# Computer Vision

## Conditional Generative Model

Tae-Hyun Oh (오태현)
전자전기공학과
POSTECH

Slide by Jongha Kim (김종하)
TAs: {Dongmin Choi , Jongha Kim, Juyong Lee, Sungbin Kim} (in alphabetic order)

boostcamp AI Tech

# 1. Conditional generative model

1.1 Conditional generative model
1.2 Conditional GAN and image translation
1.3 Example: Super resolution

# 2. Image translation GANs

2.1 Pix2Pix
2.2 CycleGAN
2.3 Perceptual loss

# 3. Various GAN applications

# 1.

# Conditional generative model

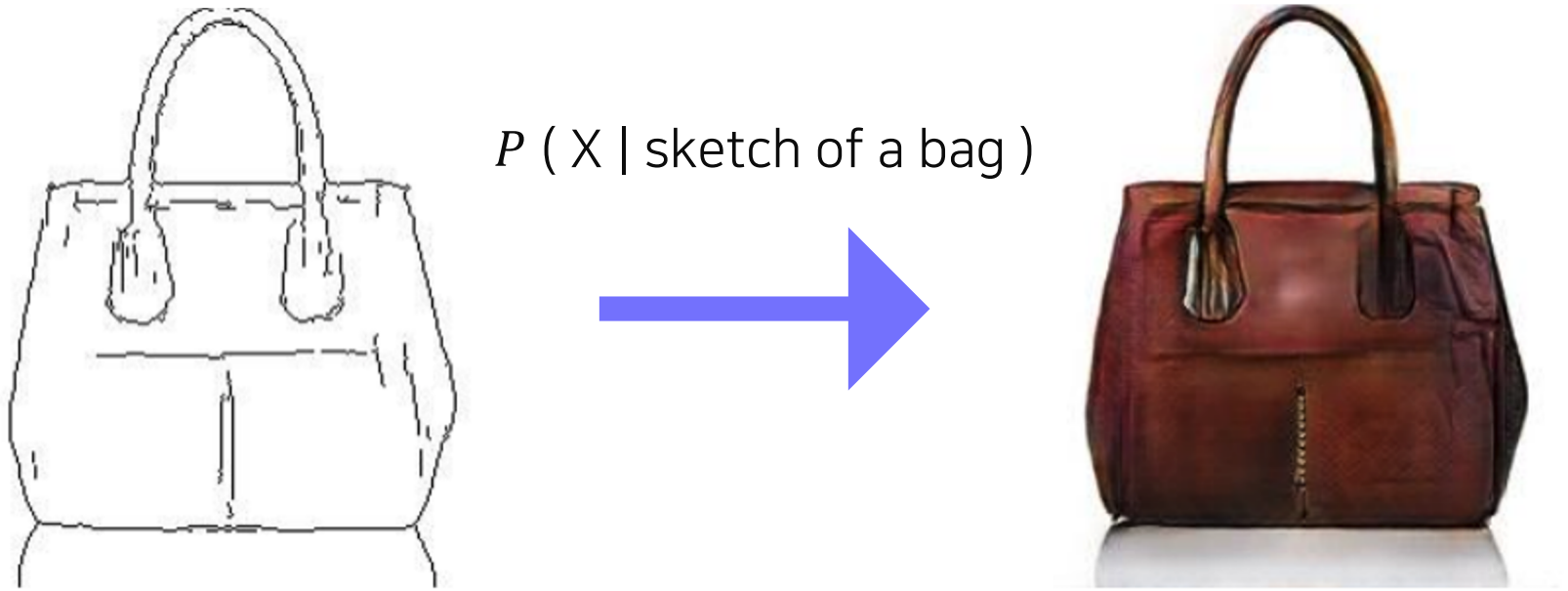
# 1.1 Conditional generative model



## Translating an image given "condition"

[Isola et al., CVPR 2017]

- We can explicitly generate an image corresponding to a given "condition"'!



# 1.1 Conditional generative model



## Generative model vs. Conditional generative model

[Isola et al., CVPR 2017]

- Generative model generates a random sample
- Conditional generative model generates a random sample under the given "condition"

### Generative Model

$P$ ( X =  )

Generating random bag images

### Conditional generative model

$P$ ( X =  | sketch =  )

Generating random bag images given a sketch

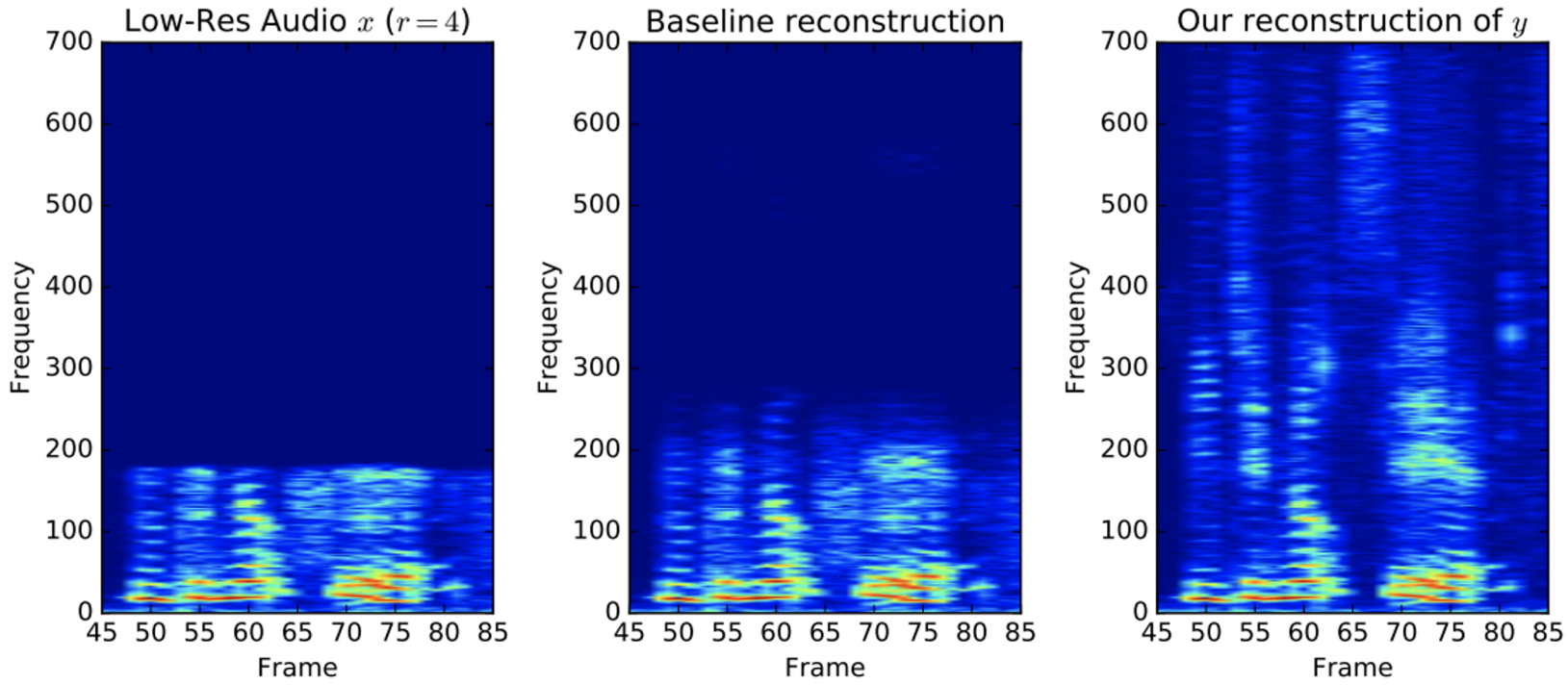
# 1.1 Conditional generative model



## Example of conditional generative model – audio super resolution

[Kuleshov et al., ICLR 2017]

- $P$(high resolution audio | low resolution audio)

