
CVPR 2017

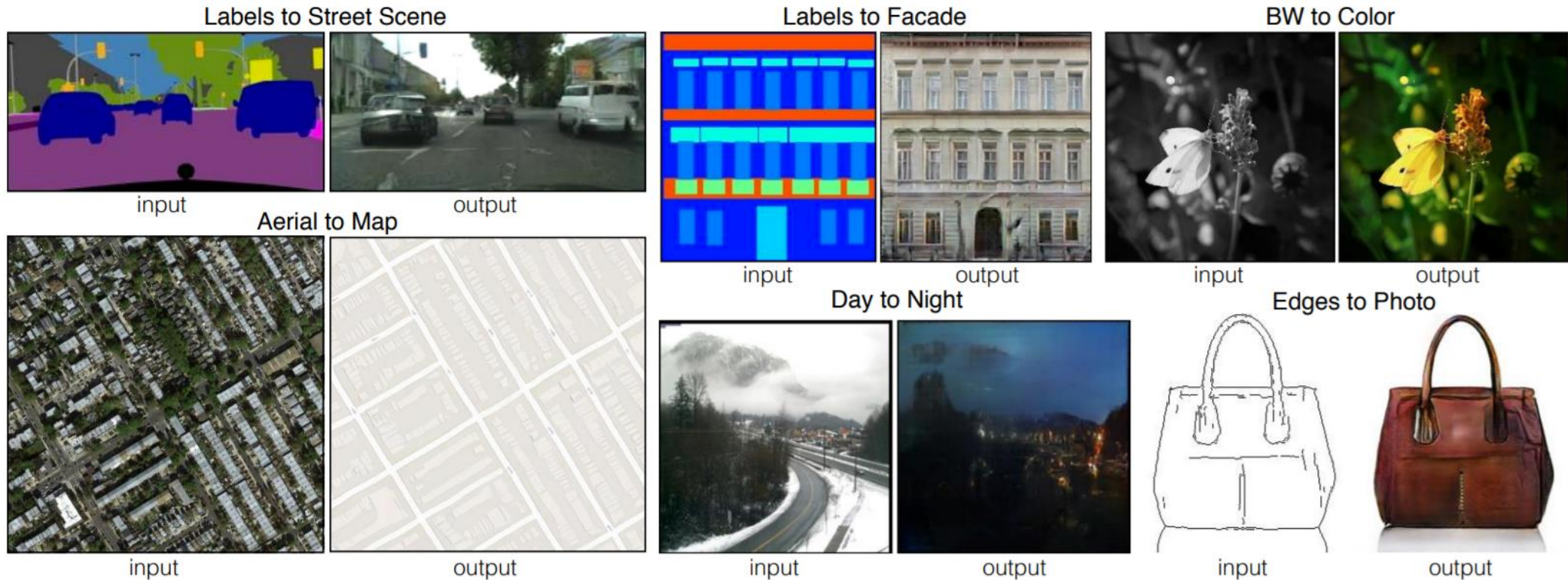
Image-to-Image Translation with Conditional Adversarial Networks

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros

Berkeley AI Research (BAIR) Laboratory, UC Berkeley

오늘 소개할 논문: Pix2Pix (CVPR 2017)

- 본 논문에서는 conditional GAN을 활용한 **간단한 image-to-image translation** 메서드를 제안합니다.
- 다양한 task에 공통적으로 적용할 수 있는 **generic approach**로 사용 가능합니다.
 - 손실 함수 및 하이퍼 파라미터의 까다로운 조정이 요구되지 않는다는 장점이 있습니다.



GAN (Generative Adversarial Networks)

논문 소개: Generative Adversarial Networks (NIPS 2014)

- GAN은 다양한 데이터를 생성할 수 있는 뉴럴 네트워크의 한 유형입니다.



