Unsupervised Image-to-Image Translation Networks

NIPS 2017

Motivation

In the unpaired data setting,

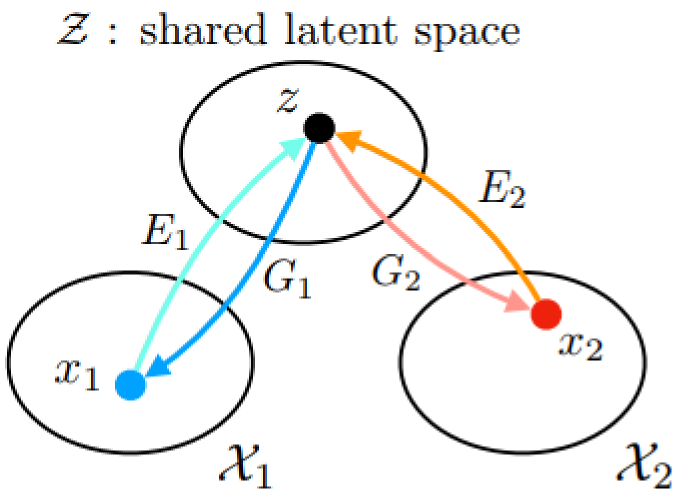
$p(zebras)$

$p(horses)$

Joint distribution $p(zebras,\ horses)$ = Infinite set!

But we want a specific result

Shared Latent Space Assumption



$$z = E_1^*(x_1) = E_2^*(x_2)$$

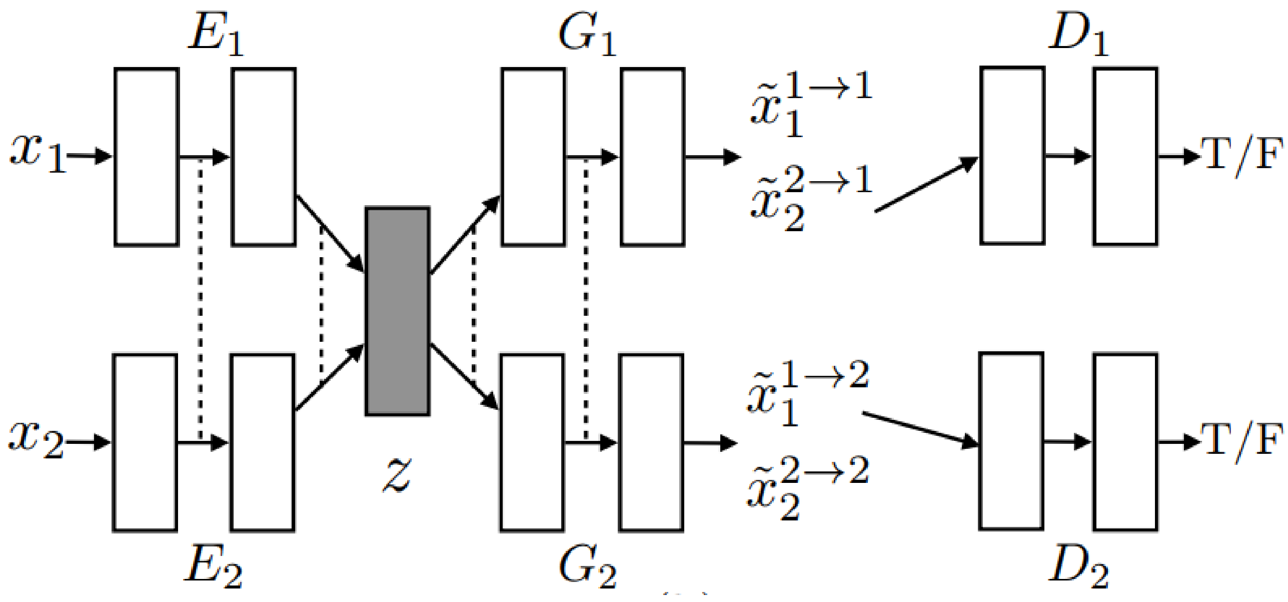
Unpaired 데이터셋이지만,
X1 도메인과 X2 도메인 내 특정 이미지끼리
Latent code z를 share 한다고 가정

