Image-to-Image Translation

Ming-Yu Liu
NVIDIA

Ting-Chun Wang
NVIDIA

A Short Introduction to GANs

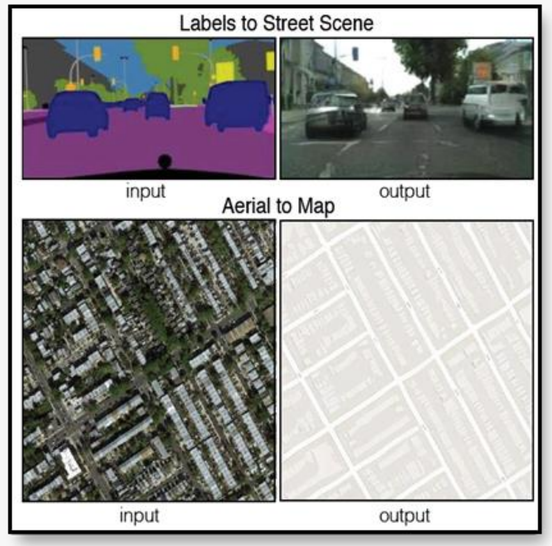
Ting-Chun Wang
NVIDIA

Recent Advances in Image Generation

Prog. GAN [Karras et al.]

BigGAN [Brock et al.]

StyleGAN [Karras et al.]



pix2pix [Isola et al.]

pix2pixHD [Wang et al.]

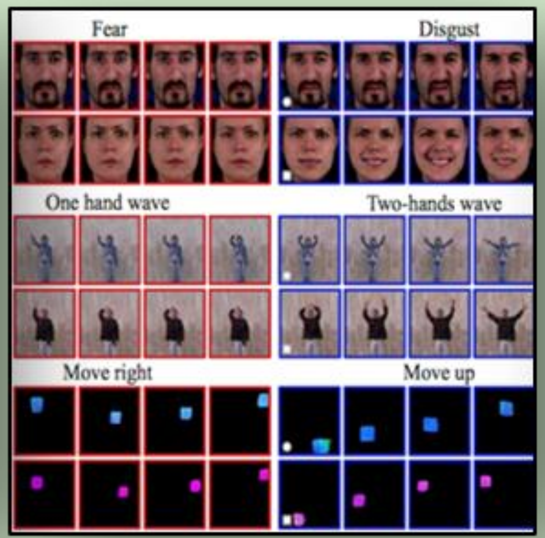
GauGAN [Park et al.]

TextGAN [Reed et al.]

StackGAN [Zhang et al.]

AttnGAN [Xu et al.]

TGAN [Saito et al.]

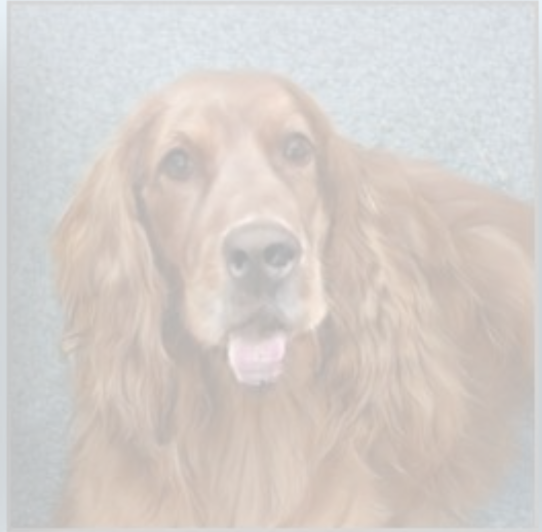


MoCoGAN [Tulyakov et al.]

vid2vid [Wang et al.]

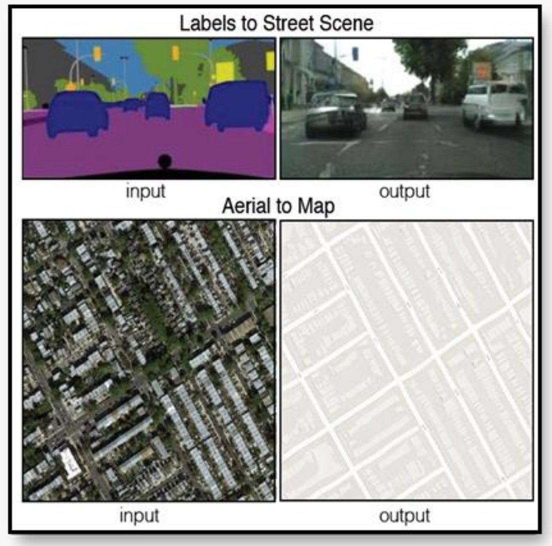
Recent Advances in Image Generation

Prog. GAN [Karras et al.]



BigGAN [Brock et al.]

StyleGAN [Karras et al.]



pix2pix [Isola et al.]

pix2pixHD [Wang et al.]

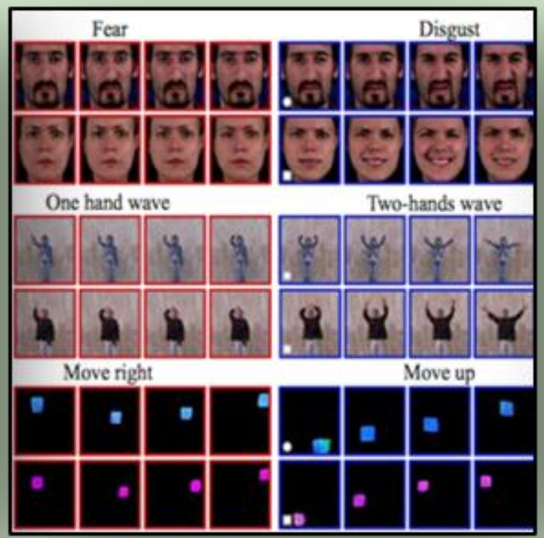
GauGAN [Park et al.]

TextGAN [Reed et al.]

StackGAN [Zhang et al.]

AttnGAN [Xu et al.]

