UNIT: Unsupervised Image-to-Image Translation Networks

한양대학교 AILAB 석사과정 엄희송

NIPS 2017

Ming-Yu Liu, Thomas Breuel, Jan Kautz (NVIDIA)

Left: input. **Right: neural network generated.** Resolution: 640x480

Left: input. **Right: neural network generated.** Resolution: 640x480

IMAGE-TO-IMAGE TRANSLATION

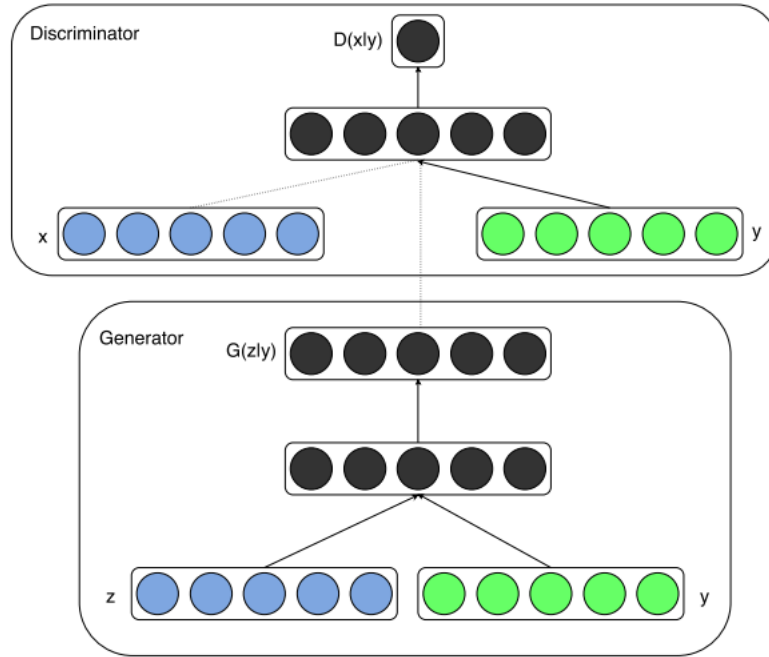"Mapping an image in one domain to a corresponding image in another domain."

Conditional GAN – pix2pix

Figure 1: Conditional adversarial net

Labels to Street Scene
input   output

Aerial to Map
input   output

Labels to Facade
input   output

BW to Color
input   output

Day to Night
input   output

Edges to Photo
input   output

Encoder-decoder

$x \rightarrow$ ... $\rightarrow y$

U-Net

$x \rightarrow$ ... $\rightarrow y$

Conditional Generative Adversarial Nets, figure 1. https://arxiv.org/abs/1411.1784
Image-to-Image Translation with Conditional Adversarial Networks, figure 1 & 3. https://arxiv.org/pdf/1611.07004