$$x_1 = G_1^*(z)$$

$$F_{1\to 2}^*(x_1) = G_2^*(E_1^*(x_1))$$

$$x_1 = F_{2\to 1}^*(F_{1\to 2}^*(x_1))$$

Implementation of Shared Latent



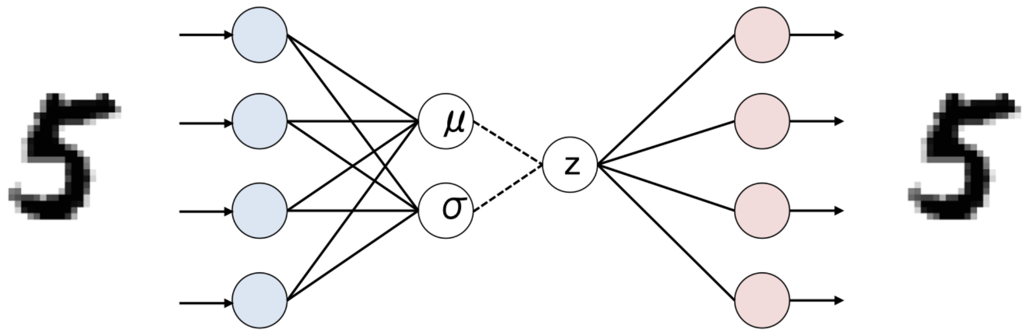
$$z \to h \begin{matrix} \nearrow & x_1 \\ \searrow & x_2 \end{matrix} \qquad G_1^* \equiv G_{L,1}^* \circ G_H^*$$

G_H => high level generation
(realization of z)
G_L => low level generation
(actual image formation)

Ex)
z = car in front, trees in back
h = car/tree occupy the following pixels
x = domain specific color (pixel value)

Method 1 (1->1)

VAE 세팅



(assumption: The latent space Z is conditionally independent and Gaussian with unit variance)

Reparameterization trick
(test 때는 sampling 없이)

$$\eta \sim \mathcal{N}(\eta|0, I)$$

$$z_1 = E_{\mu,1}(x_1) + \eta$$

```
hiddens = self.encode(images)
if self.training == True:
    noise = Variable(torch.randn(hiddens.size()).cuda(hiddens.data.get_device()))
    images_recon = self.decode(hiddens + noise)
else:
    images_recon = self.decode(hiddens)
```

Loss

$$\mathcal{L}_{\text{VAE}_1}(E_1, G_1) = \lambda_1 \text{KL}(q_1(z_1|x_1)||p_\eta(z)) - \lambda_2 \mathbb{E}_{z_1 \sim q_1(z_1|x_1)}[\log p_{G_1}(x_1|z_1)]$$

$$\mathcal{L}_{\text{VAE}_2}(E_2, G_2) = \lambda_1 \text{KL}(q_2(z_2|x_2)||p_\eta(z)) - \lambda_2 \mathbb{E}_{z_2 \sim q_2(z_2|x_2)}[\log p_{G_2}(x_2|z_2)].$$

$$p_\eta(z) = \mathcal{N}(z|0, I)$$

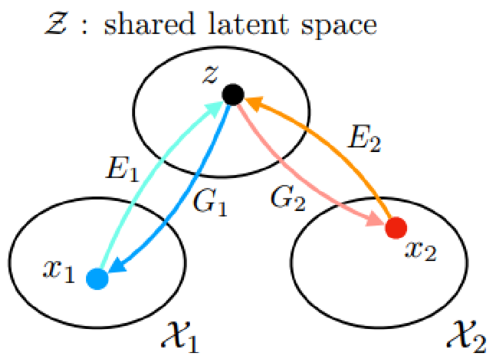
$$q_1(z_1|x_1) \equiv \mathcal{N}(z_1|E_{\mu,1}(x_1), I)$$