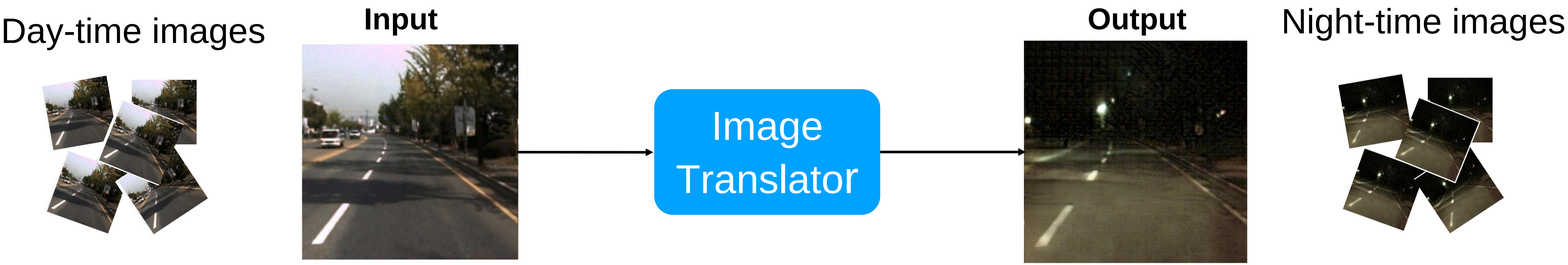
TGAN [Saito et al.]



MoCoGAN [Tulyakov et al.]

vid2vid [Wang et al.]

Image-to-Image Translation (I2I)



- Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be two different image domains
  - e.g. day-time image domain & night-time image domain
- Let $x_1 \in \mathcal{X}_1$
- I2I: the problem of translating x_1 to a *corresponding* image $x_2 \in \mathcal{X}_2$
  - Correspondence can mean different things in different contexts

Examples and Use Cases

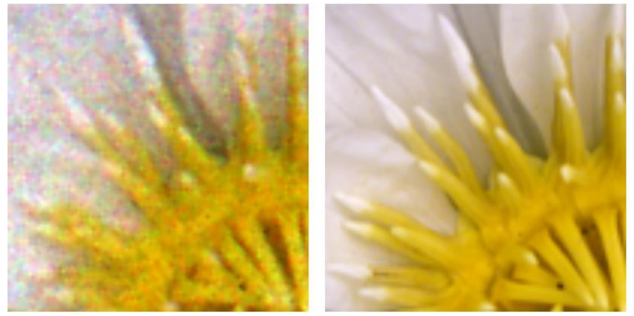
Low-res to high-res

Blurry to sharp

Thermal to color

Synthetic to real

LDR to HDR



Noisy to clean

Image to painting

Day to night

Summer to winter

- Bad weather to good weather
- Greyscale to color
- …

Prior Works

- Image translation has been studied for decades
- Different approaches have been exploited, including
  - Filtering-based
  - Optimization-based
  - Dictionary learning-based
  - Deep learning-based
  - GAN-based

Prior Works

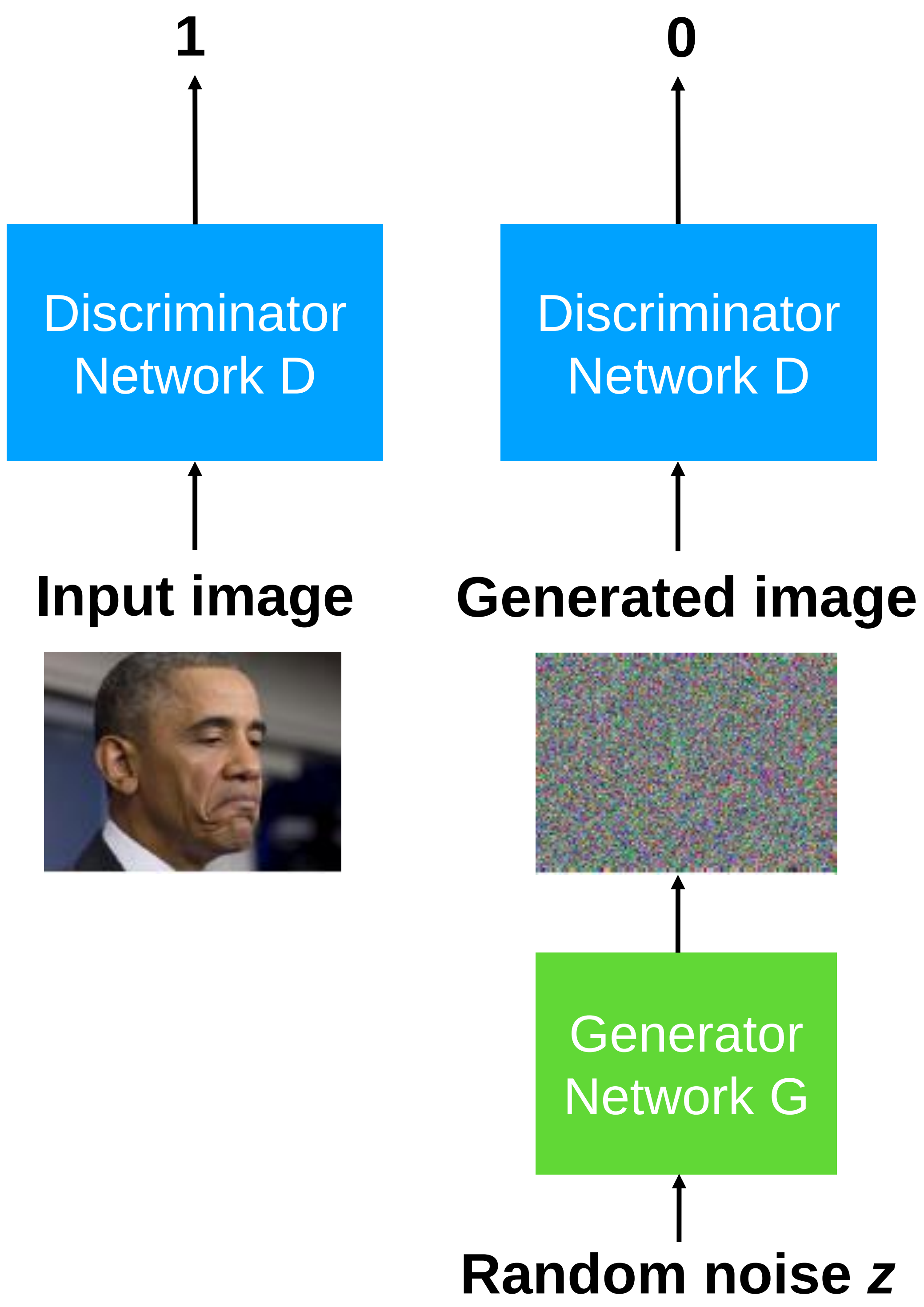
- Image translation has been studied for decades
- Different approaches have been exploited, including
  - Filtering-based
  - Optimization-based
  - Dictionary learning-based
  - Deep learning-based
  - GAN-based