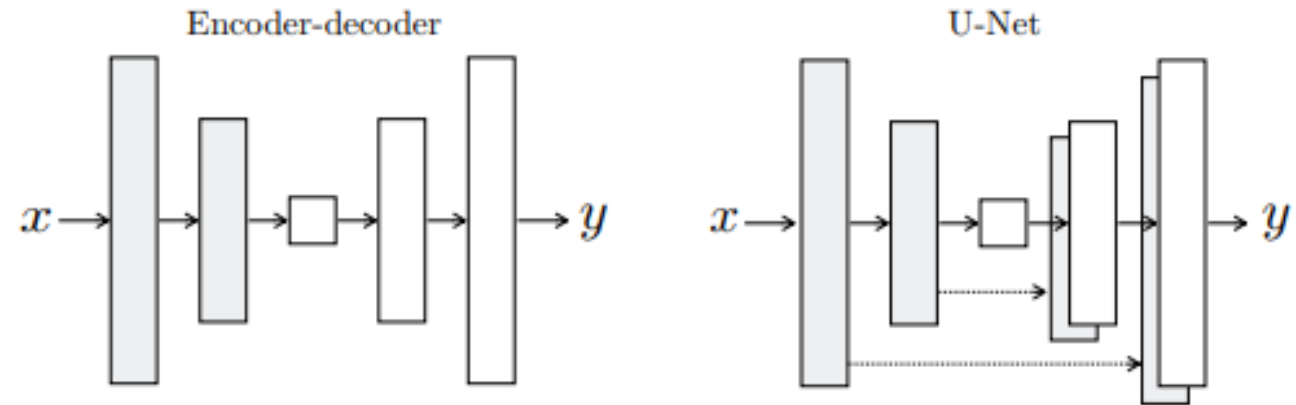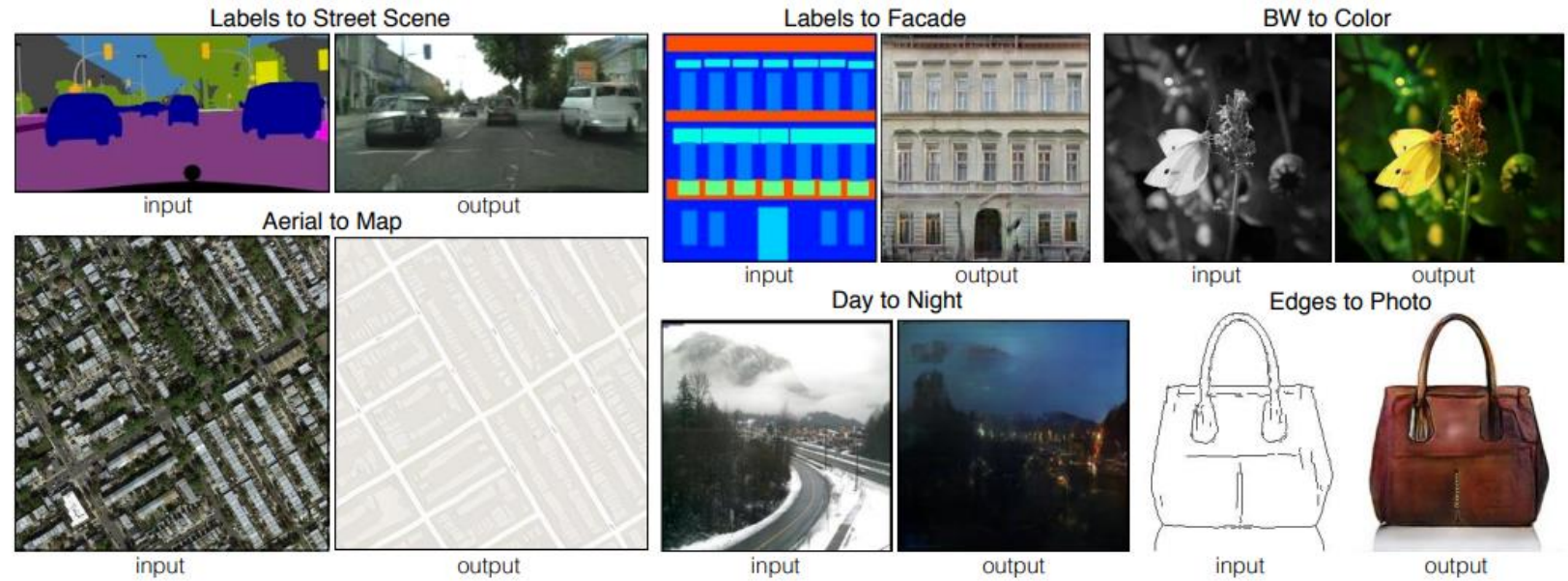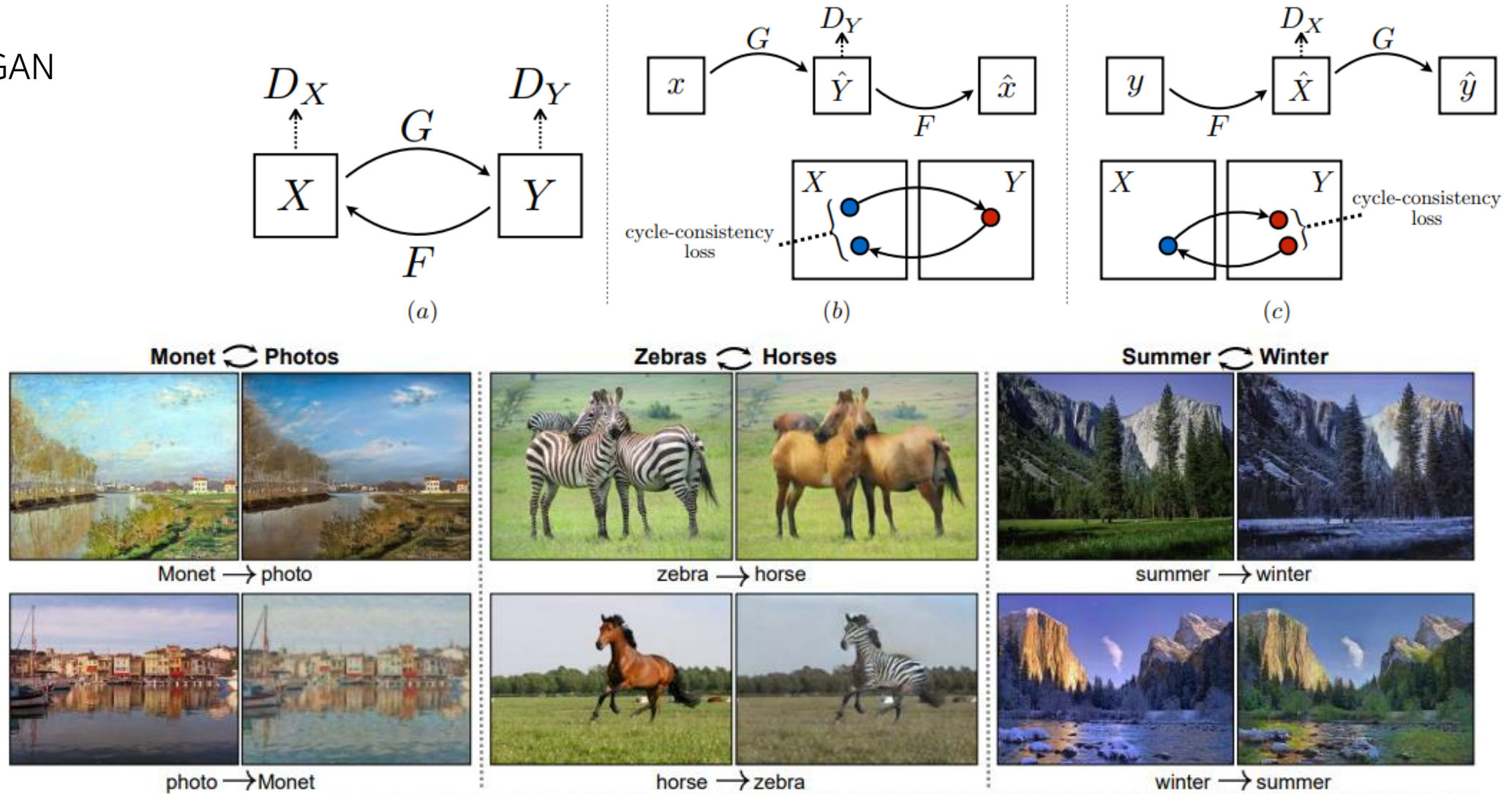# IMAGE-TO-IMAGE