Fig. Visualization of samples from the model

- Not cherry-picked
- Not memorized the training set
- Images represent sharp



Fig. Digits obtained by linearly interpolating between coordinates in z space of the model

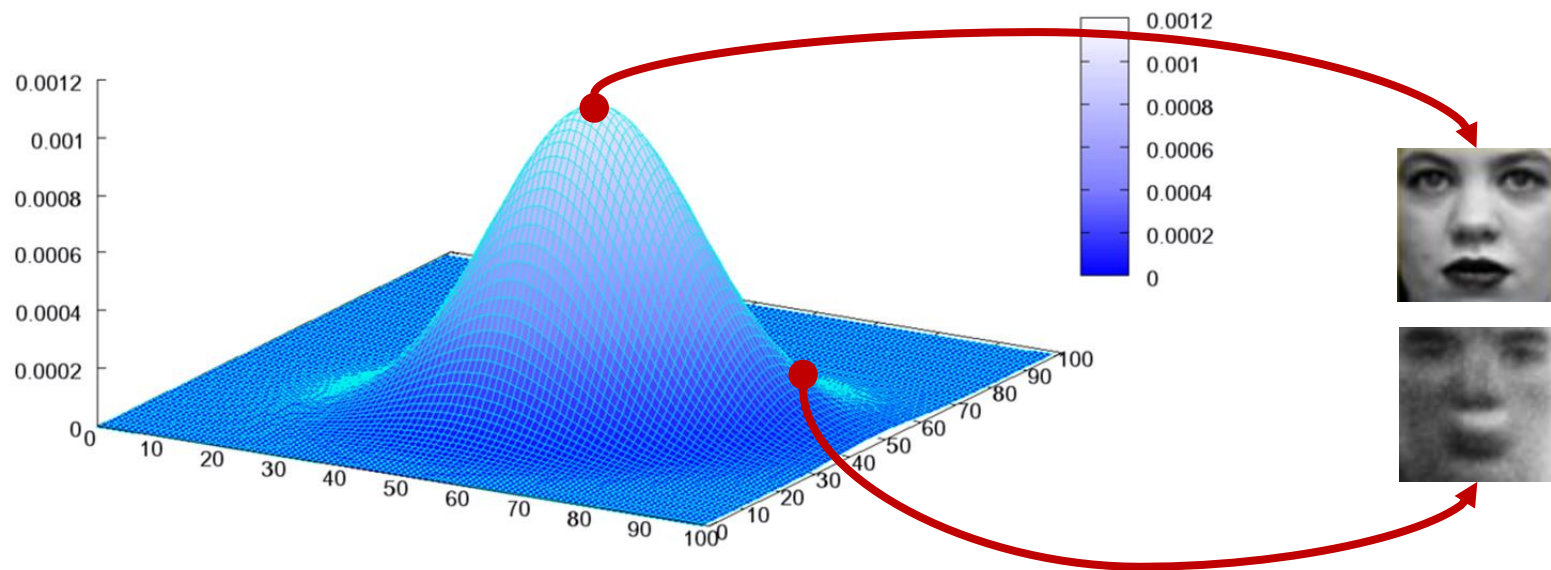
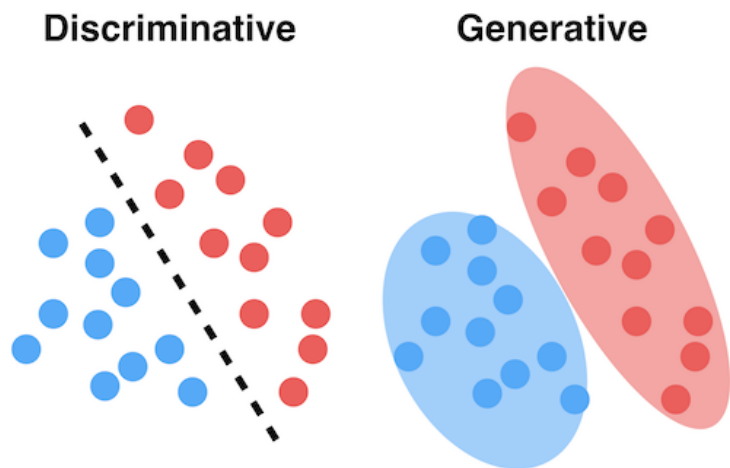
*Generative Adversarial Networks (NIPS 2014)

생성 모델 (Generative Models)

- 생성 모델은 실존하지 않지만 있을 법한 이미지를 생성할 수 있는 모델을 의미합니다.