Why are GANs useful for I2I?

~~**Hand-crafted features**~~ → Deep Networks

~~**Hand-crafted objective function**~~ → Generative Adversarial Networks

Generative Adversarial Networks (GANs)



- Forget about designing a perceptual loss
- Let’s train a new network to differential real and fake images

Goodfellow et al. NIPS 2014

GAN Objective

Solving a minimax problem

$$\min_G \max_D \quad E_{x \sim p_X}[\log D(x)] + E_{z \sim p_Z}[\log(1 - D(G(z)))]$$

GAN Objective

Solving a minimax problem

For discriminator *D*:

$$\min_{G}\max_{D} \quad E_{x\sim p_X}[\log \underbrace{D(x)}_{\text{real samples}}] + E_{z\sim p_Z}[\log(1 - \underbrace{D(\underbrace{G(z)}_{\text{generated samples}})}))]$$

1

1 0

GAN Objective

Solving a minimax problem

For Generator G:

$$\min_G \max_D \quad E_{x\sim p_X}[\log D(x)] + E_{z\sim p_Z}[\log(1 - D(G(z))]$$

0 1

GAN Objective

Solving a minimax problem

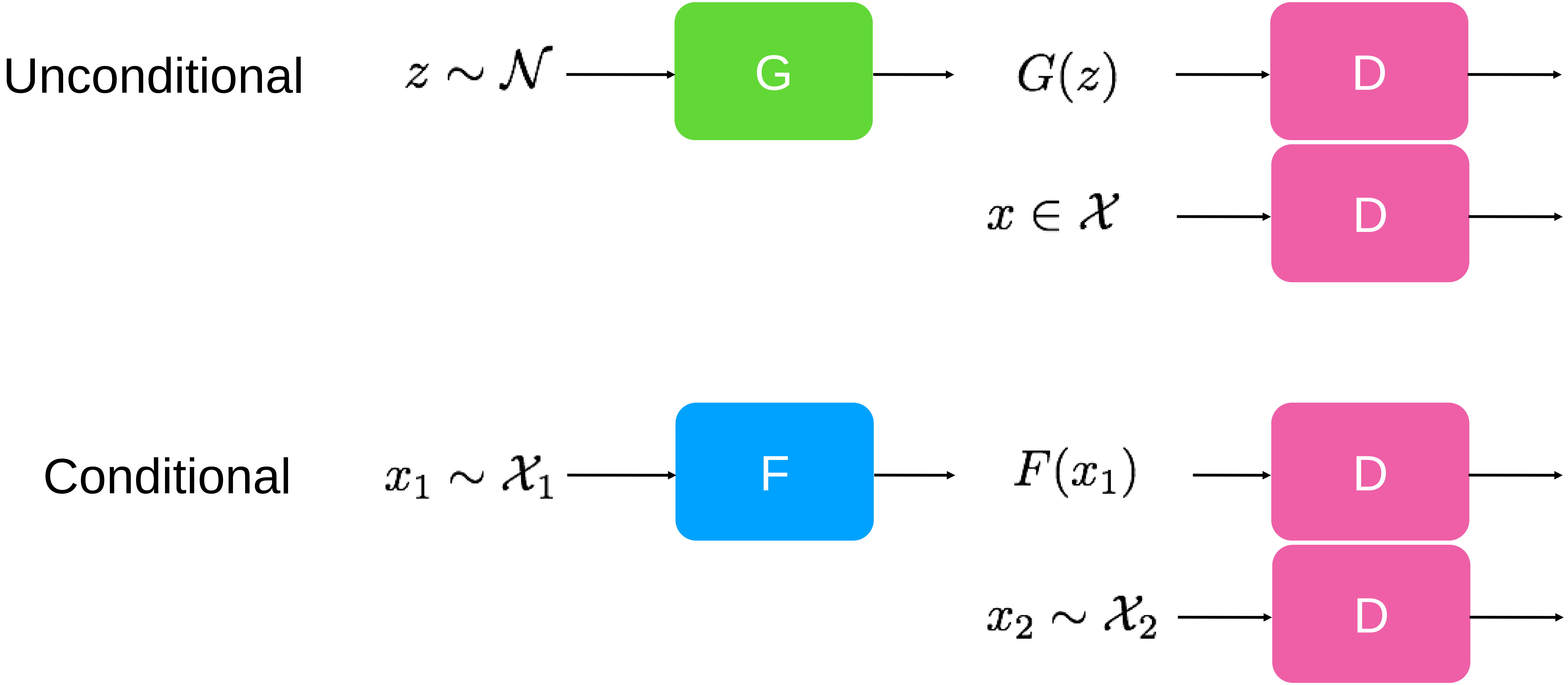
$$\min_G \max_D \quad E_{x \sim p_X}[\log D(x)] + E_{z \sim p_Z}[\log(1 - D(G(z))]$$

