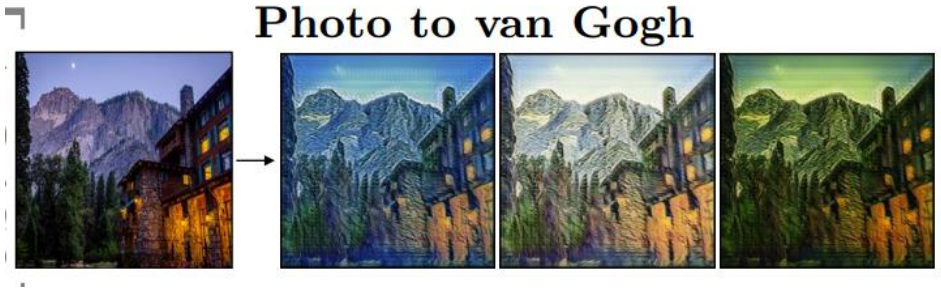
DiveRse Image-to-image Translation from unpaired data

2018 ECCV oral

Motivation

1. Image to Image translatio을 하기 위한 서로 다른 도메인의 paired data를 구하는 것이 어렵다. (unsupervised setting의 필요성)

2. 애초에 one to one mapping이 아닌 경우도 많다. (Multimodal한 아웃풋을 뽑아내는 당위성)



Method	Pix2Pix [18]	CycleGAN [48]	UNIT [27]	BicycleGAN [49]	Ours
Unpaired	-	✓	✓	-	✓
Multimodal	-	-	-	✓	✓