Generative Model (produce) → An image that does not exist but is likely to exist

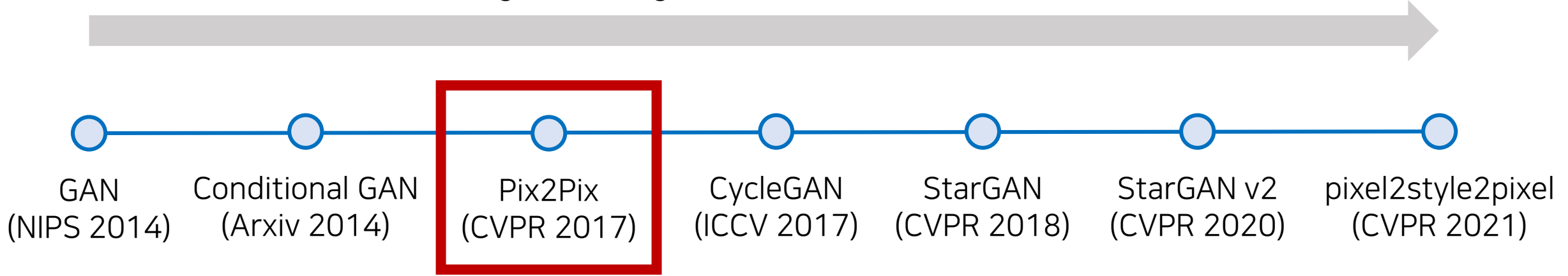
- A statistical model of the joint probability distribution
- An architecture to generate new data instances



생성 모델 (Generative Model)의 목표와 Image-to-Image Translation 소개

- 이미지 데이터의 분포를 근사하는 모델 G를 만드는 것이 생성 모델의 목표입니다.
- 모델 G가 잘 동작한다는 의미는 원래 이미지들의 분포를 잘 모델링할 수 있다는 것을 의미합니다.
 - 2014년에 제안된 **Generative Adversarial Networks (GAN)**이 대표적입니다.
- GAN이 생성 모델 중 SOTA가 되면서 매우 다양한 논문들이 파생되었습니다.
 - 오늘 다루는 Pix2Pix (Image-to-Image Translation) 또한 GAN을 기반으로 설계된 아키텍처입니다.

Image-to-Image Translation 기술의 발전 과정



Generative Adversarial Networks (GAN) 구조 살펴보기

- 생성자(generator)와 판별자(discriminator) 두 개의 네트워크를 활용한 생성 모델입니다.
- 다음의 목적 함수(objective function)를 통해 생성자는 이미지 분포를 학습할 수 있습니다.

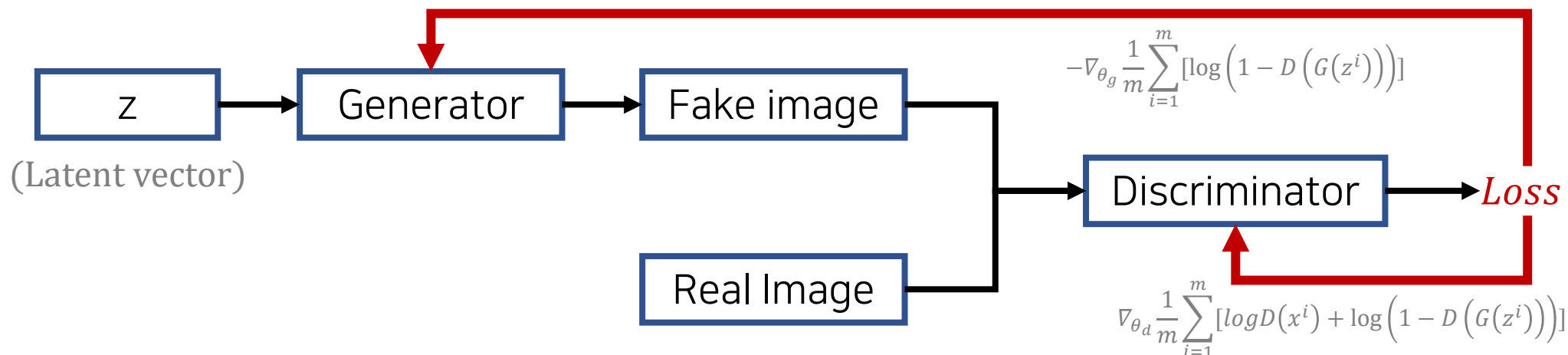
$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

