StarGAN V1, V2

StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation

StarGAN v2: Diverse Image Synthesis for Multiple Domains

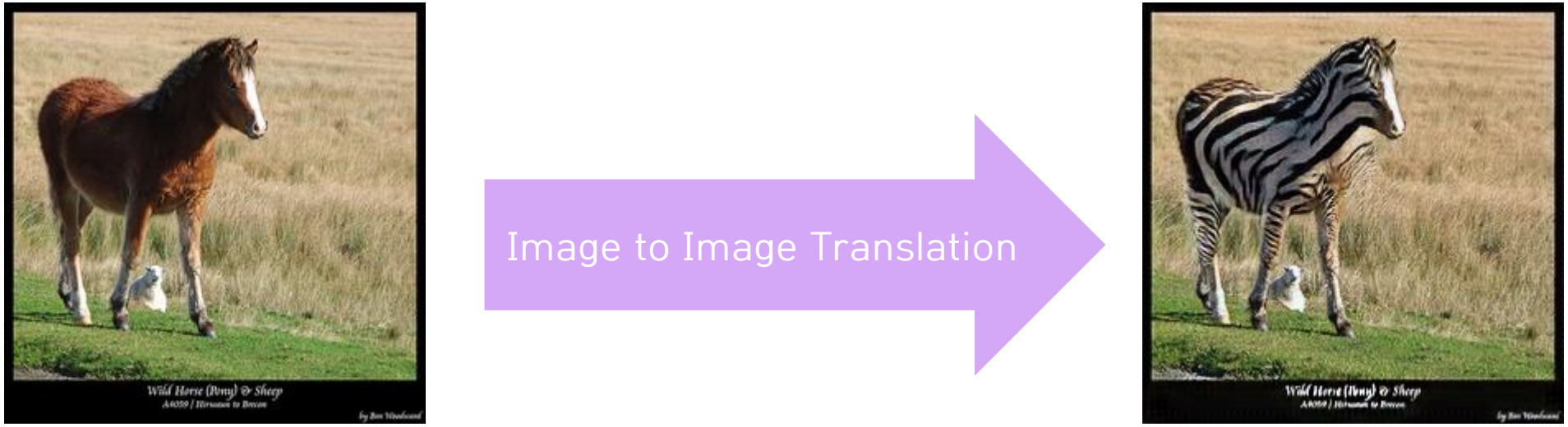
석사과정 김 진용

index

Background

Image-to-Image Translation 이란?

https://www.youtube.com/watch?v=Ko31fYGT20Y&t=407s



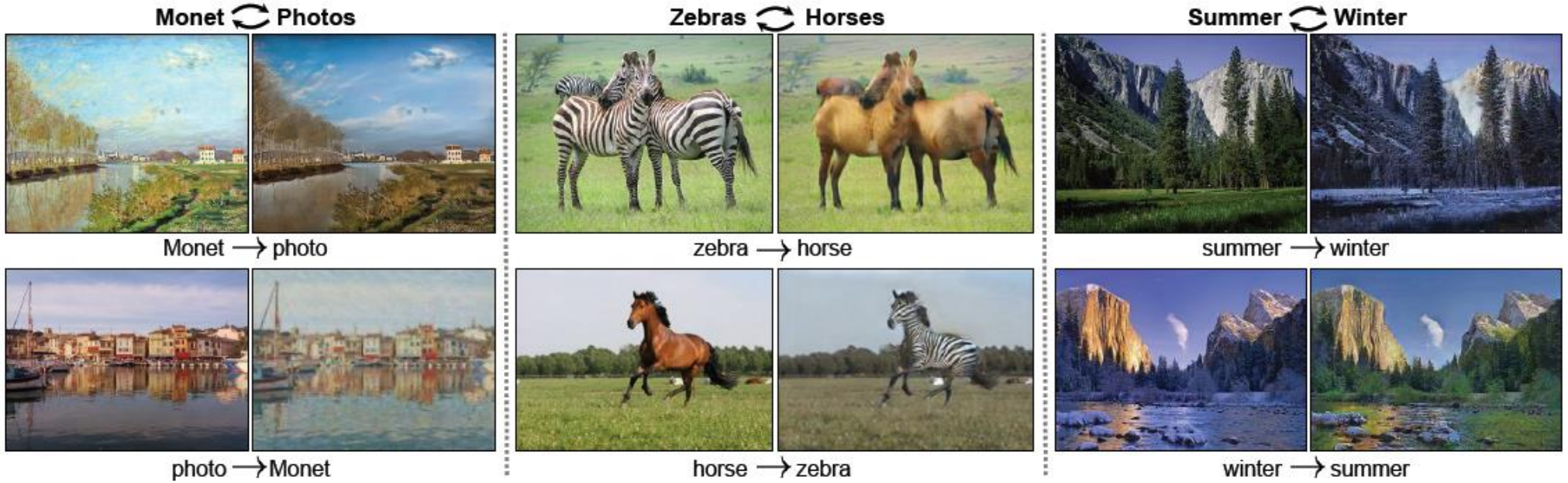


Image-to-Image Translation 이란?

https://www.youtube.com/watch?v=Ko31fYGT20Y&t=407s

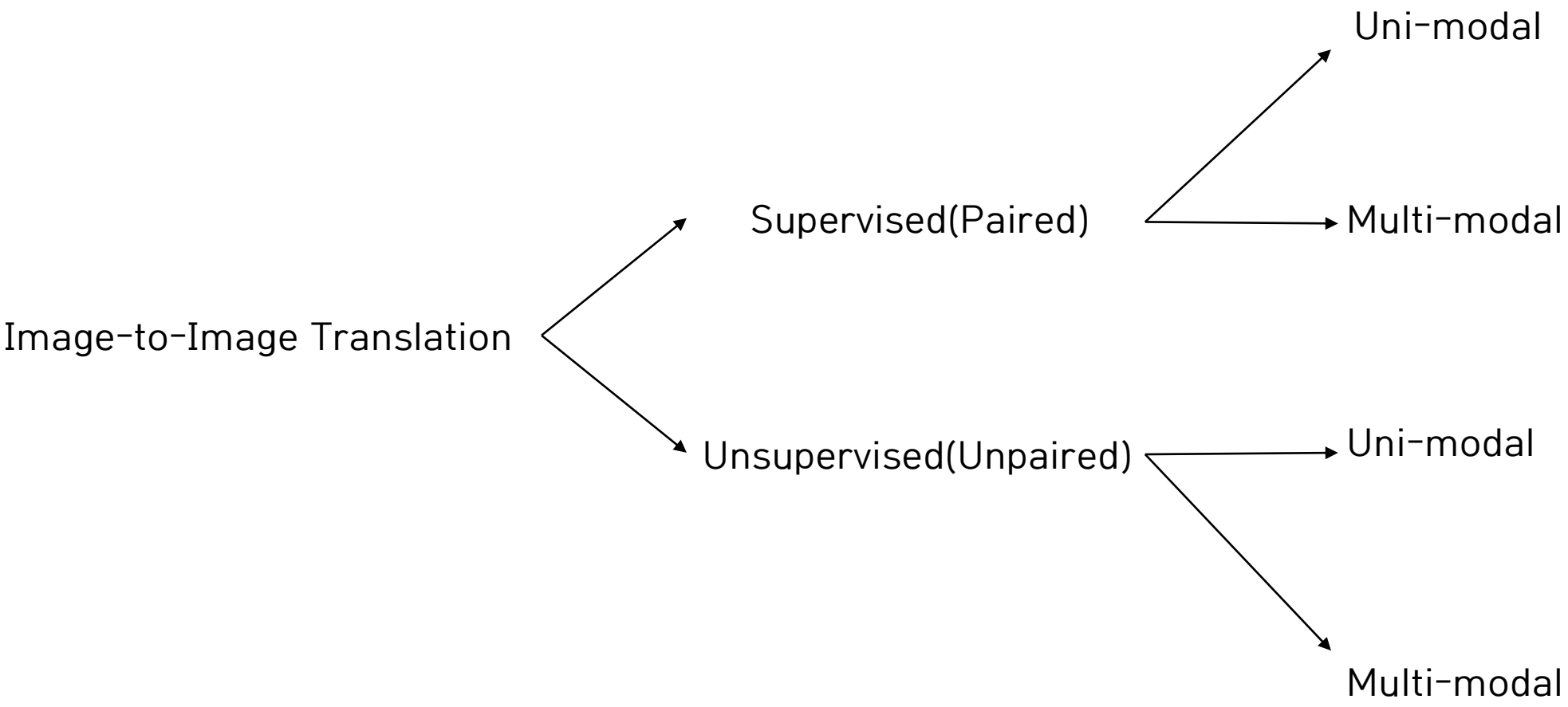


Image-to-Image Translation 이란?