What to do

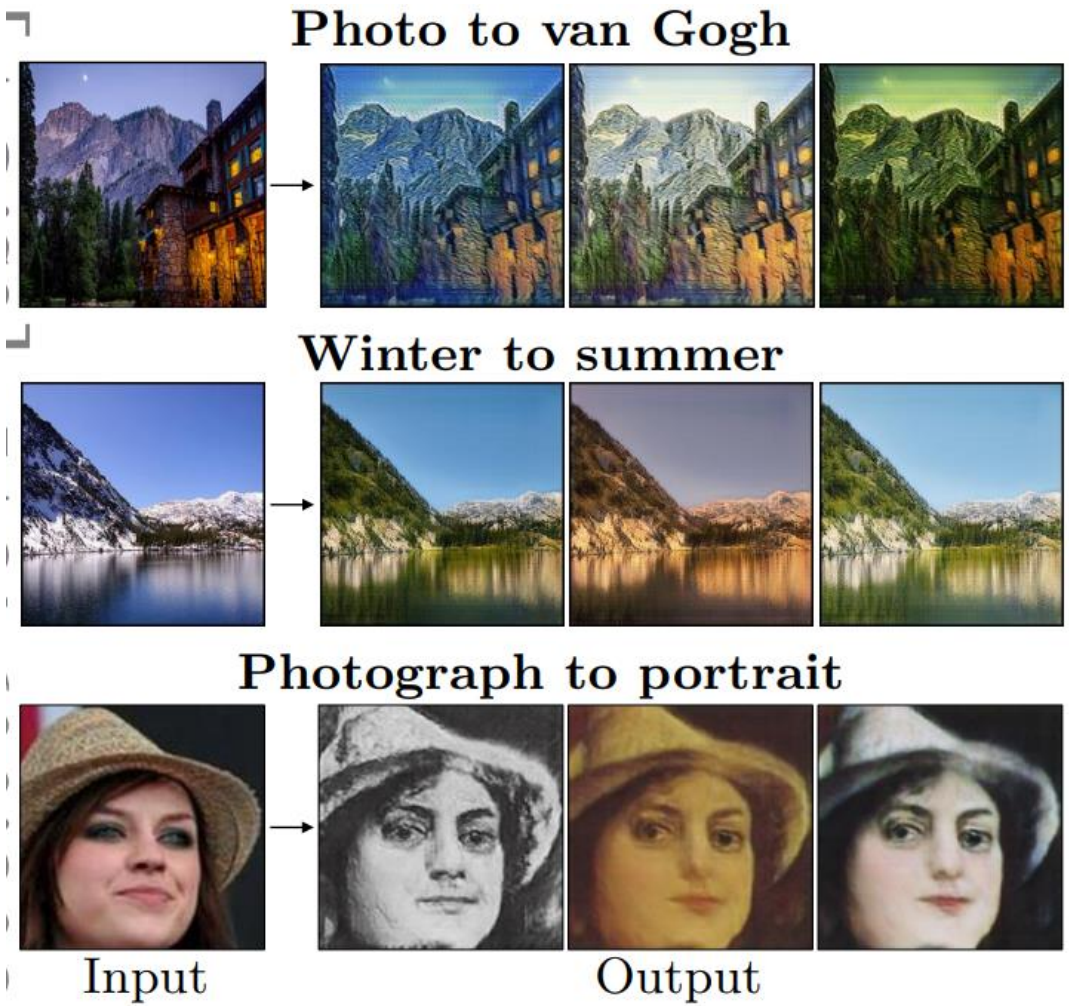
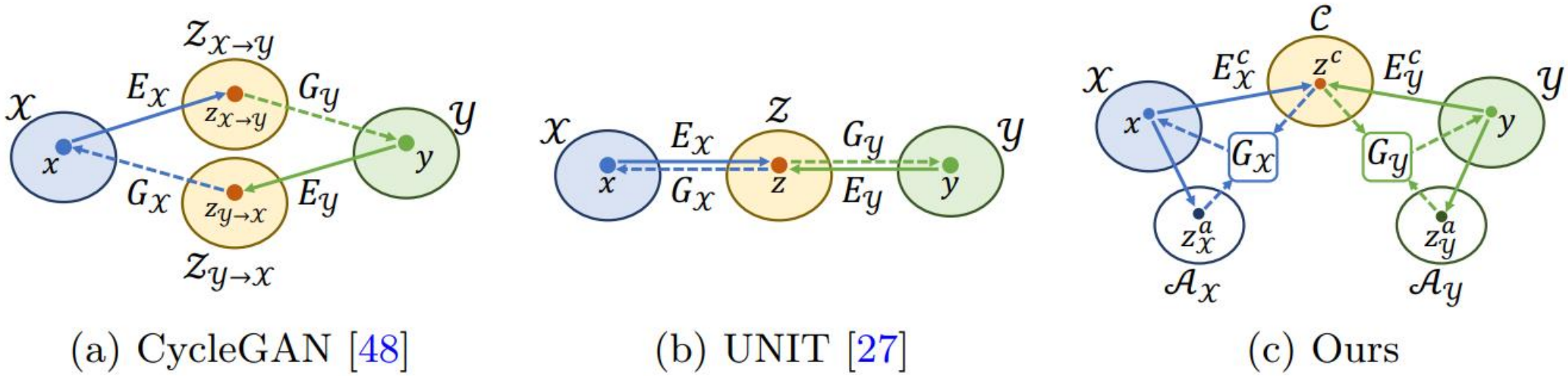


Photo to van Gogh
Winter to summer
Photograph to portrait
Input
Output


Content
Attribute
Generated
Input
Output

Method (high level)