Generator

$G(z)$: new data instance

Discriminator

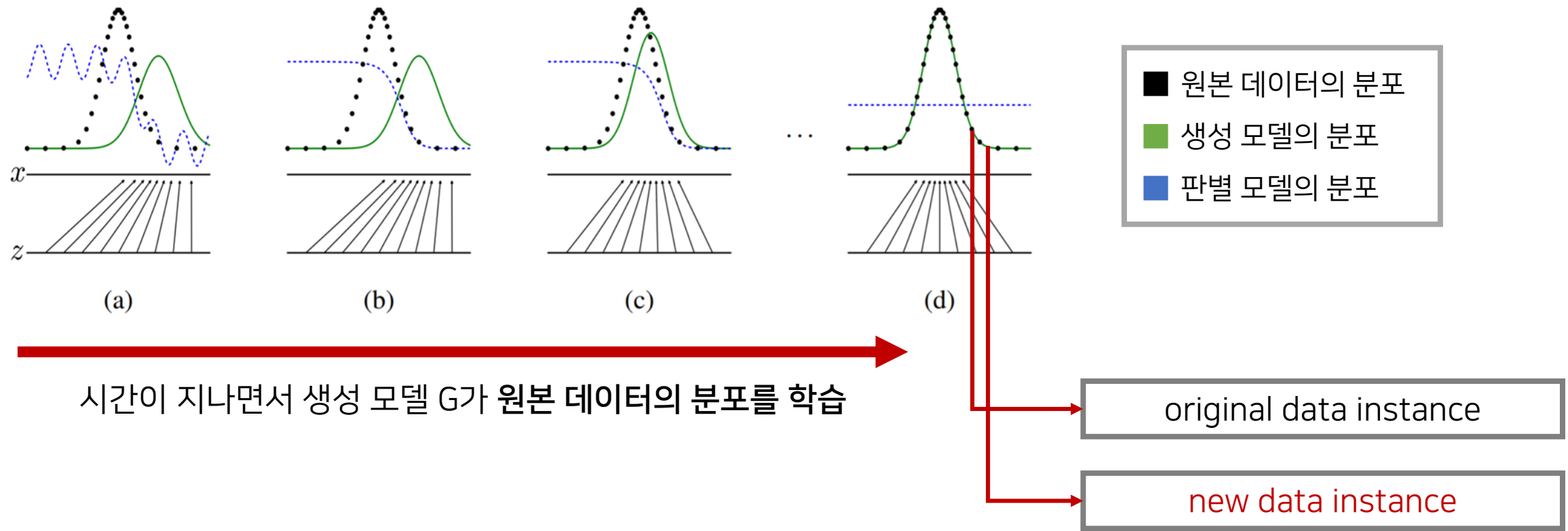
$D(x)$ = Probability: a sample came from the real distribution (Real: 1 ~ Fake: 0)



GAN의 수렴 과정

- 공식의 목표(Goal of Formulation)

- $P_g \rightarrow P_{data}, D(G(z)) \rightarrow 1/2$ ($G(z)$ is not distinguishable by D)

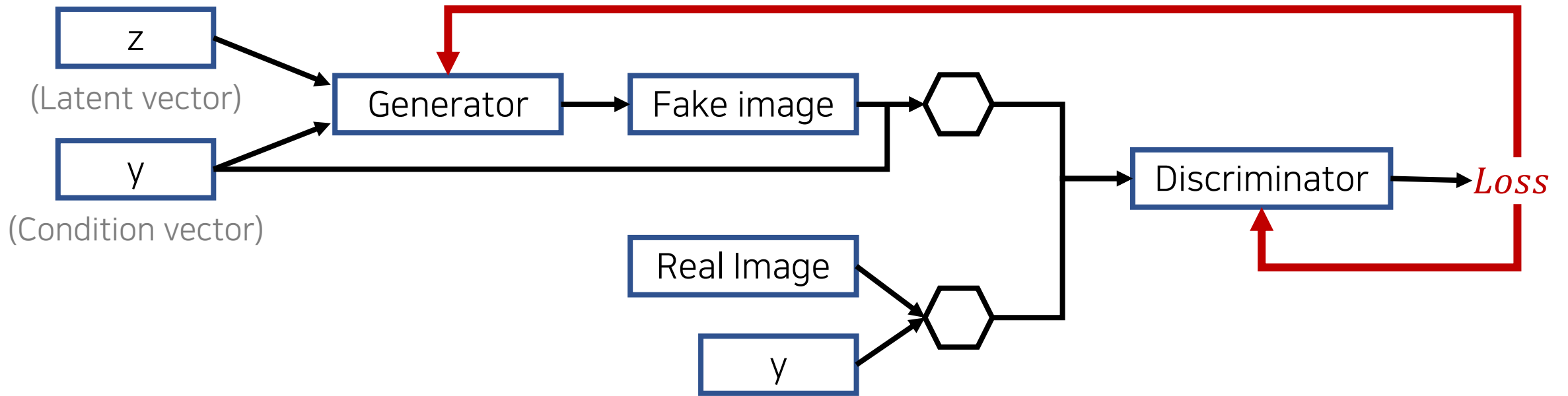


Conditional GAN (cGAN)