Method 2 (1->2)

아직까지는 Latent space는 공유하나, $z = E_1^*(x_1) = E_2^*(x_2)$ 가 보장되지 않음

$$\mathcal{L}_{\text{CC}_1}(E_1, G_1, E_2, G_2) = \lambda_3 \text{KL}(q_1(z_1|x_1)||p_\eta(z)) + \lambda_3 \text{KL}(q_2(z_2|x_1^{1\to2}))||p_\eta(z)) - \lambda_4 \mathbb{E}_{z_2\sim q_2(z_2|x_1^{1\to2})}[\log p_{G_1}(x_1|z_2)]$$

$$\mathcal{L}_{\text{CC}_2}(E_2, G_2, E_1, G_1) = \lambda_3 \text{KL}(q_2(z_2|x_2)||p_\eta(z)) + \lambda_3 \text{KL}(q_1(z_1|x_2^{2\to1}))||p_\eta(z)) - \lambda_4 \mathbb{E}_{z_1\sim q_1(z_1|x_2^{2\to1})}[\log p_{G_2}(x_2|z_1)].$$



Method 3 (GAN objective)

Latent space가 아니라, fake_x2와 real_x2의 distribution을 최대한 가깝게 하기 위한 method

$$\mathcal{L}_{\text{GAN}_1}(E_1, G_1, D_1) = \lambda_0 \mathbb{E}_{x_1 \sim P_{\mathcal{X}_1}}[\log D_1(x_1)] + \lambda_0 \mathbb{E}_{z_2 \sim q_2(z_2|x_2)}[\log(1 - D_1(G_1(z_2)))] \quad (5)$$

$$\mathcal{L}_{\text{GAN}_2}(E_2, G_2, D_2) = \lambda_0 \mathbb{E}_{x_2 \sim P_{\mathcal{X}_2}}[\log D_2(x_2)] + \lambda_0 \mathbb{E}_{z_1 \sim q_1(z_1|x_1)}[\log(1 - D_2(G_2(z_1)))]. \quad (6)$$

Overall loss

$$\min_{E_1, E_2, G_1, G_2} \max_{D_1, D_2} \mathcal{L}_{\text{VAE}_1}(E_1, G_1) + \mathcal{L}_{\text{GAN}_1}(E_1, G_1, D_1) + \mathcal{L}_{\text{CC}_1}(E_1, G_1, E_2, G_2)$$
$$\mathcal{L}_{\text{VAE}_2}(E_2, G_2) + \mathcal{L}_{\text{GAN}_2}(E_2, G_2, D_2) + \mathcal{L}_{\text{CC}_2}(E_2, G_2, E_1, G_1).$$

Result

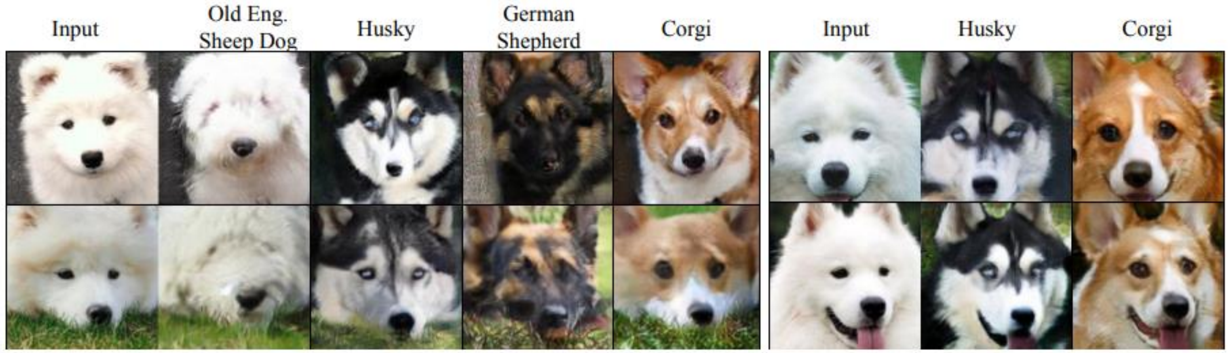


Figure 4: Dog breed translation results.