https://www.youtube.com/watch?v=Ko31fYGT20Y&t=407s

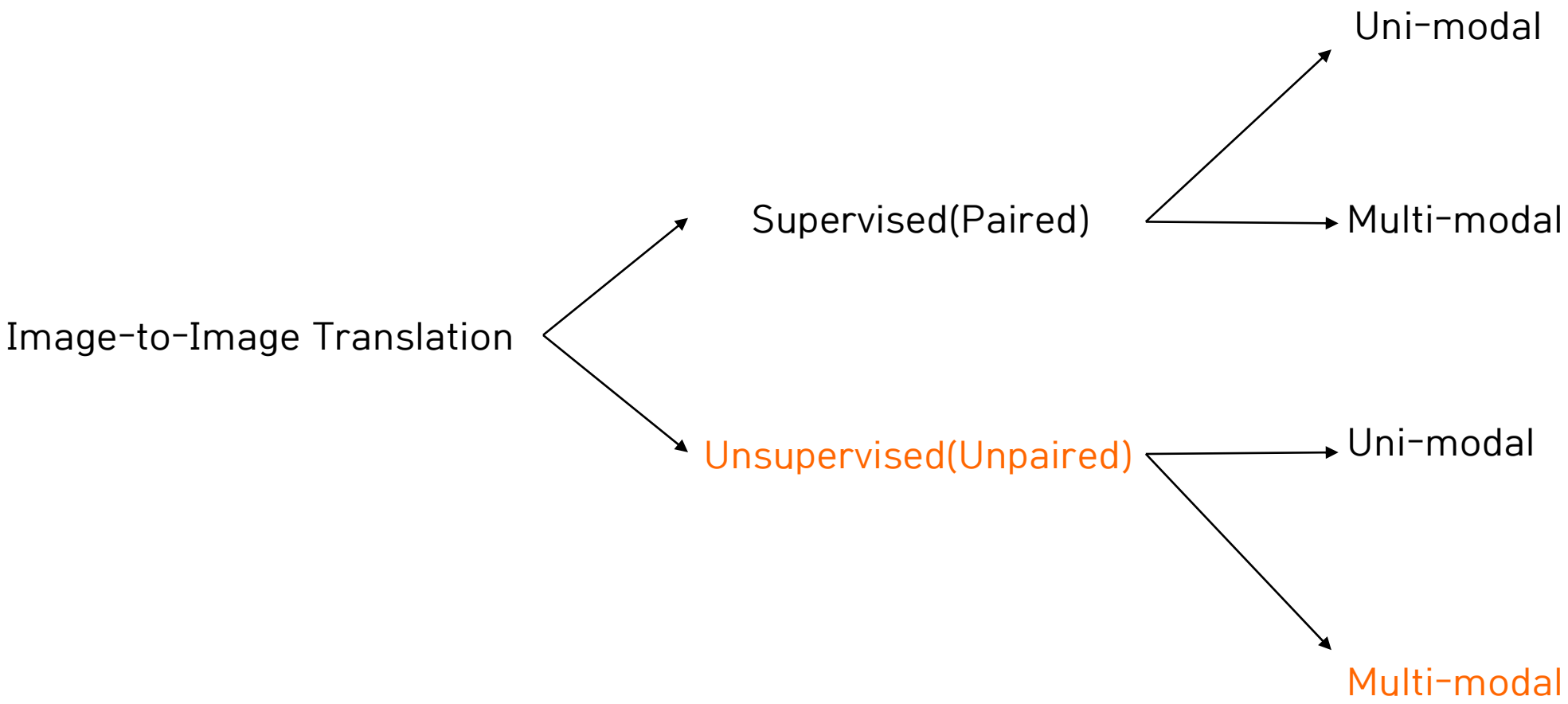


Image-to-Image Translation 이란?

Unpaired dataset의 이점

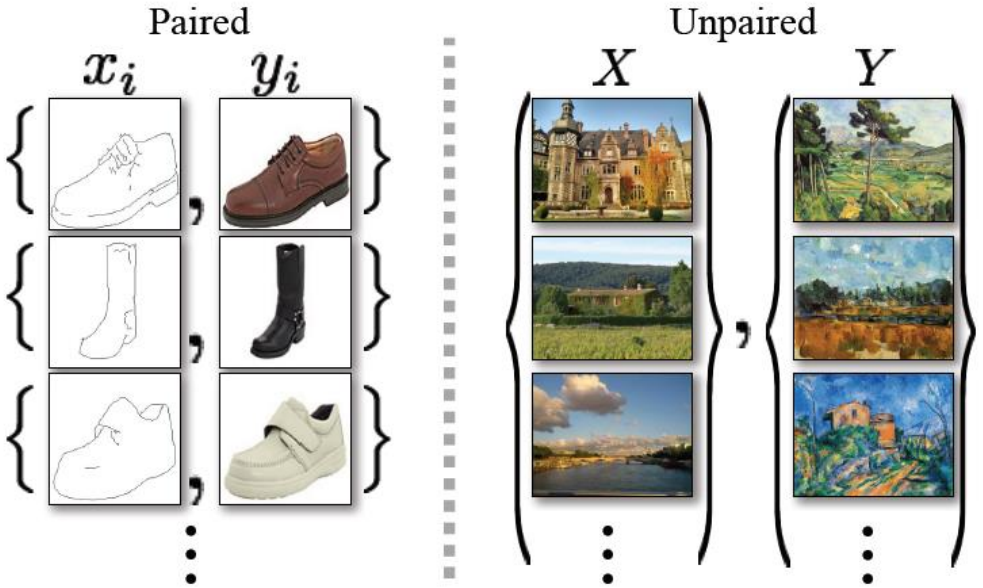


Figure 2: *Paired* training data (left) consists of training examples $\{x_i, y_i\}_{i=1}^N$, where the correspondence between x_i and y_i exists [22]. We instead consider *unpaired* training data (right), consisting of a source set $\{x_i\}_{i=1}^N$ $(x_i \in X)$ and a target set $\{y_j\}_{j=1}$ $(y_j \in Y)$, with no information provided as to which x_i matches which y_j.

- 비교적 구하기 쉽다.
- pair를 맞출 필요 없기 때문에 cost가 낮다.

Image-to-Image Translation 이란?

Multi-modal

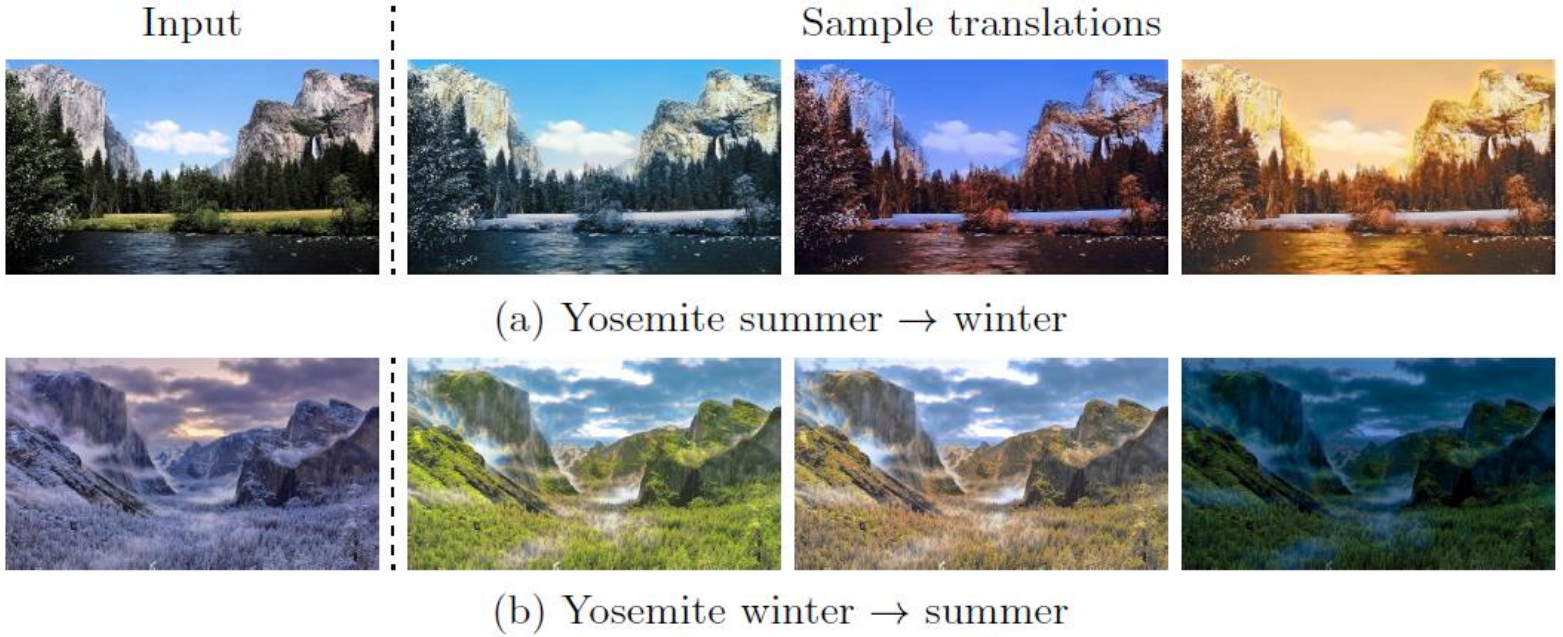


Fig. 8. Example results on Yosemite summer ↔ winter (HD resolution).

1개의 Input -> N개의 Output

Image-to-Image Translation 이란?