Conditional GAN (cGAN)

- 데이터의 모드(mode)를 제어할 수 있도록 조건(condition) 정보를 함께 입력하는 모델입니다.
 - 아래 수식 부분은 원본 논문에서 그대로 인용하였습니다.

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\log D(x|y)] + E_{z \sim p_z(z)} [\log (1 - D(G(z|y)))]$$



*Conditional Generative Adversarial Nets (2014)

Conditional GAN (cGAN)

- 데이터의 모드(mode)를 제어할 수 있도록 조건(condition) 정보를 함께 입력하는 모델입니다.



Image-to-Image Translation

Image-to-Image Translation: Pix2Pix의 개요

- The task of **I2I translation** is to change a particular aspect of a given image to another.
- 대표적인 image-to-image translation 아키텍처로 Pix2Pix가 있습니다.
 - Pix2Pix는 학습 과정에서 이미지 x 자체를 조건(condition)으로 입력받는 cGAN의 한 유형입니다.
 - Pix2Pix은 픽셀(pixel)들을 입력으로 받아 픽셀(pixel)들을 예측한다는 의미를 가집니다.

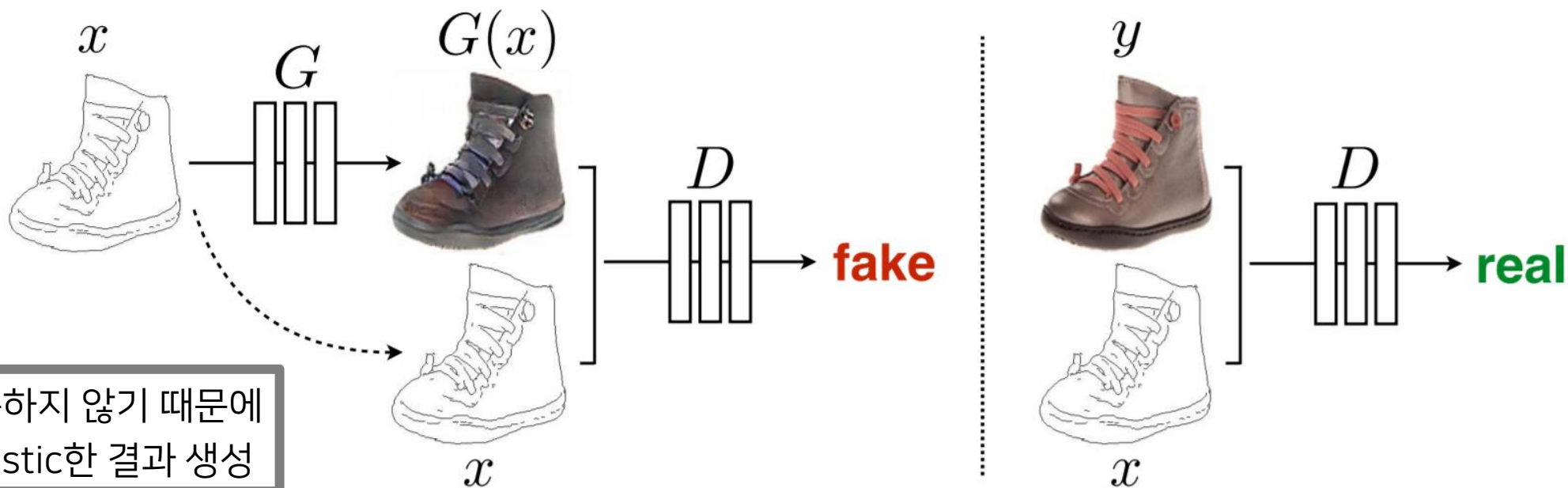
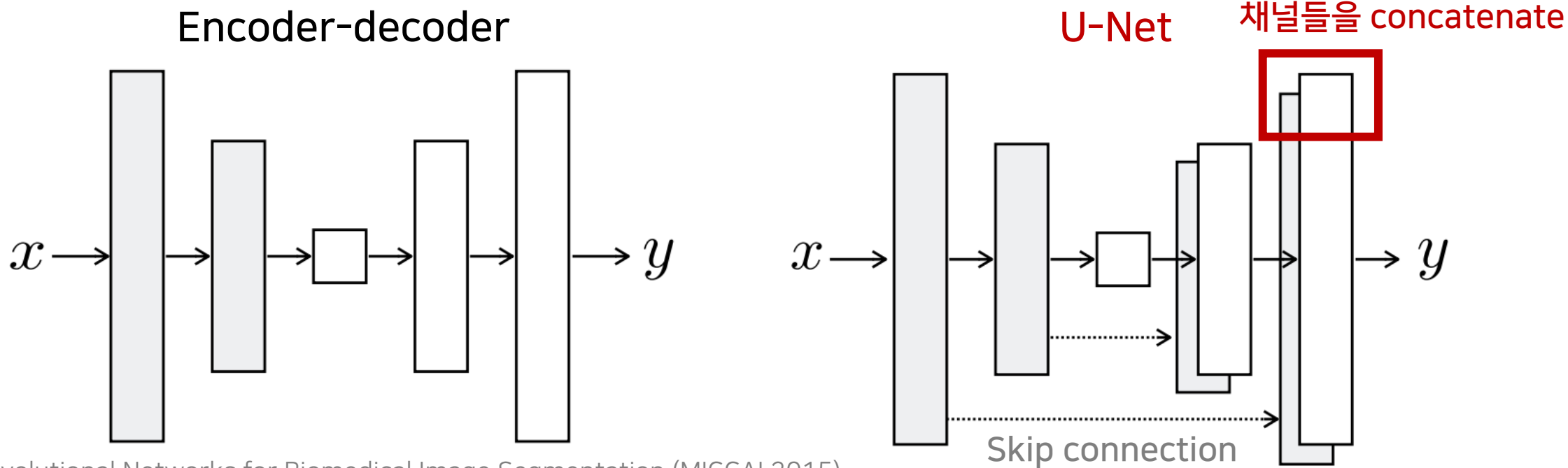