TRANSLATION
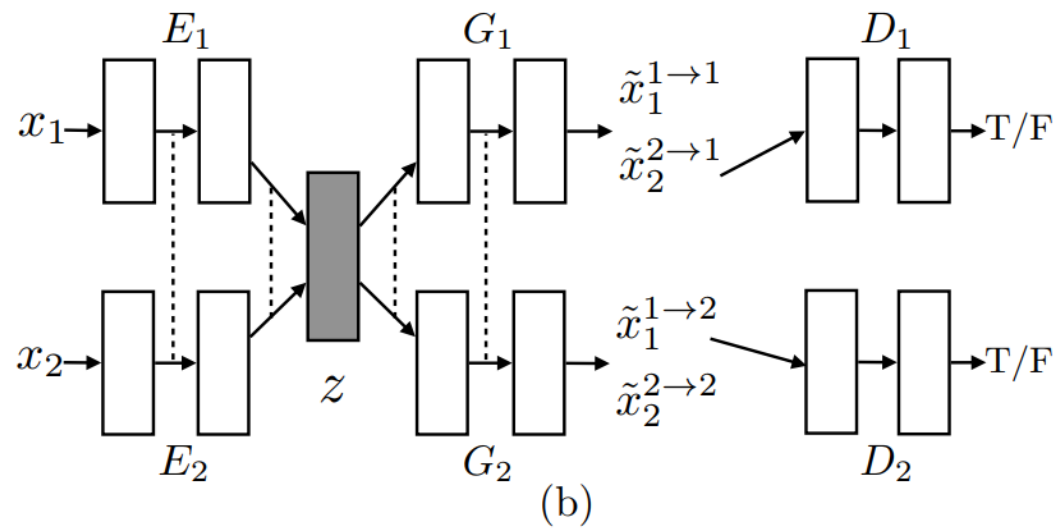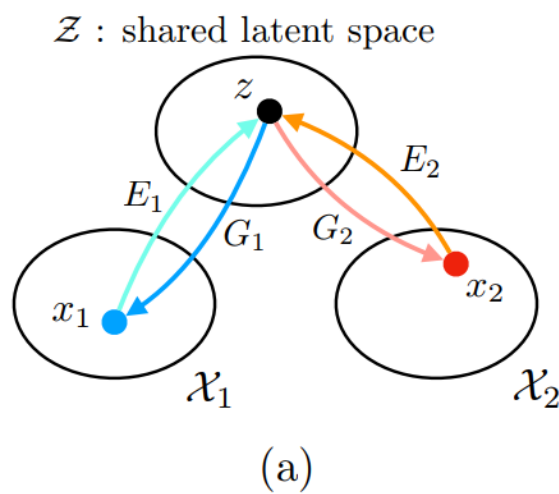
CycleGAN

Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Network, figure 1. https://arxiv.org/pdf/1703.10593

(a)

(b)

E: Encoder, G: Generator (encoder-decoder 에서 decoder 부분), D: Discriminator

$\tilde{x}_1^{1\to1}$, $\tilde{x}_2^{2\to2}$: self-reconstructed images

$\tilde{x}_1^{2\to1}$, $\tilde{x}_2^{1\to2}$: domain-translated images

목적함수

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

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

(a)

(b)

목적함수

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

$$\boxed{\mathcal{L}_{\text{VAE}_2}(E_2,G_2)} + \mathcal{L}_{\text{GAN}_2}(E_1,G_2,D_2) + \mathcal{L}_{\text{CC}_2}(E_2,G_2,E_1,G_1)$$
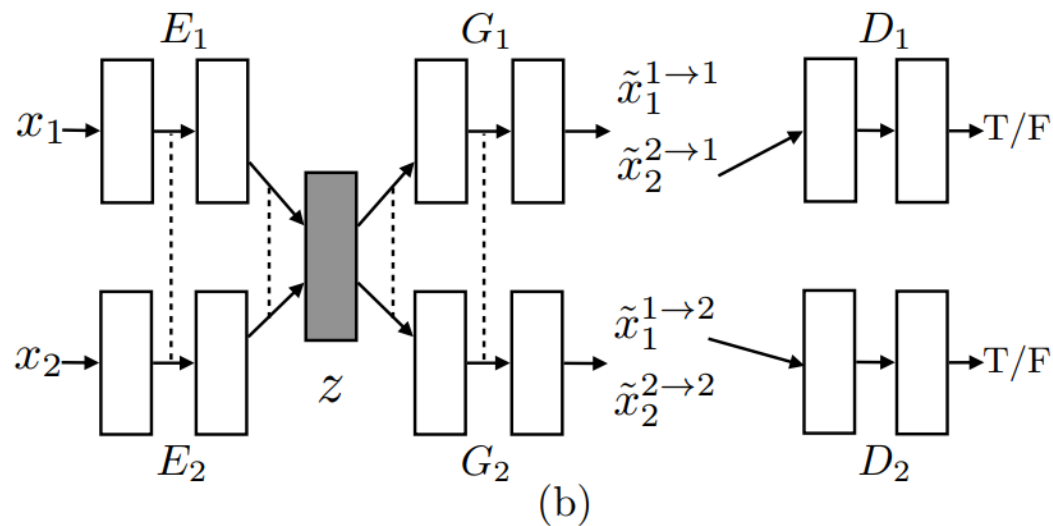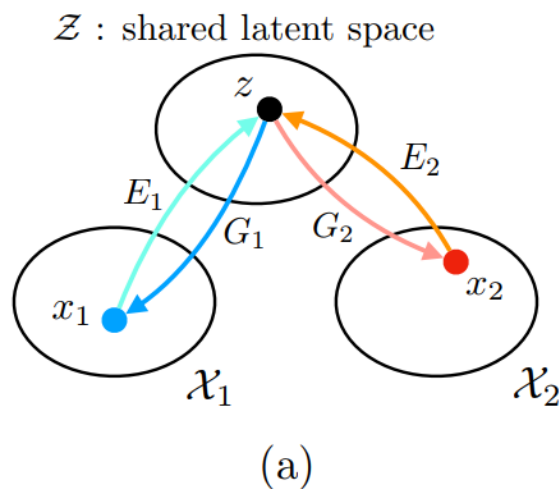
2 쌍의 VAE(Variational Auto-Encoder)를 사용