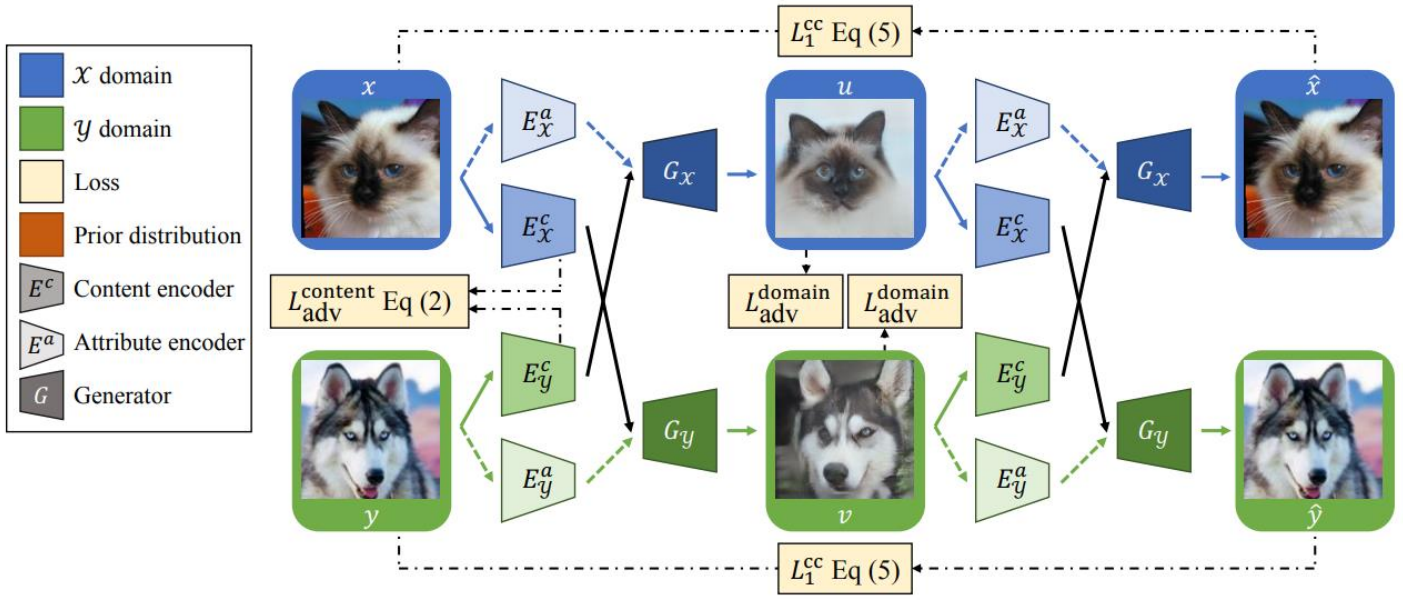


(a) CycleGAN [48] (b) UNIT [27] (c) Ours

- Domain invariant한 feature를 content space에 임배딩
  ($z^c = A\ content\ feature\ of\ a\ domian\ x$ => domain notation이 없음에 유의)

- Domain의 속성을 갖고 있는 feature를 attribute space에 임배딩
  ($z_x^a = A\ attribute\ feature\ of\ a\ domian\ x$)

Method (details)



(a) Training with unpaired images

$$(E_{\mathcal{X}}^c : \mathcal{X} \to \mathcal{C})$$
$$(E_{\mathcal{X}}^a : \mathcal{X} \to \mathcal{A}_{\mathcal{X}})$$
$$(G_{\mathcal{X}} : \{\mathcal{C}, \mathcal{A}_{\mathcal{X}}\} \to \mathcal{X})$$

content encoders $\{E_{\mathcal{X}}^c, E_{\mathcal{Y}}^c\}$ generators $\{G_{\mathcal{X}}, G_{\mathcal{Y}}\}$

attribute encoders $\{E_{\mathcal{X}}^a, E_{\mathcal{Y}}^a\}$ domain discriminators $\{D_{\mathcal{X}}, D_{\mathcal{Y}}\}$

Domain Invariant한 Content space

1. Weight sharing of the last few layers of the E_x^c, E_y^c.
2. Weight sharing of the first few layers of the G_x, G_y.
3. Content Discriminator D^c