CycleGAN

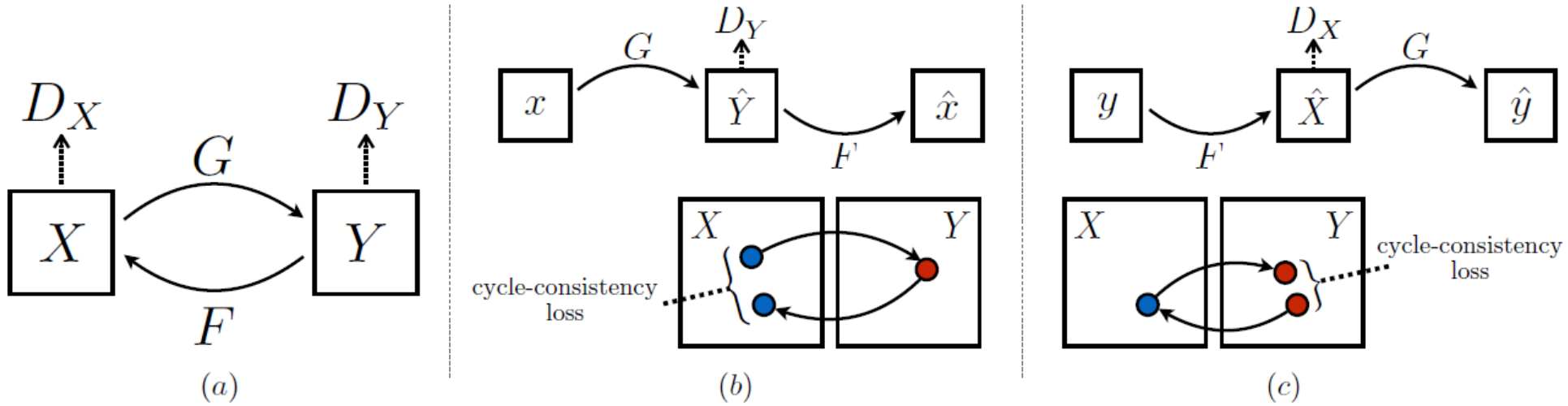


Figure 3: (a) Our model contains two mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and associated adversarial discriminators D_Y and D_X. D_Y encourages G to translate X into outputs indistinguishable from domain Y, and vice versa for D_X and F. To further regularize the mappings, we introduce two *cycle consistency losses* that capture the intuition that if we translate from one domain to the other and back again we should arrive at where we started: (b) forward cycle-consistency loss: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, and (c) backward cycle-consistency loss: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$

G : X -> Y ,F : Y - > X 맵핑을 통해 G(X) , F(Y) 가짜 이미지를 만들어냄

Image-to-Image Translation 이란?

CycleGAN

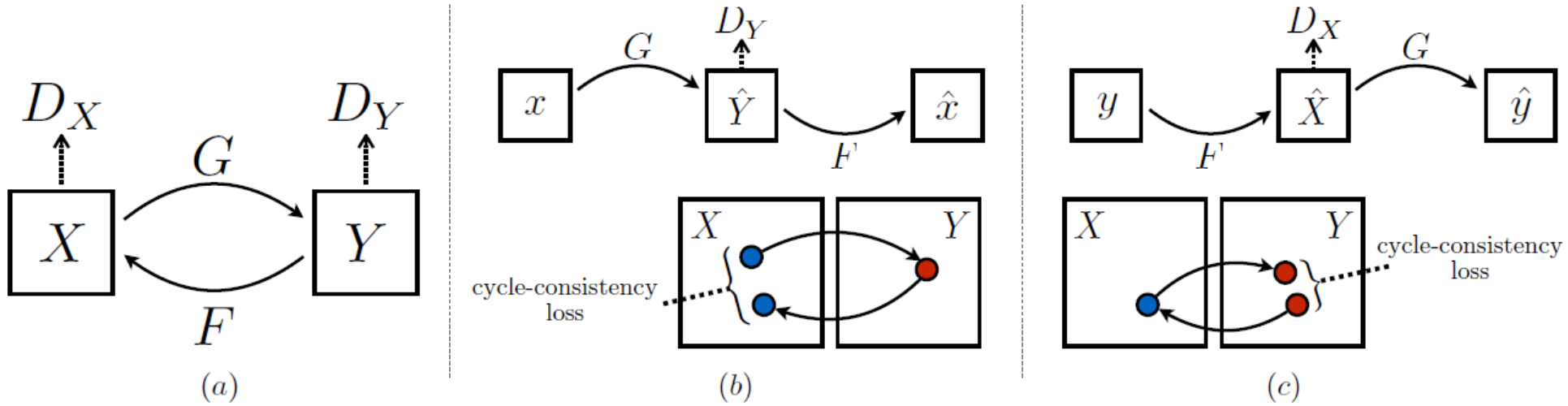
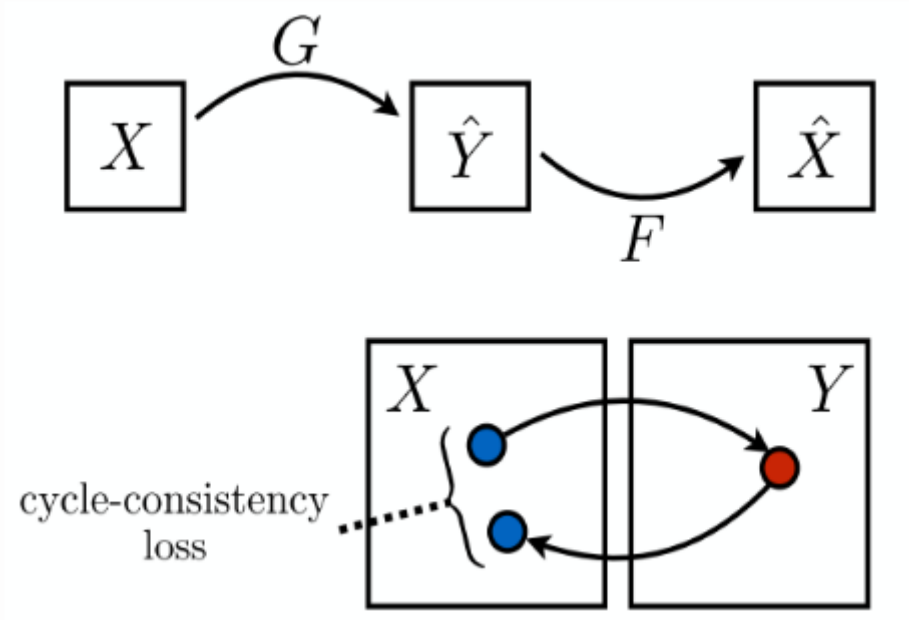


Figure 3: (a) Our model contains two mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and associated adversarial discriminators D_Y and D_X. D_Y encourages G to translate X into outputs indistinguishable from domain Y, and vice versa for D_X and F. To further regularize the mappings, we introduce two *cycle consistency losses* that capture the intuition that if we translate from one domain to the other and back again we should arrive at where we started: (b) forward cycle-consistency loss: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, and (c) backward cycle-consistency loss: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$

만약, G()에서 Y로 넘어가는 것만을 감안한다면 D_y 만을 속이려고 할 것이고
의미없는 매핑이 될 수 있다.
Solution -> **Cycle Consistency**

Image-to-Image Translation 이란?

CycleGAN



$$Loss_{x \to y} = \mathbb{E}_y[log(D_y(y))] + \mathbb{E}_x[log(1 - D_y(G(x)))] + \mathbb{E}_x[\|F(G(x)) - x\|_1]$$

$$Loss_{y \to x} = \mathbb{E}_x[log(D_x(x))] + \mathbb{E}_y[log(1 - D_x(F(y)))] + \mathbb{E}_y[\|G(F(y)) - y\|_1]$$

Cycle Consistency(순환 일관성)을 주어 생성자 2개를 거쳐 본래 도메인으로 돌아옴을 목표로 함
즉, F(G(X)) = X 가 목표

Image-to-Image Translation 이란?

CycleGAN

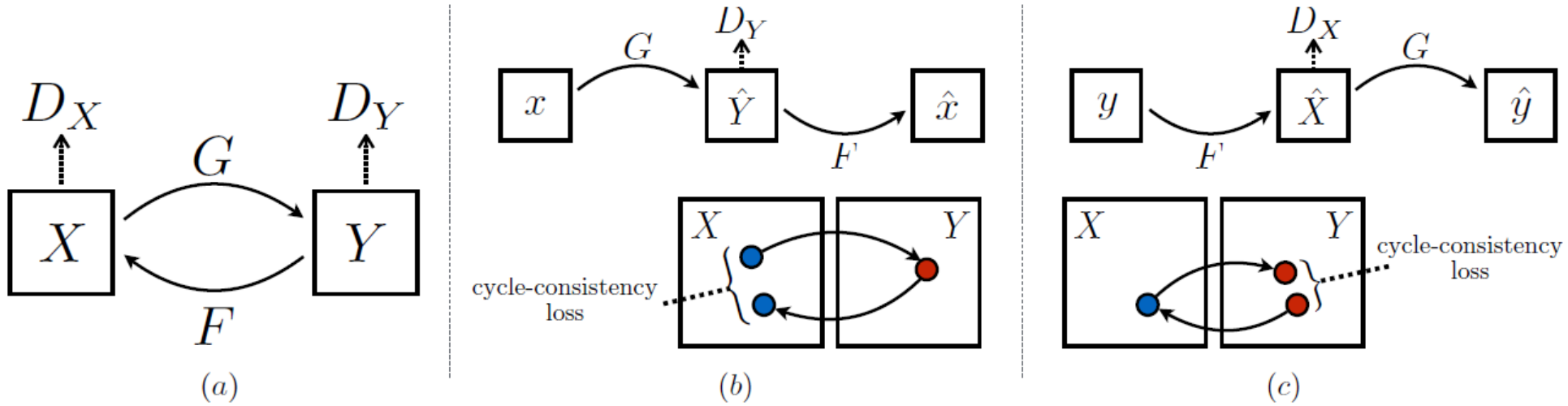


Figure 3: (a) Our model contains two mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and associated adversarial discriminators D_Y and D_X. D_Y encourages G to translate X into outputs indistinguishable from domain Y, and vice versa for D_X and F. To further regularize the mappings, we introduce two *cycle consistency losses* that capture the intuition that if we translate from one domain to the other and back again we should arrive at where we started: (b) forward cycle-consistency loss: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, and (c) backward cycle-consistency loss: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$

그러나! 치명적 약점

여러개의 이미지(Multi-modal) 생성 시 target label 만큼 학습해줘야 해서 매우매우 불리하다.

