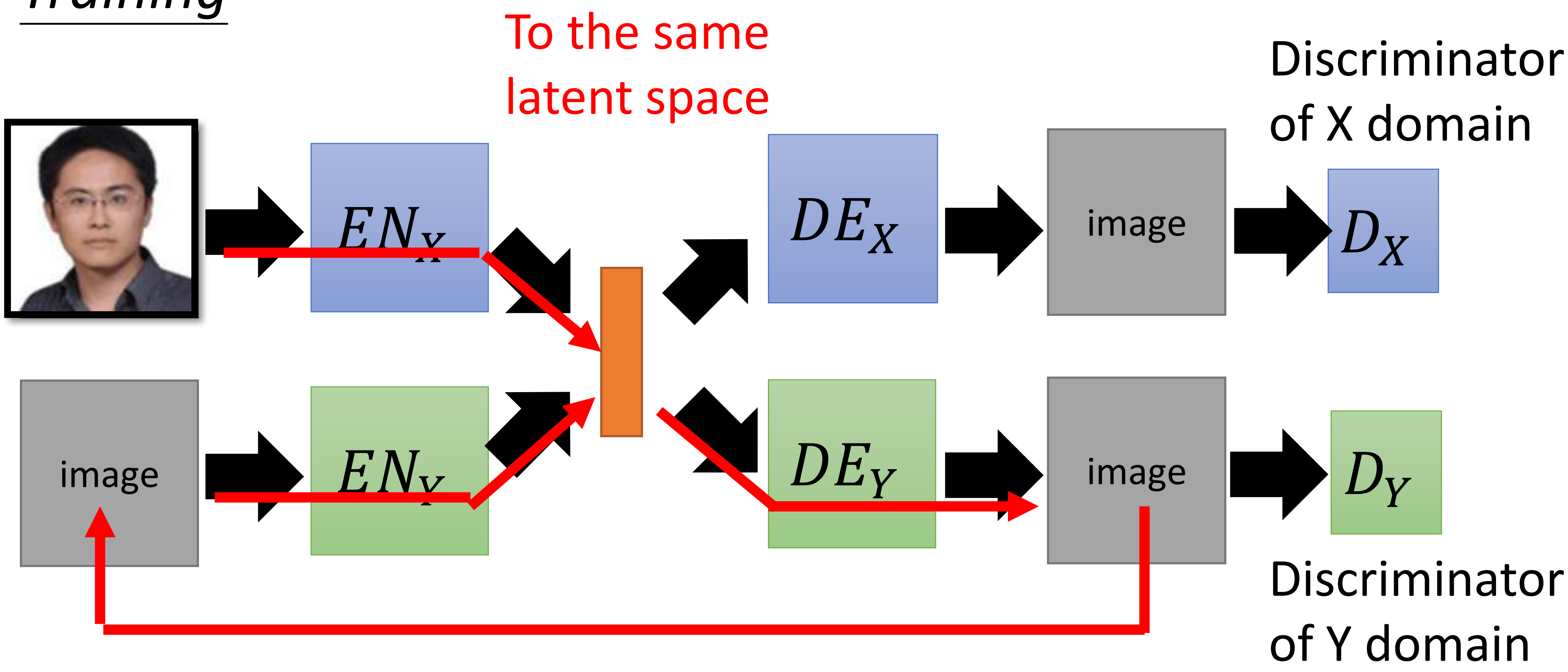


Cycle Consistency:

Used in ComboGAN [Asha Anoosheh, et al., arXiv, 017]

Projection to Common Space

Training



Semantic Consistency:

Used in DTN [Yaniv Taigman, et al., ICLR, 2017] and XGAN [Amélie Royer, et al., arXiv, 2017]

Outline

National Taiwan University

Part I: General Introduction of Generative Adversarial Network (GAN)

Part II: Applications to Natural Language Processing

Part III: Applications to Speech Processing

Unsupervised Conditional Generation

Image Style Transfer

photos

Not Paired

Vincent van Gogh's paintings

Text Style Transfer

It is good.
It's a good day.
I love you.

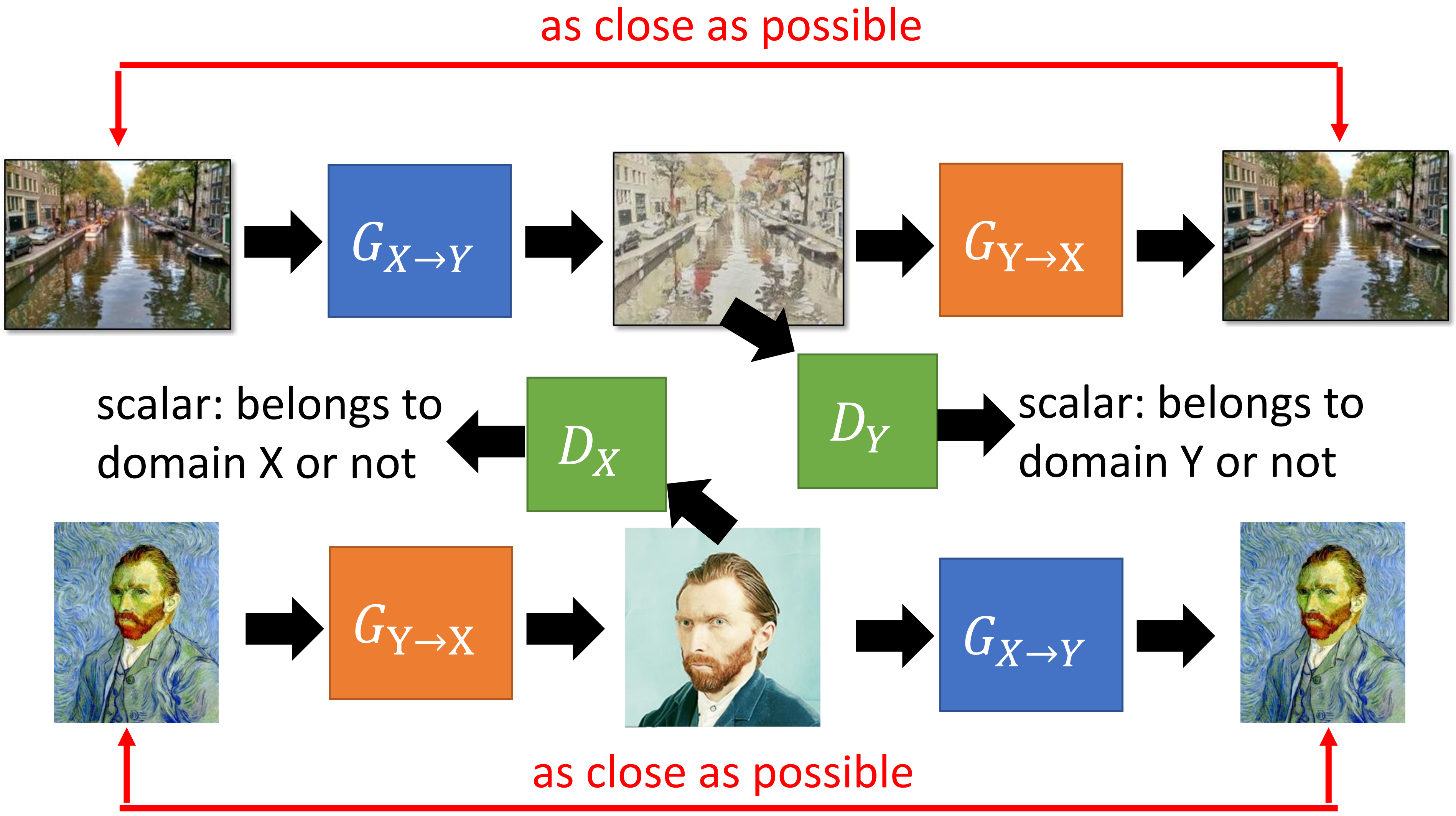
positive

Not Paired

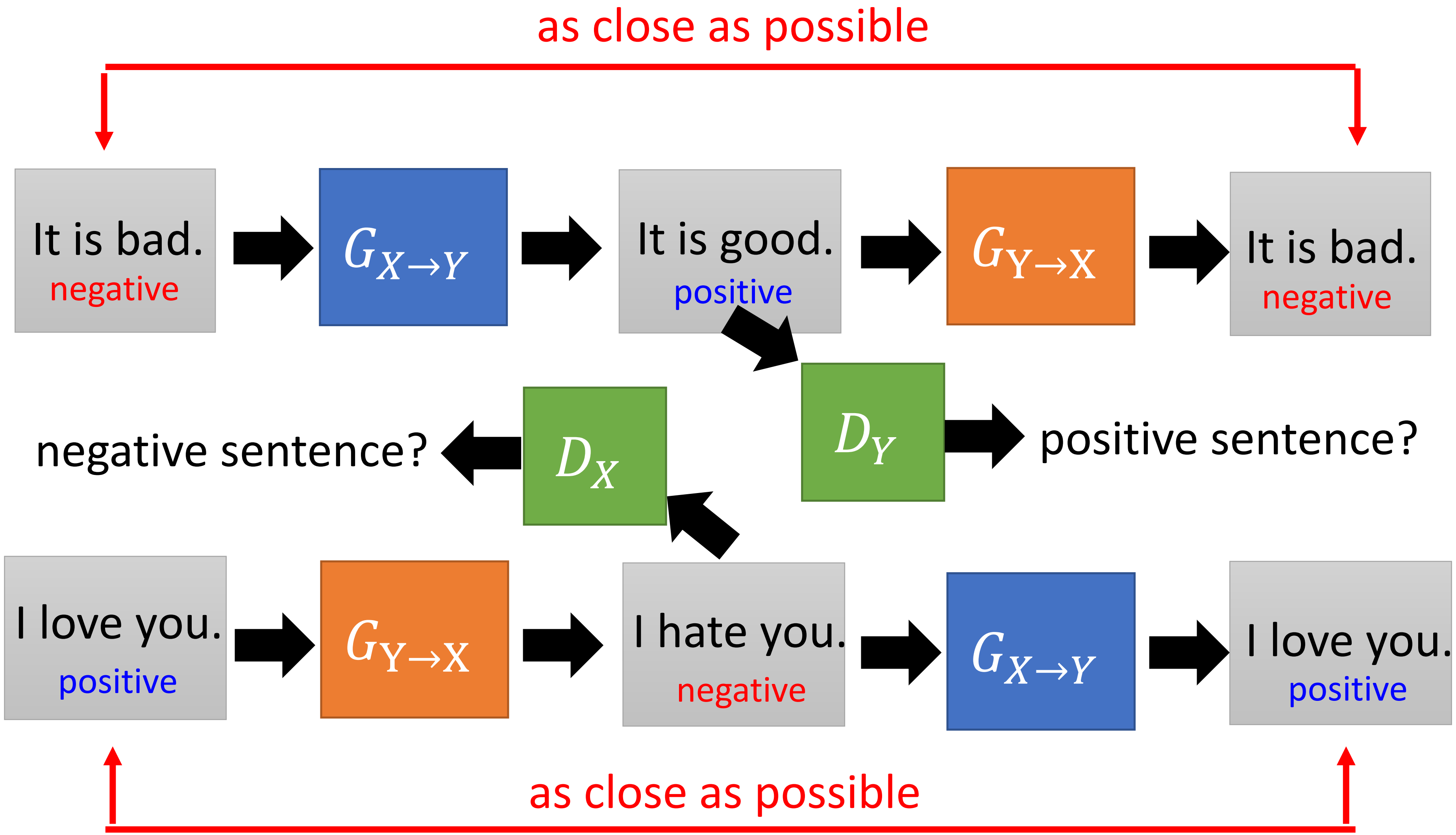
It is bad.
It's a bad day.
I don't love you.

negative

Cycle GAN

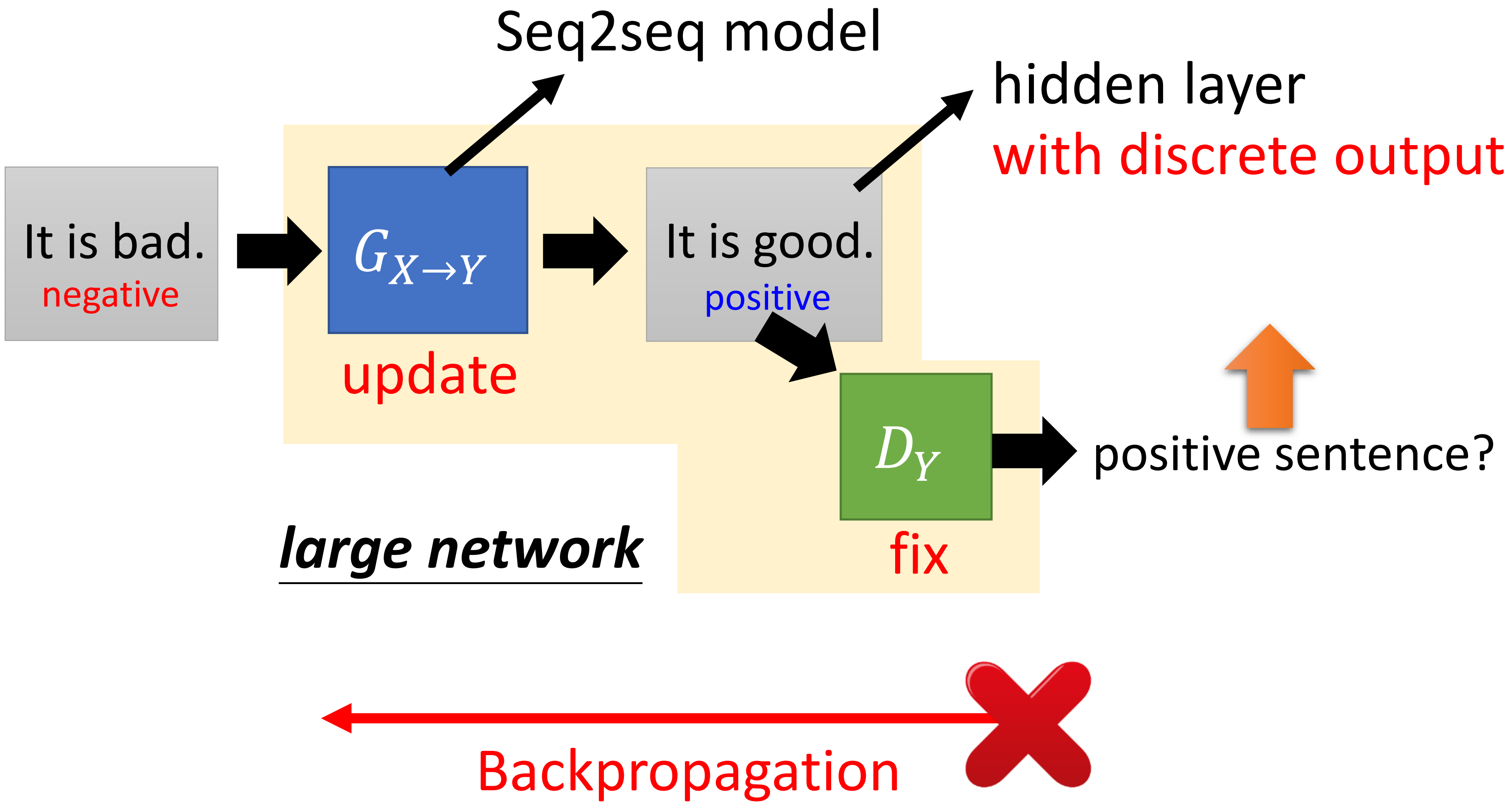

as close as possible
$G_{X \to Y}$
$G_{Y \to X}$
scalar: belongs to domain X or not
D_X
D_Y
scalar: belongs to domain Y or not
$G_{Y \to X}$
$G_{X \to Y}$
as close as possible