Training: done by alternating two stochastic gradient updates

Update G: $\max_G E_{z \sim p_{\mathcal{N}}}[\log D(G(z))]$

Update D: $\max_D E_{x \sim p_{\mathcal{X}}}[\log D(X)] + E_{z \sim p_{\mathcal{N}}}[\log(1 - D(G(z)))]$

Unconditional vs. Conditional GANs



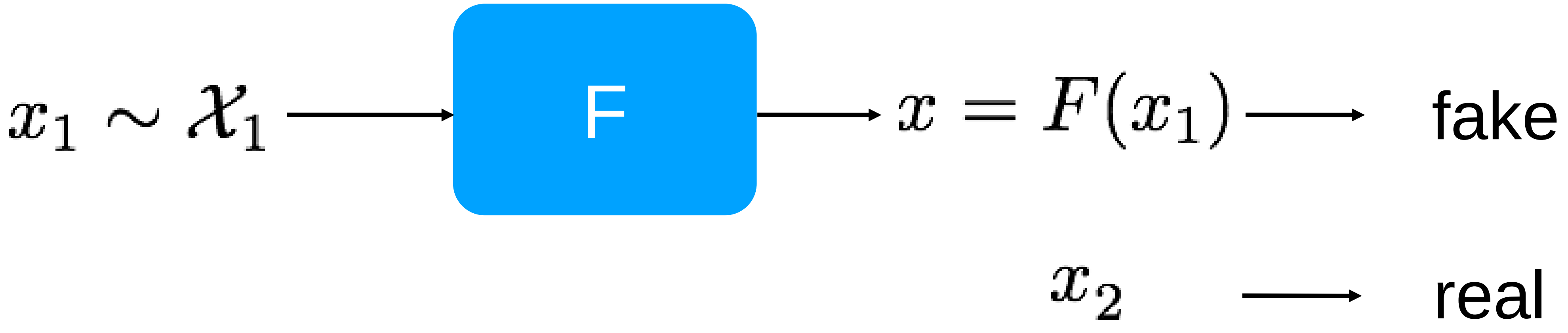
Conditional GAN for Image Translation

- Conditional GAN loss alone is <u>insufficient</u> for image translation
  - No guarantee the translated image is related to the source image
  - Generator can just completely ignore source images

- This can be easily fixed in the supervised setting
  - Where ground truth image pairs before/after translation are available

$$\text{Dataset} = \{(\underbrace{x_1^{(1)}}_{\text{input}}, \underbrace{x_2^{(1)}}_{\text{output}}), (x_1^{(2)}, x_2^{(2)}), ..., (x_1^{(N)}, x_2^{(N)})\}$$

Supervised Image-to-Image Translation

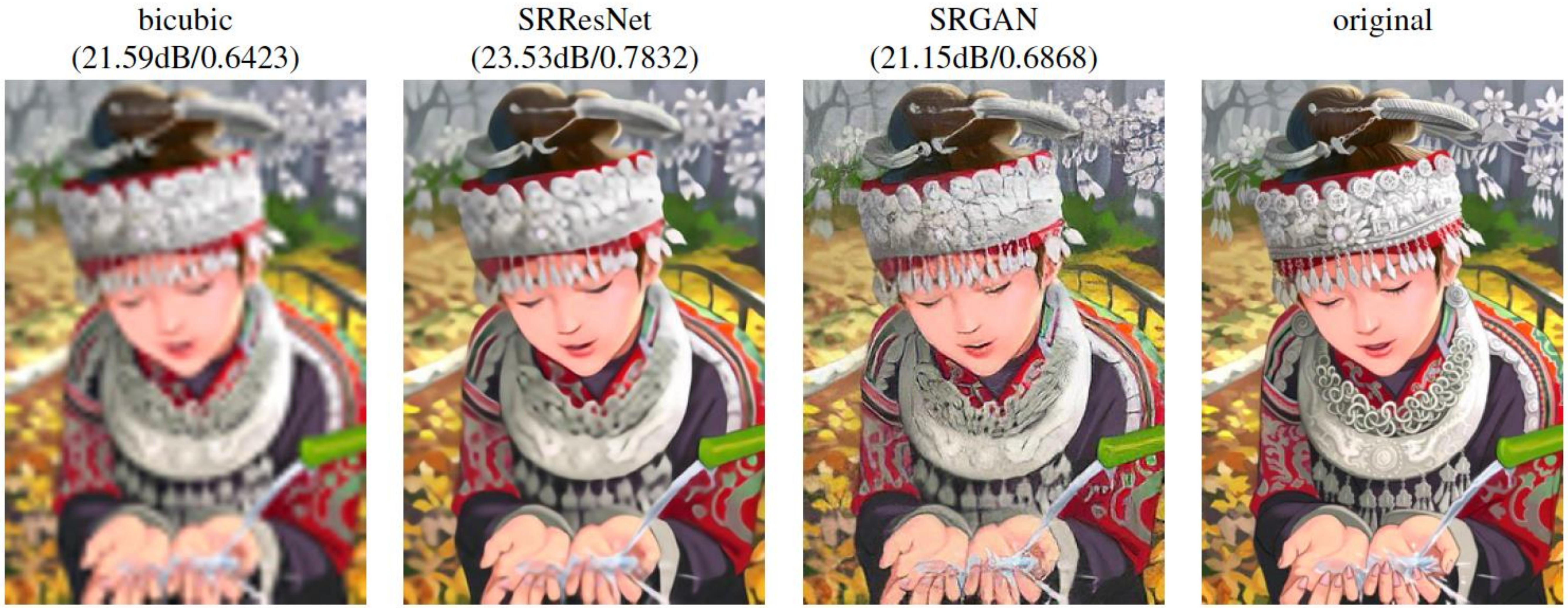


- Supervisedly relating $x = F(x_1^{(i)})$ to $x_2^{(i)}$
  - Ledig et al (CVPR'17): Adding content loss

$$||x - x_2^{(i)}||_2 + ||\mathrm{VGG}(x) - \mathrm{VGG}(x_2^{(i)})||_2$$

SRGAN

C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi “Photo-realistic image superresolution using a generative adversarial networks “, CVPR 2017