VAE 1: $x_1 \rightarrow z_1 \rightarrow x_1'$

VAE 2: $x_2 \rightarrow z_2 \rightarrow x_2'$

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

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

두 VAE가 같은 분포를 갖는 z로
encoding 하는 것을 학습시킴.

목적함수

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

$$\boxed{\mathcal{L}_{\text{VAE}_2}(E_2,G_2)} + \mathcal{L}_{\text{GAN}_2}(E_1,G_2,D_2) + \mathcal{L}_{\text{CC}_2}(E_2,G_2,E_1,G_1)$$
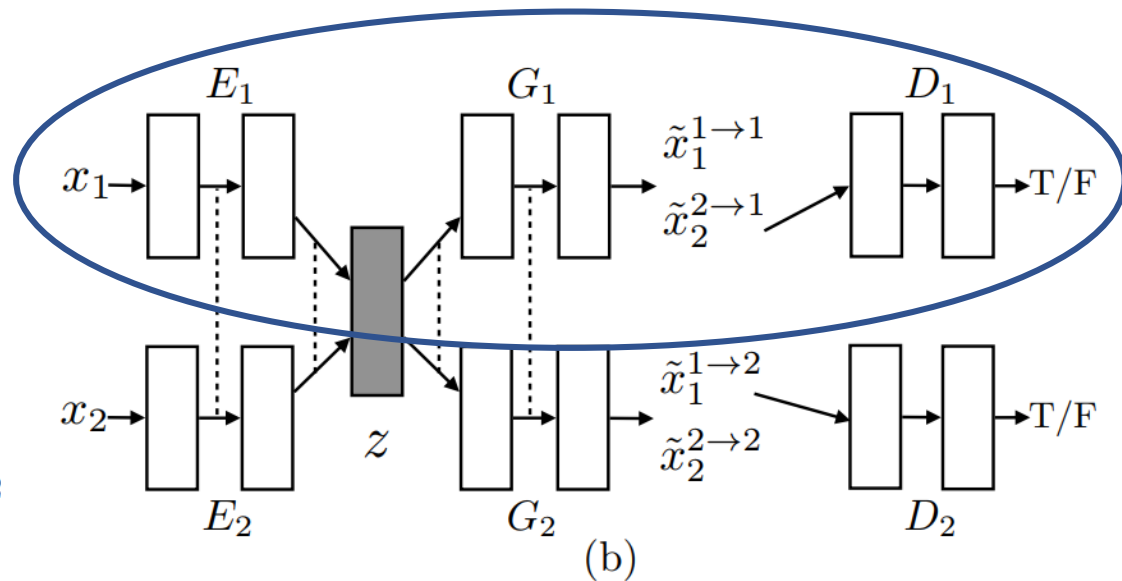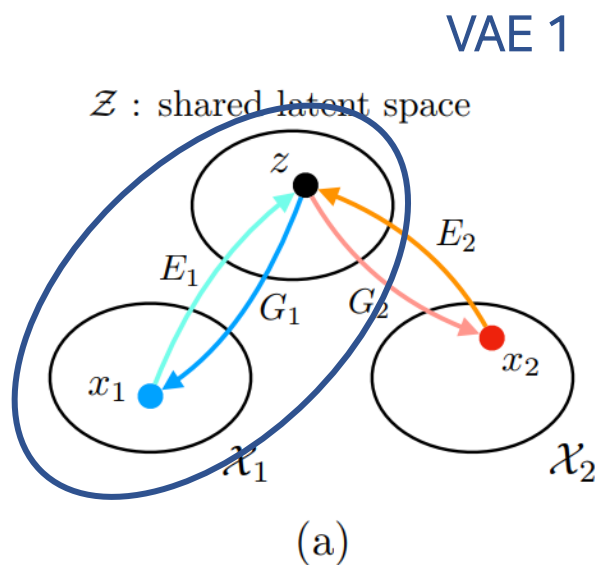
2 쌍의 VAE(Variational Auto-Encoder)를 사용