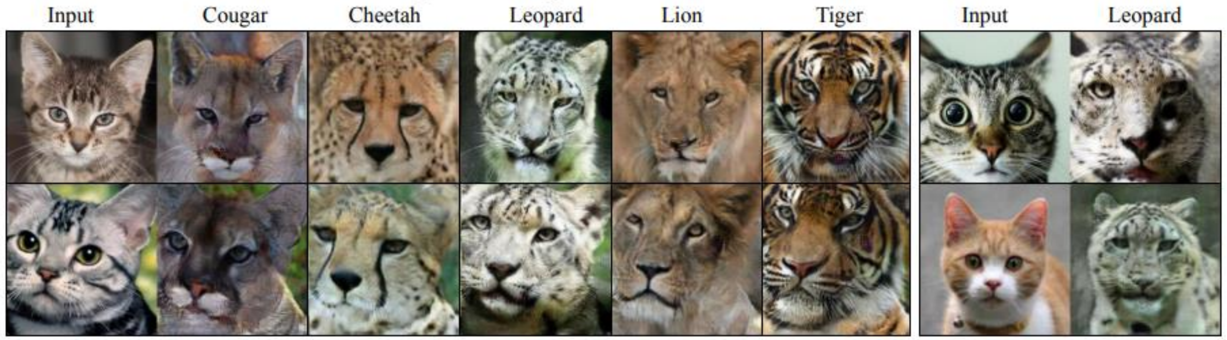


Figure 5: Cat species translation results.

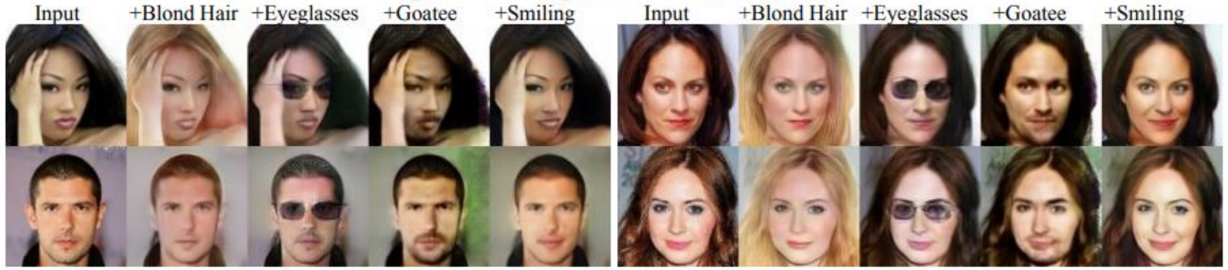


Figure 6: Attribute-based face translation results.