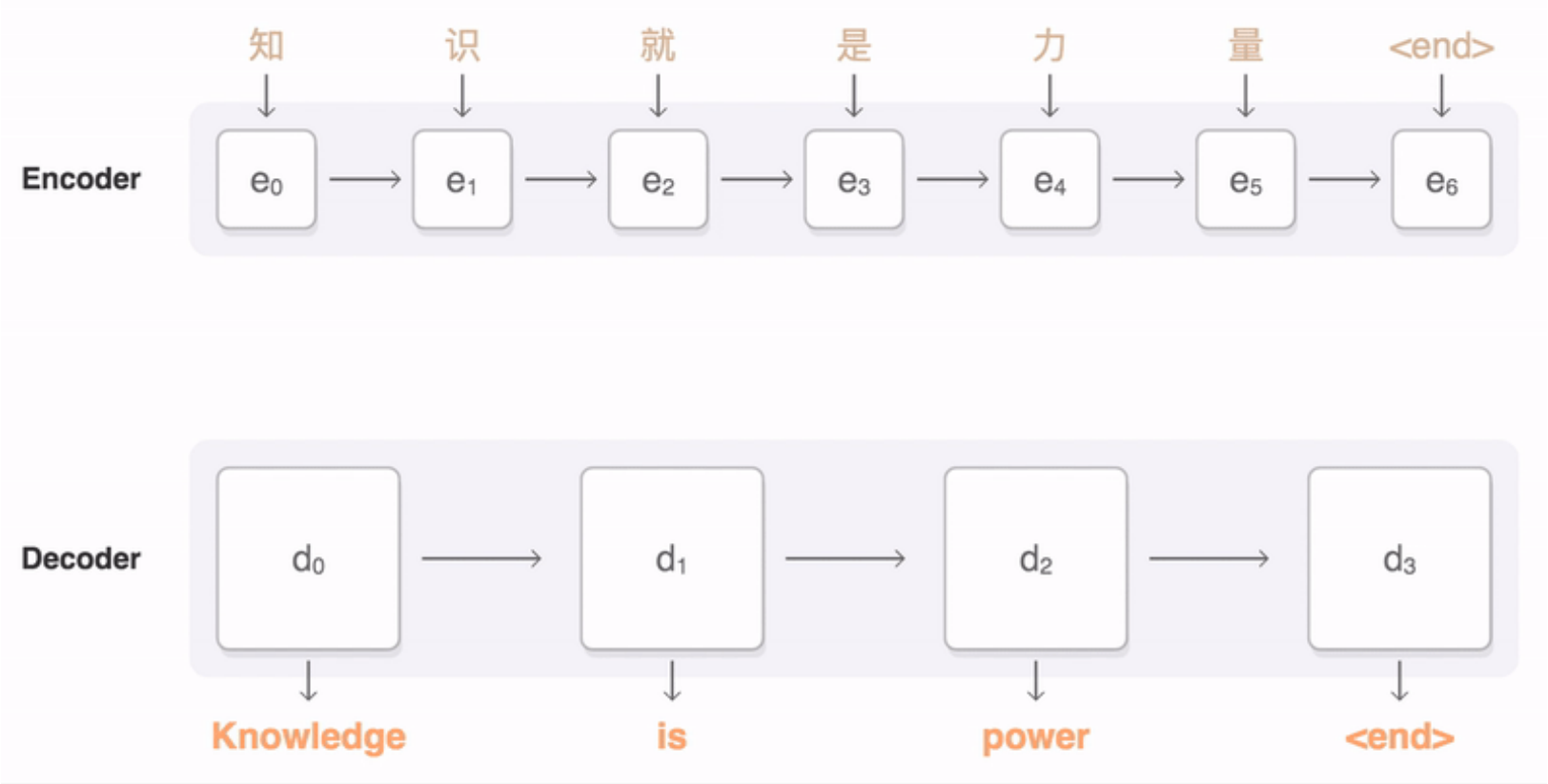


# 1.1 Conditional generative model



## Example of conditional generative model – machine translation

- $P$(English sentence | Chinese sentence)




[https://google.github.io/seq2seq/]

# 1.1 Conditional generative model



## Example of conditional generative model – article generation with the title

[Brown et al., arXiv 2020]

- $P$(A full article | An article's title and subtitle)

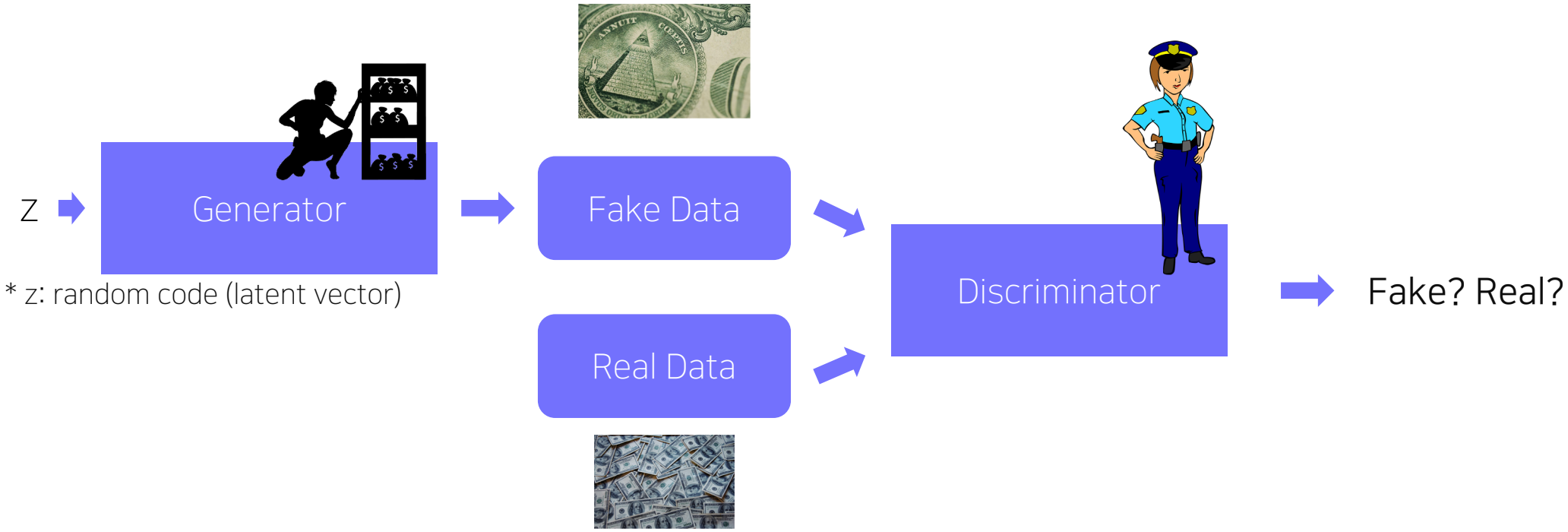
Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

# 1.1 Conditional generative model



## Recap: Generative Adversarial Network

- "Criminal" (Generator) crafts, and "Police" (Discriminator) detects counterfeit

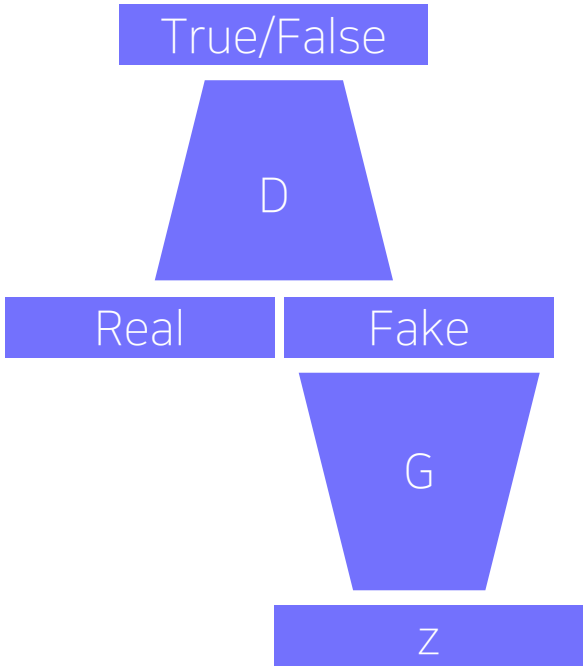


# 1.1 Conditional generative model

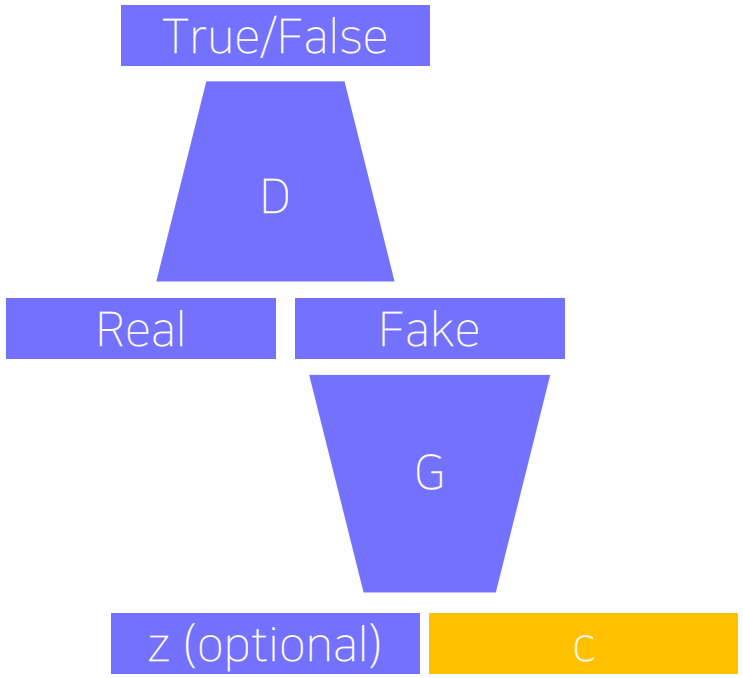


## (Basic) GAN vs. Conditional GAN

D : Discriminator / G: Generator




GAN




Conditional GAN

# 1.2 Conditional GAN and image translation



## Image-to-Image translation

[Zhu et al., ICCV 2017]

- Translating an image into another image
- Many applications: Style transfer, Super resolution, Colorization ⋯!

# Style Transfer example – Fortnite gameplay into PUBG


[https://www.youtube.com/watch?v=xkLtgwWxrec]

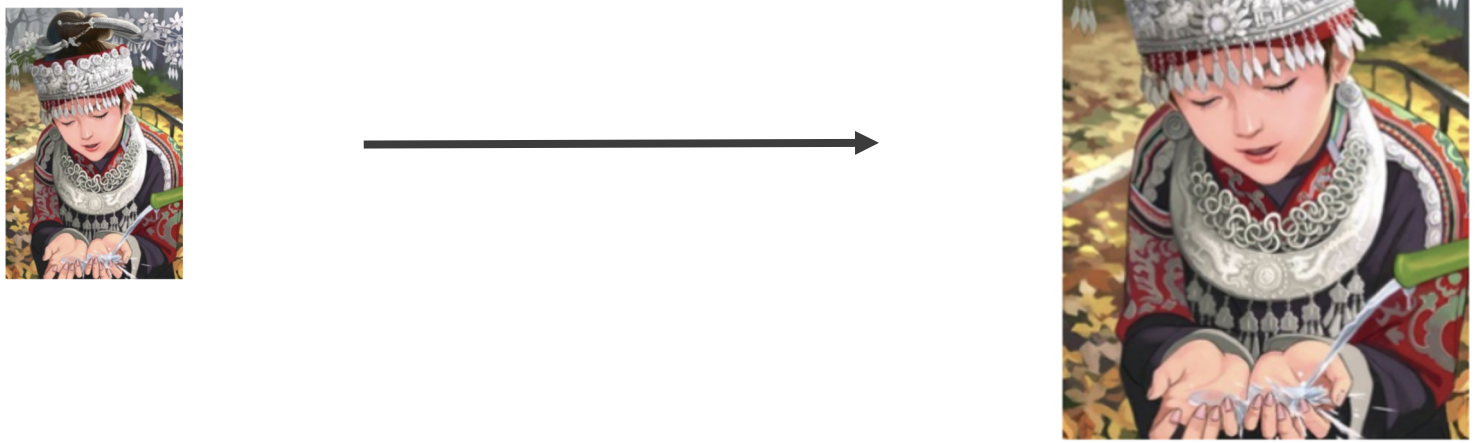
# 1.3 Example: Super resolution



## Example of conditional GAN - low resolution to high resolution

[Ledig et al., CVPR 2017]