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”

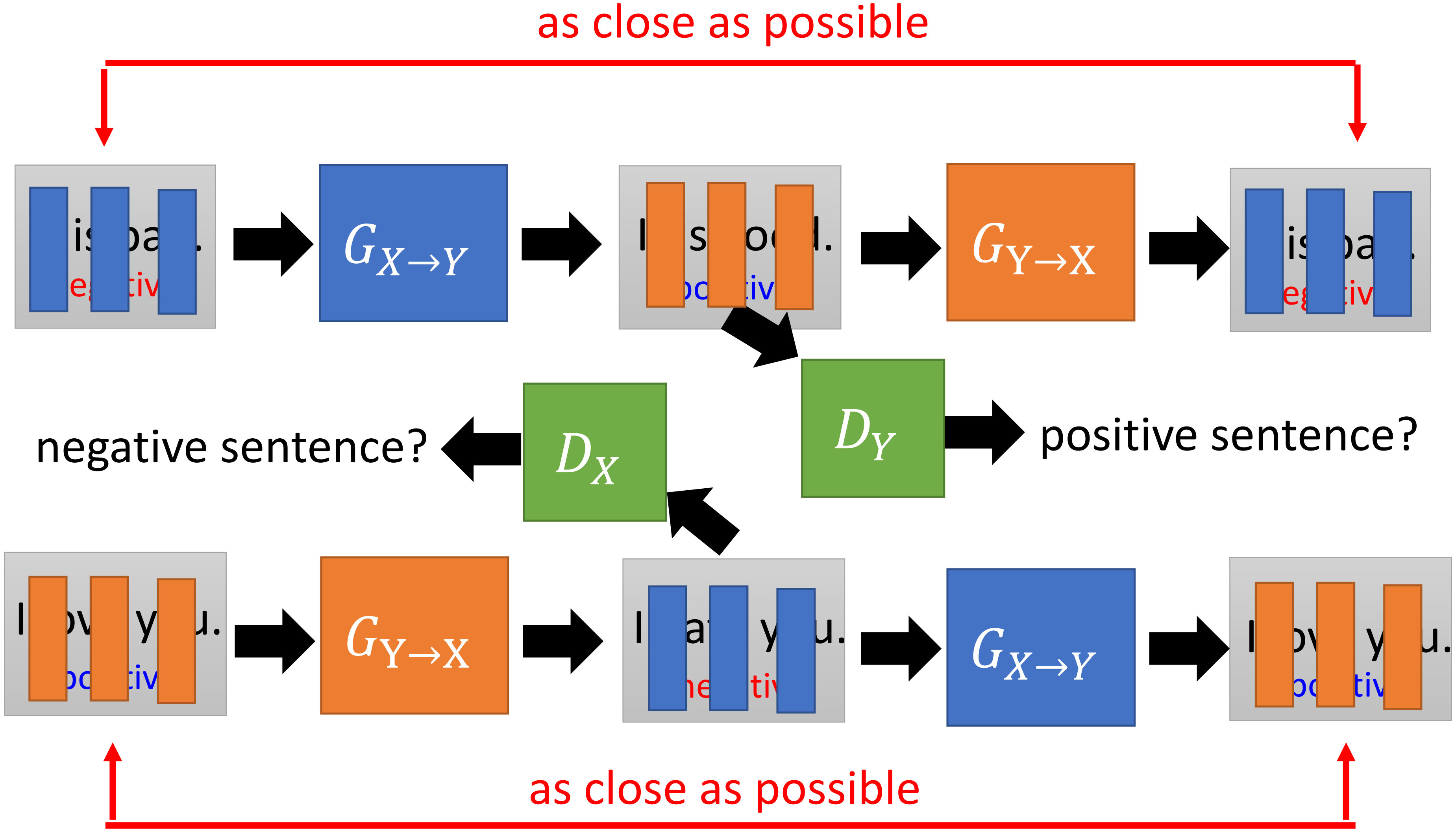
- [Yu, et al., AAAI, 2017][Li, et al., EMNLP, 2017][Tong Che, et al, arXiv, 2017][Jiaxian Guo, et al., AAAI, 2018][Kevin Lin, et al, NIPS, 2017][William Fedus, et al., ICLR, 2018]

Cycle GAN

Discrete?

Word embedding

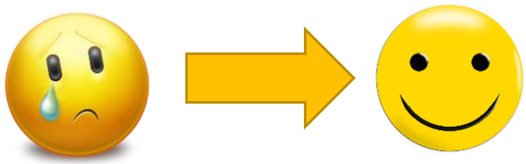
[Lee, et al., ICASSP, 2018]



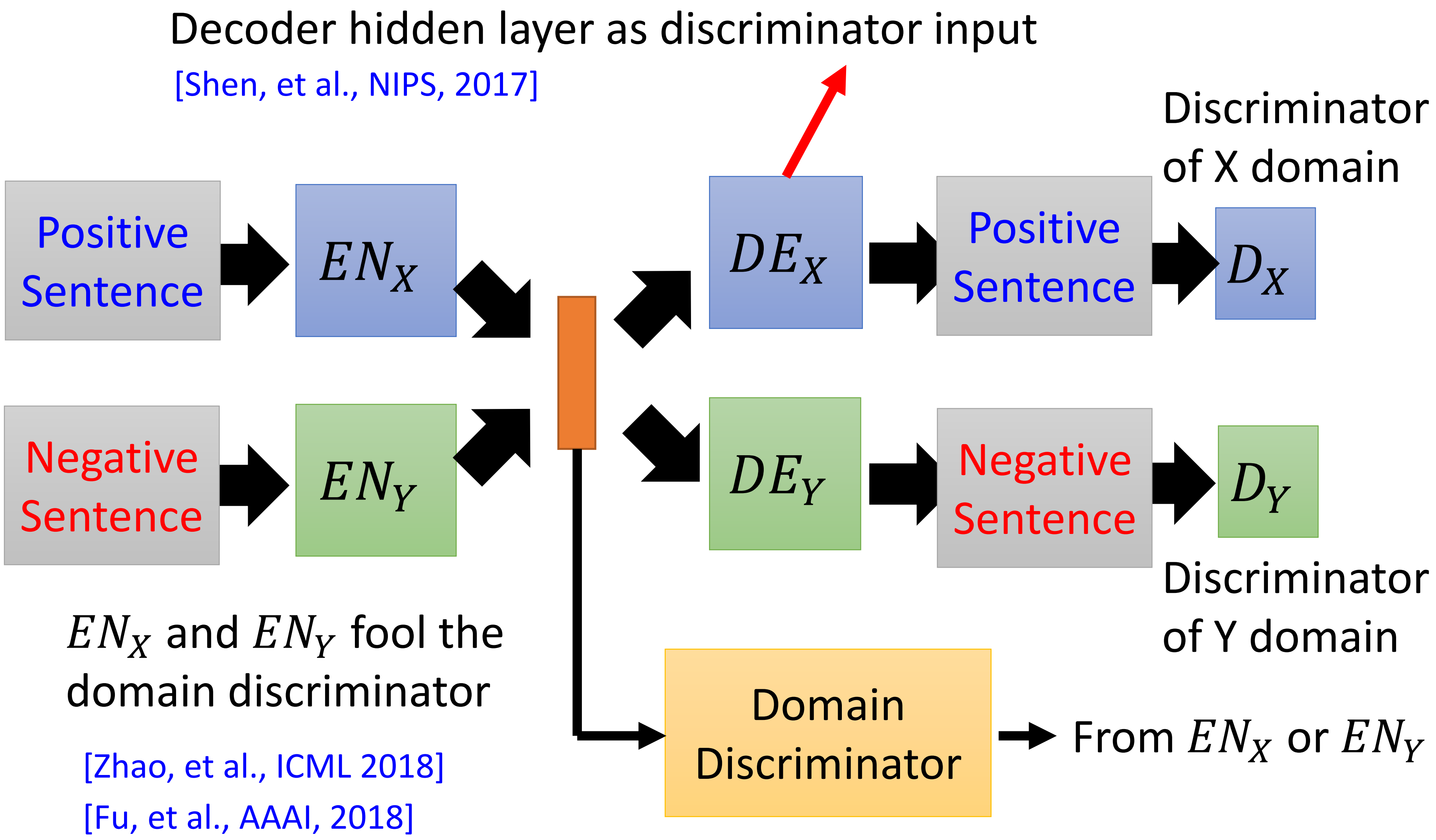
Thinks Yau-Shian Wang for providing the results.

Cycle GAN

- 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



Projection to Common Space



Unsupervised
Conditional Generation

Image Style Transfer

photos

Not Paired

Vincent van Gogh’s paintings

Text Style Transfer

document

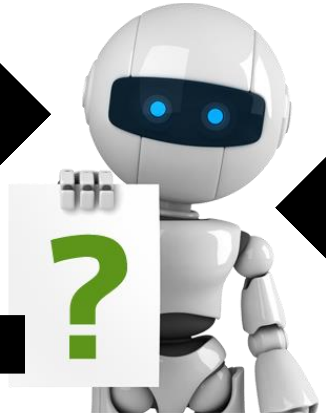
Not Paired

summary

This is **unsupervised abstractive summarization**.

Abstractive Summarization

- Now machine can do **abstractive summary** by seq2seq (write summaries in its own words)



seq2seq

summary
(in its own words)