StarGAN v1

StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation

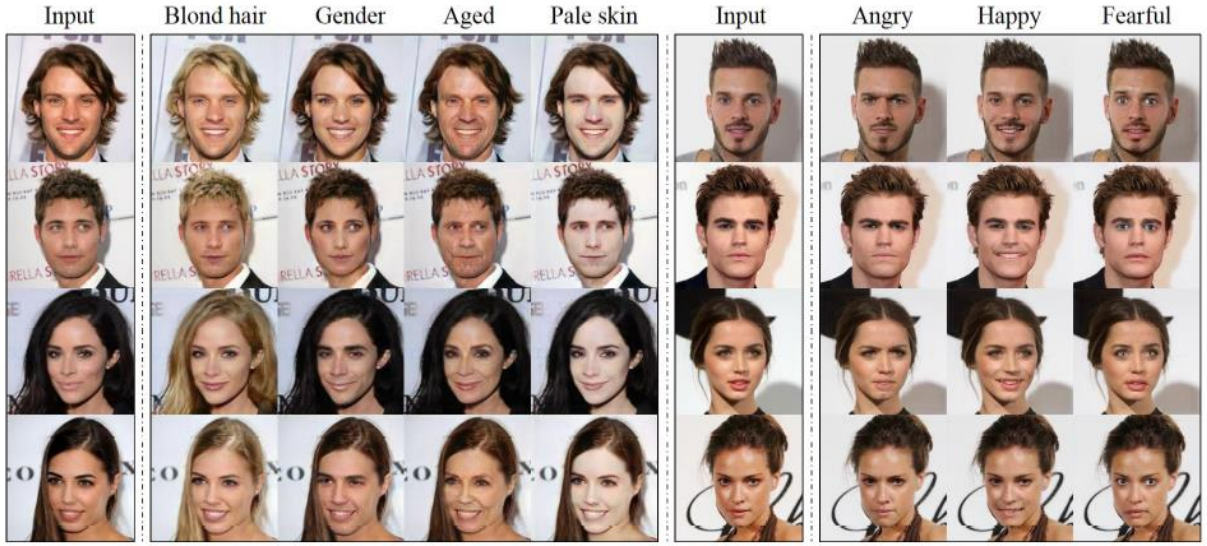
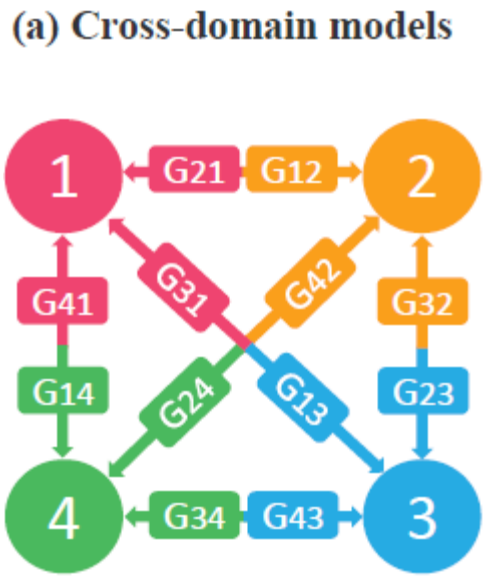


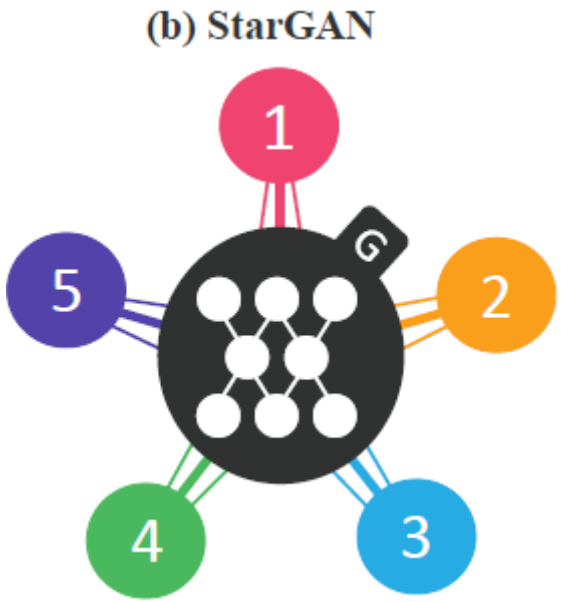
Figure 1. Multi-domain image-to-image translation results on the CelebA dataset via transferring knowledge learned from the RaFD dataset. The first and sixth columns show input images while the remaining columns are images generated by StarGAN. Note that the images are generated by a single generator network, and facial expression labels such as angry, happy, and fearful are from RaFD, not CelebA.



CycleGAN으로 Multi-modal의 이미지를 생성하기 위해서는 여러개의 Generator를 학습해야한다.
예를들어 4개의 attribute를 생성하는 Generator를 학습한다고 하면 총 12개를 학습해야함

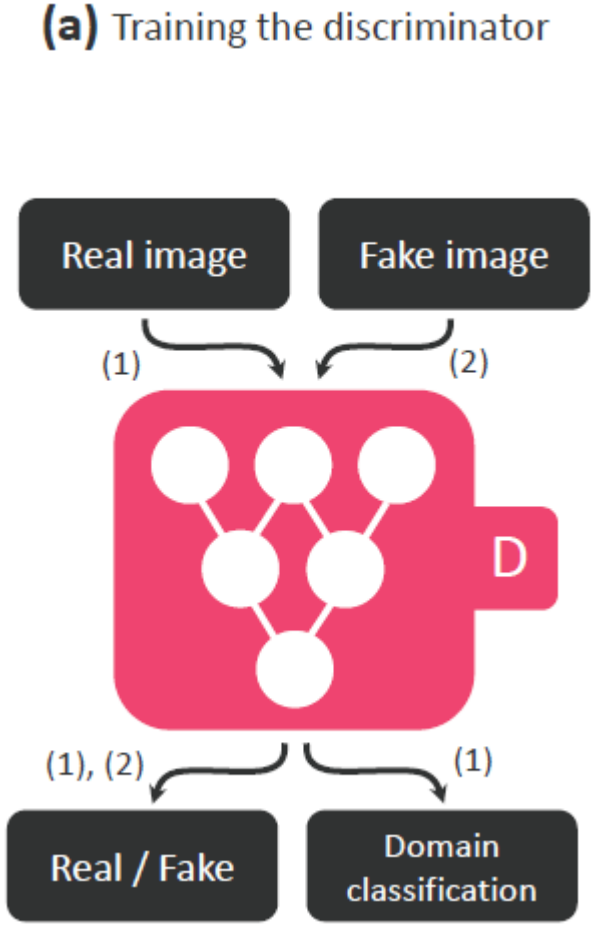
-> K(K-1)

AI Lab.
ARTIFICIAL INTELLIGENCE LAB.



이 문제를 해결하기 위해 1개의 Generator로 여러 Domain을 학습해서 생성할 수 있는 StarGAN 제안

Proposed Mehtod : StarGAN

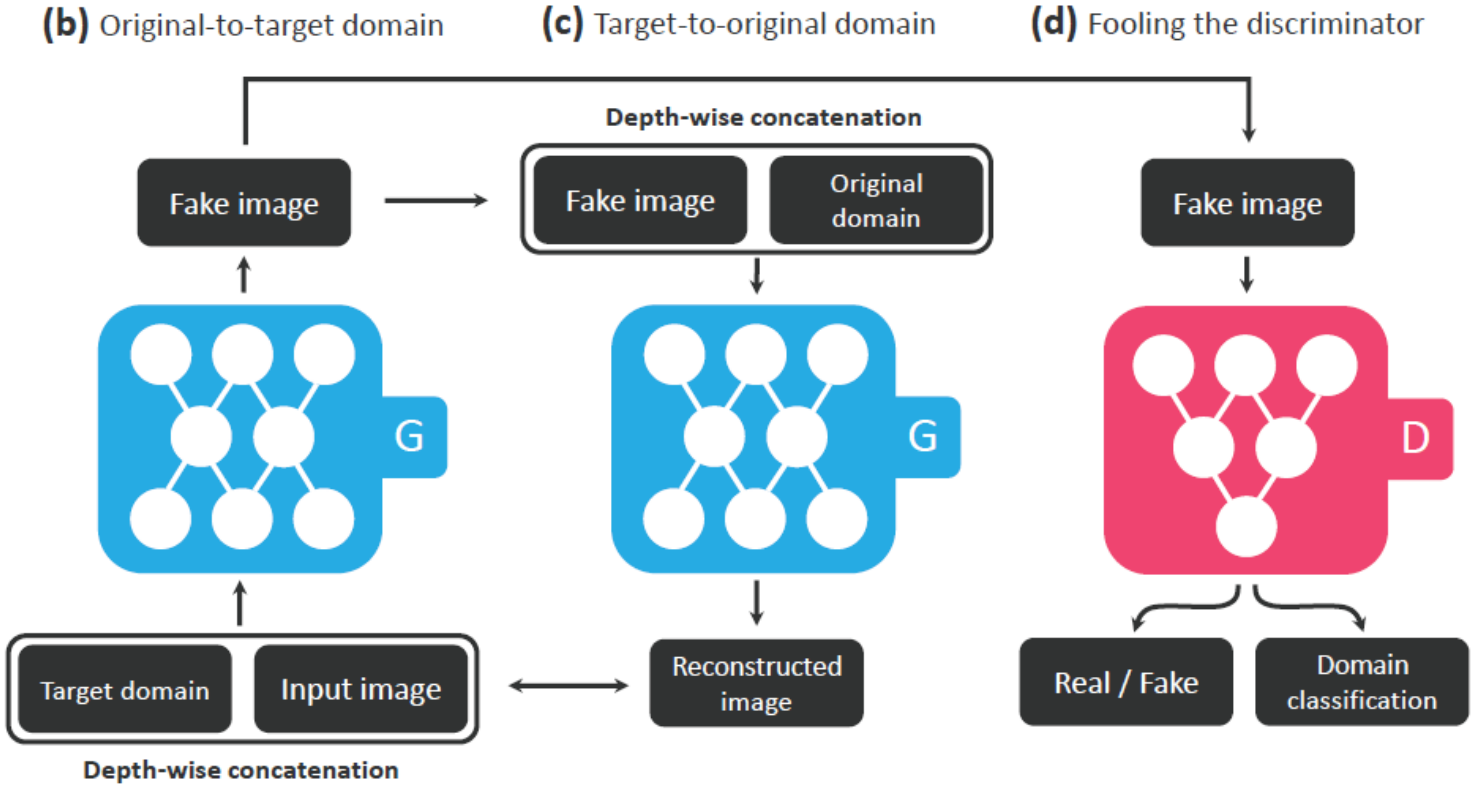


- 기존 GAN의 Discriminator와 마찬가지로 판별을 의도로 Real과 Fake를 구분하기 위한 학습

- 다른점은 Real/Fake 뿐만 아니라 Domain에 대한 Classification도 진행함.