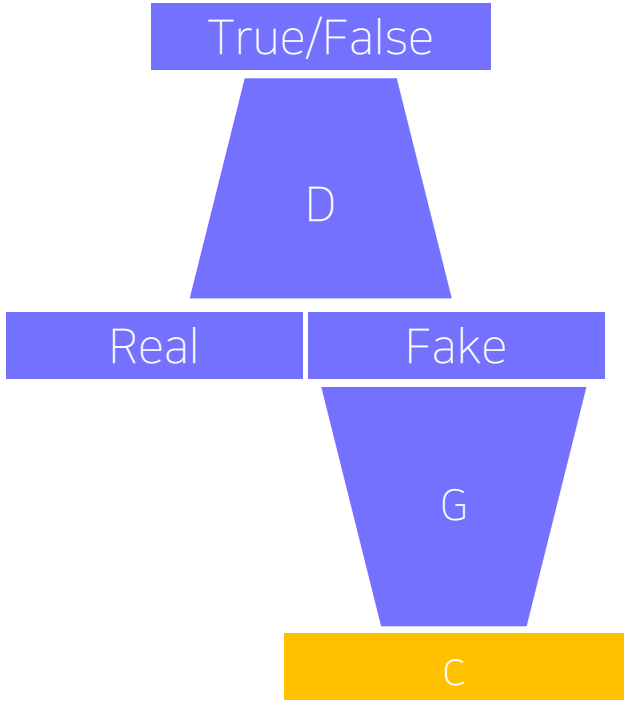
- Input: Low resolution image
  Output: High resolution image corresponding to the input image
- An example of conditional GAN



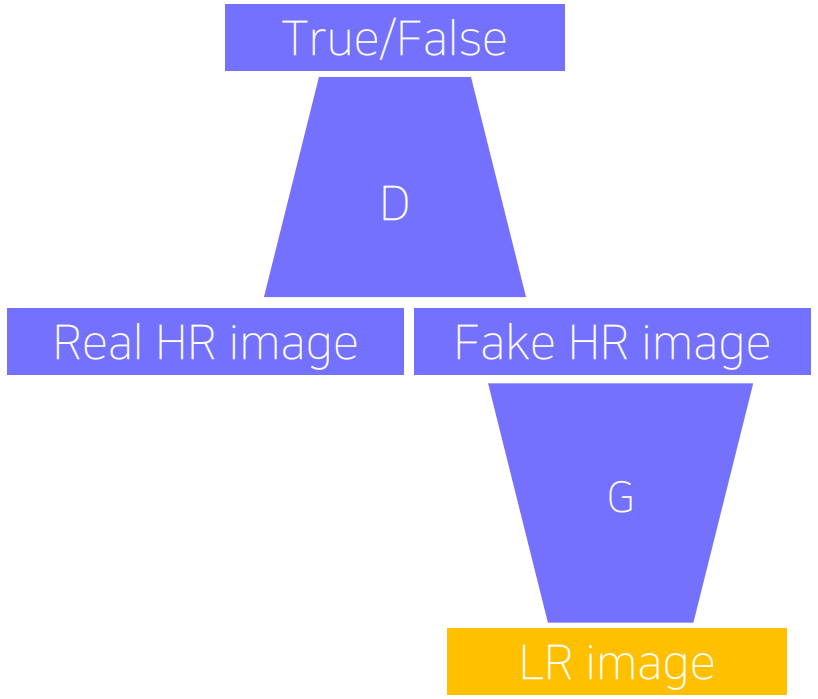
# 1.3 Example: Super resolution



Super Resolution GAN as an example of conditional GAN




Conditional GAN

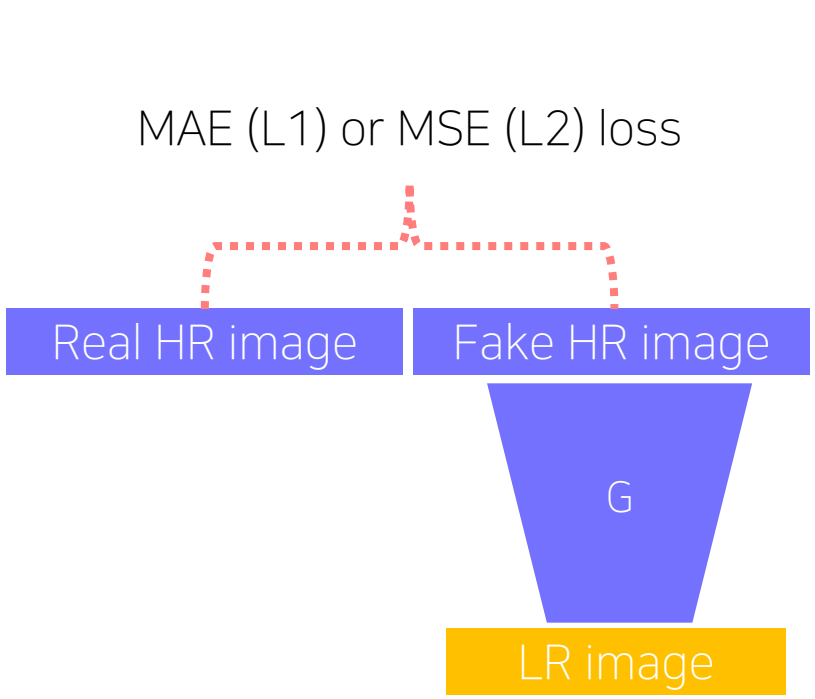



Super Resolution GAN

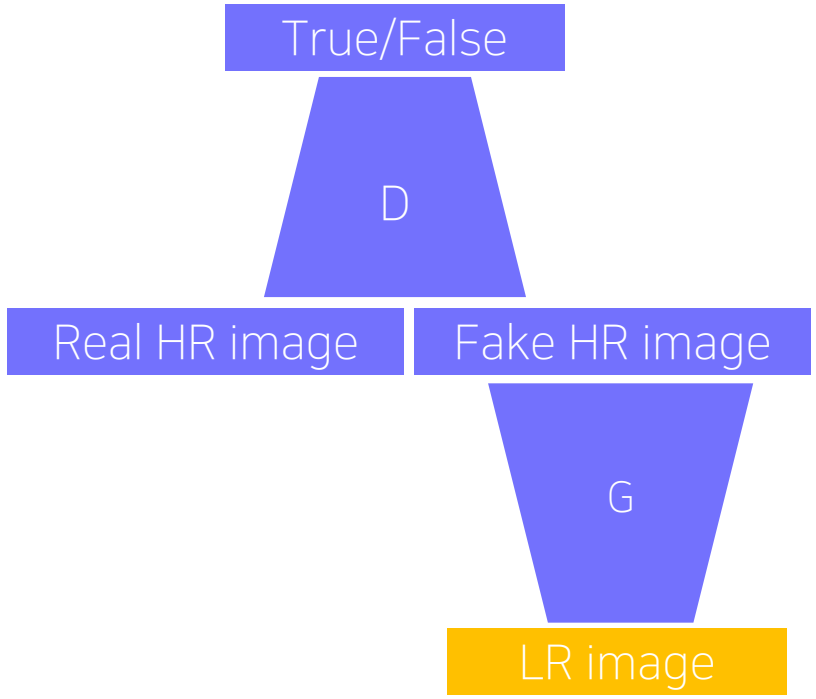
# 1.3 Example: Super resolution



## Difference between regression and conditional GAN for SR




Naïve Regression model




Super Resolution GAN

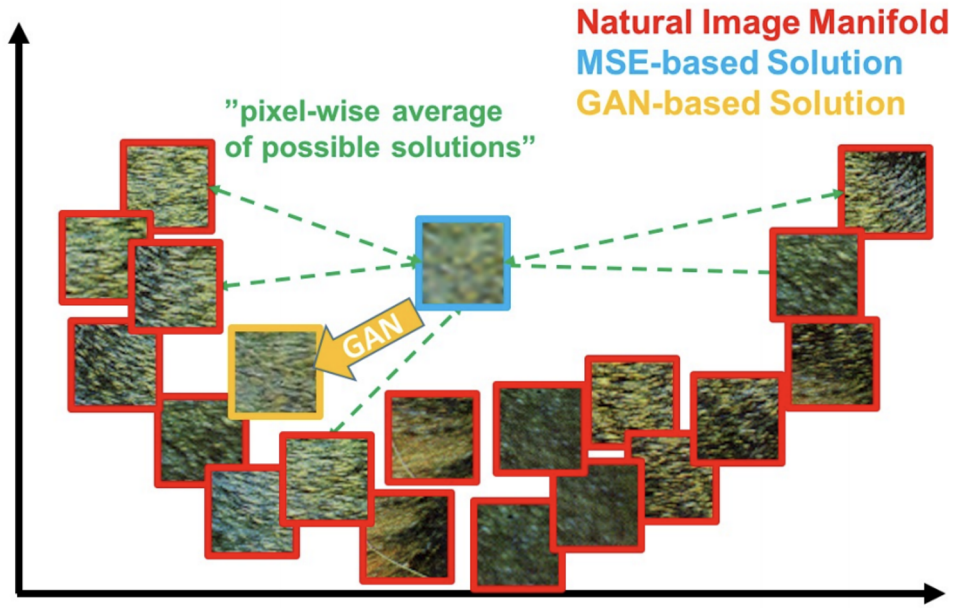
# 1.3 Example: Super resolution



## Comparison of MAE, MSE and GAN losses in an image manifold

[Ledig et al., CVPR 2017]

- MAE/MSE measure the pixel intensity difference but many similar patches exist
  - MAE/MSE produce a safe average-looking image
- GAN loss implicitly compares whether it has been seen as real or fake




$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

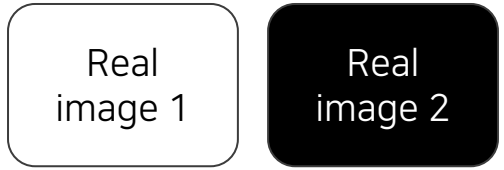
# 1.3 Example: Super resolution



## Meaning of “averaging answers” by an example

- Conditions
  - Task : Colorizing the given image
  - Real image has only two colors, “black” or “white”

Real images are like..