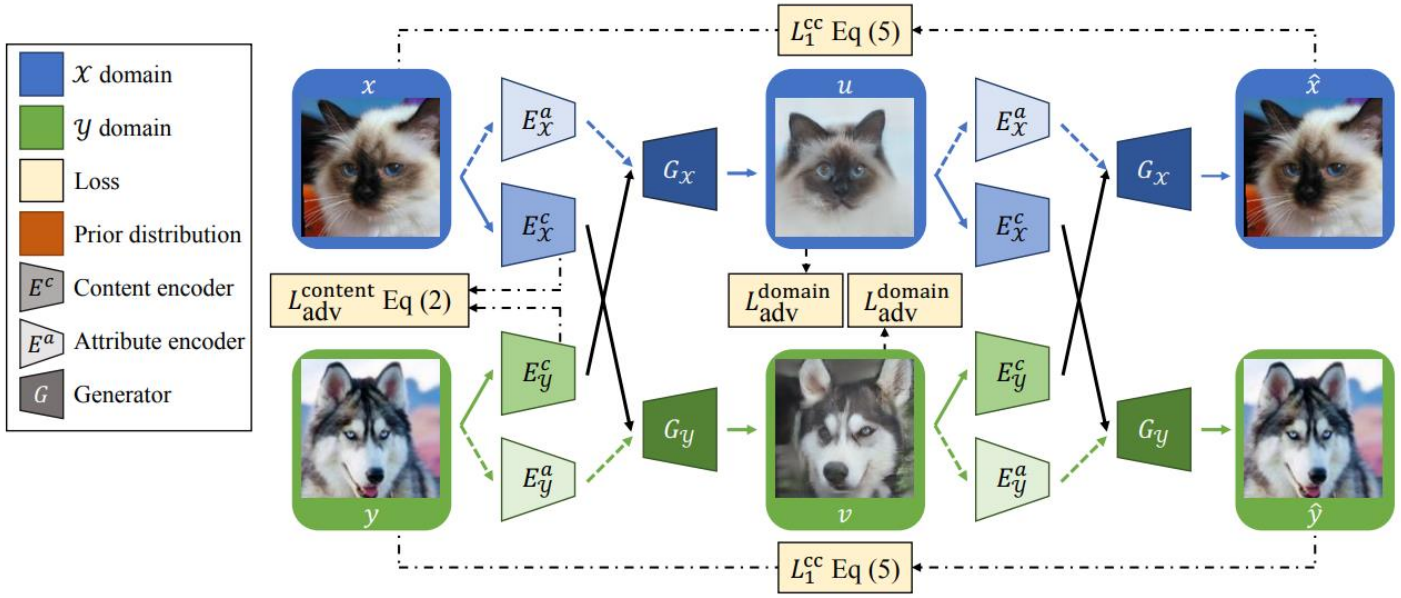
*$\arg \min_G \max_D$

코드상 D 학습: $\mathbb{E}_x\left[\log D^c\left(E_x^c(x)\right)\right] + \mathbb{E}_y\left[\log(1 - D^c\left(E_y^c(y)\right))\right]$

G 학습: $\mathbb{E}_x\left[\log(\frac{1}{2} - D^c\left(E_x^c(x)\right))\right] + \mathbb{E}_y\left[\log(\frac{1}{2} - D^c\left(E_y^c(y)\right))\right]$

- 목적: x domain의 content feature와 y domain의 content feature 간의 js divergence를 줄여 content space를 domain invariant하게 만들자

Reconstruction loss



(a) Training with unpaired images

$$L_1^{cc}(G_{\mathcal{X}}, G_{\mathcal{Y}}, E_{\mathcal{X}}^c, E_{\mathcal{Y}}^c, E_{\mathcal{X}}^a, E_{\mathcal{Y}}^a) = \mathbb{E}_{x,y}[\|G_{\mathcal{X}}(E_{\mathcal{Y}}^c(v), E_{\mathcal{X}}^a(u)) - x\|_1 + \|G_{\mathcal{Y}}(E_{\mathcal{X}}^c(u), E_{\mathcal{Y}}^a(v)) - y\|_1],$$

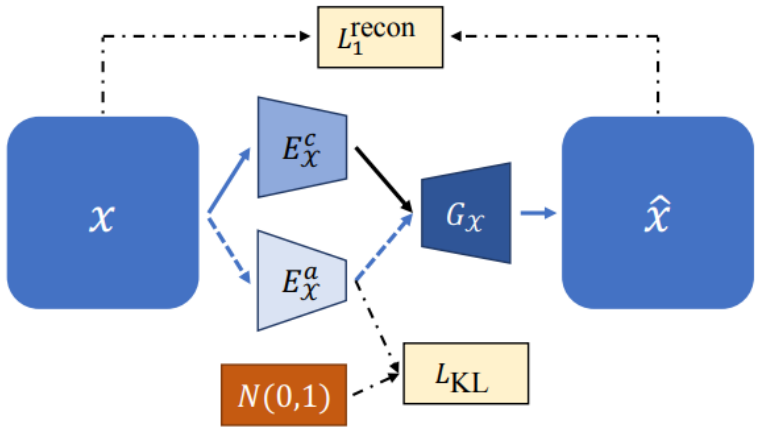
Other loss1

- Domain adversarial loss
- Self reconstruction loss (Auto encoder 같이)