summary 1

summary 2

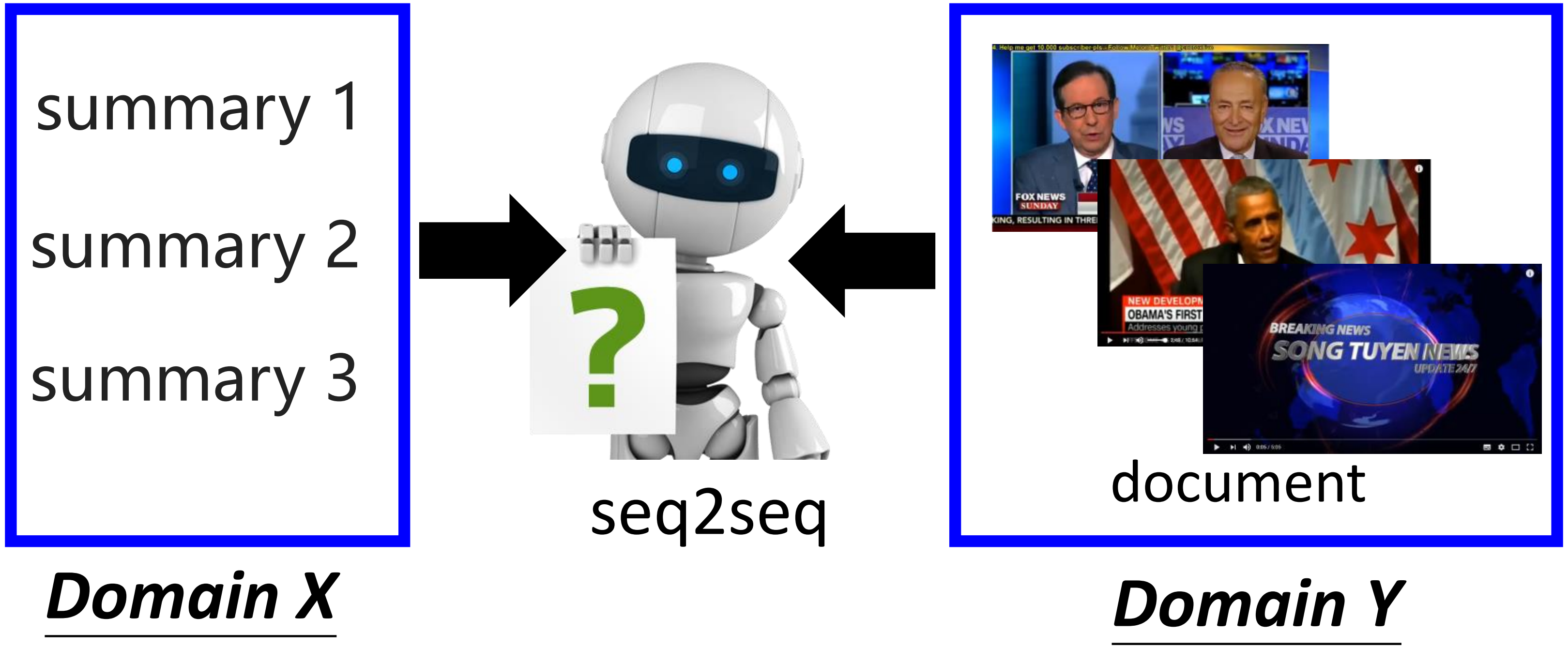
summary 3

Training Data

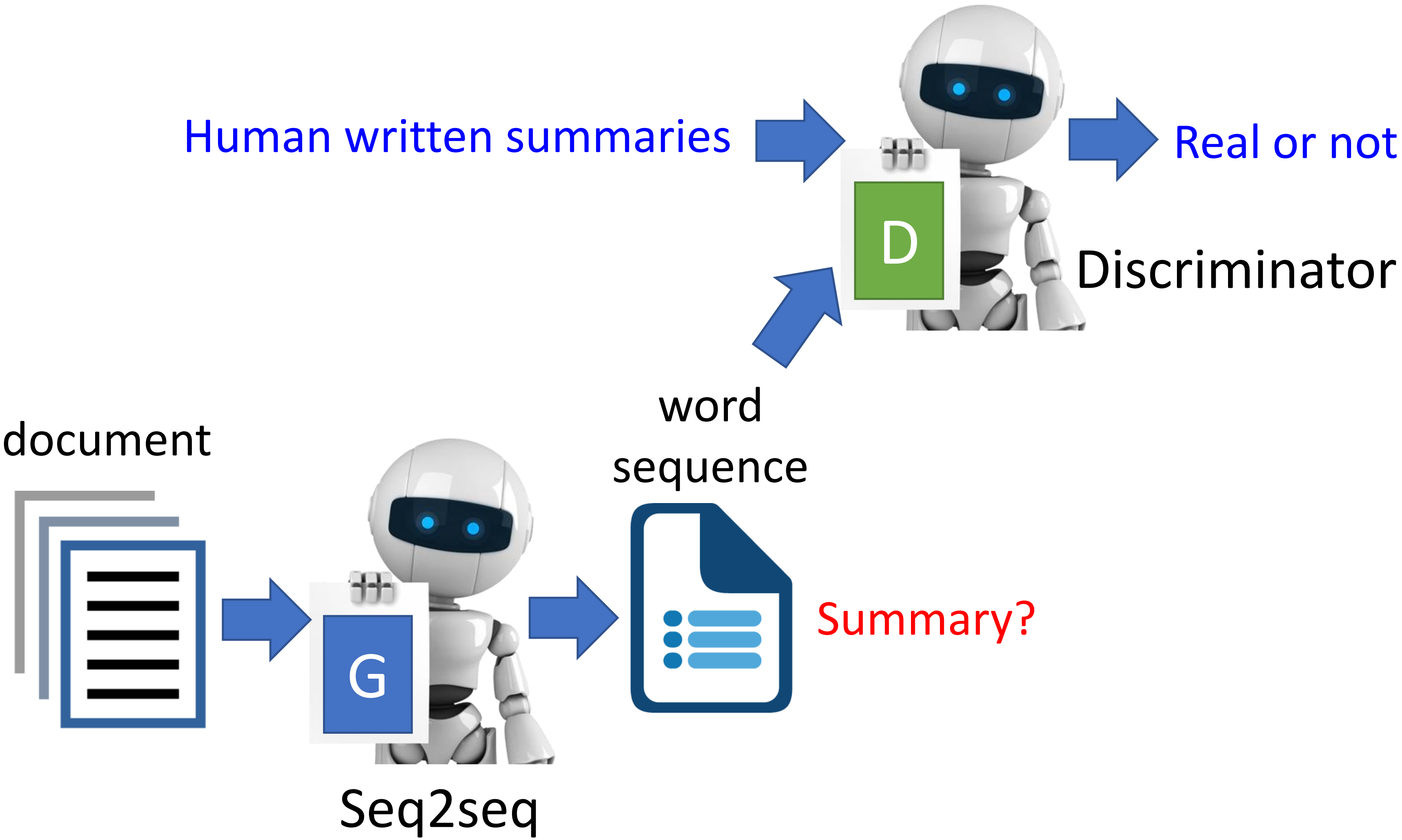
Supervised: We need lots of labelled training data.

Unsupervised Abstractive Summarization

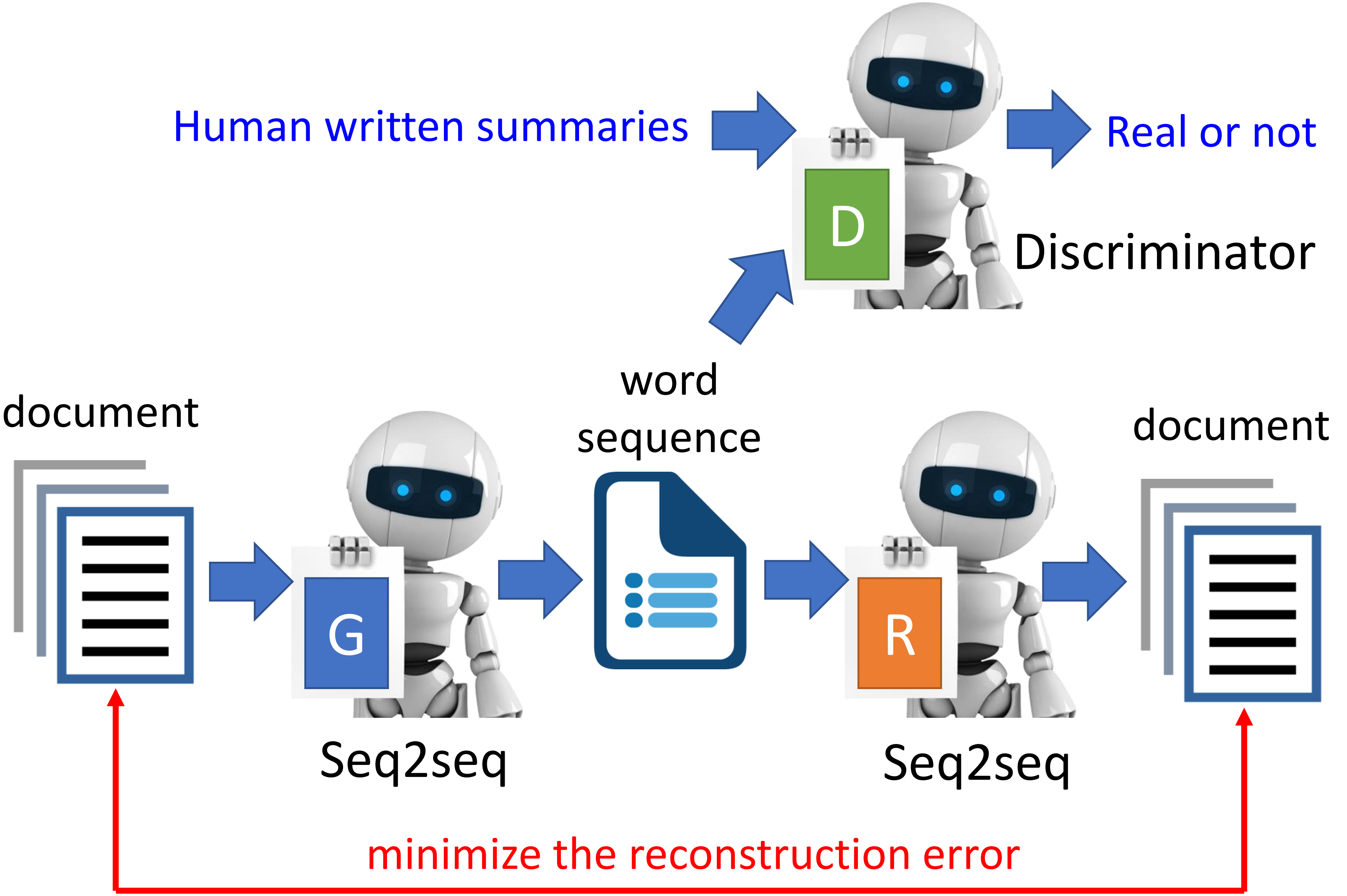
- Now machine can do **abstractive summary** by seq2seq (write summaries in its own words)



Unsupervised Abstractive Summarization



Unsupervised Abstractive Summarization

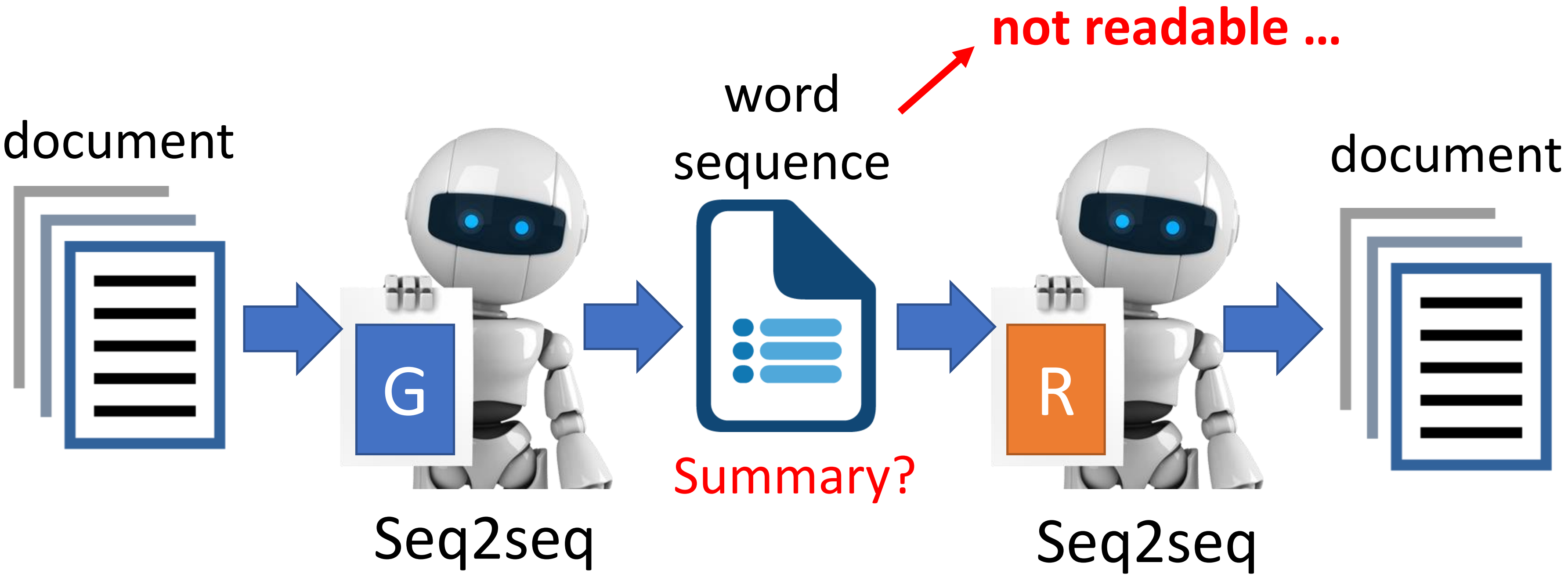


Unsupervised Abstractive Summarization

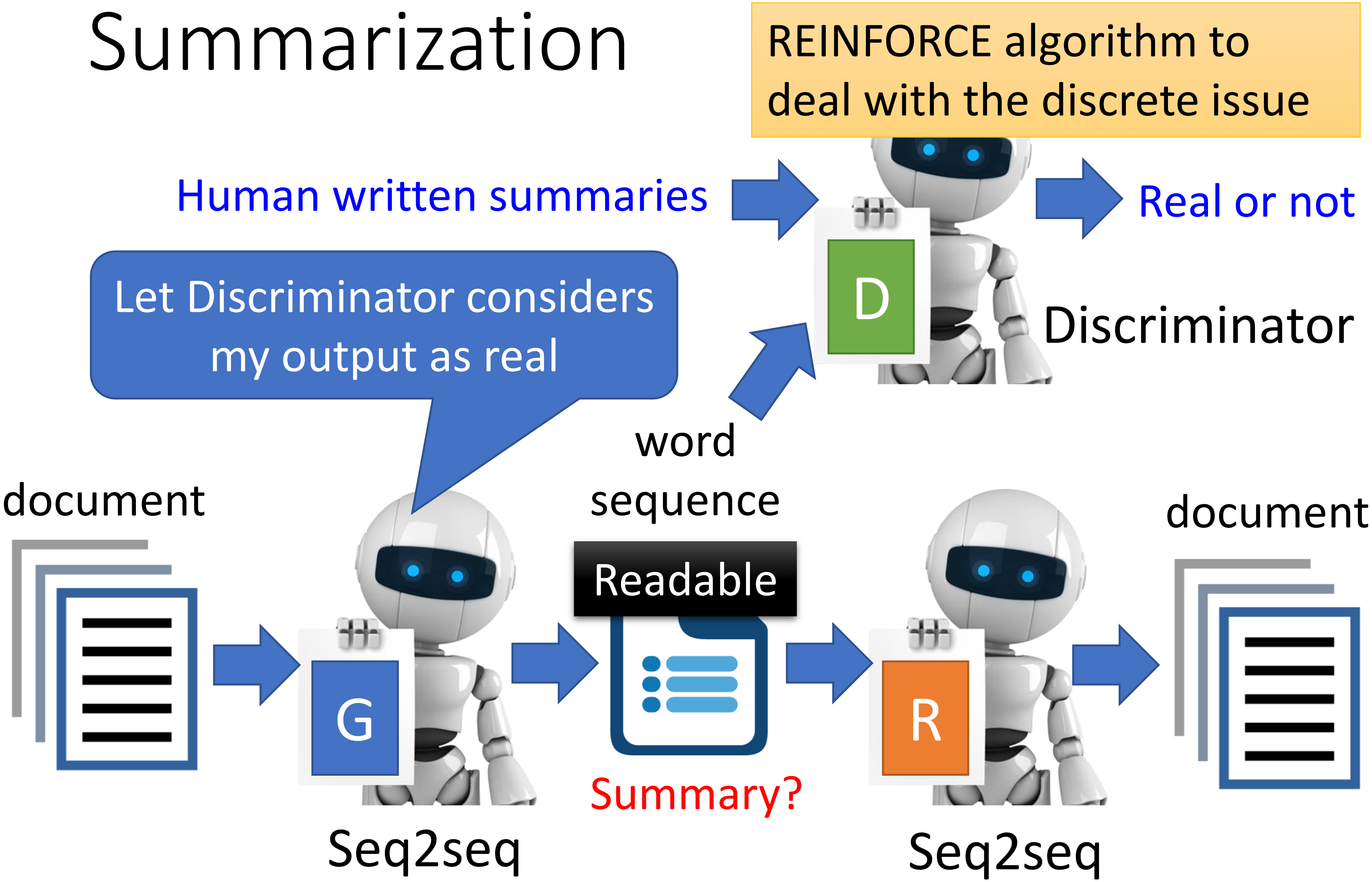
Only need a lot of documents to train the model

This is a ***seq2seq2seq auto-encoder***.

Using a sequence of words as latent representation.