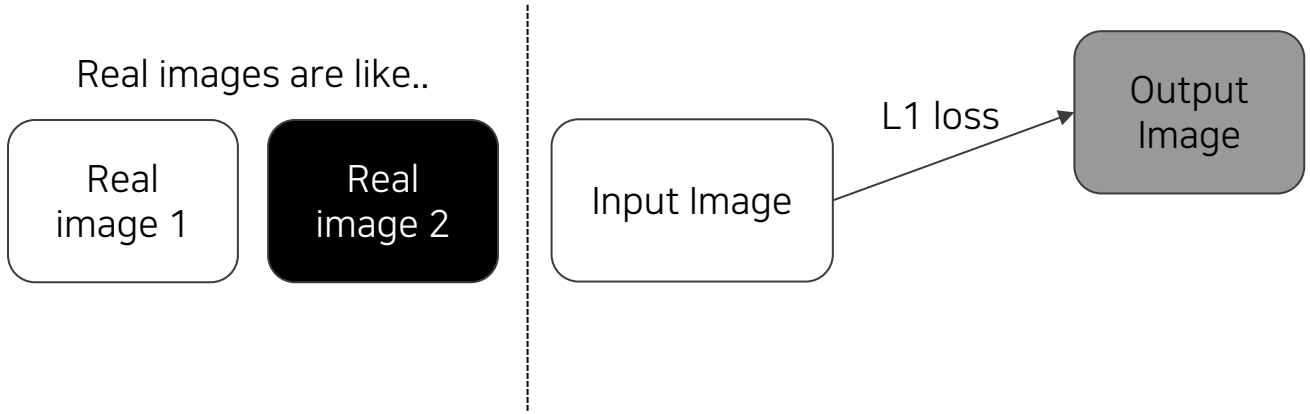


# 1.3 Example: Super resolution



## Meaning of “averaging answers” by an example

- Conditions
  - Task : Colorizing the given image
  - Real image has only two colors, “black” or “white”
- L1 loss generates “gray” output, an average of “black” and “white” (possible solutions)

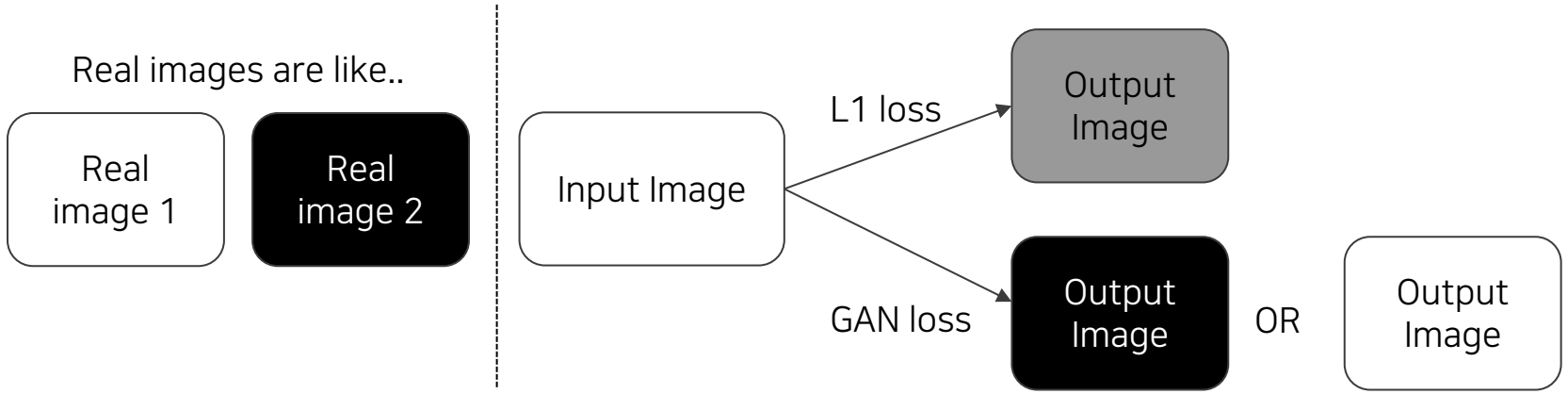


# 1.3 Example: Super resolution



## Meaning of “averaging answers” by an example

- Conditions
  - Task : Colorizing the given image
  - Real image has only two colors, “black” or “white”
- L1 loss generates “gray” output, an average of “black” and “white” (possible solutions)
- GAN loss generates “black” or “white” output, since gray image is caught by discriminator



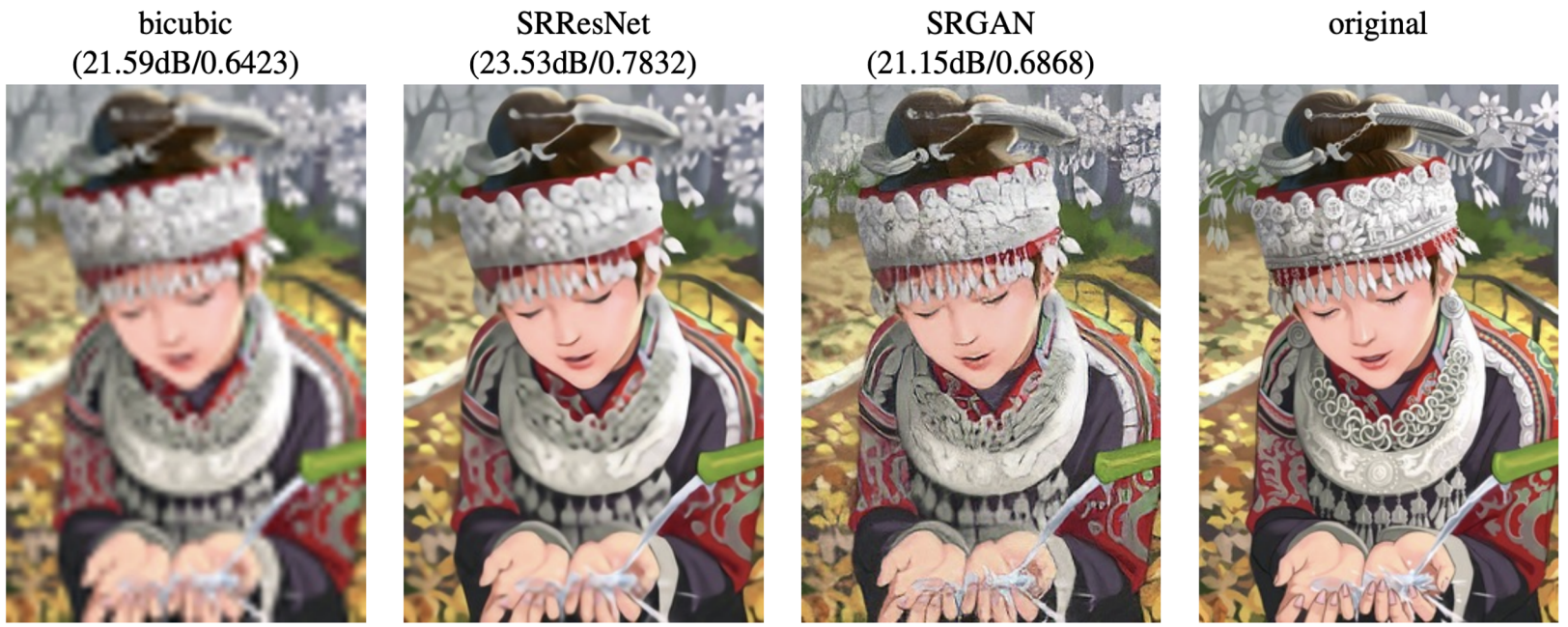
# 1.3 Example: Super resolution



## GAN loss for Super Resolution (SRGAN)

[Ledig et al., CVPR 2017]

- SRGAN generates more “realistic” and sharp images than SRResNet (MSE loss)