- Multi Domain이 가능하게 된 이유

Proposed Mehtod : StarGAN



- Generator의 학습은 Original to target(X -> Y)와 Target to Original(G(X) -> X) 두 가지 이루어짐 (Cycle Consistency)

- Image 뿐만 아니라 Domain도 들어가기 때문에 guided한 특성이 없지 않음.

Proposed Mehtod : StarGAN

Objective Full function

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls}\,\mathcal{L}^r_{cls},$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls}\,\mathcal{L}^f_{cls} + \lambda_{rec}\,\mathcal{L}_{rec},$$

Domain Classification Loss

$$\mathcal{L}^r_{cls} = \mathbb{E}_{x,c'}[-\log D_{cls}(c'|x)],$$ Real : Original Domain | Original Image

$$\mathcal{L}^f_{cls} = \mathbb{E}_{x,c}[-\log D_{cls}(c|G(x,c))].$$ Fake : Target Domain | Fake Image

Experiments : StarGAN

CelebA dataset

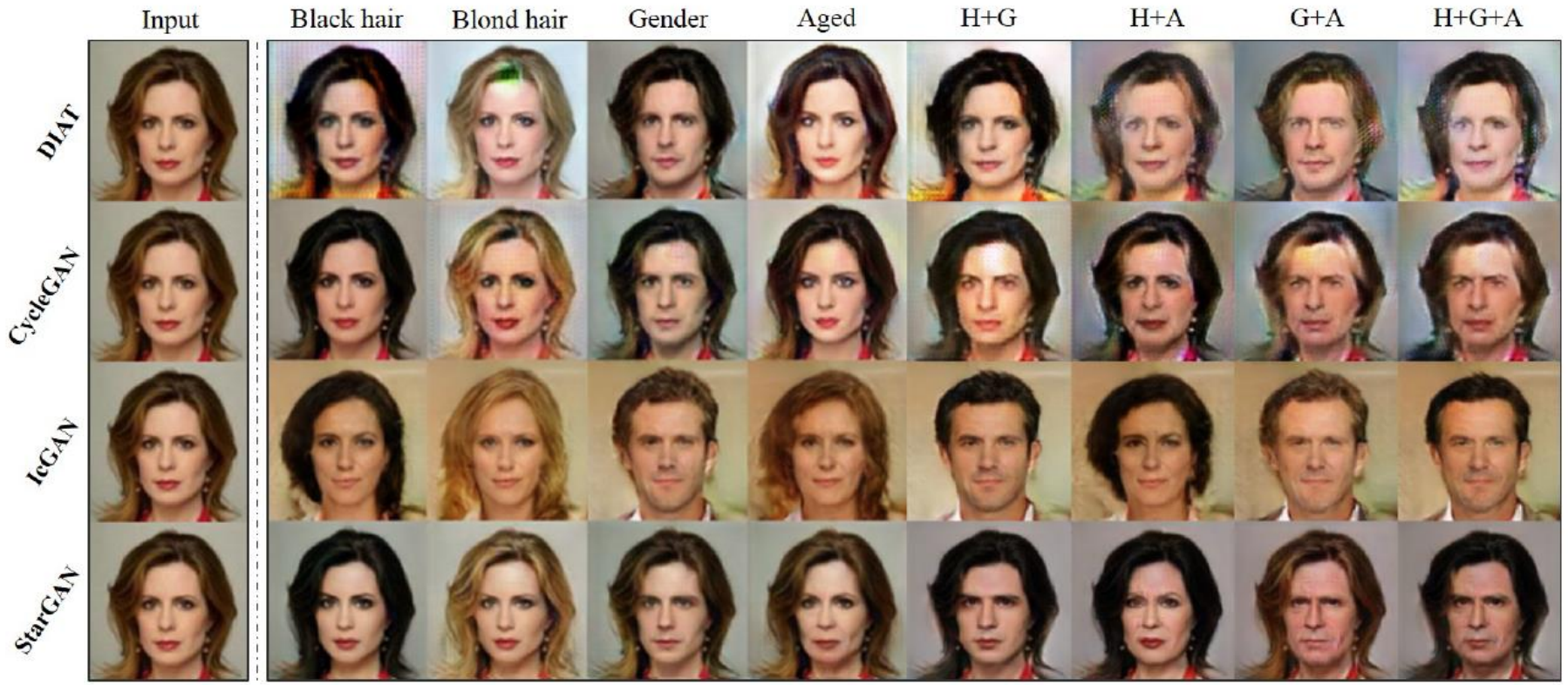


Figure 4. Facial attribute transfer results on the CelebA dataset. The first column shows the input image, next four columns show the single attribute transfer results, and rightmost columns show the multi-attribute transfer results. H: Hair color, G: Gender, A: Aged.

Experiments : StarGAN

RaFD dataset

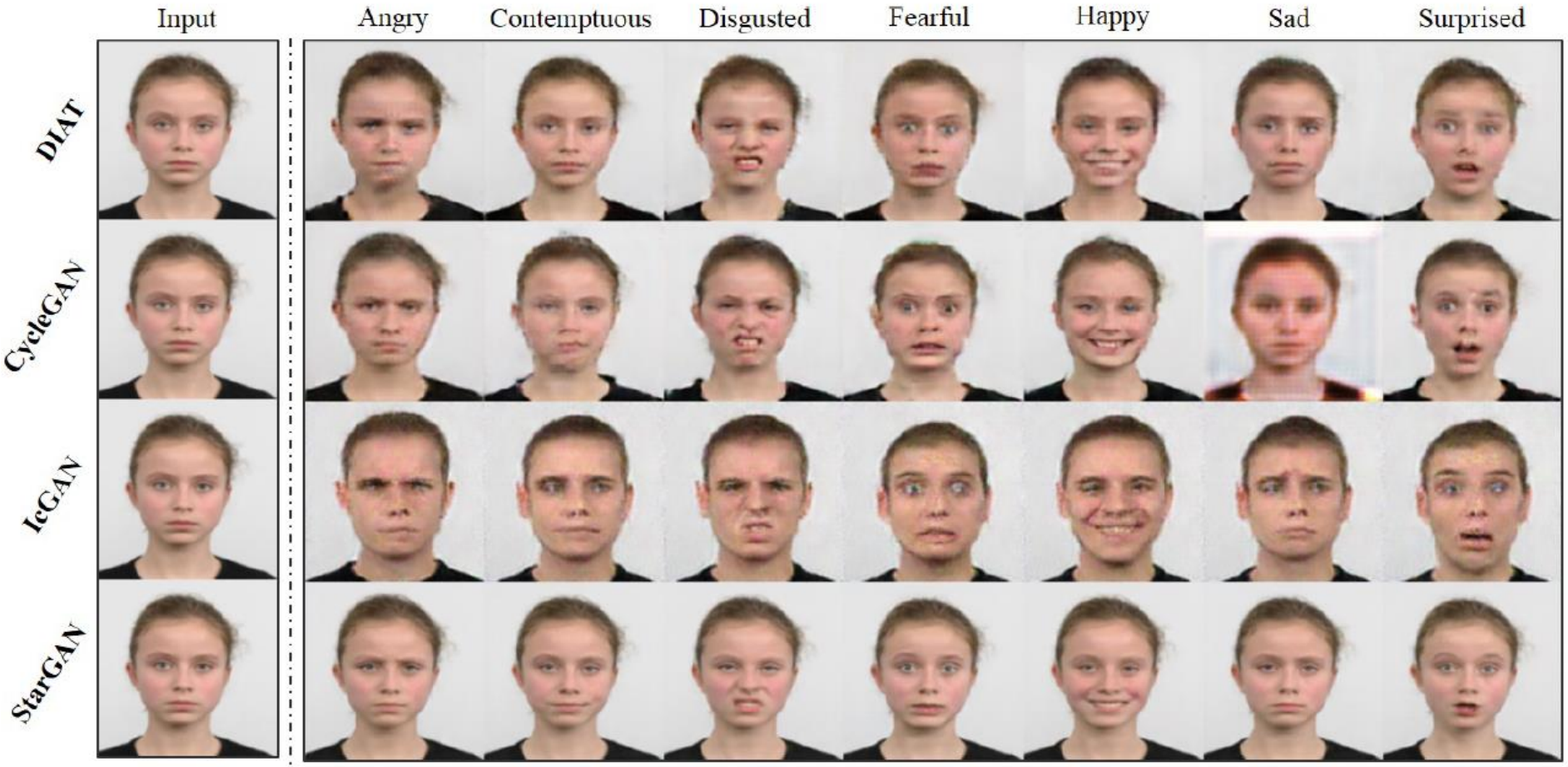


Figure 5. Facial expression synthesis results on the RaFD dataset.