Image-to-Image Translation: Pix2Pix의 아키텍처

- Pix2Pix는 이미지를 조건(condition)으로 입력받아, 이미지를 출력으로 내보냅니다.
- 이를 효과적으로 처리할 수 있는 U-Net 기반의 네트워크 아키텍처를 사용합니다.
 - Input and output both are renderings of the same underlying structure.
 - There is a great deal of low-level information shared between the input and output.



*U-Net: Convolutional Networks for Biomedical Image Segmentation (MICCAI 2015)

Image-to-Image Translation: Pix2Pix의 손실 함수

- GAN은 기본적으로 다른 생성 모델에 비하여 blurry한 결과가 나오는 문제가 적은 편입니다.
- GAN의 성능을 더 향상시키기 위해 L1 손실(loss) 함수를 함께 사용합니다. (ground-truth와 유사한 결과)
 - Euclidean distance is minimized by averaging all plausible outputs, which causes blurring.
 - L2 손실을 이용할 때보다 L1 손실을 이용했을 때 흐림(blurring) 현상이 덜 발생합니다.

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

현실적인 이미지를 만들도록 실제 정답과 유사하도록

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

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

Image-to-Image Translation: Pix2Pix의 Discriminator

- Pix2Pix의 discriminator는 convolutional PatchGAN 분류 모델을 사용합니다.
 - 이미지 전체에 대하여 판별하지 않고, 이미지 내 패치 단위로 진짜/가짜 여부를 판별합니다.
- In order to model high-frequencies, it is sufficient to restrict our attention to the structure in **local image patches**.
 - PatchGAN only penalizes structure at the scale of patches.
 - This discriminator tries to classify if each $N \times N$ patch in an image is real or fake.
 - **장점**: fewer parameters, runs faster, can be applied to arbitrarily large images.

*Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks

Image-to-Image Translation: Pix2Pix의 성능 평가 (U-Net 아키텍처의 장점)

- U-Net 구조와 함께 본 논문에서 제안한 $\text{loss}(L1 + \text{cGAN})$ 를 사용할 때 가장 우수한 결과를 보입니다.



Fig. Adding skip connections to an encoder-decoder to create a “U-Net” results in much higher quality results.