$$\hat{y} = G_{\mathcal{Y}}(E^c_{\mathcal{Y}}(y), E^a_{\mathcal{Y}}(y))$$

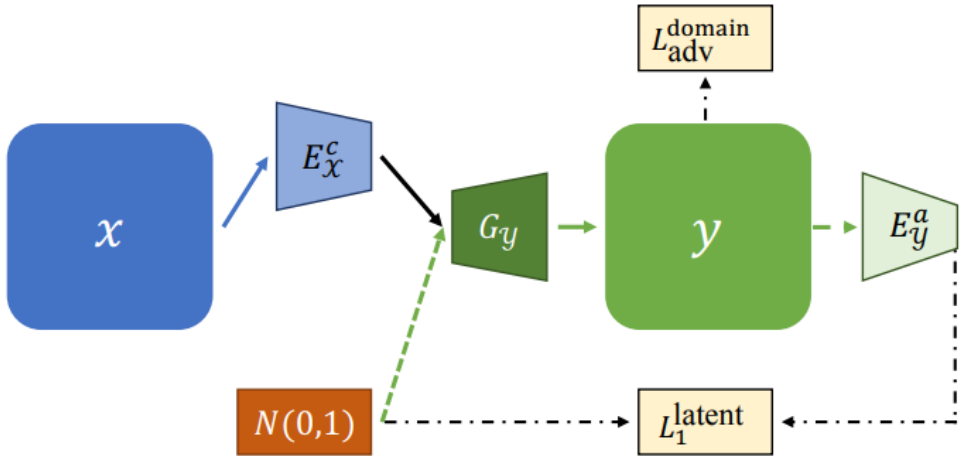
- KL loss (sampling을 위한 attribute feature regularization)

$$L_{\mathrm{KL}} = \mathbb{E}[D_{\mathrm{KL}}((z_a)\|N(0,1))]$$



Other loss2

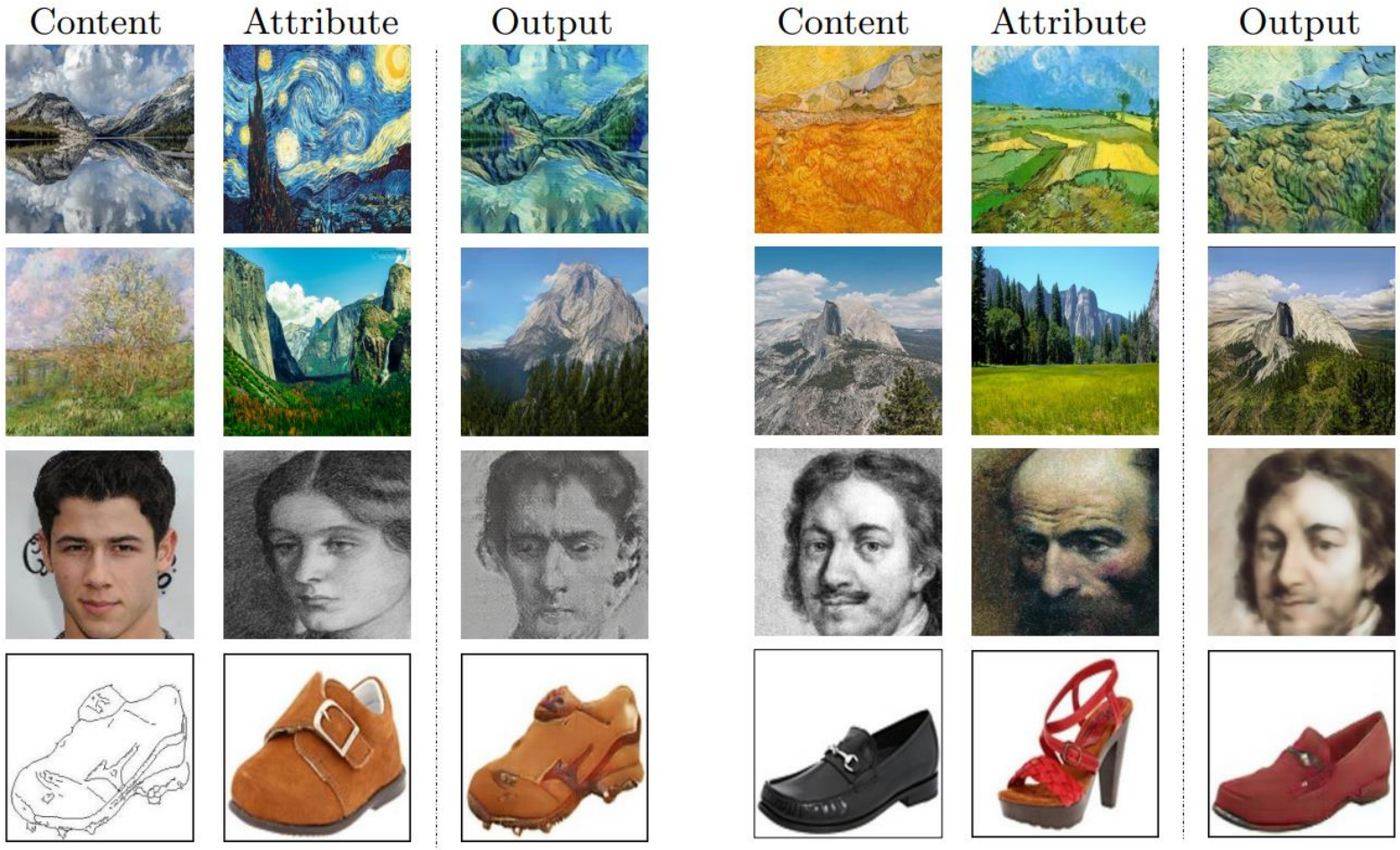
- Latent regression loss(KL과 같은 목적, cross domain)