VAE 1: $x_1 \to z_1 \to x_1'$
VAE 2: $x_2 \to z_2 \to x_2'$

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

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

두 VAE가 같은 분포를 갖는 z로
encoding 하는 것을 학습시킴.

VAE 1

$\mathcal{Z}$: shared latent space

(a)

(b)

목적함수

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

$$\boxed{\mathcal{L}_{\text{VAE}_2}(E_2,G_2)} + \mathcal{L}_{\text{GAN}_2}(E_1,G_2,D_2) + \mathcal{L}_{\text{CC}_2}(E_2,G_2,E_1,G_1)$$
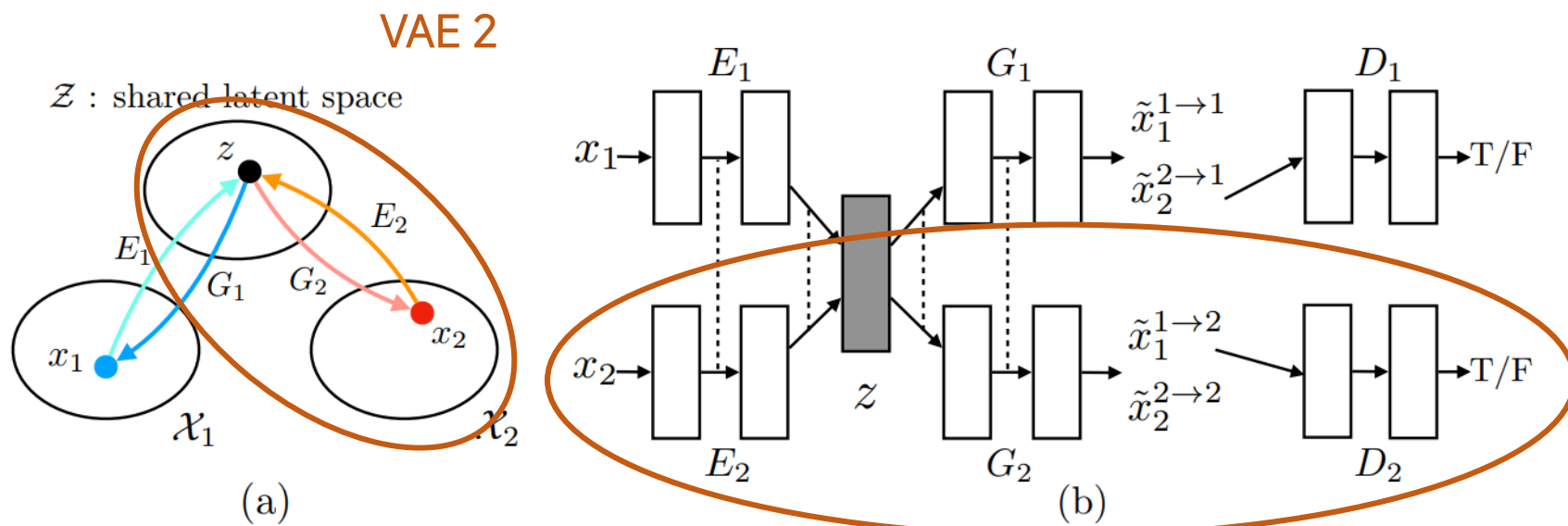
2 쌍의 VAE(Variational Auto-Encoder)를 사용