Unsupervised Abstractive Summarization



Experimental results

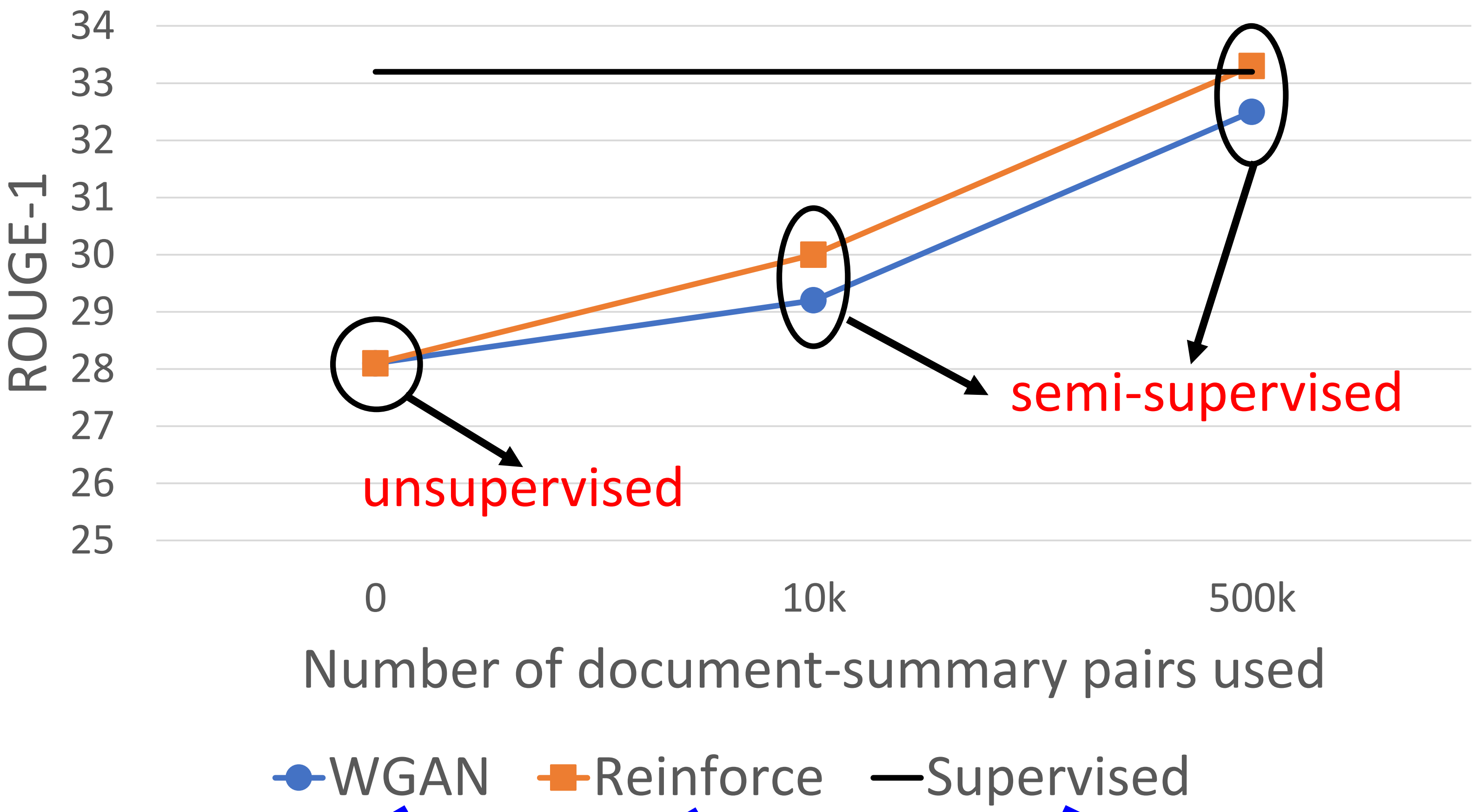
English Gigaword (Document title as summary)

	ROUGE-1	ROUGE-2	ROUGE-L
Supervised	33.2	14.2	30.5
Trivial	21.9	7.7	20.5
Unsupervised (matched data)	28.1	10.0	25.4
Unsupervised (no matched data)	27.2	9.1	24.1

- Matched data: using the title of English Gigaword to train Discriminator
- No matched data: using the title of CNN/Diary Mail to train Discriminator

Semi-supervised Learning

Using matched data



Approaches to deal with the discrete issue.

3.8M pairs are used.

Outline

National Taiwan University

Part I: General Introduction of Generative Adversarial Network (GAN)

Part II: Applications to Natural Language Processing

Part III: Applications to Speech Processing

Unsupervised
Conditional Generation

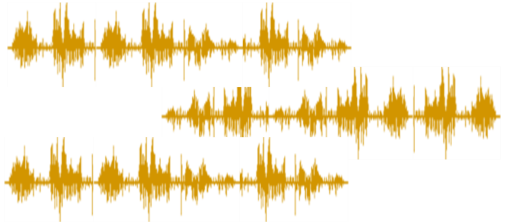
Image Style Transfer

Not Paired

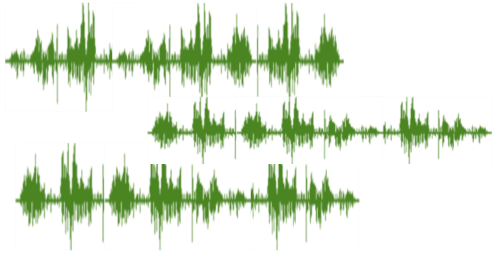
photos

Vincent van Gogh's
paintings

Speech Style Transfer



Not Paired



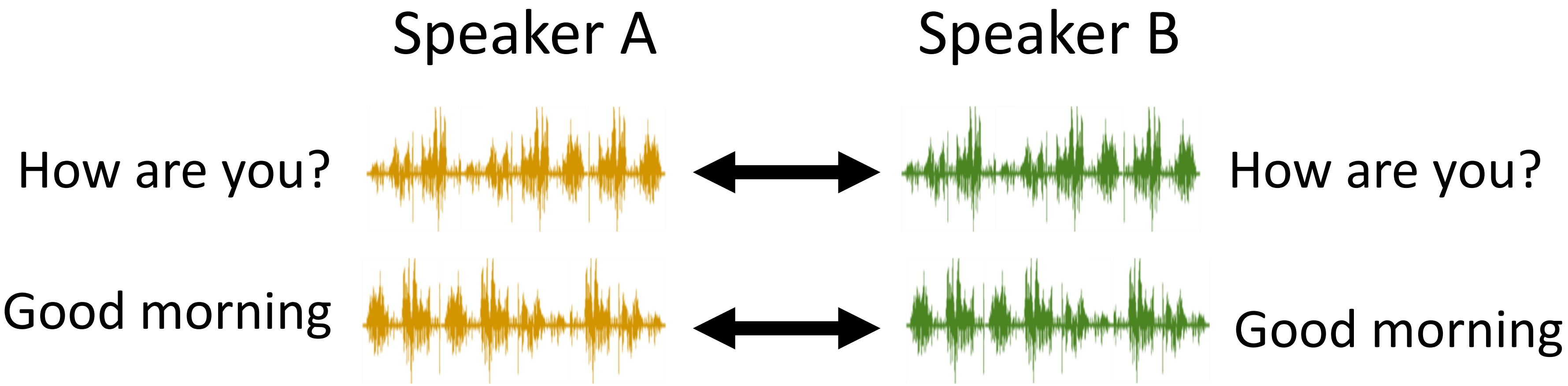
Speaker A

Speaker B

This is **unsupervised voice conversion**.

Voice Conversion

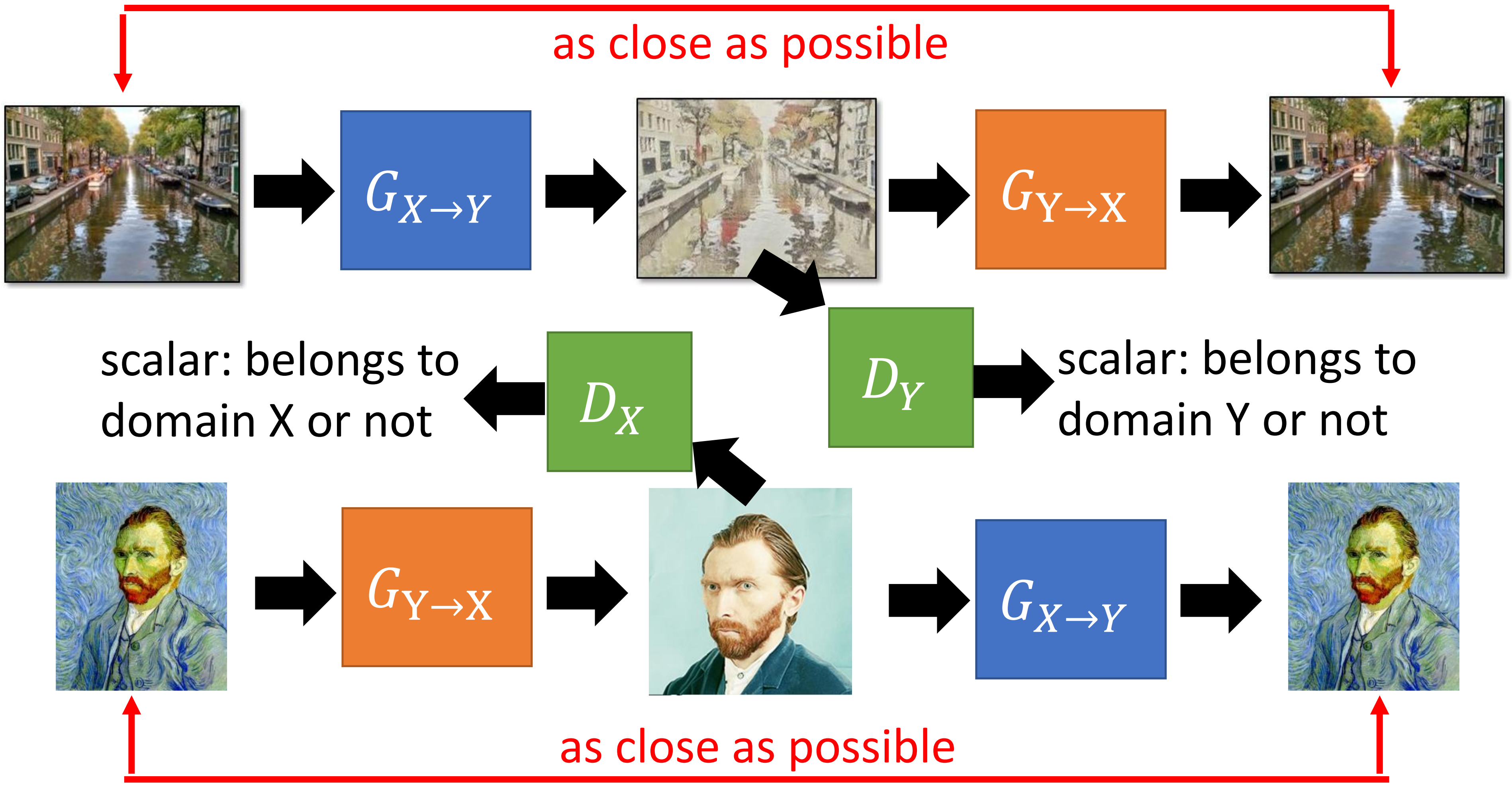
In the past



With GAN

Speakers A and B are talking about completely different things.