$$\min_{G,E^c,E^a} \max_{D,D^c} \quad \lambda_{\text{adv}}^{\text{content}} L_{\text{adv}}^{\text{c}} + \lambda_1^{\text{cc}} L_1^{\text{cc}} + \lambda_{\text{adv}}^{\text{domain}} L_{\text{adv}}^{\text{domain}} + \lambda_1^{\text{recon}} L_1^{\text{recon}}$$

$$+\lambda_1^{\text{latent}} L_1^{\text{latent}} + \lambda_{\text{KL}} L_{\text{KL}}$$

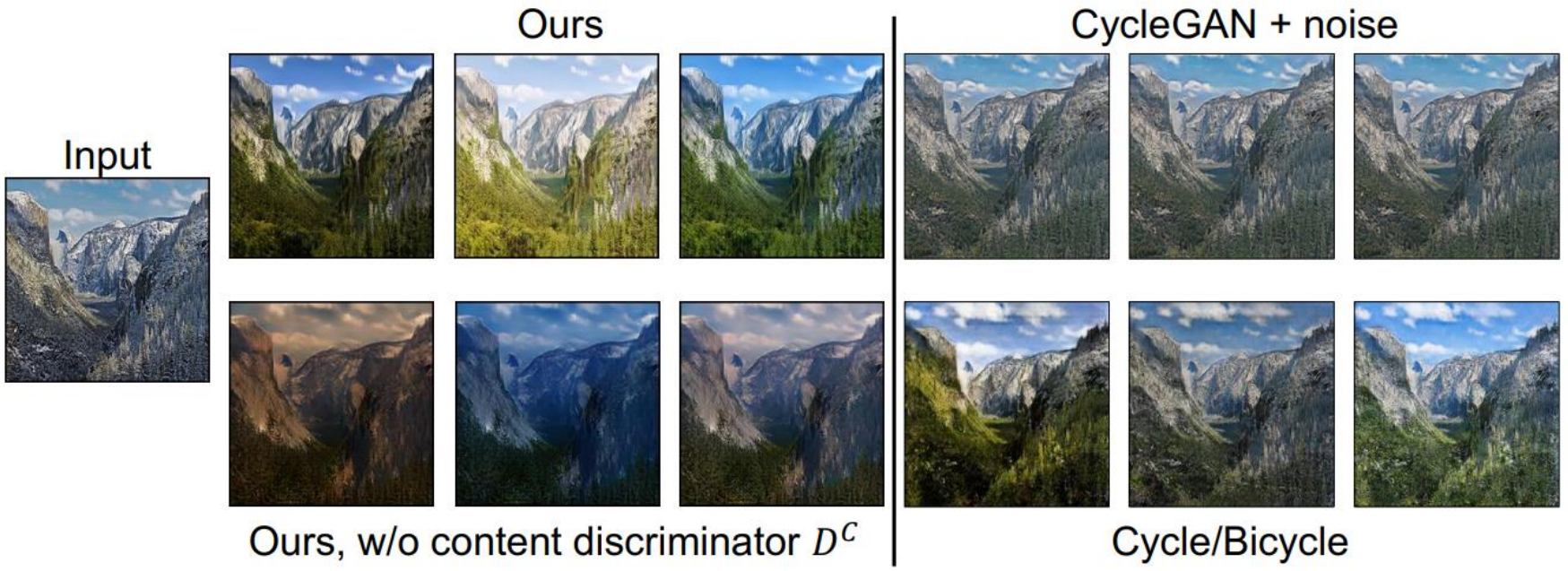
Qualitative result



(a) Inter-domain attribute transfer

(b) Intra-domain attribute transfer

겨울 -> 여름 데이터
Multimodal함에 대한 비교



Embedding space에
Attribute feature가 continuous
하게 잘 임배딩 되었다를
보여주는 ..

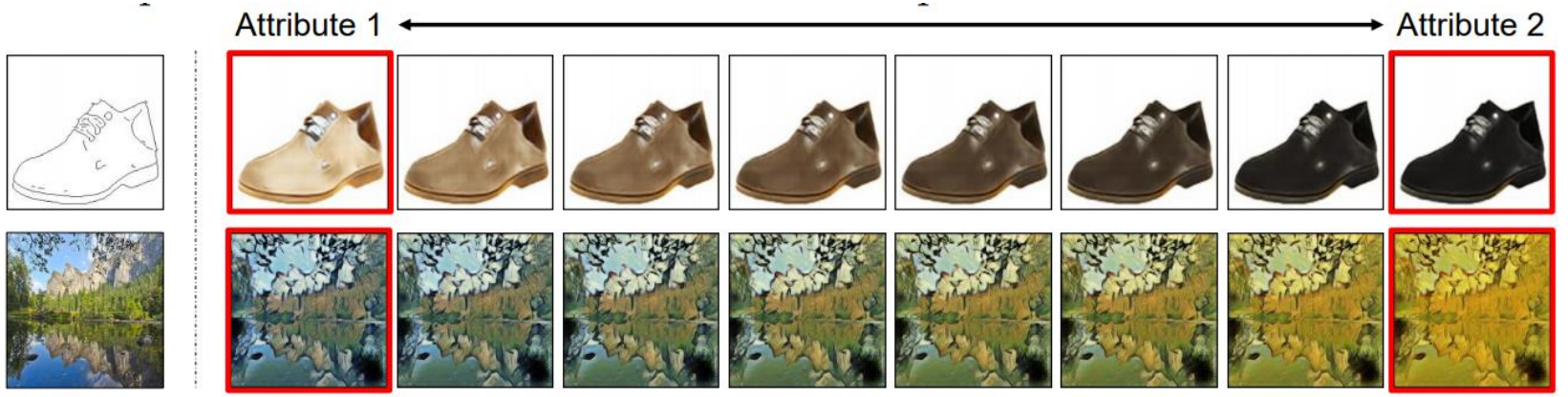
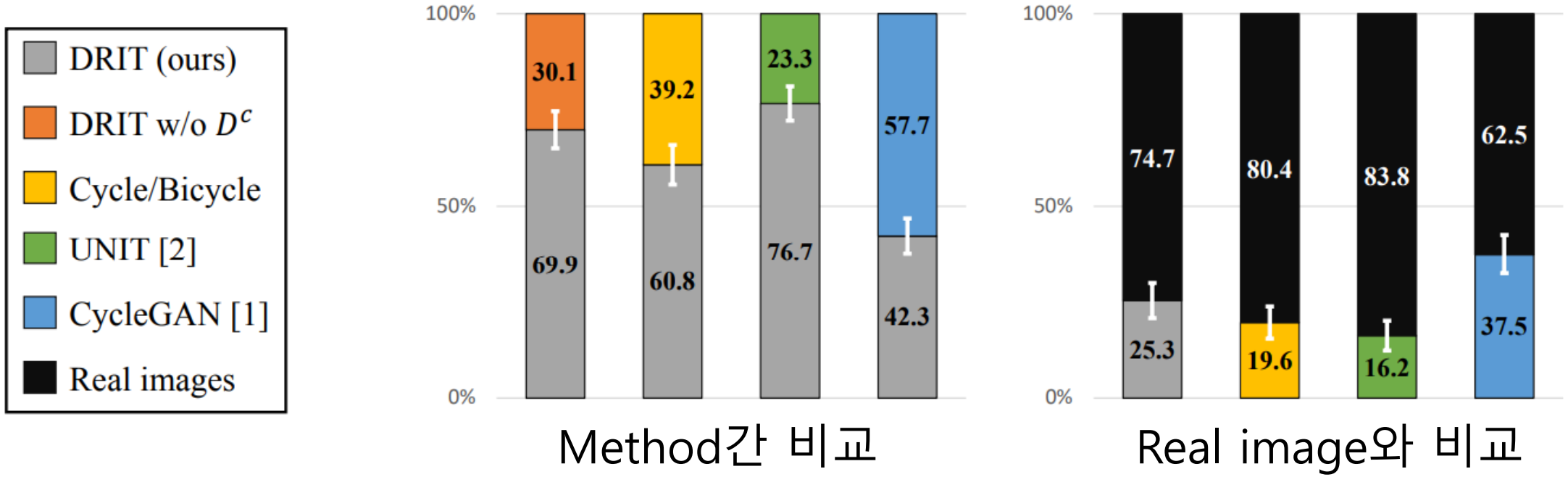