Experiments : StarGAN

AMT preference result

Method	Hair color	Gender	Aged
DIAT	9.3%	31.4%	6.9%
CycleGAN	20.0%	16.6%	13.3%
IcGAN	4.5%	12.9%	9.2%
StarGAN	**66.2%**	**39.1%**	**70.6%**

Table 1. AMT perceptual evaluation for ranking different models on a single attribute transfer task. Each column sums to 100%.

Method	H+G	H+A	G+A	H+G+A
DIAT	20.4%	15.6%	18.7%	15.6%
CycleGAN	14.0%	12.0%	11.2%	11.9%
IcGAN	18.2%	10.9%	20.3%	20.3%
StarGAN	**47.4%**	**61.5%**	**49.8%**	**52.2%**

Table 2. AMT perceptual evaluation for ranking different models on a multi-attribute transfer task. H: Hair color; G: Gender; A: Aged.

다른 모델들보다 높은 선호도 나타남.

Experiments : StarGAN

Classification acc & Parameter (Complexity)

Method	Classification error	# of parameters
DIAT	4.10	52.6M × 7
CycleGAN	5.99	52.6M × 14
IcGAN	8.07	67.8M × 1
StarGAN	**2.12**	**53.2M × 1**
Real images	0.45	-

Table 3. Classification errors [%] and the number of parameters on RaFD dataset.

여러 도메인임에도 모델이 하나만 사용되기 때문에 획기적으로 줄어든 parameter를 볼 수 있음

https://arxiv.org/abs/1711.09020

StarGAN v2

StarGAN v2: Diverse Image Synthesis for Multiple Domains

StarGAN V2 : intro

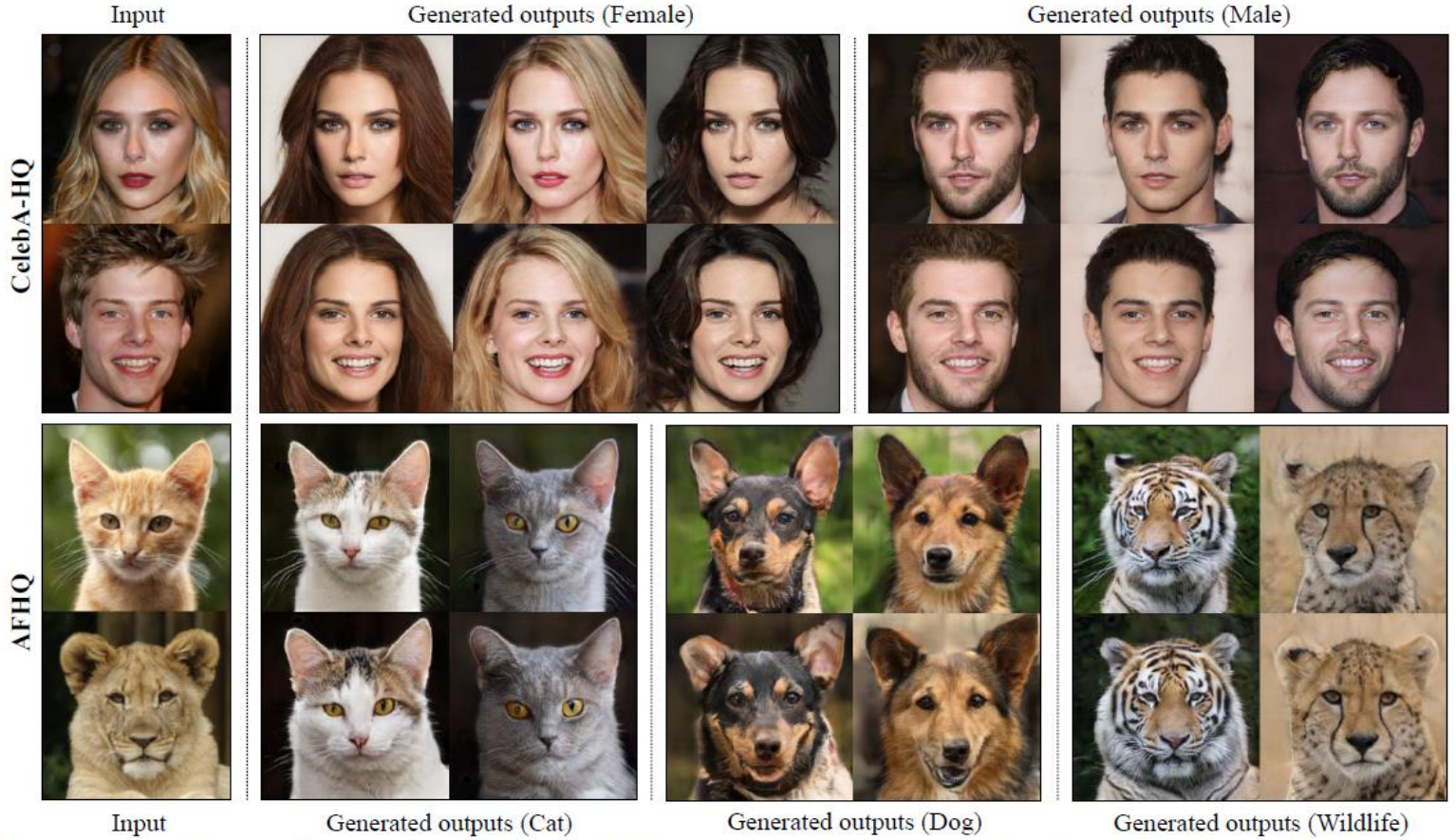
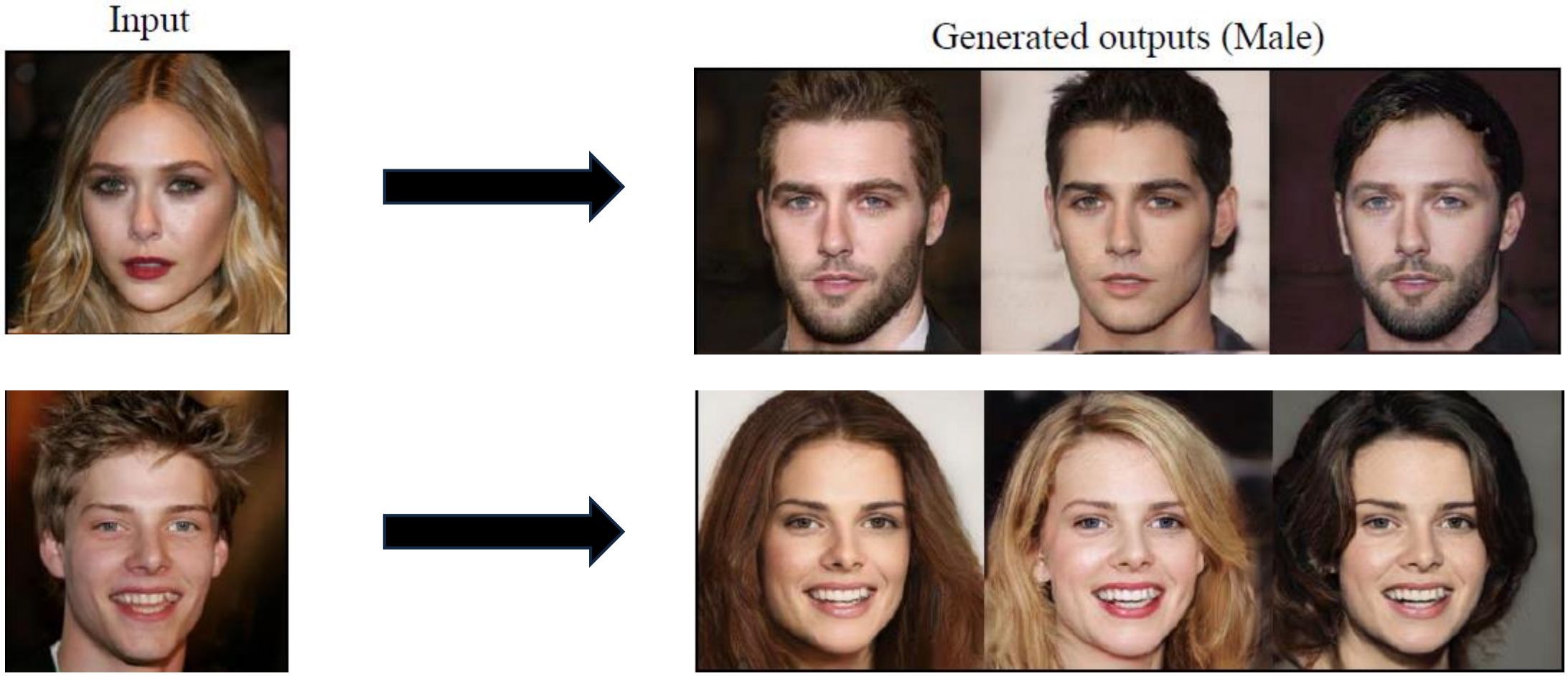


Figure 1. Diverse image synthesis results on the CelebA-HQ dataset and the newly collected animal faces (AFHQ) dataset. The first column shows input images while the remaining columns are images synthesized by StarGAN v2.