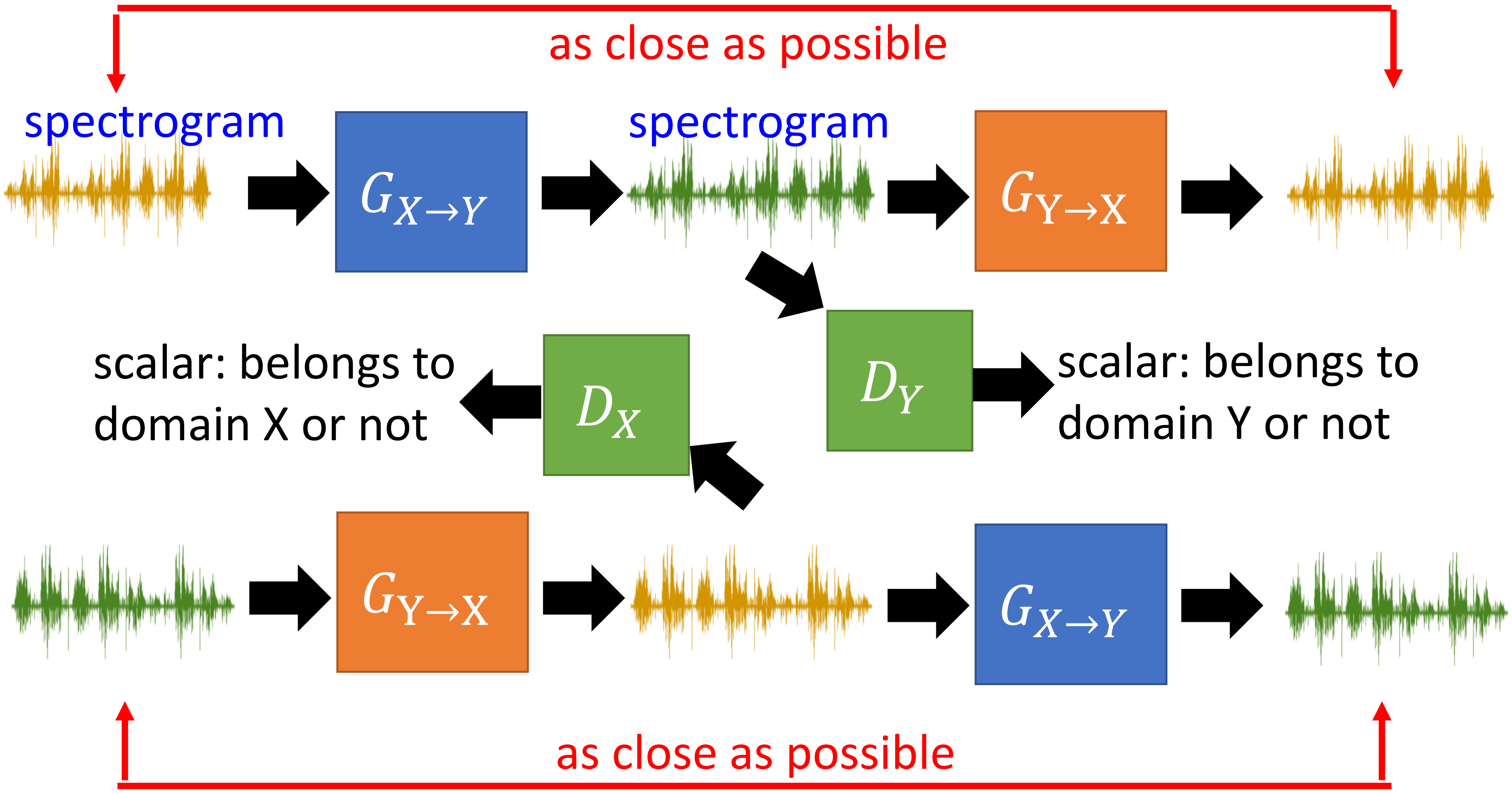
Cycle GAN


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

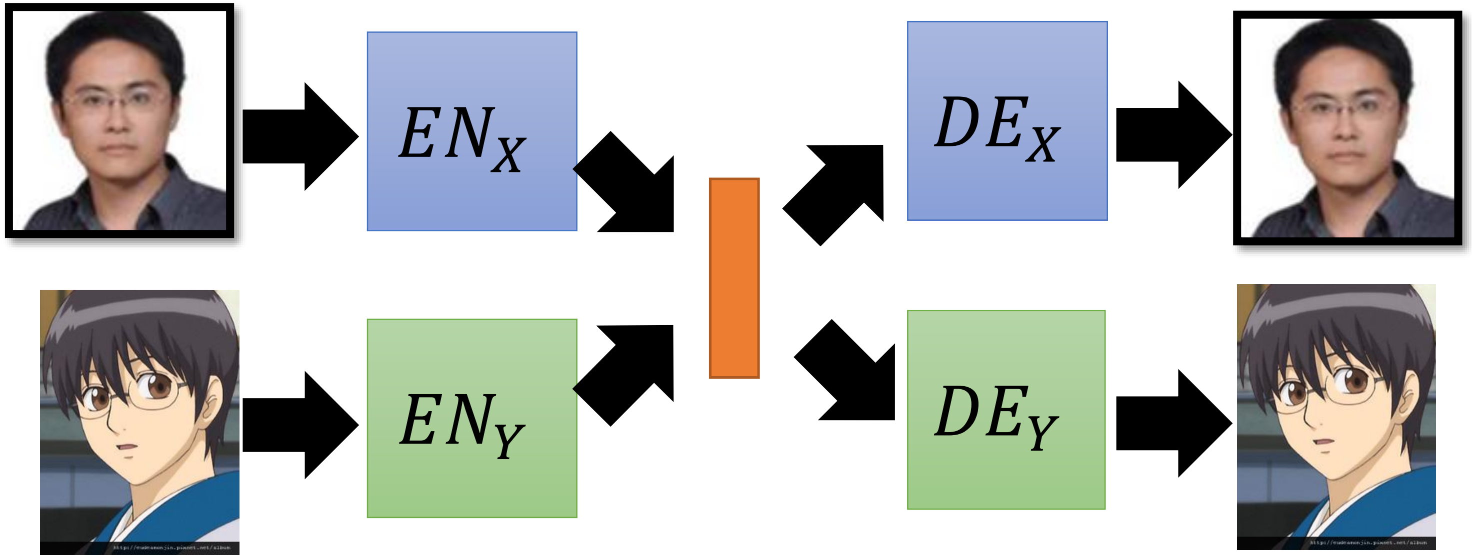
Cycle GAN for Voice Conversion

X: Speaker A, Y: Speaker B

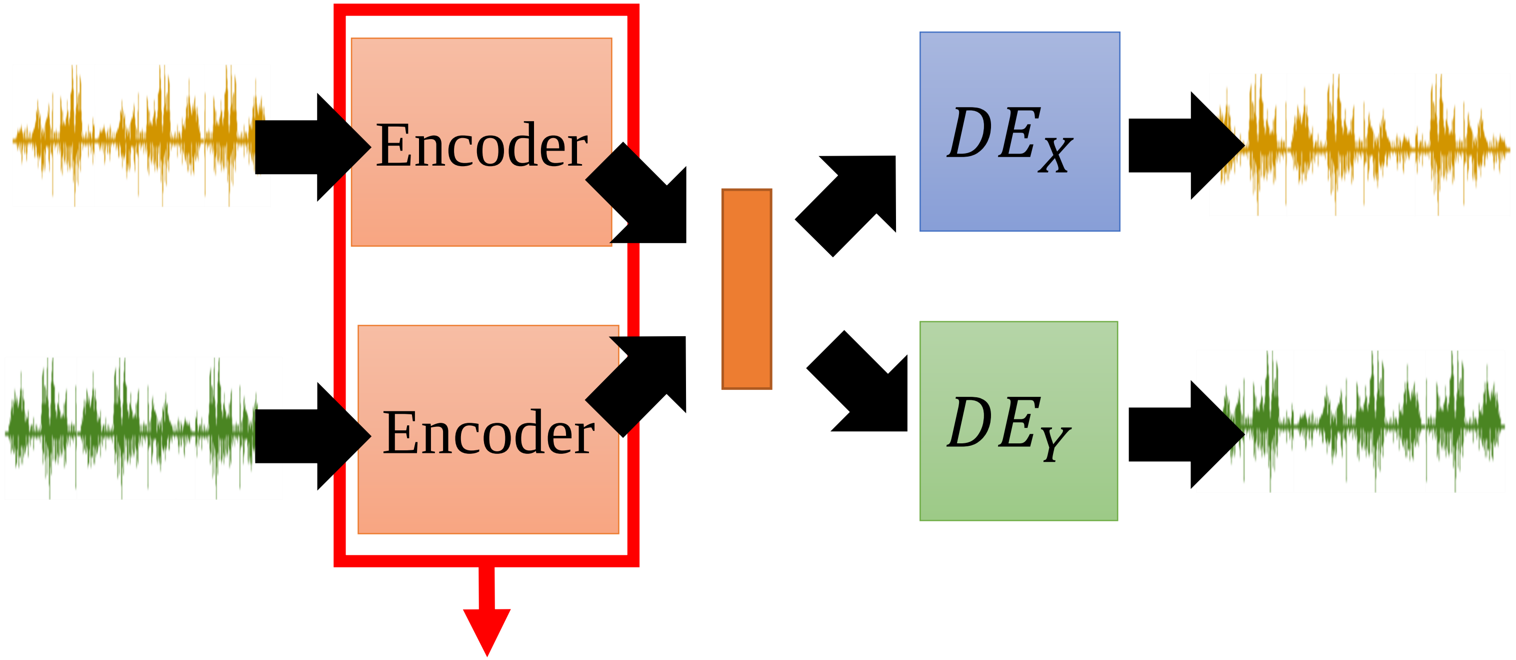
[Takuhiro Kaneko, et. al, arXiv, 2017][Fuming Fang, et. al, ICASSP, 2018][Yang Gao, et. al, ICASSP, 2018]



Projection to Common Space



Projection to Common Space

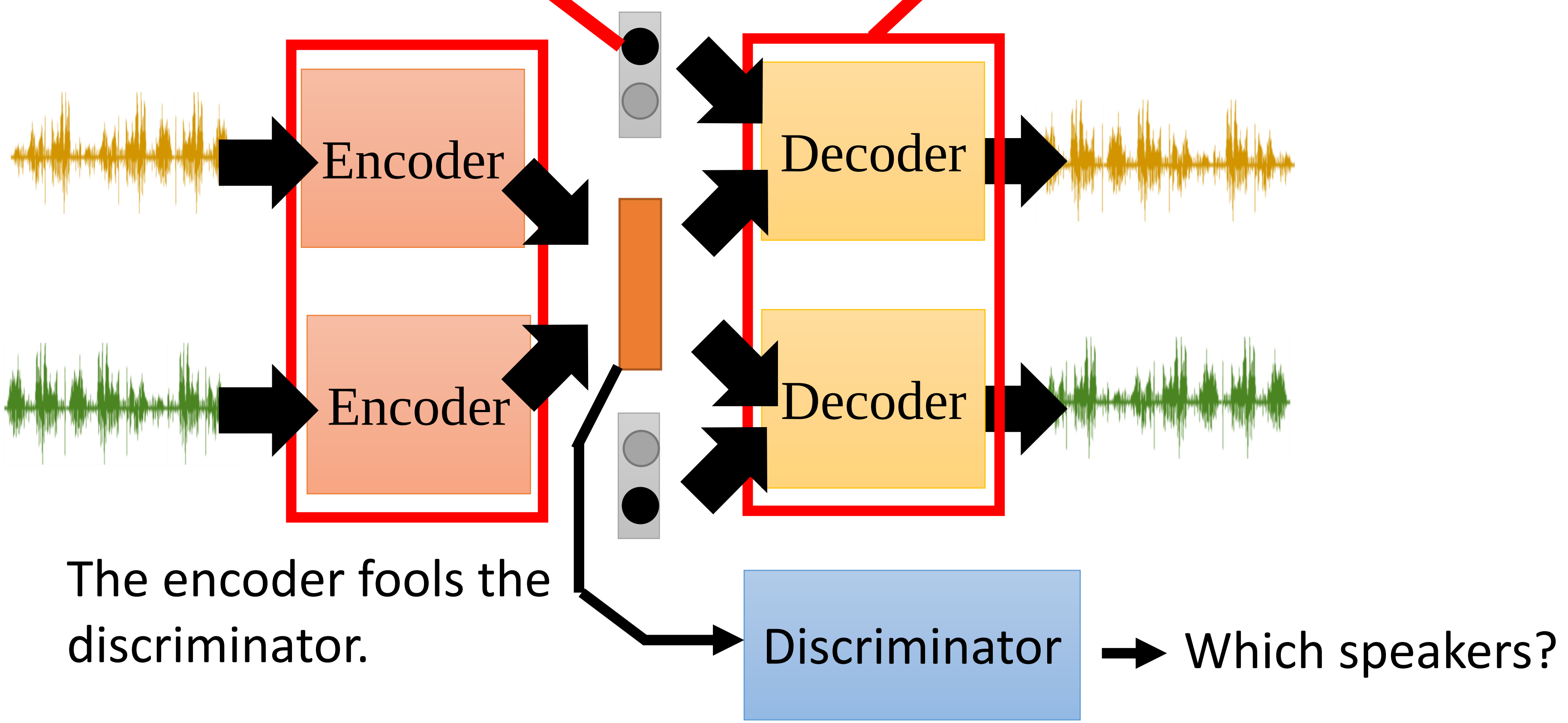


- All the speakers share the same encoder.
- The model can deal with the speakers never seen during training.

Projection to Common Space

Use a vector (one-hot) to represent speaker identity.

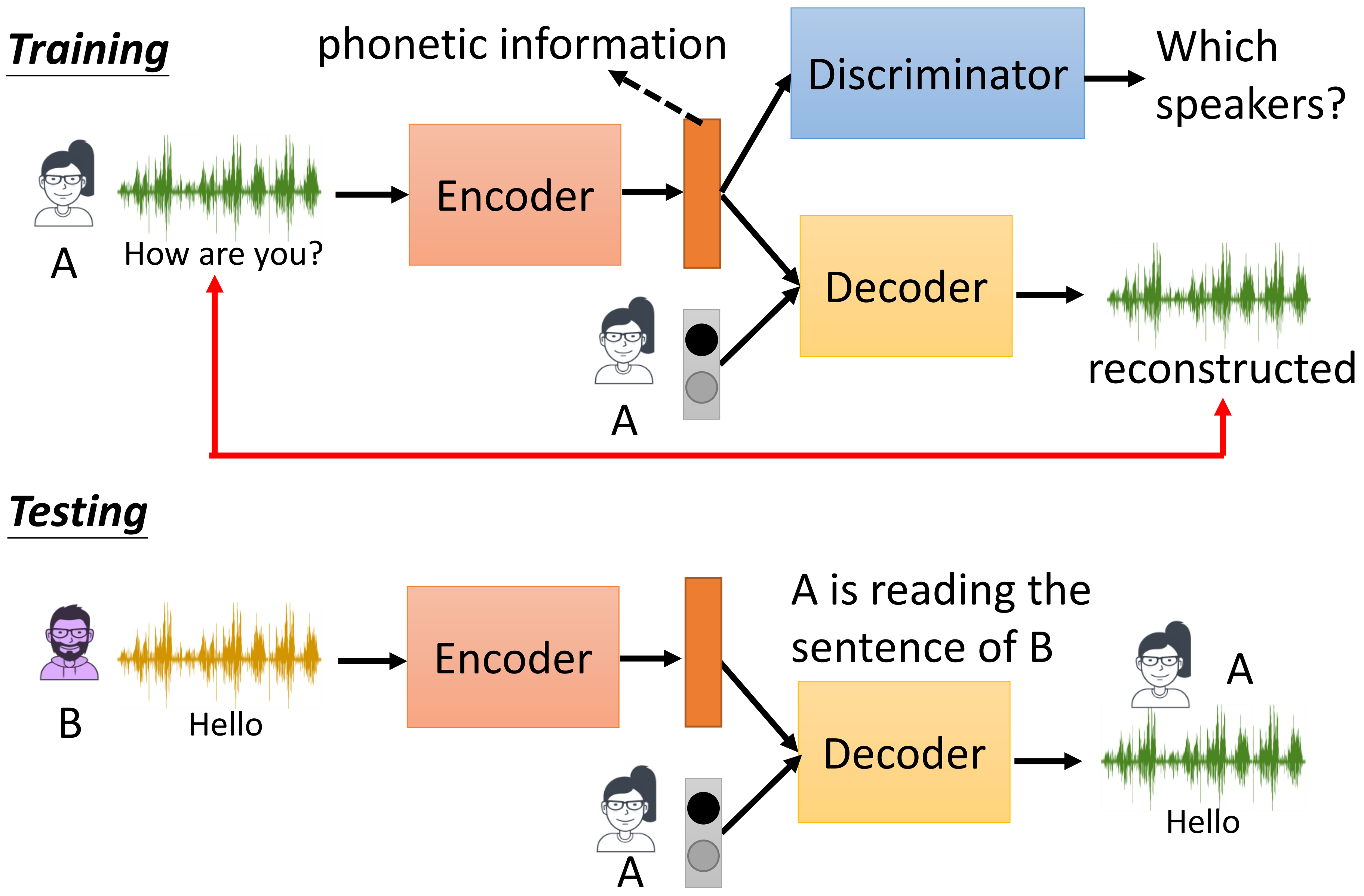
All the speakers also share the same decoder.