# 2.

# Image translation GANs

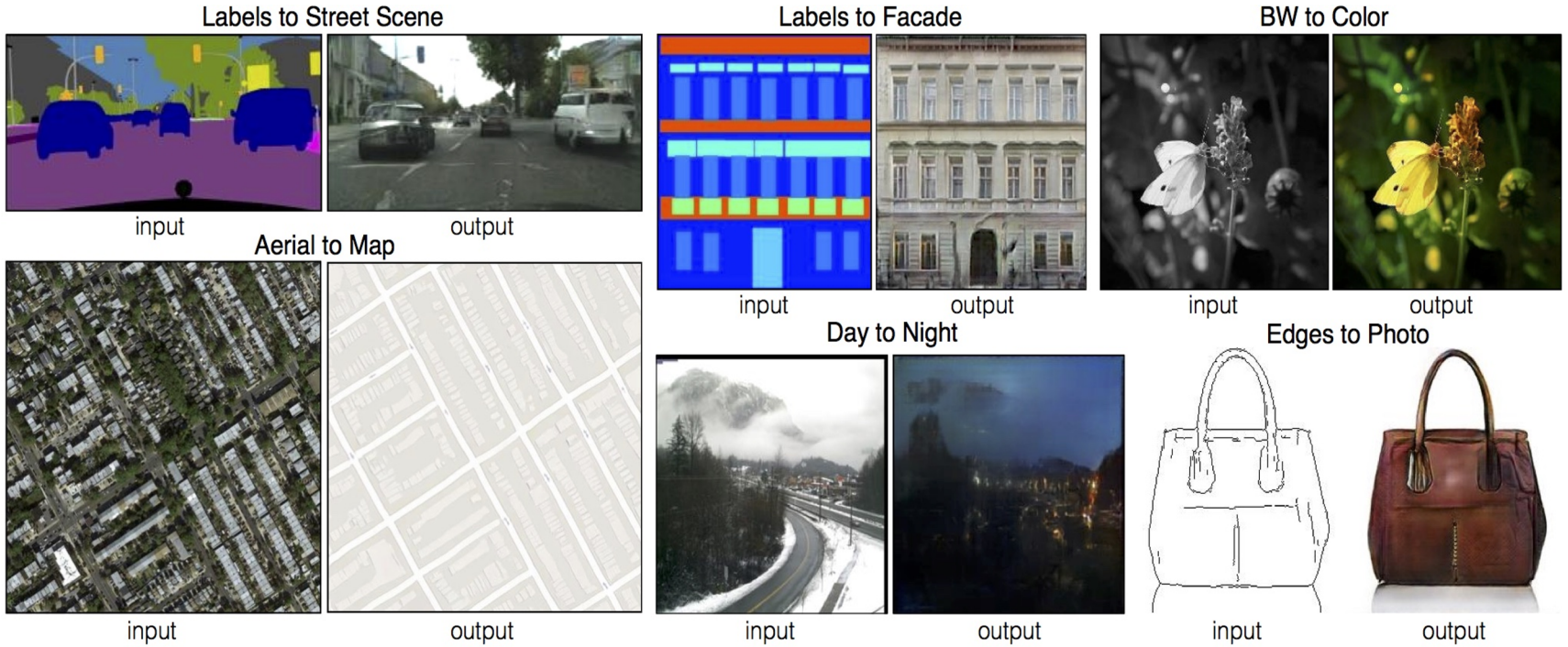
# 2.1 Pix2Pix



## Task definition

[Isola et al., CVPR 2017]

- Translating an image to a corresponding image in another domain (e.g., style)
- Example of a conditional GAN where the condition is given as an input image



# 2.1 Pix2Pix



## Loss function of Pix2Pix

[Isola et al., CVPR 2017]

- If only using L1 loss, blurry images are generated
- GAN (adversarial) loss induces more realistic outputs close to real distribution

### Total loss (GAN loss + L1 loss)

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G)$$

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))]$$

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1]$$

# 2.1 Pix2Pix

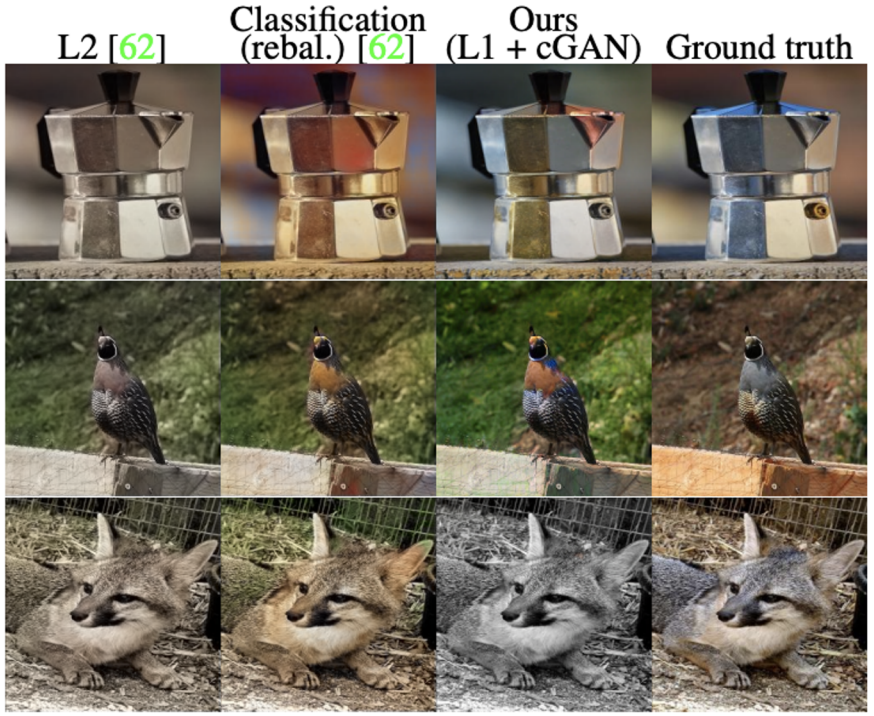


## Role of GAN loss in Pix2Pix

[Isola et al., CVPR 2017]

- Pix2Pix generates realistic images by using both GAN loss and L1 loss


Semantic map to photo




Colorization

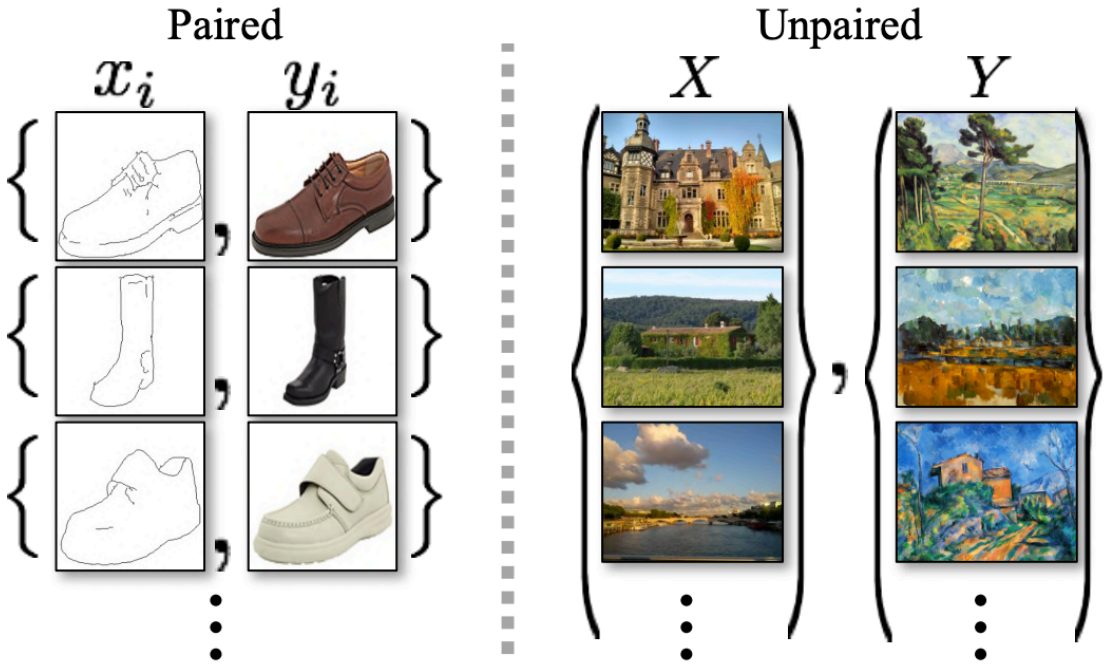
# 2.2 CycleGAN



## CycleGAN

[Zhu et al., ICCV 2017]

- In Pix2Pix, we need "pairwise data" to learn translation between two domains (supervised)
  - E.g., "sketch" and "real image" pairs are required to translate a sketch into a real image
  - However, it is hard or impossible to get a pairwise dataset