1. Conclusion 내용:
The translation model is unimodal due to
The Gaussian latent space assumption
이 가정을 하는 당위성이 뭔지를 모르겠음.

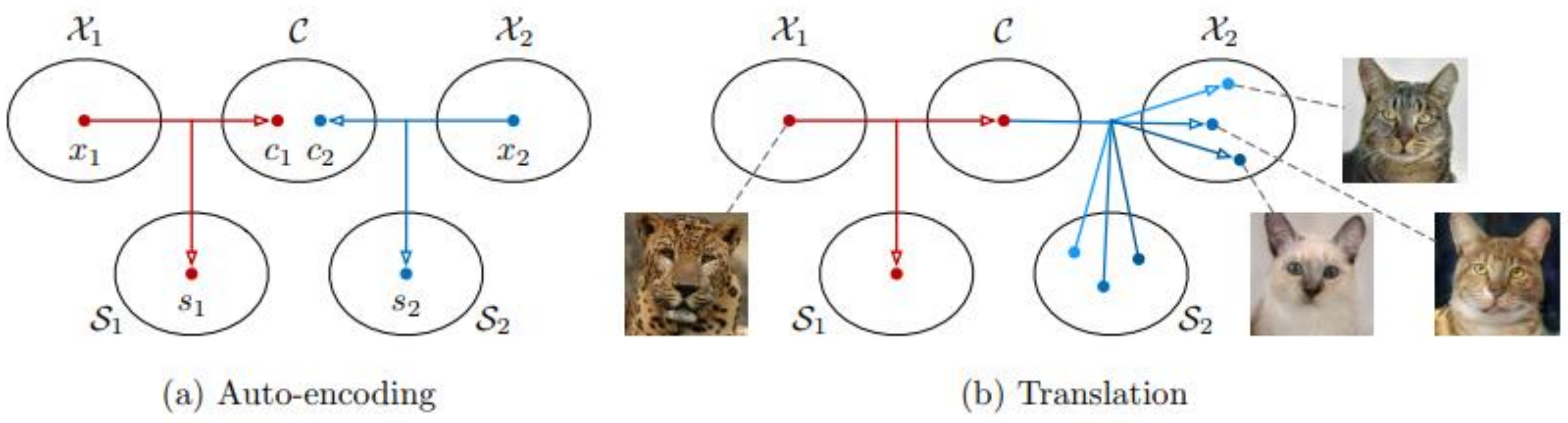
추후 multimodality를 가능하게 하는 모델을
만들겠다 함.
⇒ 애초에 후속 페이지를 염두에 두고
작성한 페이퍼인듯.

Multimodal Unsupervised Image-to-Image Translation

ECCV 2018

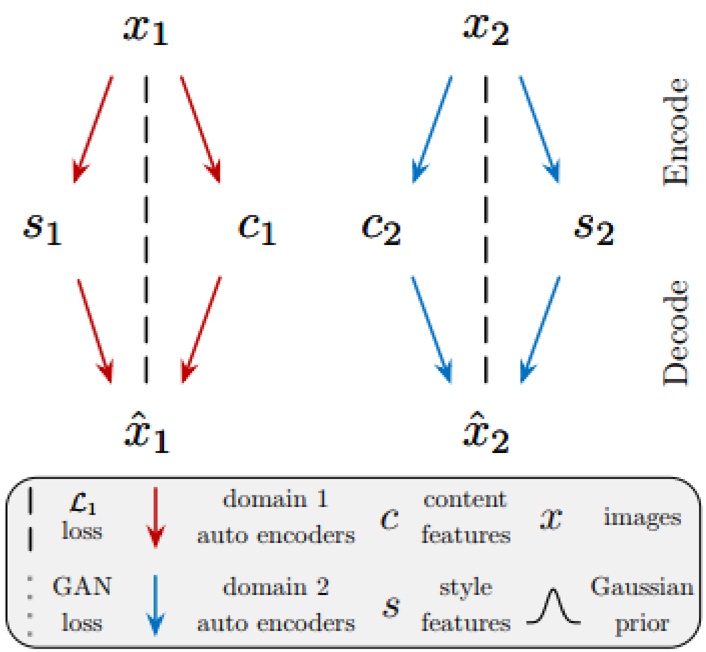
Idea

- Shared latent space (From UNIT)
- Separated latent space (Content, Style)



- Multimodal output (many to many setting is reasonable)
- Exemplar

Method 1: Within-domain (AE)