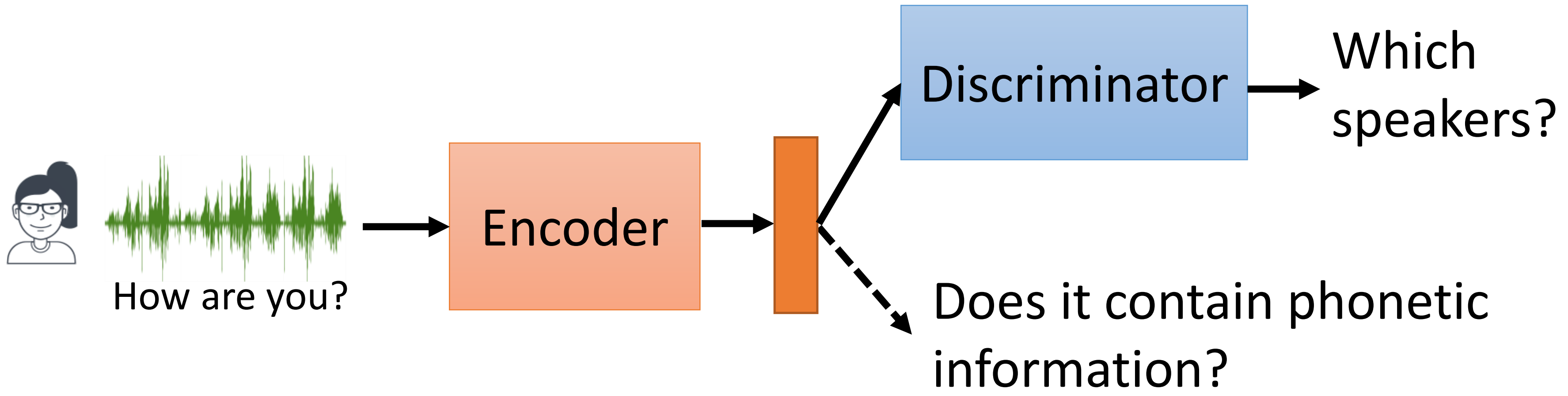


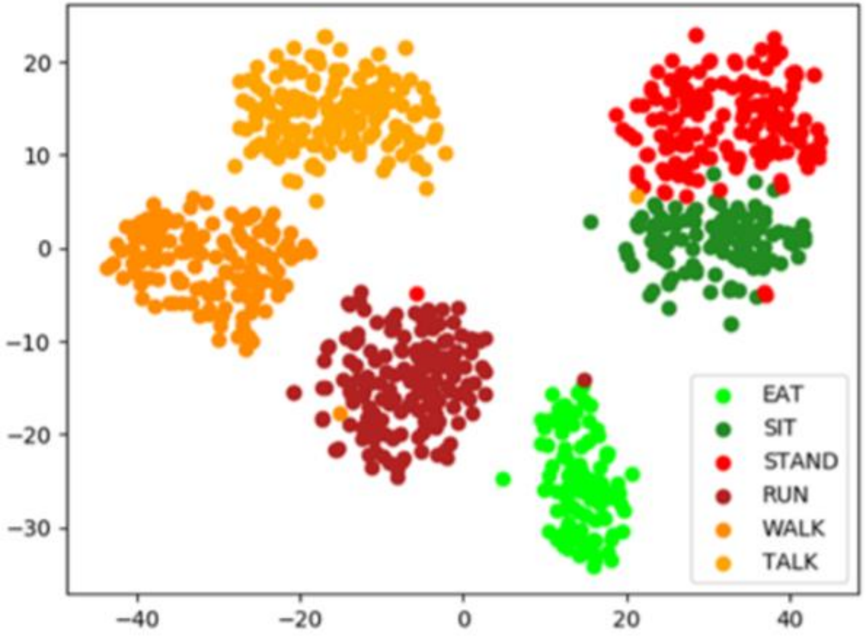
The encoder fools the discriminator.

We hope that encoder can extract the phonetic information while removing the speaker information.