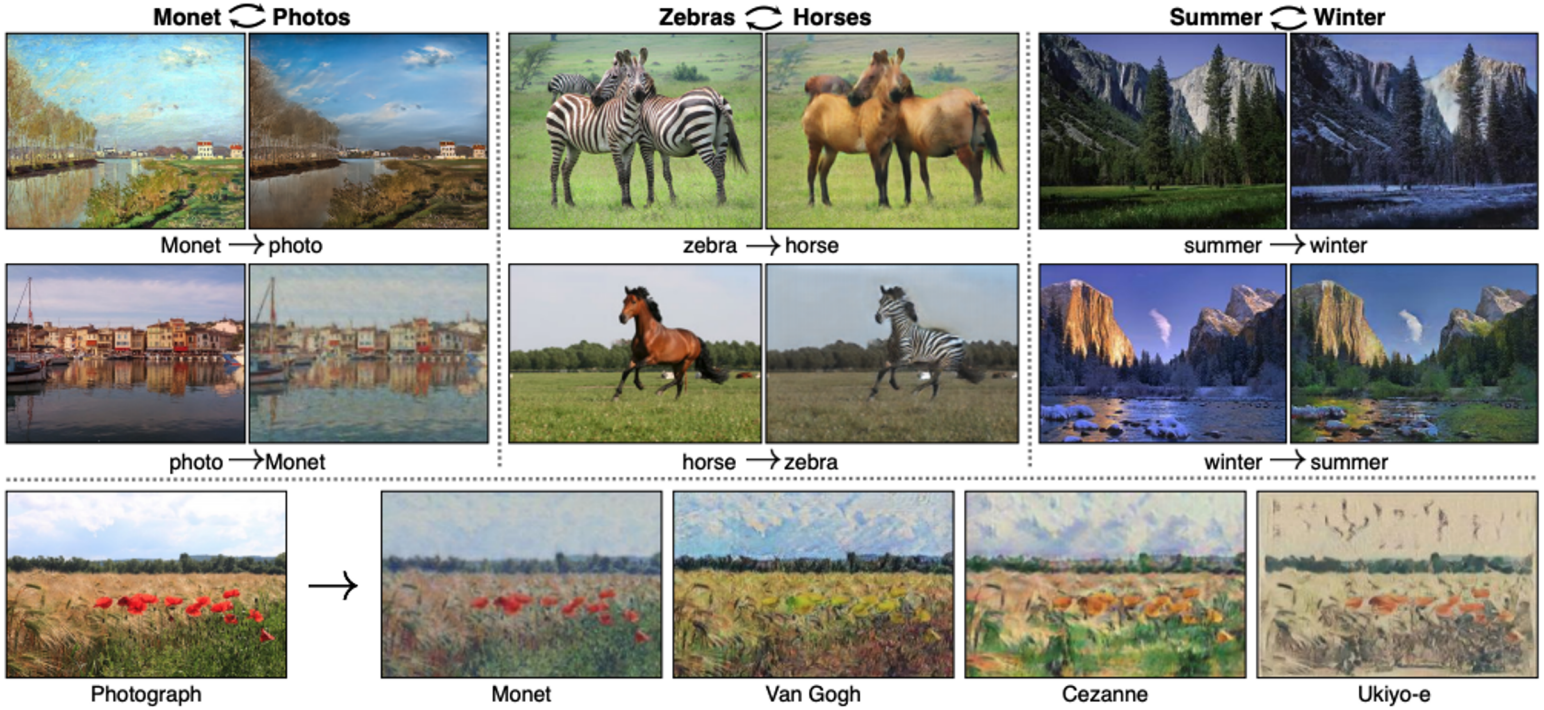
# 2.2 CycleGAN



## CycleGAN

[Zhu et al., ICCV 2017]

- CycleGAN enables the translation between domains with non-pairwise datasets
- Do not require direct correspondences between (Monet portrait, real photo)



# 2.2 CycleGAN

## Loss function of CycleGAN

[Zhu et al., ICCV 2017]

- CycleGAN loss = GAN loss (in both direction) + Cycle-consistency loss

$$L_{GAN}(X \rightarrow Y) + L_{GAN}(Y \rightarrow X) + L_{cycle}(G, F)$$

where G/F are generators

- GAN loss: Translate an image in domain A to B, and vice versa

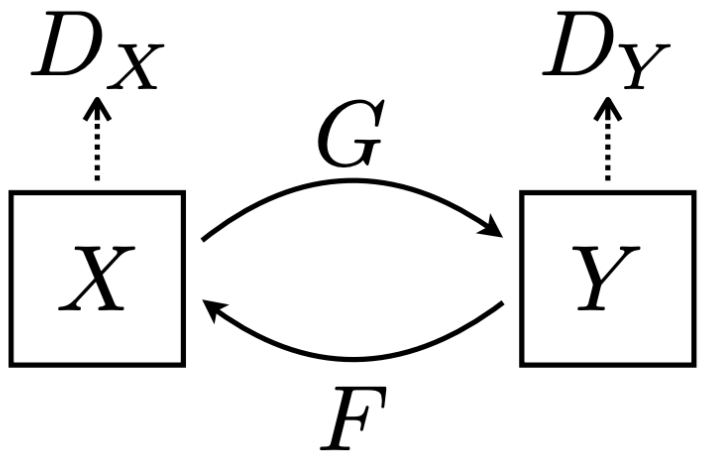
- Cycle-consistency loss: Enforce the fact that an image and its manipulated one by going back-and-forth should be same

# 2.2 CycleGAN

## GAN loss in CycleGAN

[Zhu et al., ICCV 2017]

- GAN loss does translation
- CycleGAN has two GAN losses, in both direction (X → Y, Y → X)
- GAN loss : $L(D_X) + L(D_Y) + L(G) + L(F)$
- $G, F$: generator
- $D_X, D_Y$: discriminator



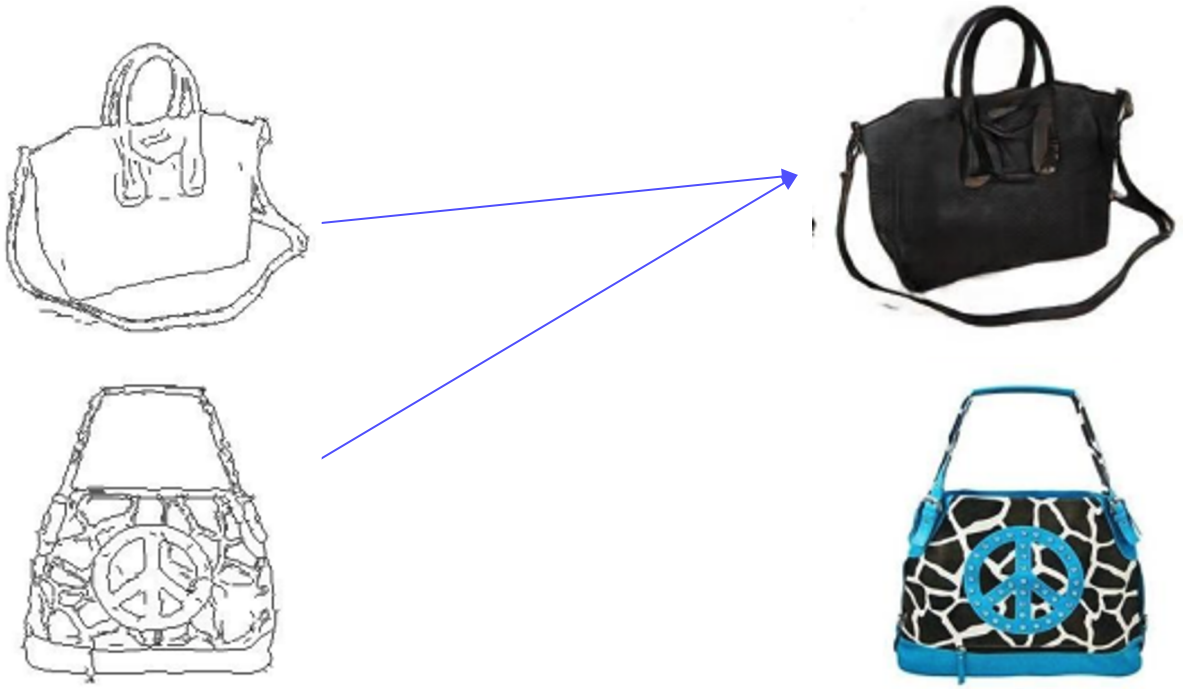
# 2.2 CycleGAN



If we solely use GAN loss …

[Zhu et al., ICCV 2017]

- (Mode Collapse) Regardless of input, the generator could always output the same one!
- Contents of input is not properly reflected in output!



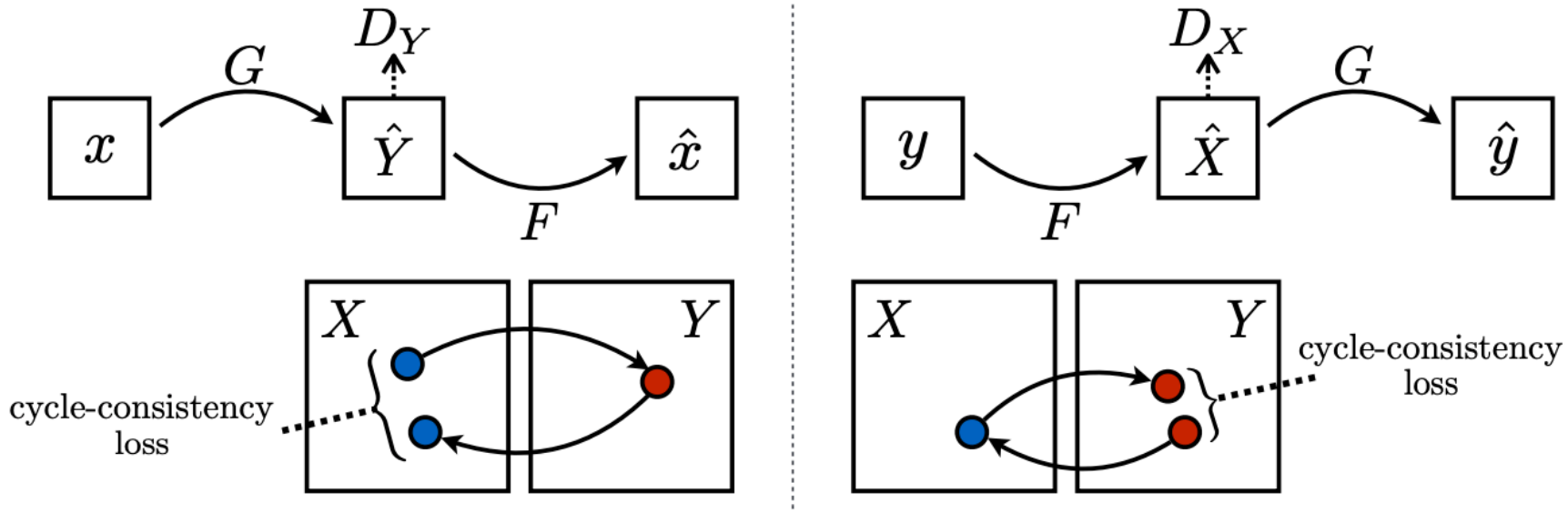
# 2.2 CycleGAN



## Solution: Cycle-consistency loss to preserve contents

[Zhu et al., ICCV 2017]

- Translate an image in X to Y, and translate its output image into X again
  - The recovered image should be same with the original image!
- Contents in an image should be preserved to do so
- No supervision (i.e., self-supervision)



# 2.3 Perceptual loss



## GAN is hard to train

GAN is hard to train (alternating training is required)

Is there another way to get a high-quality image without GAN?