Supervised Image-to-image Translation

$x_1 \sim \mathcal{X}_1 \longrightarrow \boxed{\text{F}} \longrightarrow x = F(x_1) \longrightarrow$ fake

$x_2 \longrightarrow$ real

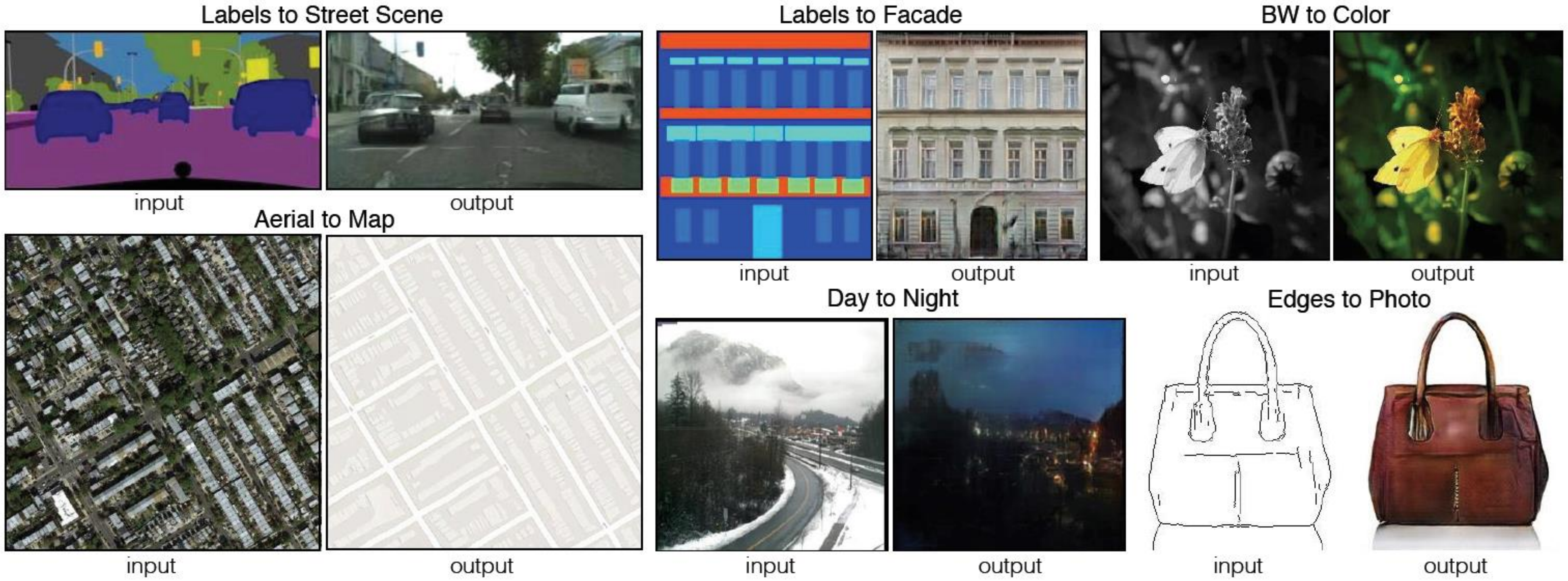
- Supervisedly relating $x = F(x_1^{(i)})$ to $x_2^{(i)}$
  - Isola et al (CVPR'17): Learning a joint distribution

Discriminator sees both input and output images

$$\max_F E_{p_{\mathcal{X}_1}}[\log(D(x_1, F(x_1)))]$$

Pix2Pix

P. Isola, J. Zhu, T. Zhou, A. Efros “Image-to-image translation with conditional generative networks“, CVPR 2017



Unsupervised Image-to-image Translation

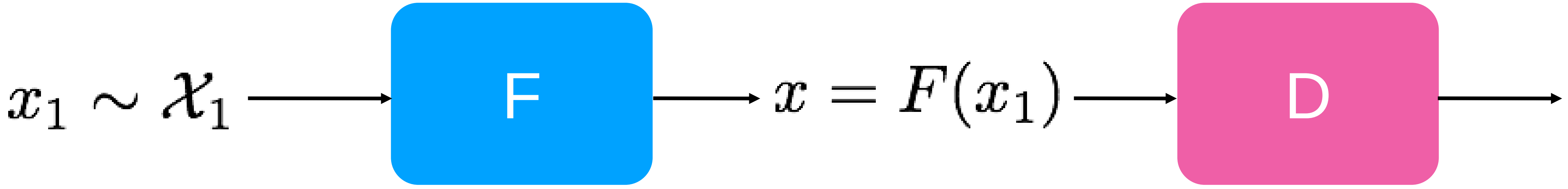
- Corresponding images could be expensive to get
- In the unsupervised setting
  - No correspondence between the two datasets