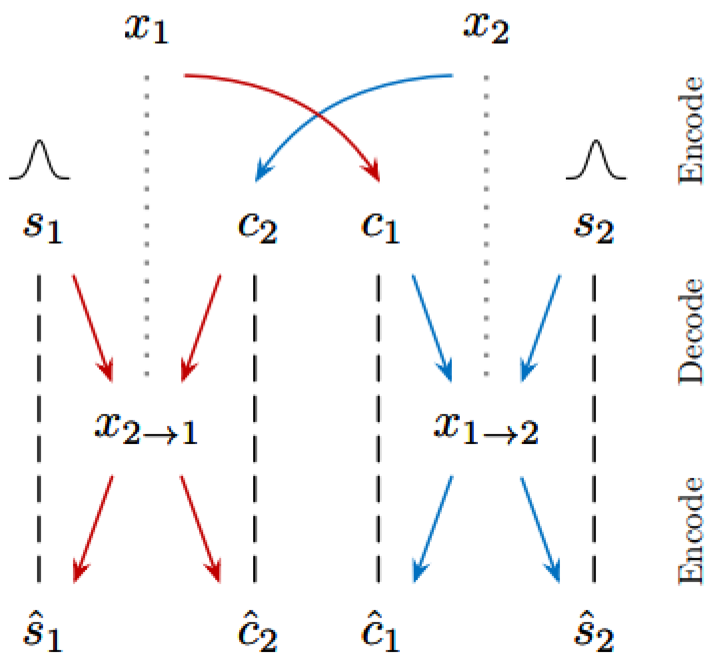


(a) Within-domain reconstruction

Image Reconstruction

$$\mathcal{L}_{\text{recon}}^{x_1} = \mathbb{E}_{x_1 \sim p(x_1)}[||G_1(E_1^c(x_1), E_1^s(x_1)) - x_1||_1]$$

Method 2 : Cross-domain(translation)



(b) Cross-domain translation

Latent Reconstruction

$$\mathcal{L}^{c_1}_{\text{recon}} = \mathbb{E}_{c_1\sim p(c_1),s_2\sim q(s_2)}[||E^c_2(G_2(c_1,s_2)) - c_1||_1] \quad (2)$$

$$\mathcal{L}^{s_2}_{\text{recon}} = \mathbb{E}_{c_1\sim p(c_1),s_2\sim q(s_2)}[||E^s_2(G_2(c_1,s_2)) - s_2||_1] \quad (3)$$

where $q(s_2)$ is the prior $\mathcal{N}(0,\mathbf{I})$, $p(c_1)$ is given by $c_1 = E^c_1(x_1)$ and $x_1 \sim p(x_1)$.

Total loss

$$\mathcal{L}_{\text{GAN}}^{x_2} = \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)}[\log(1 - D_2(G_2(c_1, s_2)))] + \mathbb{E}_{x_2 \sim p(x_2)}[\log D_2(x_2)]$$

$$\min_{E_1, E_2, G_1, G_2} \max_{D_1, D_2} \mathcal{L}(E_1, E_2, G_1, G_2, D_1, D_2) = \mathcal{L}_{\text{GAN}}^{x_1} + \mathcal{L}_{\text{GAN}}^{x_2} + \lambda_x(\mathcal{L}_{\text{recon}}^{x_1} + \mathcal{L}_{\text{recon}}^{x_2}) + \lambda_c(\mathcal{L}_{\text{recon}}^{c_1} + \mathcal{L}_{\text{recon}}^{c_2}) + \lambda_s(\mathcal{L}_{\text{recon}}^{s_1} + \mathcal{L}_{\text{recon}}^{s_2})$$

\+ perceptual loss(instance norm(real_A), instance norm(fake_B))
=> content가 올바로 유지되고 있는가

Question

1. 우리는 N(0,1)에 맞춰주는 세팅 (KL div.) 없이 style을 encoding하였다. 어떻게 노말에서 샘플링이 가능한가?