Image-to-Image Translation: Pix2Pix의 성능 평가 (U-Net 아키텍처의 장점)

- U-Net 구조와 함께 본 논문에서 제안한 $\text{loss}(\text{L1} + \text{cGAN})$ 를 사용할 때 가장 우수한 결과를 보입니다.

Loss	Per-pixel acc.	Per-class acc.	Class IOU
Encoder-decoder (L1)	0.35	0.12	0.08
Encoder-decoder (L1+cGAN)	0.29	0.09	0.05
U-net (L1)	0.48	0.18	0.13
U-net (L1+cGAN)	0.55	0.20	0.14

Table. FCN-scores for different generator architectures (and objectives), evaluated on Cityscapes labels↔photos.

Image-to-Image Translation: Pix2Pix의 성능 평가 (본 논문에서 제안한 손실 함수의 장점)



- L1 loss만 사용하는 경우 blurry한 결과를 보입니다.
- cGAN loss만 사용하는 경우 sharp하지만 visual artifacts가 존재하는 문제가 있습니다.
- 두 개의 loss를 적절히 섞어 사용할 때 그러한 artifact가 적으면서도 우수한 결과가 나오는 것을 확인할 수 있습니다.

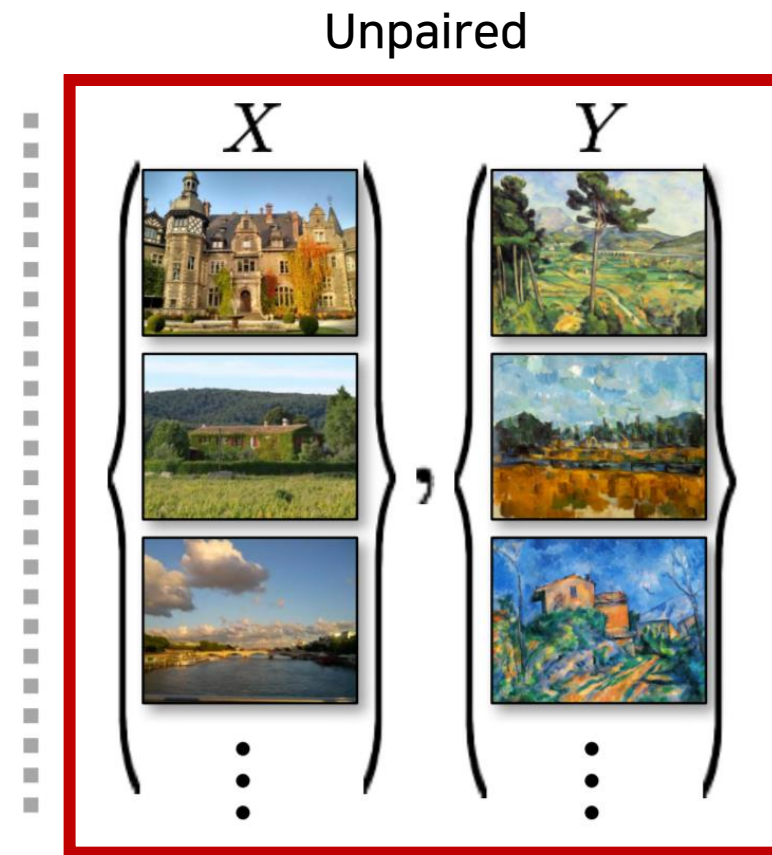
Image-to-Image Translation: Pix2Pix의 성능 평가 (본 논문에서 제안한 손실 함수의 장점)

Loss	Per-pixel acc.	Per-class acc.	Class IOU
L1	0.42	0.15	0.11
GAN	0.22	0.05	0.01
cGAN	0.57	0.22	0.16
L1+GAN	0.64	0.20	0.15
L1+cGAN	0.66	0.23	0.17
Ground truth	0.80	0.26	0.21

Table. FCN-scores for different losses, evaluated on Cityscapes labels↔photos.

Image-to-Image Translation: Pix2Pix의 한계점

- Pix2Pix는 서로 다른 두 도메인 X , Y 의 데이터를 한 쌍으로 묶어 학습을 진행합니다.
 - 다만 colorization과 같은 태스크에서는 데이터셋을 구성하기 쉬우나 그렇지 않은 경우도 존재합니다.



한 쌍으로 묶이지 않은(unpaired)
데이터 셋에 대해서도 적용이 가능할까요?



CycleGAN을 이용해 해결 가능