$$\text{Dataset}_1 = \{x_1^{(n_1)}, x_1^{(n_2)}, ..., x_1^{(n_N)}\}$$

$$\text{Dataset}_2 = \{x_2^{(m_1)}, x_2^{(m_2)}, ..., x_2^{(m_M)}\}$$

- Need additional constraints/assumptions for learning the translation

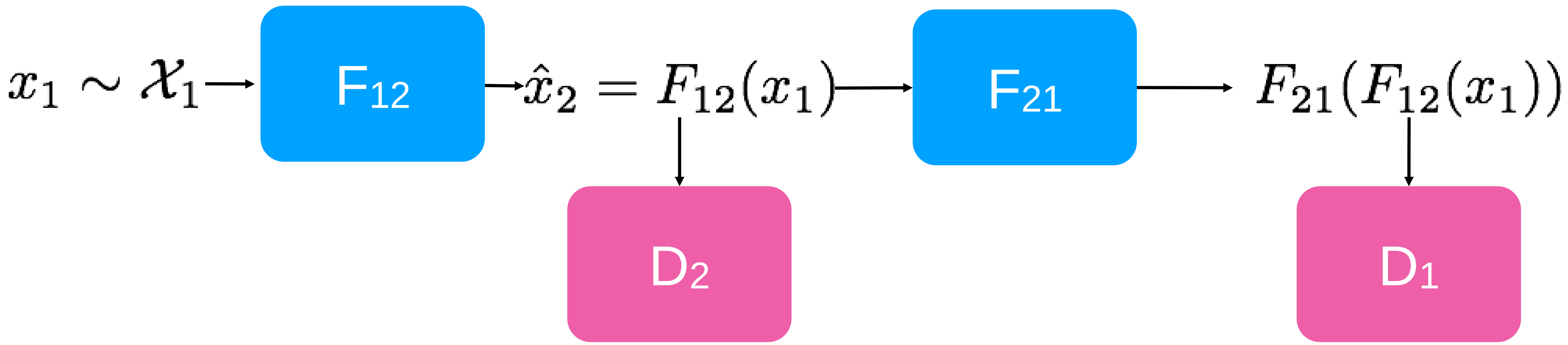
SimGAN



- Srivastava et al. (CVPR’17): adding cross-domain content loss

$$\max_F E_{p_{\mathcal{X}_1}} [\log D(F(x_1)) - \lambda ||F(x_1) - x_1||_1]$$

Cycle Constraint



- Learning a two-way translation
- DiscoGAN by Kim et al. (ICML'17)
- CycleGAN by Zhu et al. (ICCV'17)

$$\max_{F_{12},F_{21}} E_{p_{\mathcal{X}_1}}[\log(D_2(F_{12}(x_1)) - \lambda||F_{21}(F_{12}(x_1)) - x_1||_p^p]+$$
$$E_{p_{\mathcal{X}_2}}[\log(D_1(F_{21}(x_2)) - \lambda||F_{12}(F_{21}(x_2)) - x_2||_p^p]$$

CycleGAN: Unsupervised Image-to-Image Translation

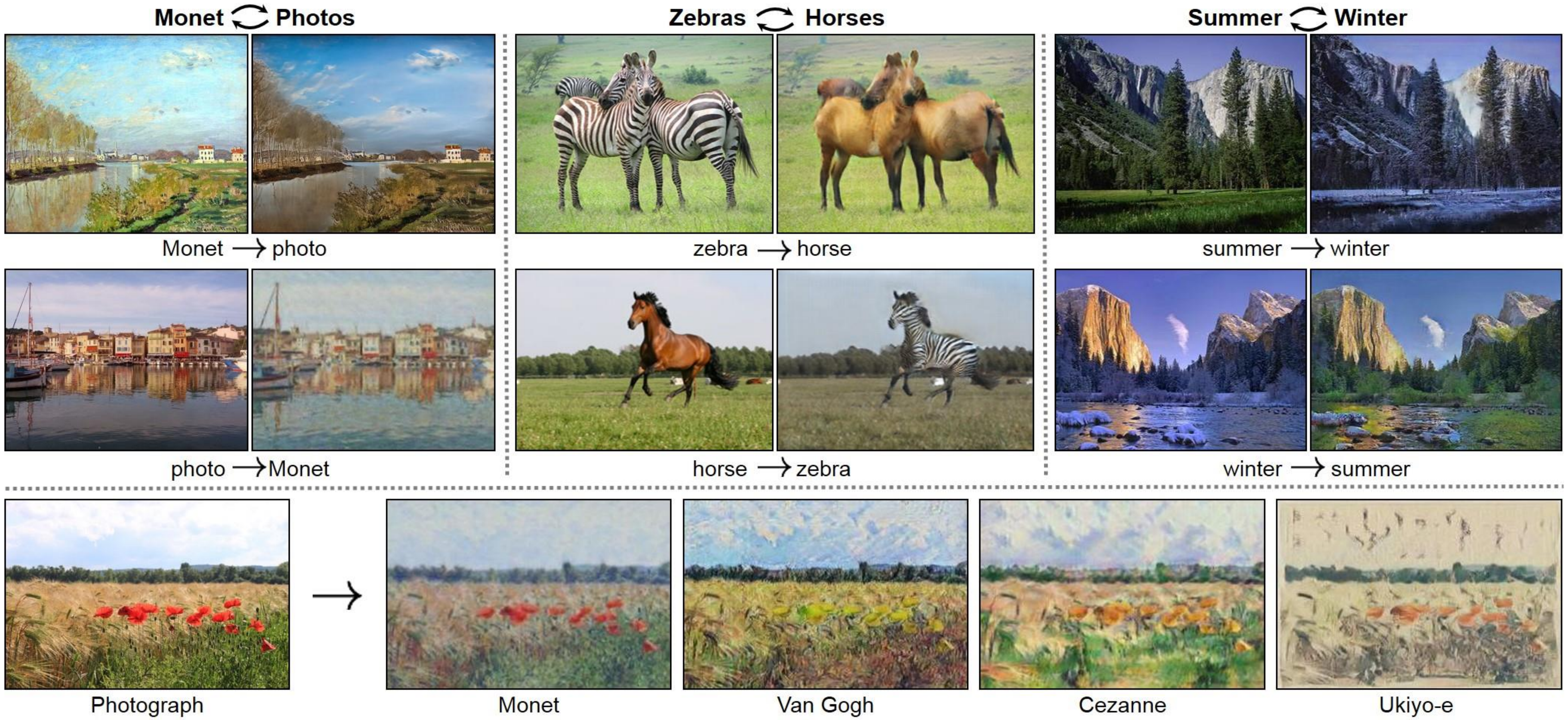


Image Translation Results

Results on Unsupervised Thermal-Image-to-RGB-Image Translation. Left: input thermal image. Right: Output color image.

Results on Unsupervised RGB-Image-to-Thermal-Image Translation. Left: input color image. Right: Output thermal image.

Results on Unsupervised Day-Image-to-Night-Image Translation. Left: input day image. Right: Output night image.

Results on Unsupervised Night-Image-to-Day-Image Translation. Left: input night image. Right: Output day image.