2. Shared Content Space라고 하는데, 정작 domain 간 서로 다른 encoder를 사용한다. 어떻게 된 것인가?

Details

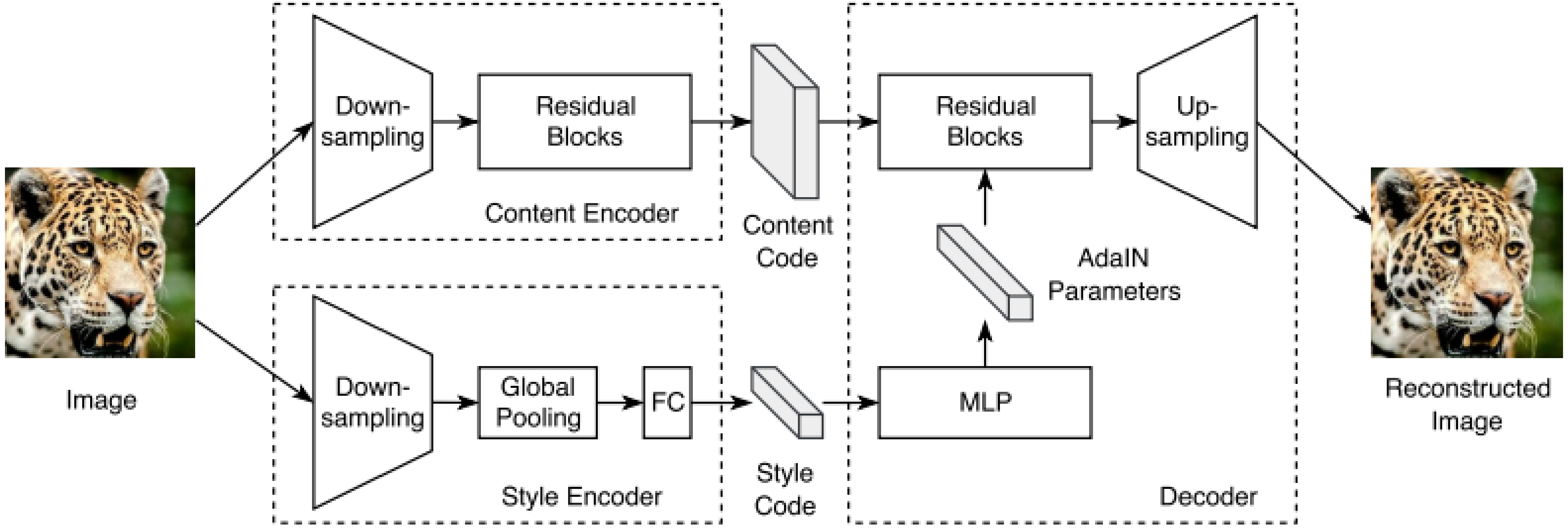
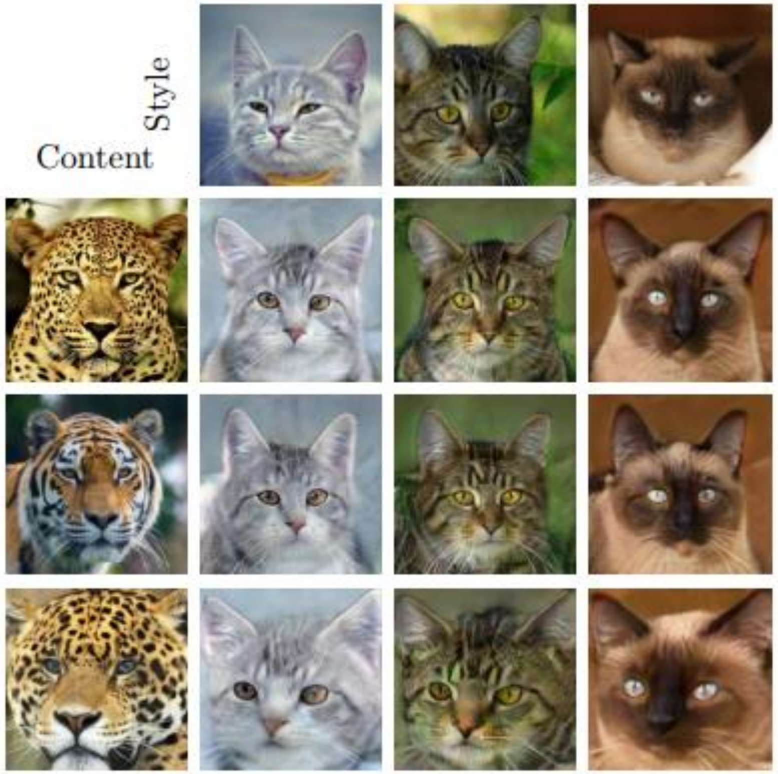