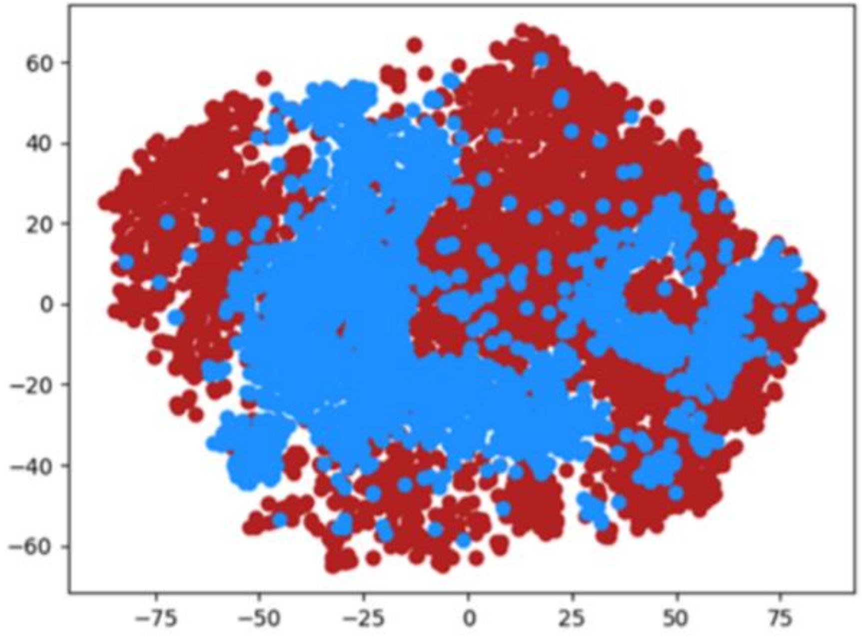
Projection to Common Space



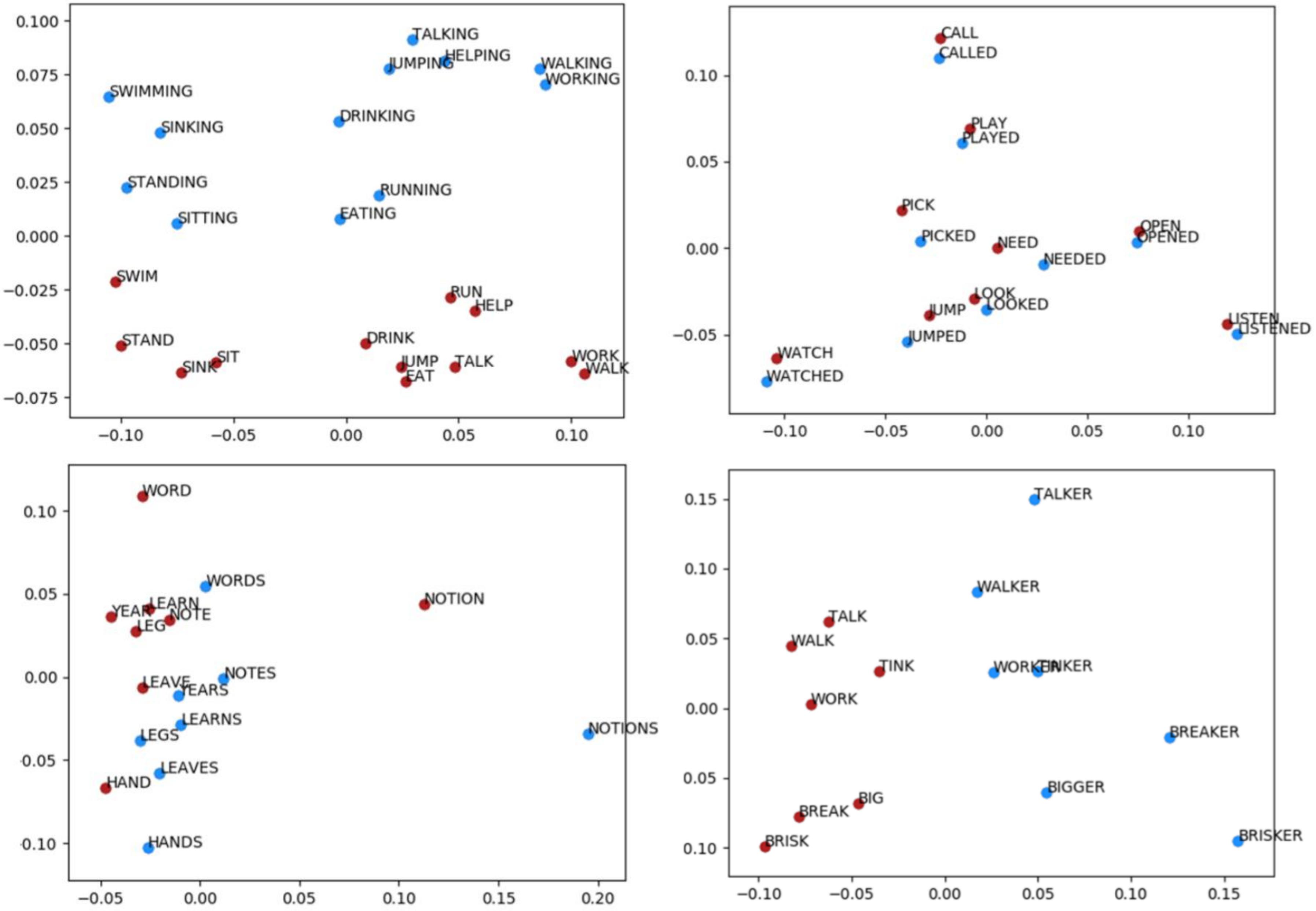




Different colors: different words

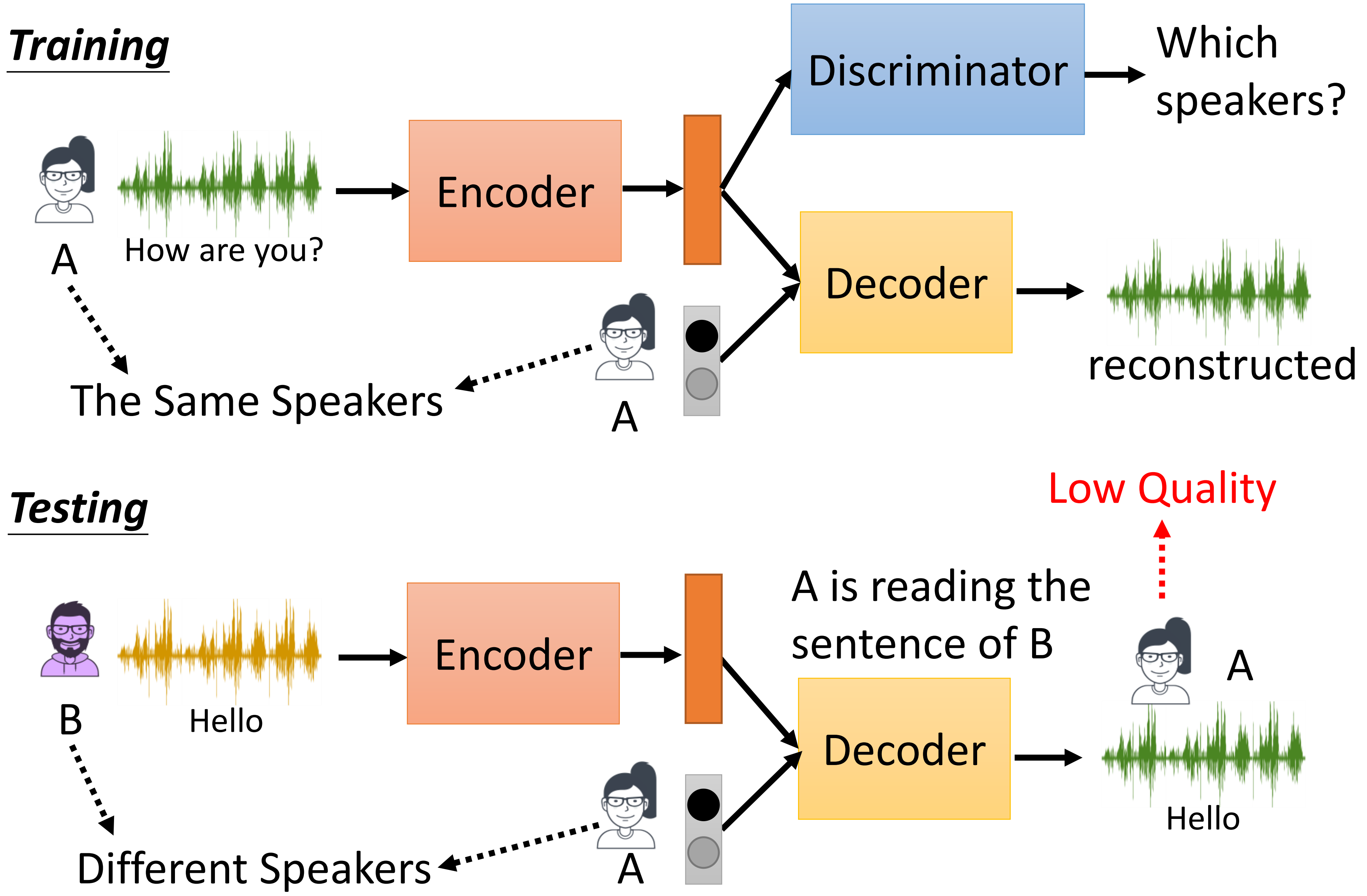


Different colors: different speakers



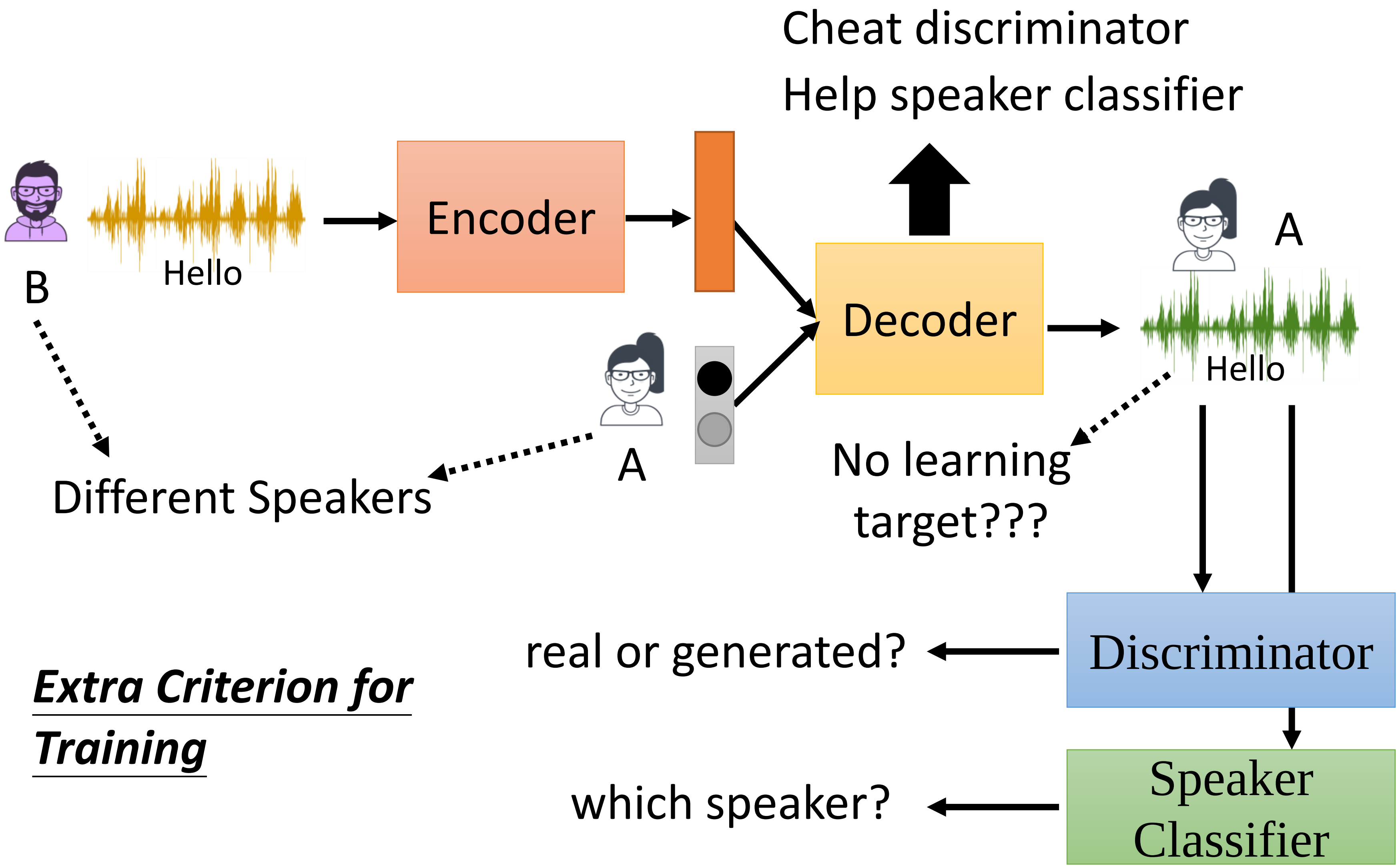
"Audio" Word to Vector

Issues



Training
Discriminator
Which speakers?
Encoder
Decoder
A
How are you?
The Same Speakers
A
reconstructed
Testing
Low Quality
A is reading the sentence of B
Encoder
Decoder
B
Hello
A
Different Speakers
A
Hello

2nd Stage Training

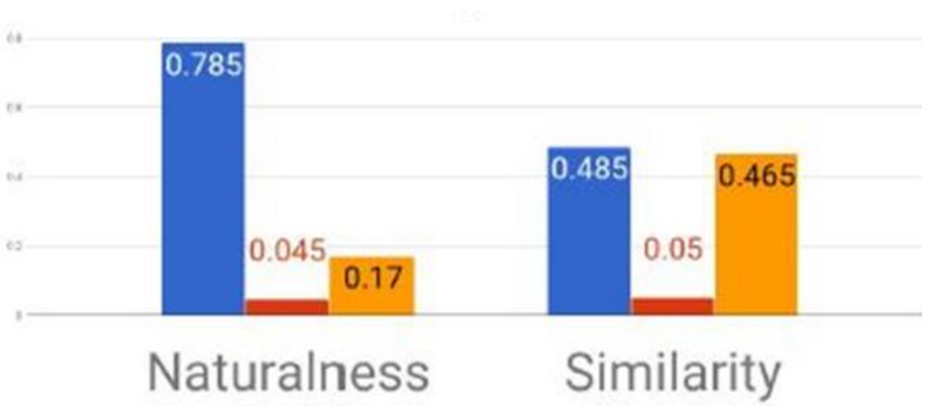


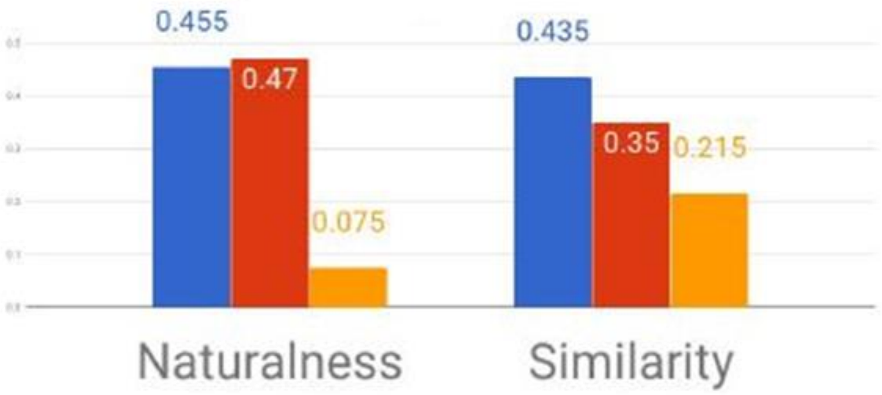
Extra Criterion for Training

[Chou et al., INTERSPEECH, 2018]

Experimental Results

- Subjective evaluations(20 speakers in VCTK)

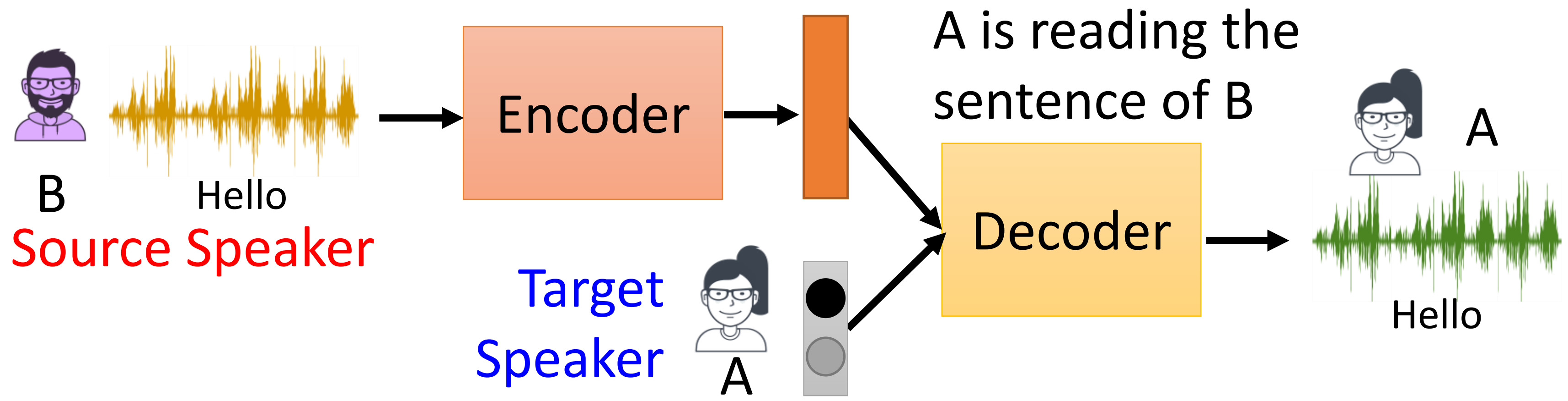




“Two stages” is better
“One stage” is better
Indistinguishable

“Projection” is better
“Cycle GAN” is better
Indistinguishable

Demo



Source:

Target:

Source to Target:

Thanks Ju-chieh Chou for providing the results.

https://jjery2243542.github.io/voice_conversion_demo/

Target Speaker

Source Speaker
(Never seen during training!)

Source to Target

Me

Me

Me

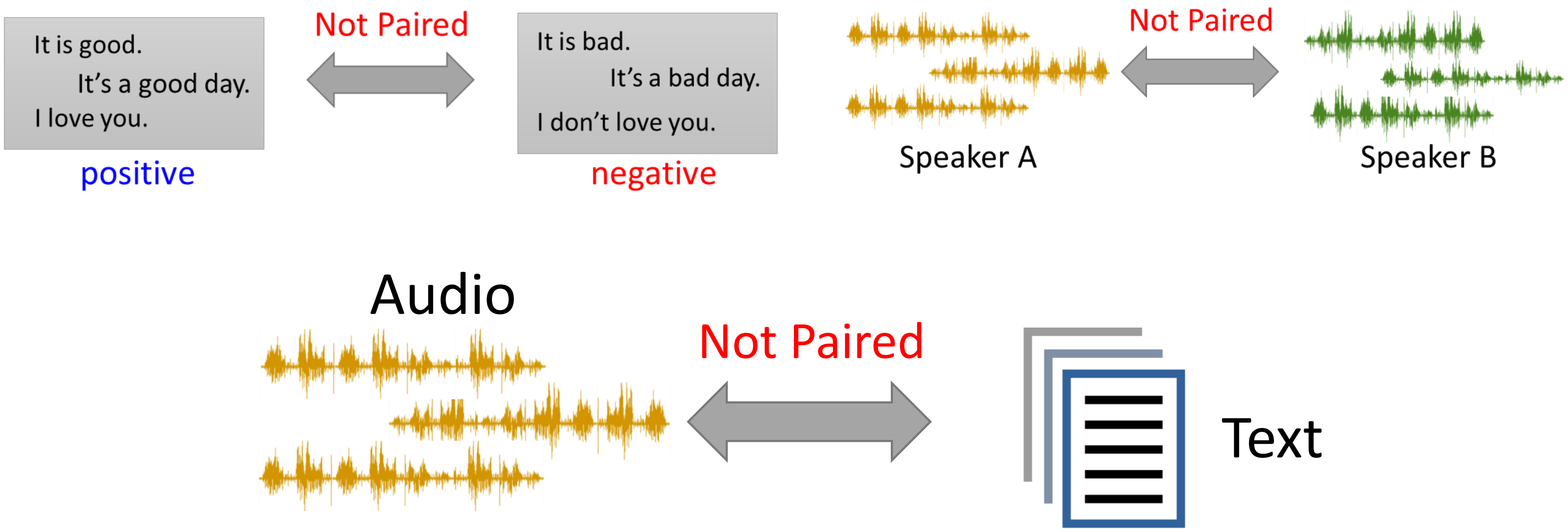
Me

(doesn't work.
Just for fun)

Thanks Ju-chieh Chou for providing the results.

https://jjery2243542.github.io/voice_conversion_demo/

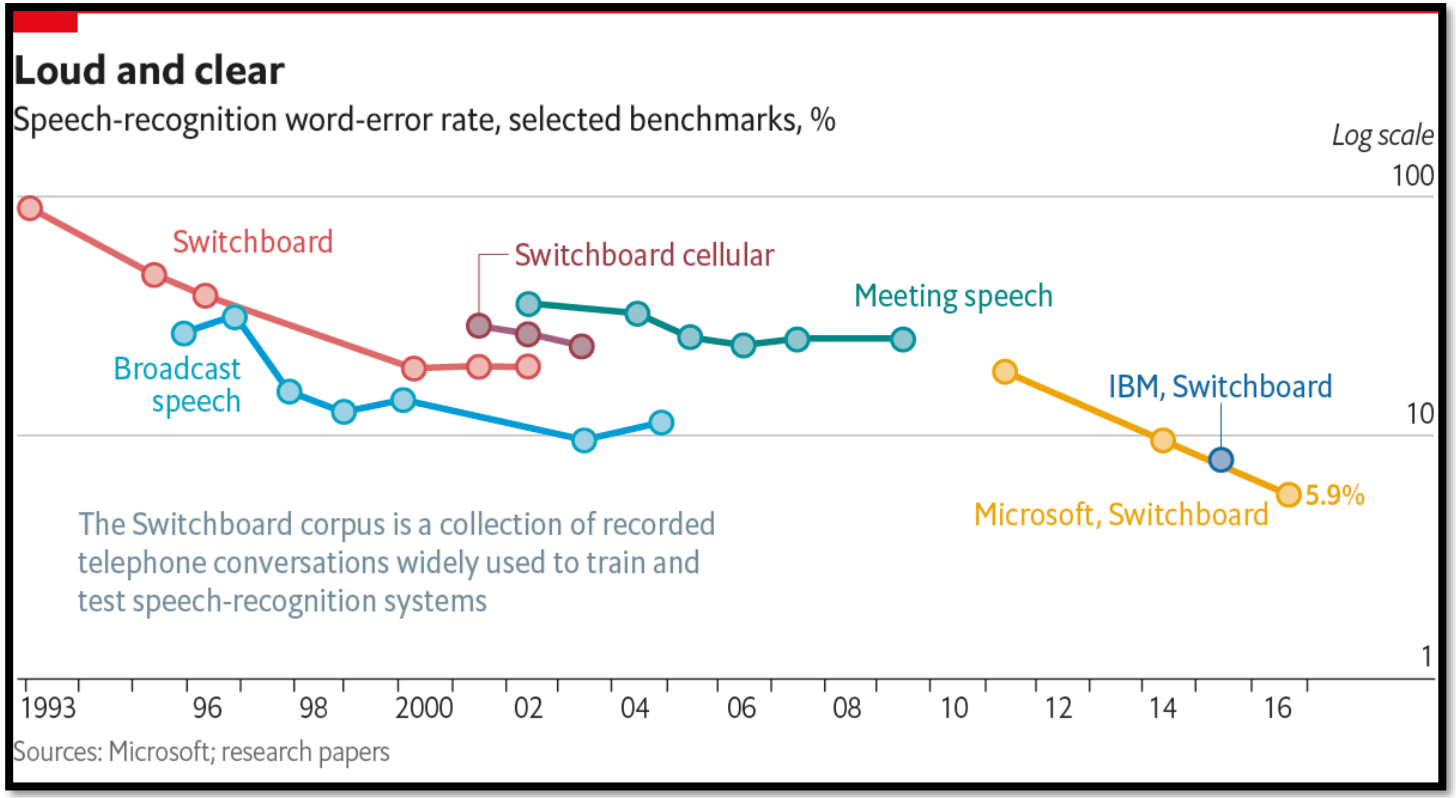
Unsupervised Conditional Generation



This is **unsupervised speech recognition**.

https://devopedia.org/images/article/102/9180.1532710057.png

Supervised Speech Recognition



(I believe you have seen similar figures before.)

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

http://www.parenting.com/article/teach-baby-to-talk

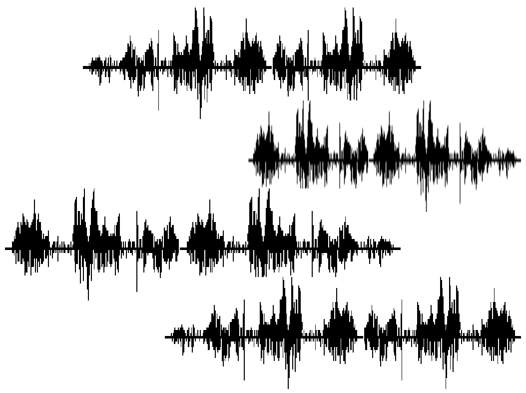
Speech Recognition in the Future

Learning human language with very little supervision

[Liu, et al., INTERSPEECH, 2018]
[Chen, et al., arXiv, 2018]

Unsupervised Speech Recognition

- Machine learns to recognize speech from unparallel speech and text.