Image Translation Results

Results on Unsupervised Sunny-Image-to-Rainy-Image Translation. Left: input sunny image. Right: Output rainy image.

Results on Unsupervised Rainy-Image-to-Sunny-Image Translation. Left: input rainy image. Right: Output sunny image.

Back View | Front View | Left View | Right View

Foggy image to clear sky image

Attribute-based Face Image Translation

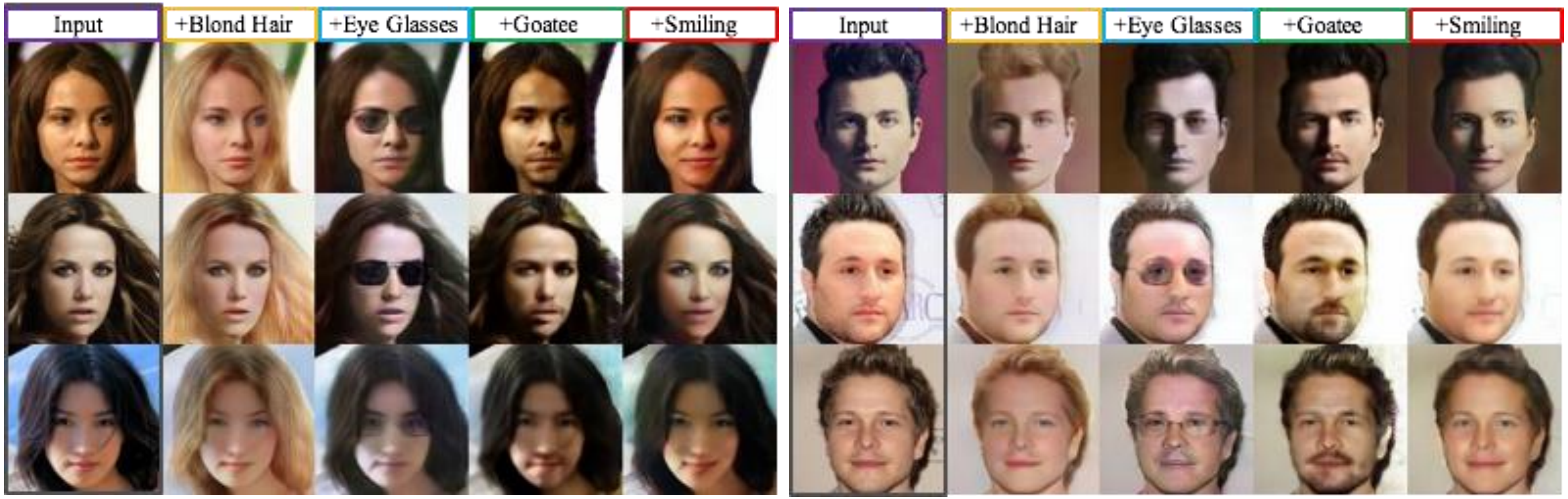


Image Translation Results

Input +Blondhair

Input +Eyeglasses

Input -Goatee

Input -Smiling

Image Translation Results

Input **+Blondhair**

Input **+Eyeglasses**

Input **+Goatee**

Input **+Smiling**

Improving GAN Training

Tricks

- Label smoothing
- Historical batches
- …

New objectives

- EBGAN
- LSGAN
- WGAN
- BEGAN
- …

Surrogate or auxiliary objective

- UnrolledGAN
- WGAN-GP
- DRAGAN
- …

Network architecture

- LAPGAN
- Stacked GAN

Wasserstein GAN

M. Arjovsky, S. Chintala, L. Bottou. "Wasserstein GAN." 2016

Replace classifier with a critic function

Discriminator

GAN $$\max_D E_{x \sim p_X}[\log D(x)] + E_{z \sim p_Z}[\log(1 - D(G(z))]$$

WGAN $$\max_D E_{x \sim p_X}[D(x)] - E_{z \sim p_Z}[D(G(z))]$$

Generator

GAN $$\max_G E_{z \sim p_Z}[\log D(G(z))]$$

WGAN $$\max_G E_{z \sim p_Z}[D(G(z))]$$

Wasserstein GAN

M. Arjovsky, S. Chintala, L. Bottou. “Wasserstein GAN.” 2016

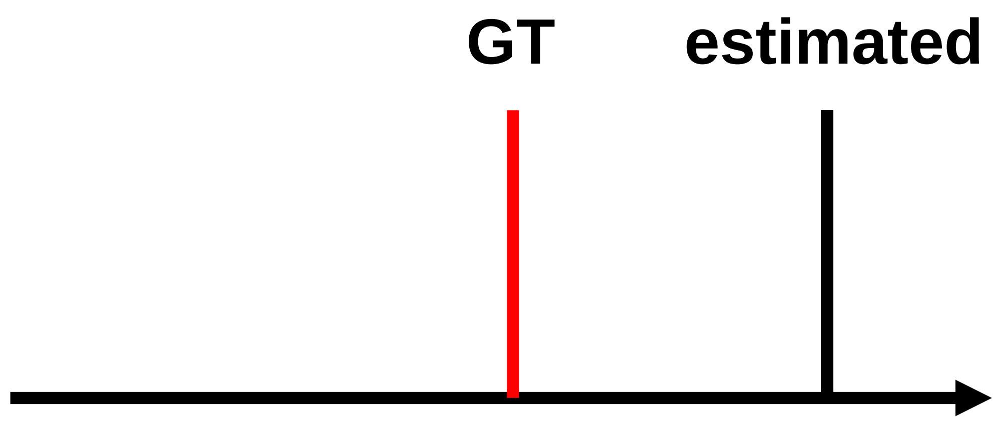
GAN: minimize Jensen-Shannon divergence between p_X and $p_{G(Z)}$