Fig. 7: **Linear interpolation between two attribute vectors.** Translation results with linear-interpolated attribute vectors between two attributes (highlighted in red).

Quantitative result1



Method간 비교

Real image와 비교

- 어느 것이 더 진짜같냐? AB test

Quantitative result2

- Diversity, LPIPS metric을 사용 (이미지 2개를 인풋으로 받아 similarity를 측정하는 pretrained model)

- 100개의 real image로부터 translation한 100개의 fake image 중 1000개 페어를 무작위 추출

*(mode collapsing 여부를 따져보는 듯)

Table 2: **Diversity.** We use the LPIPS metric [47] to measure the diversity of generated images on the Yosemite dataset.

Method	Diversity
real images	.448 ± .012
DRIT	**.424** ± .010
DRIT w/o D^c	.410 ± .016
UNIT [27]	.406 ± .022
CycleGAN [48]	<u>.413</u> ± .008
Cycle/Bicycle	.399 ± .009

Quantitative result3

Table 3: **Reconstruct error**. We use the edge-to-shoes dataset to measure the quality of our attribute encoding. The reconstruction error is $\|y - G_{\mathcal{Y}}(E^c_{\mathcal{X}}(x), E^a_{\mathcal{Y}}(y))\|_1$. * BicycleGAN uses *paired* data for training.

Method	Reconstruct error
BicycleGAN [49]*	**0.0945**
DRIT	<u>0.1347</u>
DRIT, w/o D^c	0.2076

X와 y의 페어가 있는 데이터셋에서,
X의 Content와 Y의 attribute를 받아 생성한 이미지가 얼마나 정답과 차이가 나는가?

Baseline: paired im2im settin의 대표격 페이퍼인 bicycleGAN

아쉬운 점

- 페이퍼의 메인 contribution인 Multimodal함에 대한 정량적 결과가 제시됐어야 함.

(하나의 real image에 대한 랜덤 샘플링 attribute n개 forwarding,
N개 이미지 간 pair를 구성하여 pairwise distance)

- 같은 목적을 가진 두개의 로스가 있으나, ablation study에 이것이 드러나지 않음 또한, ablation study가 오직 content discriminator만 다룸

- MUNIT을 몰랐을리가 없는데 (attribute embed size: 8)

아예 언급을 안함. Method의 차이점과 성능 차이를 비교했으면 더 좋지 않았을까..