Domain : 우리가 눈으로 보고 그룹을 지을 수 있는 것 (e.g. gender)
Style : 각 이미지가 가지고있는 Unique appearance (e.g. hair, bread, makeup)

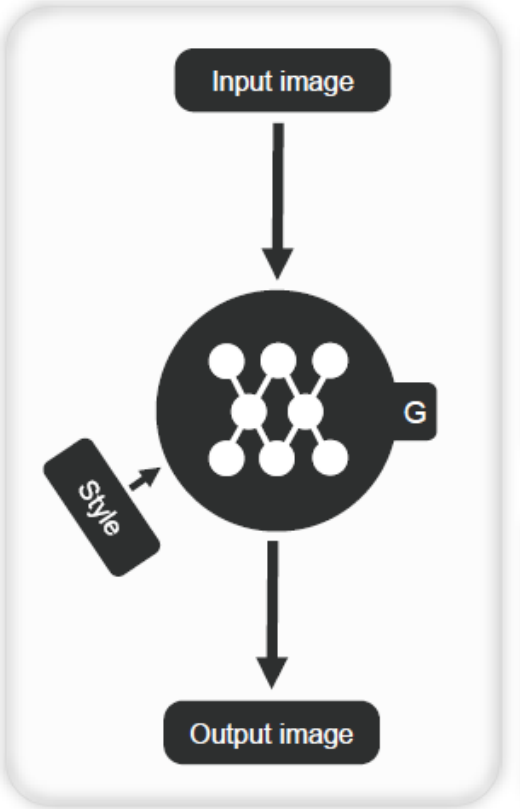
StarGAN V2 : intro



Input (남, 녀)을 바뀐 domain(바뀐 성별) 내에서 다양한 style(머리카락, 수염 등)

StarGAN V2 : Proposed model

Generator



(a) Generator

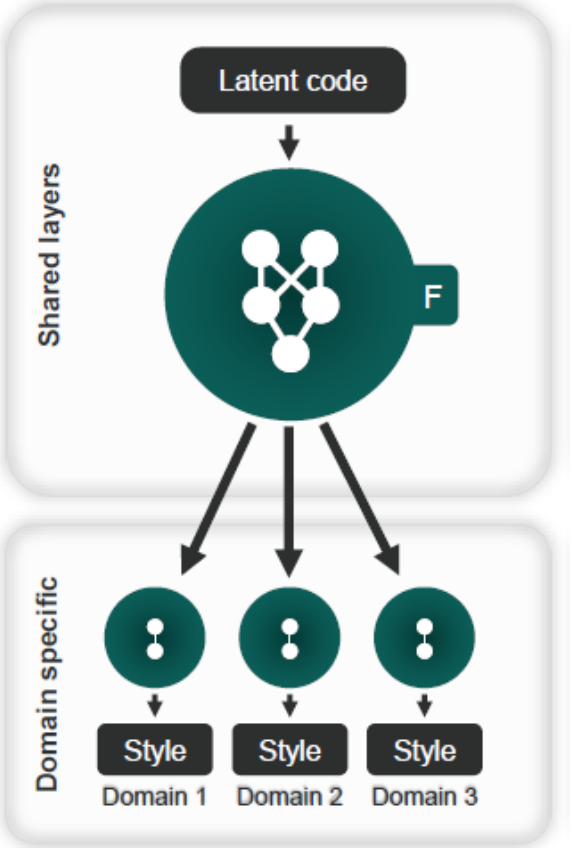
Input image **x**는 Style code **s**가 반영되어 G(x,s)가 output으로 나타남

Style code s는 Mapping network와 Style encoder에서 추출됨

style은 적용시키기 위해 Adaptive instance normalization(AdaIN) 사용

StarGAN V2 : Proposed model

Generator



(b) Mapping network

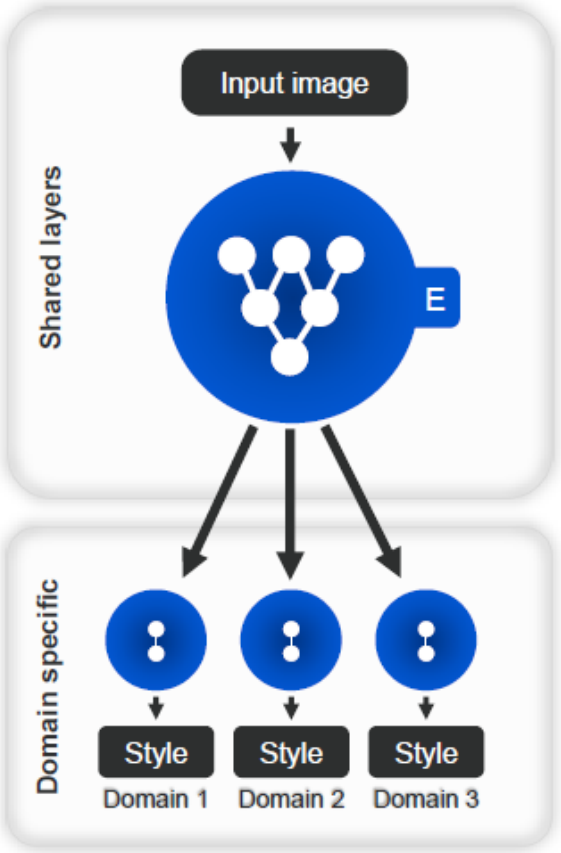
$$s = F_y(z)$$

latent code z (random noise) 를 domain y 에 대한 정보를 F를 통해서 Style code s 를 produce

우리의 multi-task architecture 은 F가 모든 도메인에 대해 스타일 표현을 효율적으로 할 수 있게 만들었다.

StarGAN V2 : Proposed model

Generator



(c) Style encoder

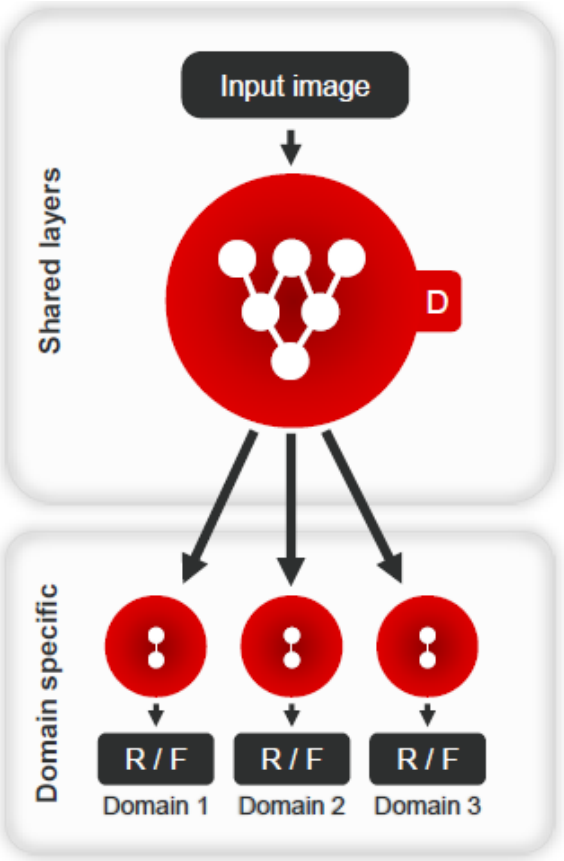
$$s = E_y(x)$$

Input image x (random noise) 에 대한 domain y 정보를 F를 통해서 Style code s 를 추출함

input image x에 대한 원래의 스타일을 반영할 수 있도록 도와줌

StarGAN V2 : Proposed model

Generator



(d) Discriminator

기존 V1 discriminator와 비슷

binary classification을 통해 x와 G(x,s) 구분하는 학습

StarGAN V2 : Proposed model

Full objective

$$\mathcal{L}_D = -\mathcal{L}_{adv},$$

$$\mathcal{L}_{F,G,E} = \mathcal{L}_{adv} + \lambda_{sty}\,\mathcal{L}_{sty} - \lambda_{ds}\,\mathcal{L}_{ds} + \lambda_{cyc}\,\mathcal{L}_{cyc},$$

StarGAN V2 : Proposed model

Adversarial loss

$$\mathcal{L}_{adv} = \mathbb{E}_{\mathbf{x},y}\left[\log D_y(\mathbf{x})\right] + \mathbb{E}_{\mathbf{x},\tilde{y},\mathbf{z}}[\log\left(1 - D_{\tilde{y}}(G(\mathbf{x},\tilde{\mathbf{s}}))\right)],$$

Generator Output G(x,s')와 real image x에 대한 adversarial loss

$$\mathcal{L}_{cyc} = \mathbb{E}_{\mathbf{x},y,\tilde{y},\mathbf{z}}\left[||\mathbf{x} - G(G(\mathbf{x},\tilde{\mathbf{s}}),\hat{\mathbf{s}})||_1\right],$$

Cycle consistency loss를 통해 **원래 이미지의 특징을 보존**

StarGAN V2 : Proposed model

Style reconstruction

$$\mathcal{L}_{sty} = \mathbb{E}_{\mathbf{x},\tilde{y},\mathbf{z}}\left[\|\tilde{\mathbf{s}} - E_{\tilde{y}}(G(\mathbf{x},\tilde{\mathbf{s}}))\|_1\right].$$

$$s = E_y(x)$$

Style Encoder가 Style을 잘 mapping 할 수 있게 학습
Style Encoder는 앞서 말했듯, Generator가 input image x에 대한 style을 제공하여 원래 이미지로 reconstruction 하게 도와줌