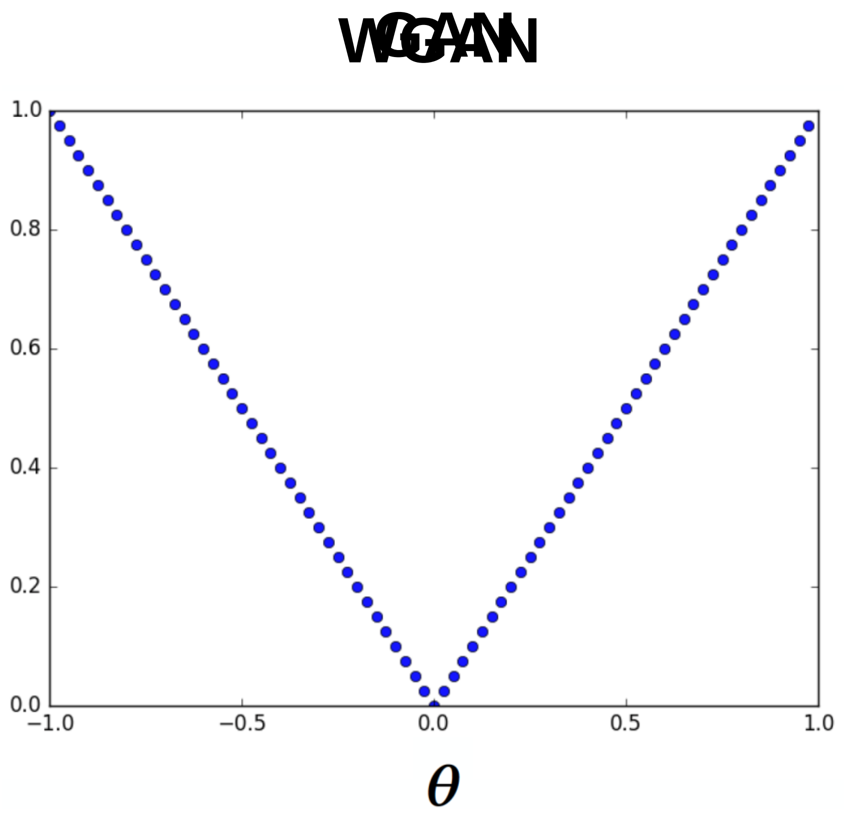
$$JS(p_X||p_{G(Z)}) = KL(p_X||\frac{p_X + p_{G(Z)}}{2}) + KL(p_{G(Z)}||\frac{p_X + p_{G(Z)}}{2})$$

WGAN: minimize earth mover distance between p_X and $p_{G(Z)}$

$$EM(p_X, p_{G(Z)}) = \inf_{\gamma \in \prod(p_X, p_{G(Z)})} E_{(x,y)\sim\gamma}[||x-y||]$$

GAN vs. WGAN





- In this example
  - GAN
    - uniform (JS) distance across all space
    - gradient = 0
  - WGAN
    - smaller (EM) distance when closer to GT
    - has gradient toward GT

Disadvantage of WGAN

- Needs to ensure discriminator is 1-Lipschitz

$$||D(x) - D(y)|| \leq K||x - y||$$

  - i.e., gradient is bounded everywhere and doesn't explode
- Realized by weight clipping

WGAN-GP

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Domoulin, A. Courville.
“Improved Training of Wasserstein GANs.” 2017

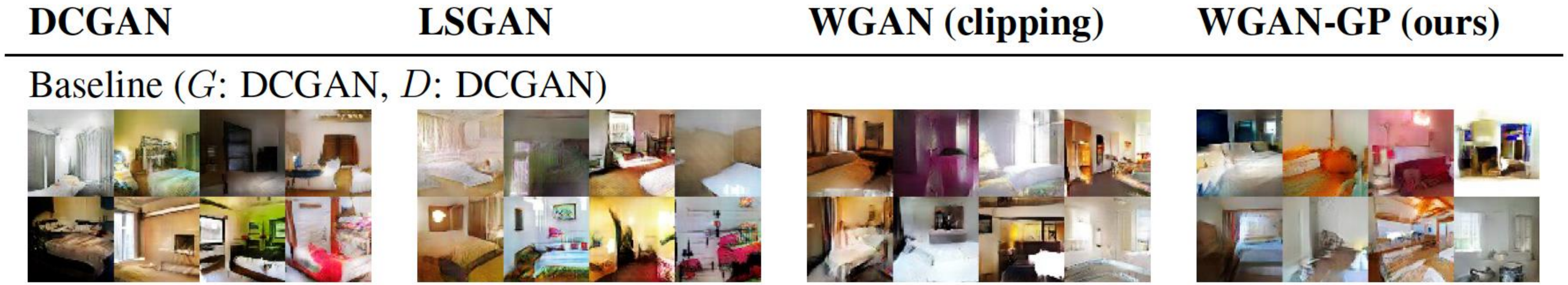
Instead of weight clipping, apply ***gradient penalty***

$$\min_G \max_D E_{x \sim p_X}[D(x)] - E_{z \sim p_Z}[D(G(Z))] + \boxed{\lambda E_{y \sim p_Y}[(||\nabla_y D(y)||_2 - 1)^2]}$$

$$y = ux + (1-u)G(z)$$

y: imaginary samples (between real and fake)

Optimal critic has unit gradient norm almost everywhere



Spectral Normalization

T. Miyato, T. Kataoka, M. Koyama, Y. Yoshida. "Spectral Normalization for Generative Adversarial Networks." *2018.*

- Lipschitz constant of a linear function is its largest singular value (*spectral norm)*

$$||Ax|| \leq K||x||$$

- Spectral normalization: Replaces every weight W with W / σ(W)
  - σ: largest singular value of W
  - Ensures discriminator gradient is always bounded
- Computing σ during training
  - Direct computation is very time consuming
  - Fast approximation using power iteration

Evaluation Metrics

- Inception Score (IS)
  - Each generated image should have a distinct label
  - Overall set of generated images should have diverse labels
  - The larger the distance between these two, the better

- Fréchet Inception Distance (FID)
  - Use inception network to extract features from images
  - Model real/fake features with two multivariate Gaussians
  - The lower the distance between these two, the better

- Human Evaluation