Fig. 3. Our auto-encoder architecture.

$$\mathrm{AdaIN}(z, \gamma, \beta) = \gamma \left(\frac{z - \mu(z)}{\sigma(z)} \right) + \beta$$

Qualitative test

(b) edges → shoes



(b) big cats → house cats

(a) edges ↔ shoes

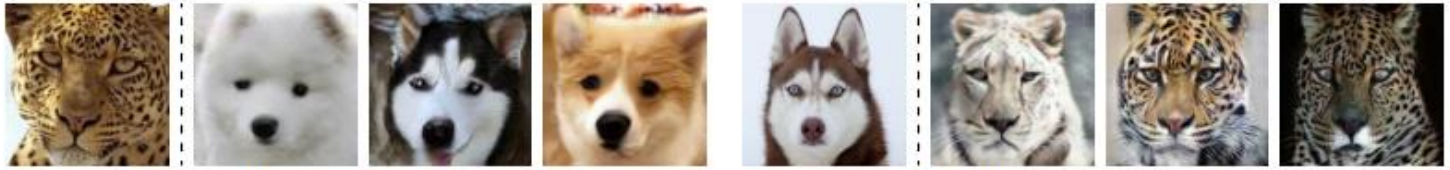
(b) edges ↔ handbags

(a) house cats → big cats

(b) big cats → house cats

(c) house cats → dogs

(d) dogs → house cats



(e) big cats → dogs

(f) dogs → big cats

Question

1. Normal에서 sampling한다고 하는데, 그 지점이 전 슬라이드 Details에서 어디일까

Ablation study

Fig. 4. Qualitative comparison on edges → shoes. The first column shows the input and ground truth output. Each following column shows 3 random outputs from a method.

Quantitative result

	edges → shoes		edges → handbags	
	Quality	Diversity	Quality	Diversity
UNIT [15]	37.4%	0.011	37.3%	0.023
CycleGAN [8]	36.0%	0.010	40.8%	0.012
CycleGAN* [8] with noise	29.5%	0.016	45.1%	0.011
MUNIT w/o $\mathcal{L}^{x}_{\text{recon}}$	6.0%	0.213	29.0%	0.191
MUNIT w/o $\mathcal{L}^{c}_{\text{recon}}$	20.7%	0.172	9.3%	0.185
MUNIT w/o $\mathcal{L}^{s}_{\text{recon}}$	28.6%	0.070	24.6%	0.139
MUNIT	50.0%	0.109	50.0%	0.175
BicycleGAN [11]†	56.7%	0.104	51.2%	0.140
Real data	N/A	0.293	N/A	0.371

	CycleGAN		CycleGAN* with noise		UNIT		MUNIT	
	CIS	IS	CIS	IS	CIS	IS	CIS	IS
house cats → big cats	0.078	0.795	0.034	0.701	0.096	0.666	0.911	0.923
big cats → house cats	0.109	0.887	0.124	0.848	0.164	0.817	0.956	0.954
house cats → dogs	0.044	0.895	0.070	0.901	0.045	0.827	1.231	1.255
dogs → house cats	0.121	0.921	0.137	0.978	0.193	0.982	1.035	1.034
big cats → dogs	0.058	0.762	0.019	0.589	0.094	0.910	1.205	1.233
dogs → big cats	0.047	0.620	0.022	0.558	0.096	0.754	0.897	0.901
Average	0.076	0.813	0.068	0.762	0.115	0.826	1.039	1.050

CIS: Multimodal (diverse outputs from a single input image)
&
IS: High-quality