# 2.3 Perceptual loss



Perceptual loss, yet another approach for achieving high quality output

- GAN loss
  - Relatively hard to train and code (Generator & Discriminator adversarially improve)
  - Do not require any pre-trained networks
  - Since no pre-trained network is required, can be applied to various applications

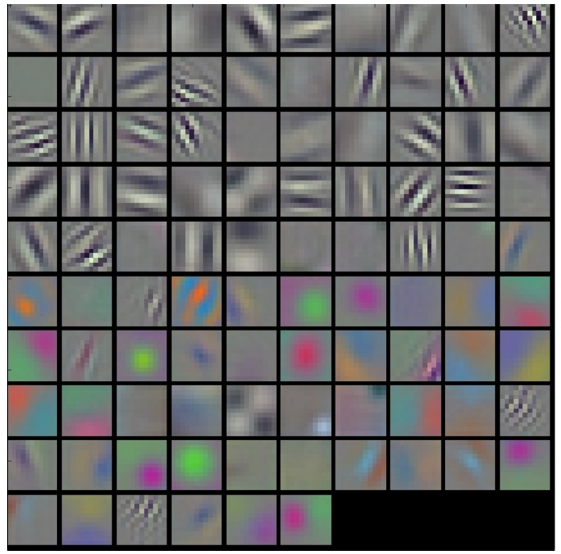
- Perceptual loss
  - Simple to train and code (trained only with simple forward & backward computation)
  - Requiring a pre-trained network to measure a learned loss

# 2.3 Perceptual loss



## Perceptual loss

- Observation: Pre-trained classifiers have filter responses similar to humans' visual perception
- By utilizing such pre-trained "perception," we may transform an image to a perceptual space



Filter visualization of an early layer of a pre-trained network

[https://cs231n.github.io/understanding-cnn/]

# 2.3 Perceptual loss

## Perceptual loss

[Johnson et al., ECCV 2016]

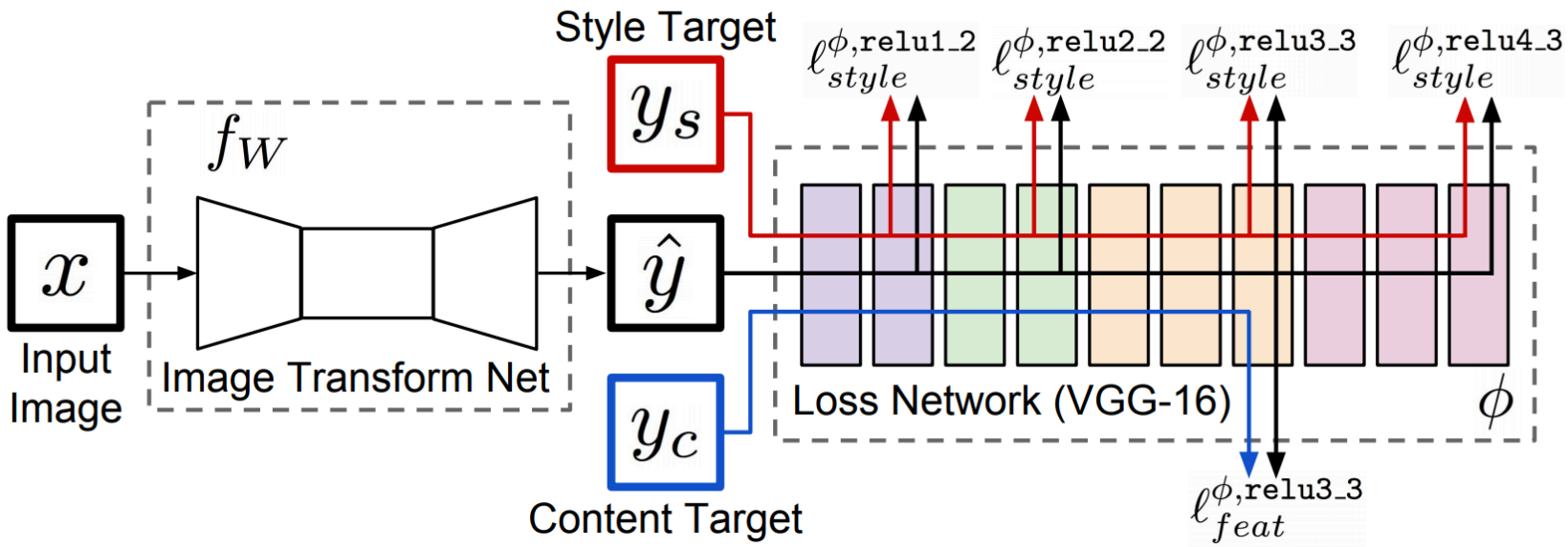

Style transfer examples with the perceptual loss

# 2.3 Perceptual loss

## Perceptual loss

[Johnson et al., ECCV 2016]

- Image Transform Net.: Output a transformed image from input
- Loss Network: Compute style and feature losses between a generated image and targets
  - Typically, the VGG model pre-trained on ImageNet is used
  - Fixed during training Image Transform Net



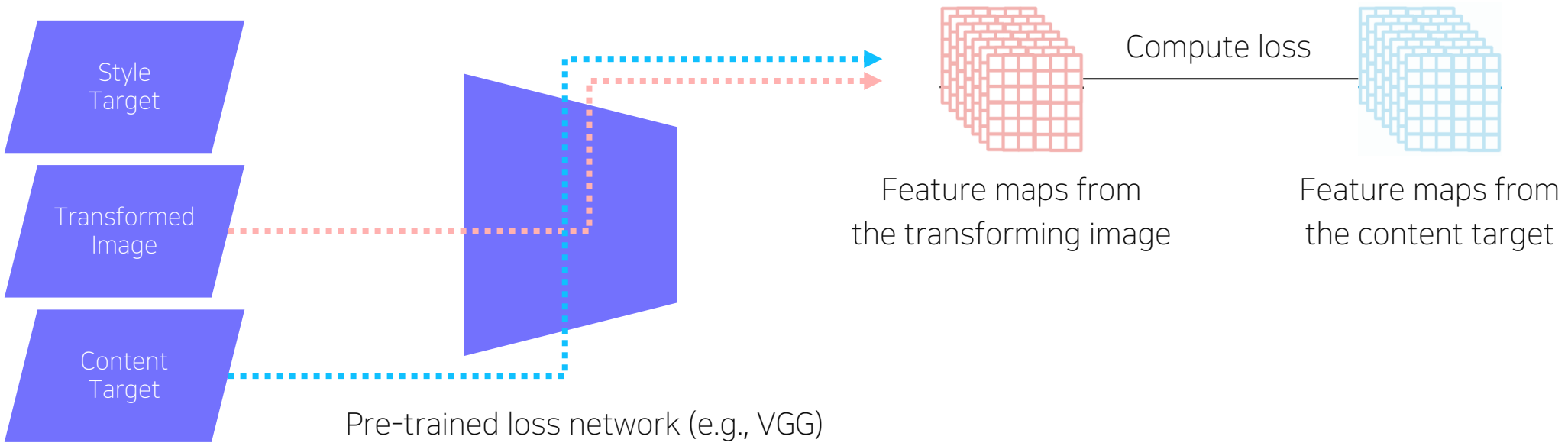
# 2.3 Perceptual loss



## Feature reconstruction loss

[Johnson et al., ECCV 2016]

- The output image and target image are fed into the loss network
- Compute L2 loss between the feature maps of output and target images



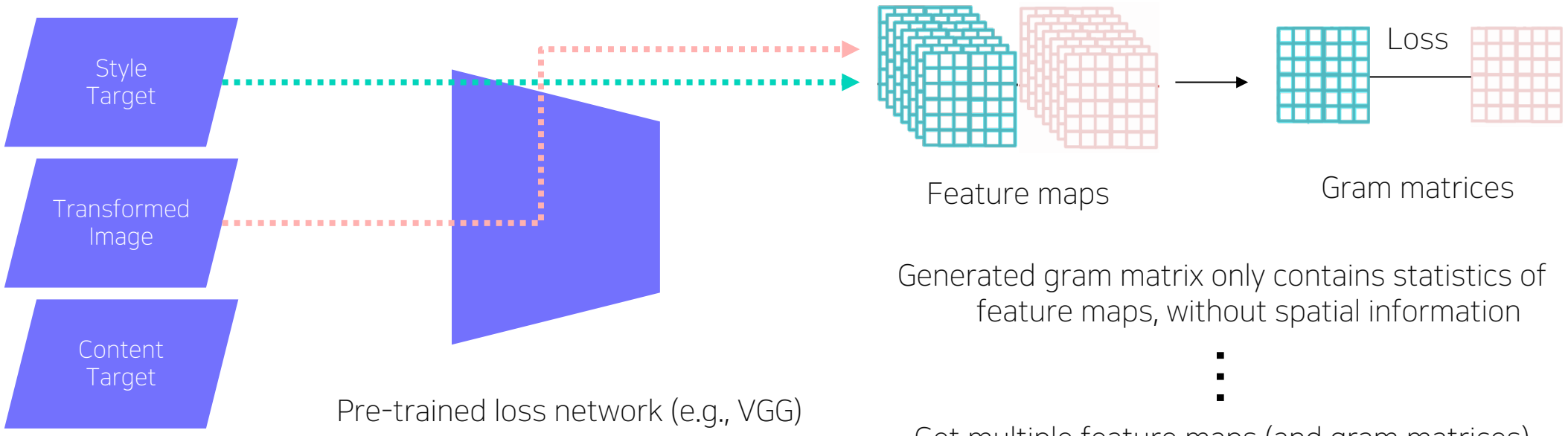
# 2.3 Perceptual loss



## Style reconstruction loss

[Johnson et al., ECCV 2016]

- Similarly, the output image and target image are fed into the loss network
- Compute L2 loss between the gram matrices generated from the feature maps