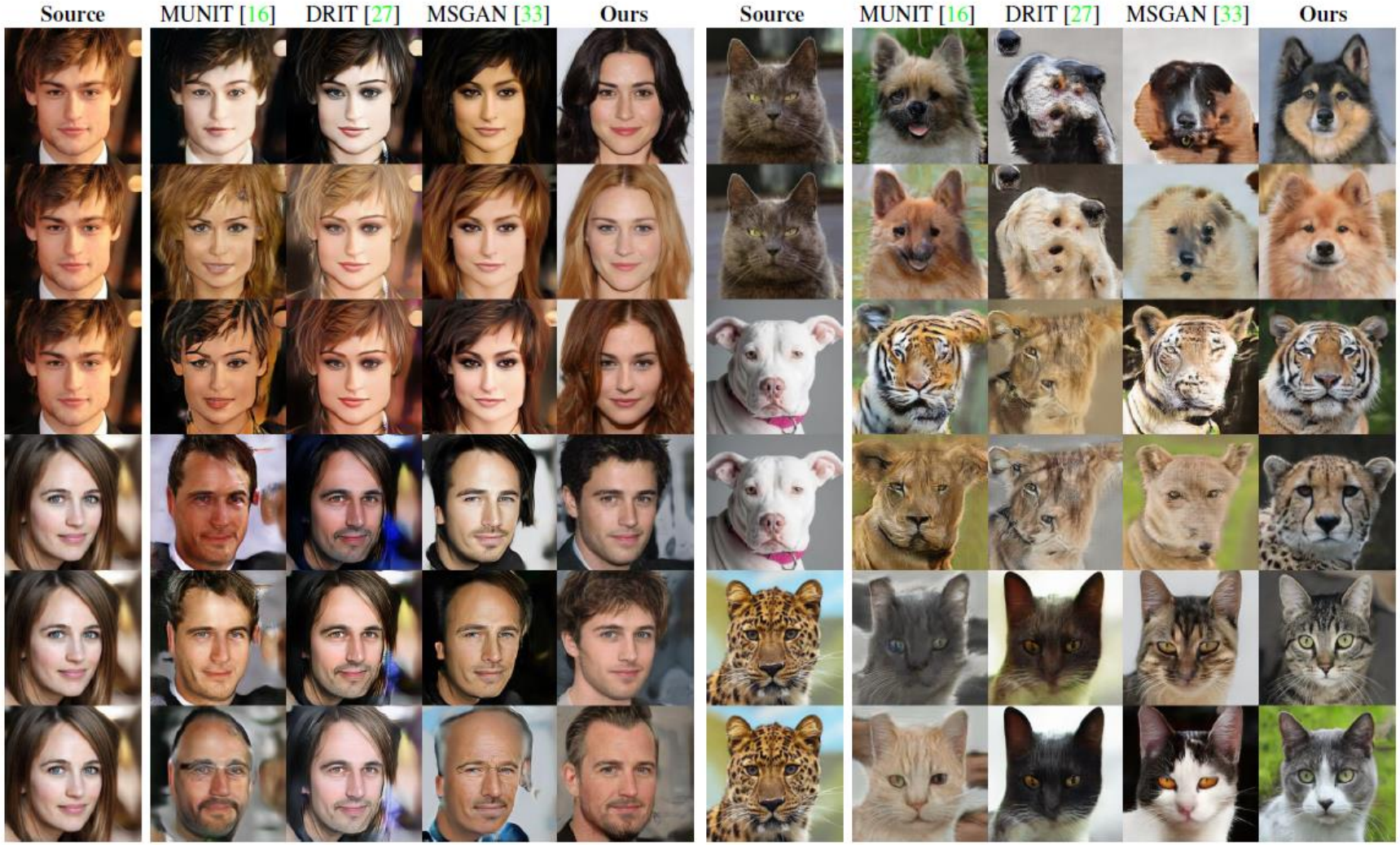
Style diversification

$$\mathcal{L}_{ds} = \mathbb{E}_{x,\tilde{y},z_1,z_2}[\|G(x,\tilde{s}_1) - G(x,\tilde{s}_2)\|_1]$$

latent vector 1,2 = z_1, z_2 를 이용해 제 3의 style(랜덤한) 을 F()에서 추출 된 s'1, s'2 이를 이용해 diversity를 주는 부분을 학습

Optimal 한 점(학습의 끝)이 없기 때문에 loss는 linearly decay했다.

StarGAN V2 : Experiments

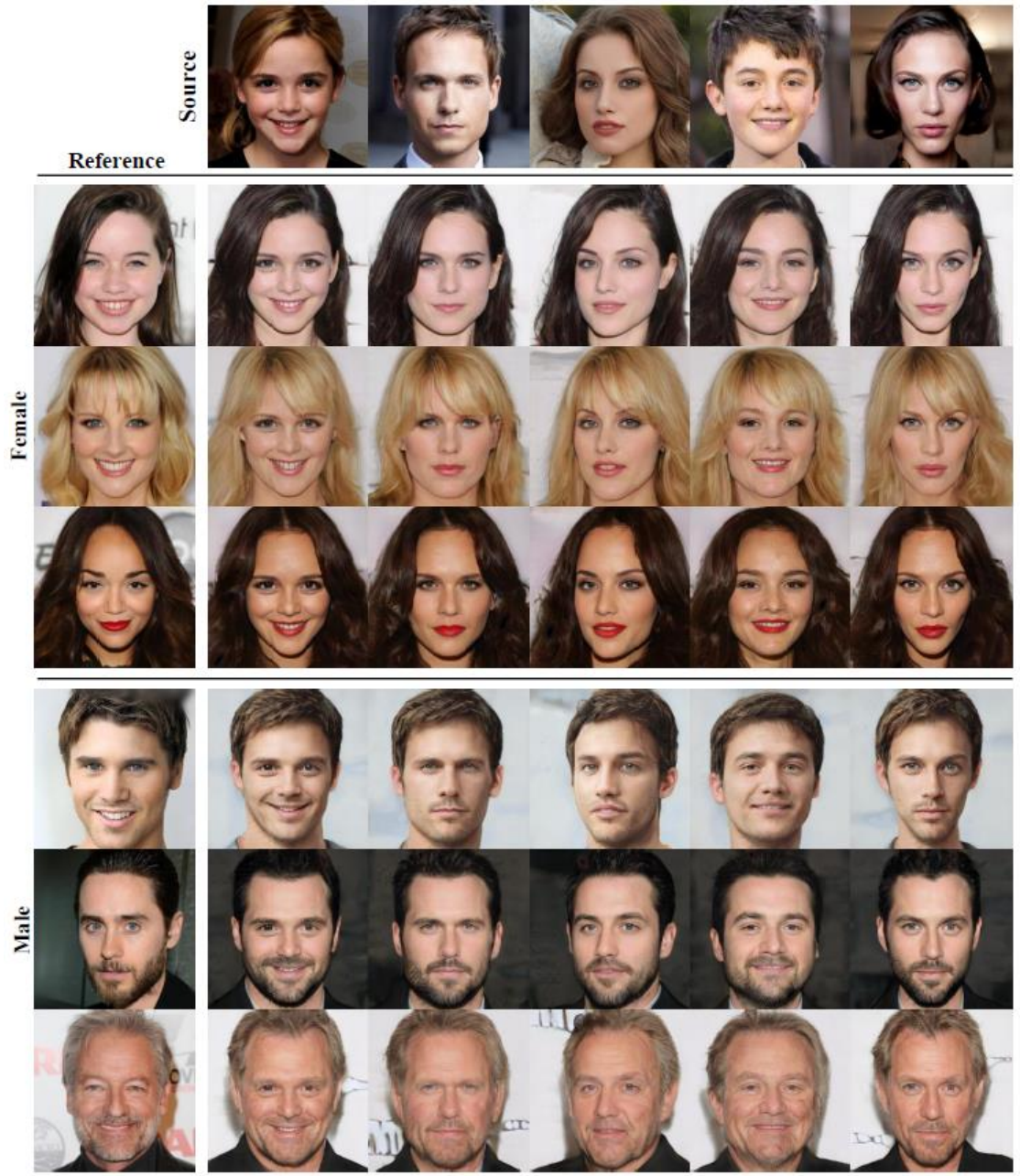


(a) Latent-guided synthesis on CelebA-HQ

(b) Latent-guided synthesis on AFHQ

Figure 5. Qualitative comparison of latent-guided image synthesis results on the CelebA-HQ and AFHQ datasets. Each method translates the source images (left-most column) to target domains using randomly sampled latent codes. **(a)** The top three rows correspond to the results of converting male to female and vice versa in the bottom three rows. **(b)** Every two rows from the top show the synthesized images in the following order: cat-to-dog, dog-to-wildlife, and wildlife-to-cat.

StarGAN V2 : Experiments



AI Lab.
ARTIFICIAL INTELLIGENCE LAB.

StarGAN V2 : Experiments

	Method	FID	LPIPS
A	Baseline StarGAN [7]	98.4	-
B	+ Multi-task discriminator	91.4	-
C	+ Tuning (*e.g.*, R_1 regularization)	80.5	-
D	+ Latent code injection	32.3	0.312
E	+ Replace (D) with style code	21.2	0.406
F	+ Diversity regularization	**18.0**	**0.428**

Table 1. Performance of various configurations on CelebA-HQ. Frechet inception distance (FID) indicates the distance between two distributions of real and generated images (lower is better), while learned perceptual image patch similarity (LPIPS) measures the diversity of generated images (higher is better).

Method	CelebA-HQ		AFHQ	
	FID	LPIPS	FID	LPIPS
MUNIT [16]	107.1	0.176	223.9	0.199
DRIT [27]	53.3	0.311	114.8	0.156
MSGAN [33]	39.6	0.312	69.8	0.375
StarGAN v2	**20.2**	**0.397**	**19.7**	**0.503**
Real images	15.1	-	13.1	-

Table 3. Quantitative comparison on reference-guided synthesis. We sample ten reference images to synthesize diverse images.

보아야 할 테이블 2개

Alblation study : LPIPS (다양성) 대폭 상승했고, FID도 많이 높아짐