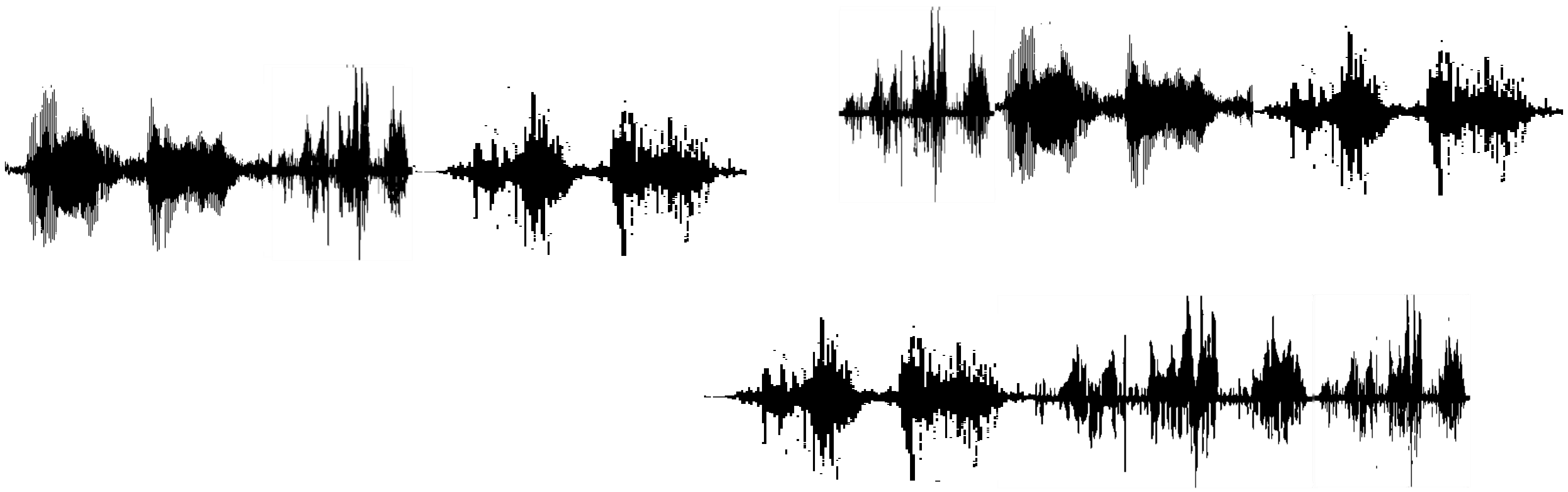
audio collection
(without text annotation)

text documents
(not parallel to audio)

This idea was too crazy to be realized in the past.

However, it becomes possible with GAN recently.

Acoustic Token Discovery

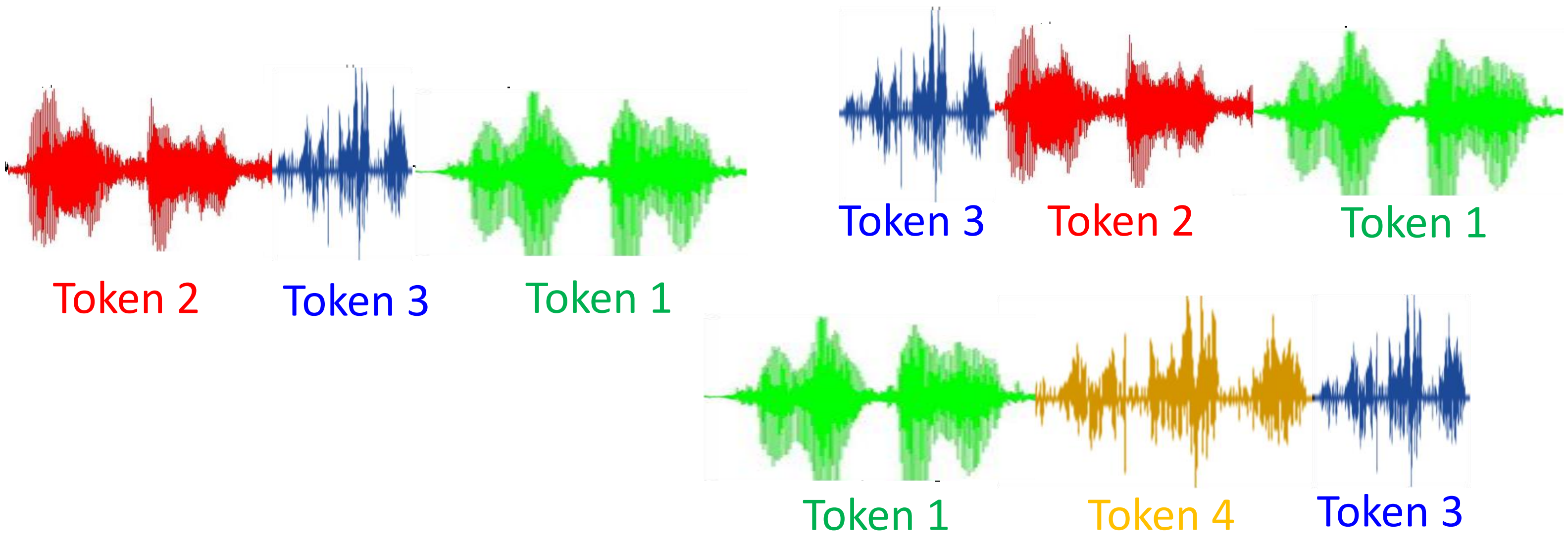


Acoustic tokens can be discovered from audio collection without text annotation.

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



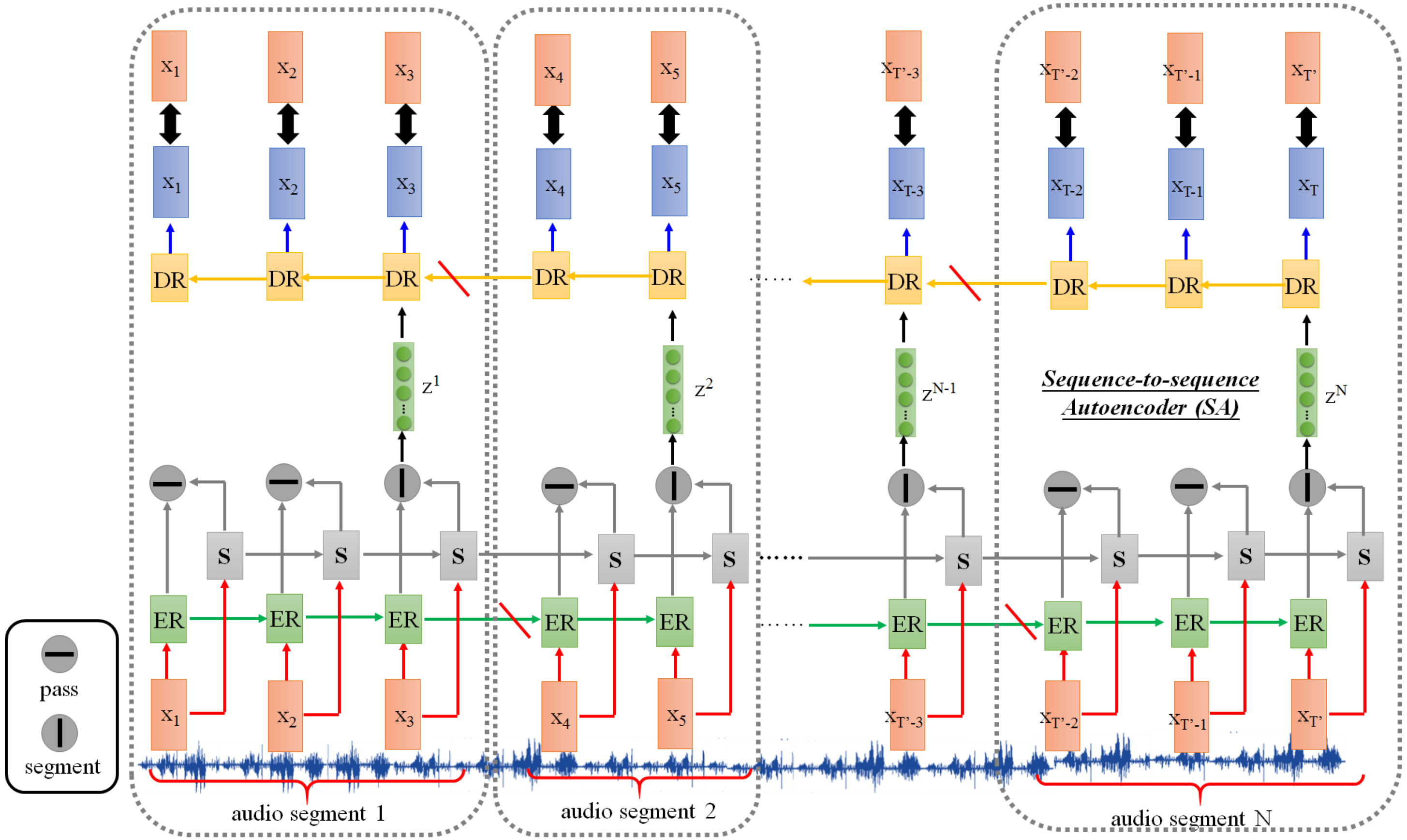
Acoustic tokens can be discovered from audio collection without text annotation.

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

p_1 p_2 p_3 p_4

p_1 p_3 p_2

p_1 p_4 p_3 p_5 p_5

p_1 p_5 p_4 p_3

Phone-level Acoustic Pattern Discovery

Cycle GAN

AY L AH V Y UW

G UH D B AY

HH AW AA R Y UW

AY M F AY N

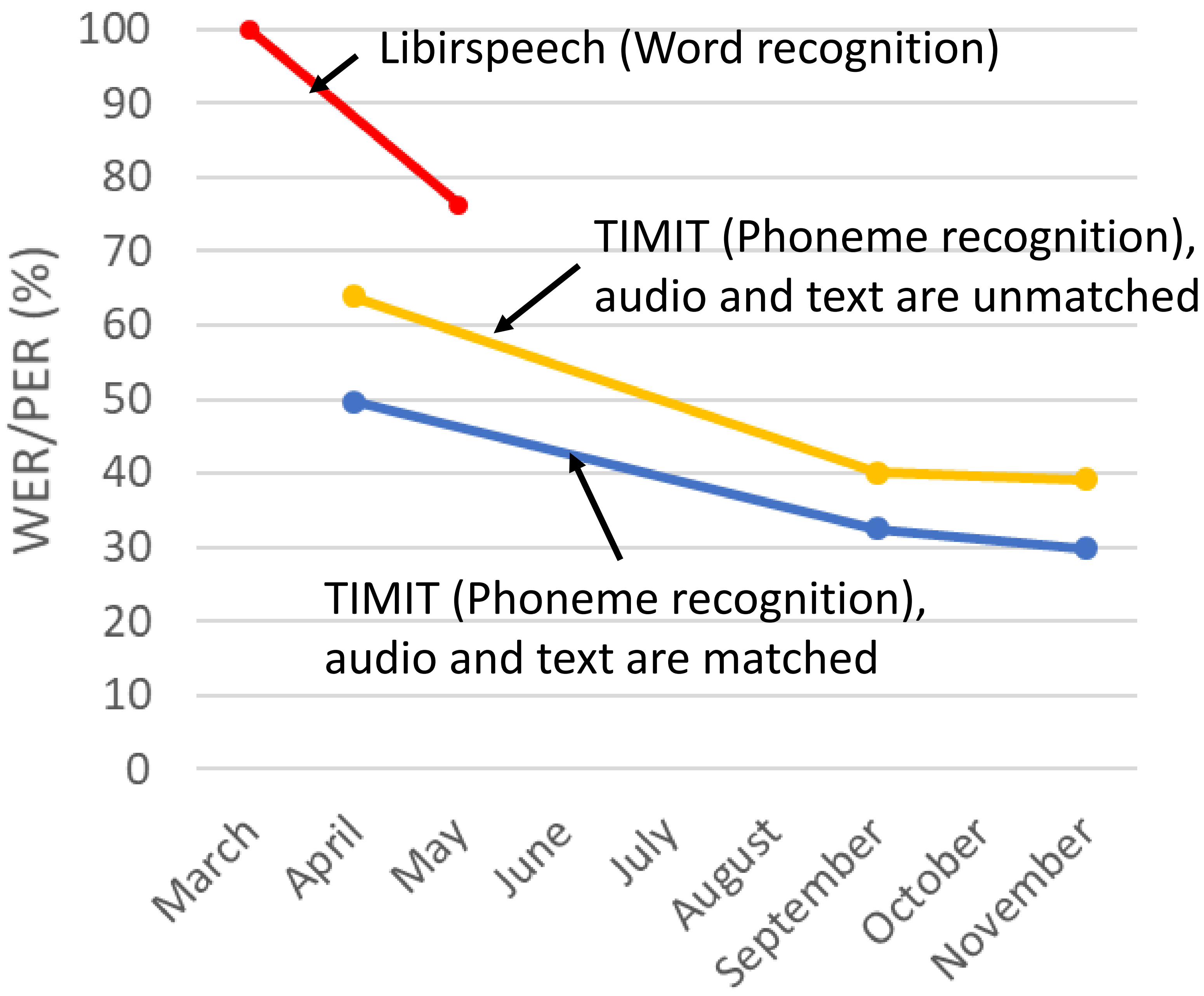
T AY W AA N

Phoneme sequences from Text

p_1 = "AY"

[Liu, et al., INTERSPEECH, 2018]

[Chen, et al., arXiv, 2018]



Libirspeech (Word recognition)
TIMIT (Phoneme recognition), audio and text are unmatched
TIMIT (Phoneme recognition), audio and text are matched
WER/PER (%)
100
90
80
70
60
50
40
30
20
10
0
March
April
May
June
July
August
September
October
November

Concluding Remarks

Part I: General Introduction of Generative Adversarial Network (GAN)

Part II: Applications to Natural Language Processing

Part III: Applications to Speech Processing

To Learn More …..

(My YouTube Channel, 30K subscribers, 2.4M total views)