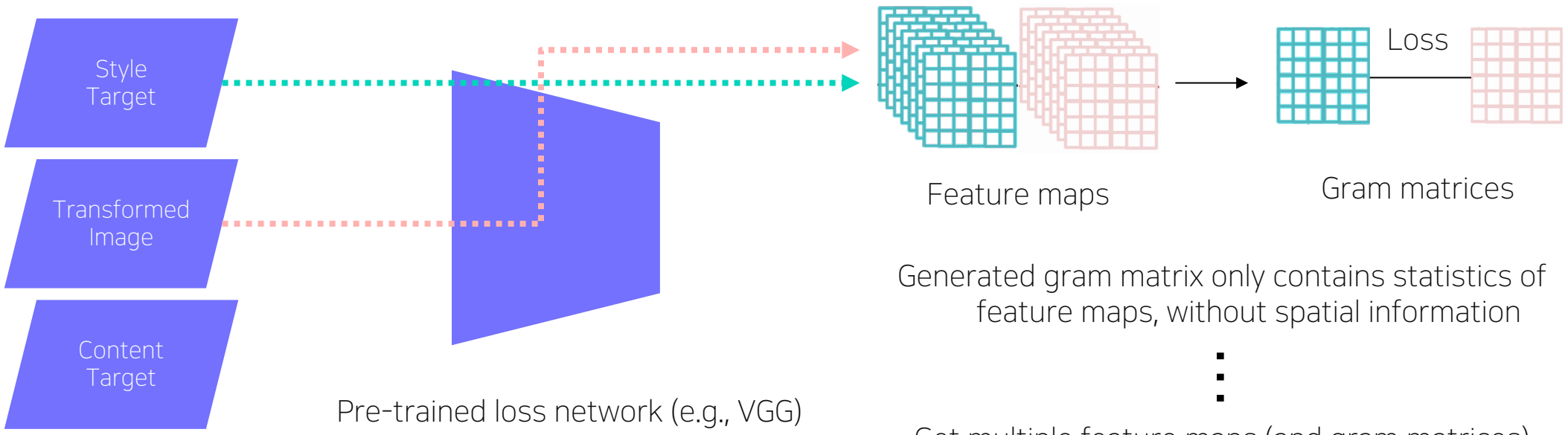
# 2.3 Perceptual loss



## Style reconstruction loss

[Johnson et al., ECCV 2016]

- Similarly, the output image and target image are fed into the loss network
- Compute L2 loss between the gram matrices generated from the feature maps



# 3.
# Various GAN applications

## Deepfake

- Converting human face or voice in video into another face or voice

Non-existing people


"Fake" president speeches

×



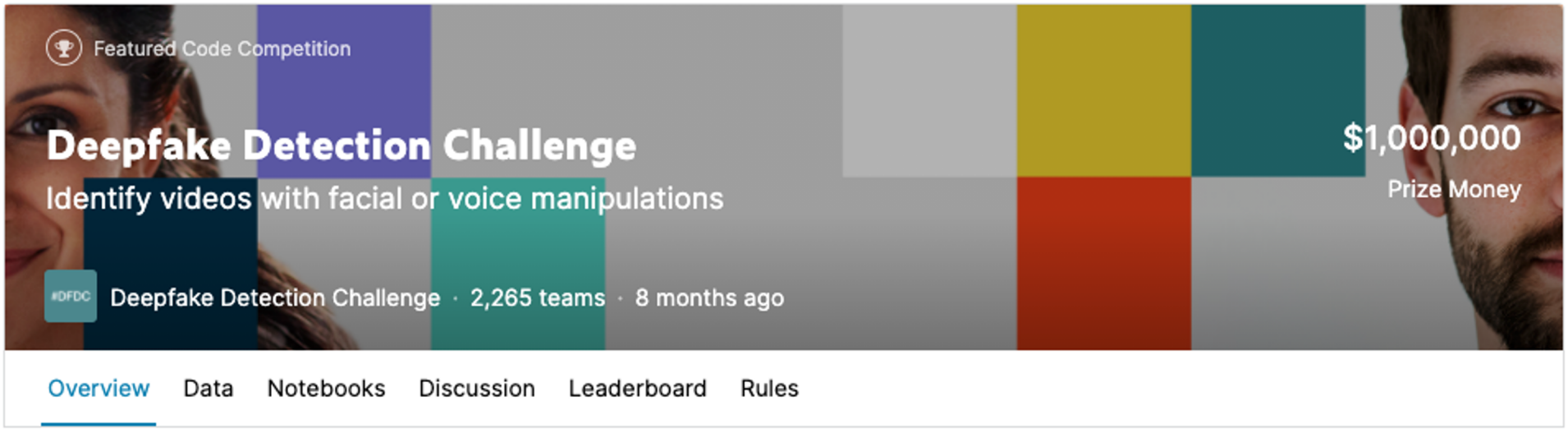




[https://www.bloomberg.com/news/videos/2018-09-26/it-s-getting-harder-to-spot-a-deep-fake-video-video]

## Ethical concerns about Deepfake

- Fake video (images) can be easily generated using GAN
- Preventing crimes using GAN is emerging as an important challenge




Kaggle Deepfake detection challenge

[https://www.kaggle.com/c/deepfake-detection-challenge]

# 3.2 Face de-identification



## Face de-identification

[Gafni et al., ICCV 2019]

- Protecting privacy by slightly modifying human face image
- Results look similar to human but hard for computer to identify them as same person

Original

De-identified

Target

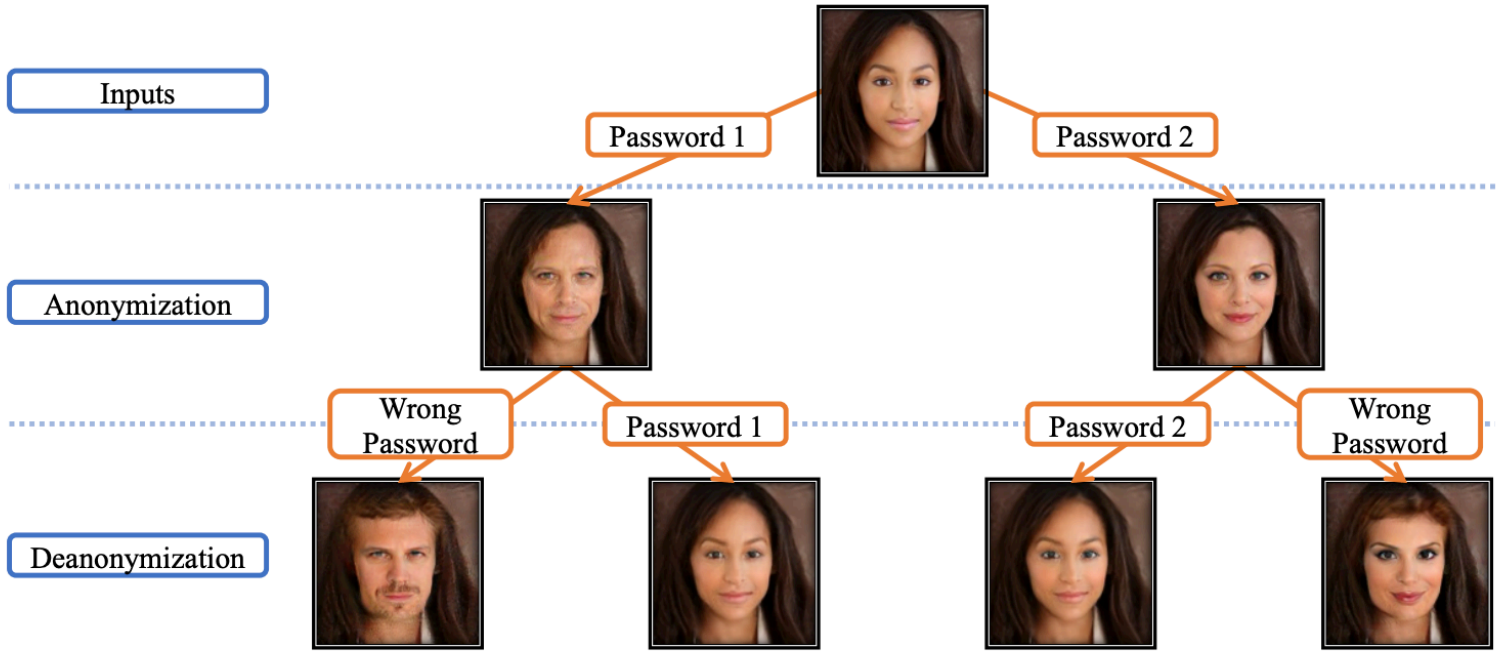
# 3.2 Face de-identification



## Face anonymization with passcode

[Gu et al., ECCV 2020]

- De-identifying human face with a specific passcode
- Only the authorized with passcode is able to decrypt and get the original image



# 3.3 Video translation (manipulation)

Pose transfer
[Liu et al., ICCV 2019]

Video-to-video translation
[Wang et al., NeurIPS 2018]

Video-to-game: controllable characters
[Gafni et al., ICLR 2018]

# Reference

## 1. Conditional generative model


- Isola et al., Image-to-Image Translation with Conditional Adversarial Networks, CVPR 2017
- Kuleshov et al., Audio Super Resolution using Neural Networks, ICLR 2017
- Brown et al., Language Models are few shot learners, arXiv 2020
- Zhu et al., Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, ICCV 2017
- Ledig et al., Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network, CVPR 2017


## 2. Image translation GANs


- Isola et al., Image-to-Image Translation with Conditional Adversarial Networks, CVPR 2017
- Zhu et al., Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, ICCV 2017
- Johnson et al., Perceptual Losses for Real-Time Style Transfer and Super-Resolution, ECCV 2016

# Reference

## 3. Various GAN applications


- Gafni et al., Live Face De-Identification in Video, ICCV 2019
- Gu et al., Password-conditioned Anonymization and Deanonymization with Face Identity Transformers, ECCV 2020
- Liu et al., Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis, ICCV 2019
- Wang et al., Video-to-Video Synthesis, NeurIPS 2018
- Gafni et al., Vid2Game: Controllable Characters Extracted from Real-World Videos, ICLR 2018

# End of Document
# Thank You.

상위 카테고리 입력란