VAE 1: $x_1 \to z_1 \to x_1'$
VAE 2: $x_2 \to z_2 \to x_2'$

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

$$\mathcal{L}_{\text{VAE}_2}(E_2,G_2) = \lambda_1 \text{KL}(q_2(z_2|x_2)||p_\eta(z)) - \lambda_2 \mathbb{E}_{z_2 \sim q_2(z_2|x_2)}[\log p_{G_2}(x_2|z_2)]$$
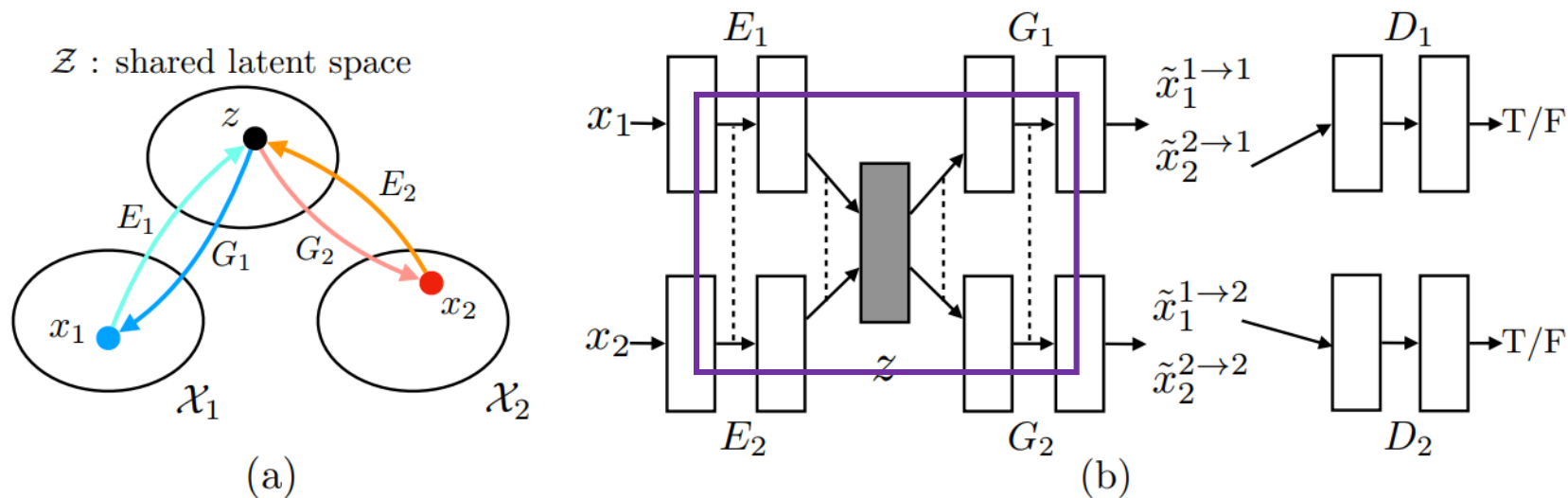
두 VAE가 같은 분포를 갖는 z로
encoding 하는 것을 학습시킴.

VAE 2

$\mathcal{Z}$: shared latent space

(a)

(b)

Weight Sharing

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

$$\boxed{\mathcal{L}_{\text{VAE}_2}(E_2, G_2)} + \mathcal{L}_{\text{GAN}_2}(E_1, G_2, D_2) + \mathcal{L}_{\text{CC}_2}(E_2, G_2, E_1, G_1)$$

Tie last few layer of E & first few layer of G (Dashed lines)

GAN training 과 같이 적용시킬 때 common latent code z 로 encoding 하도록 학습이 가능하다.
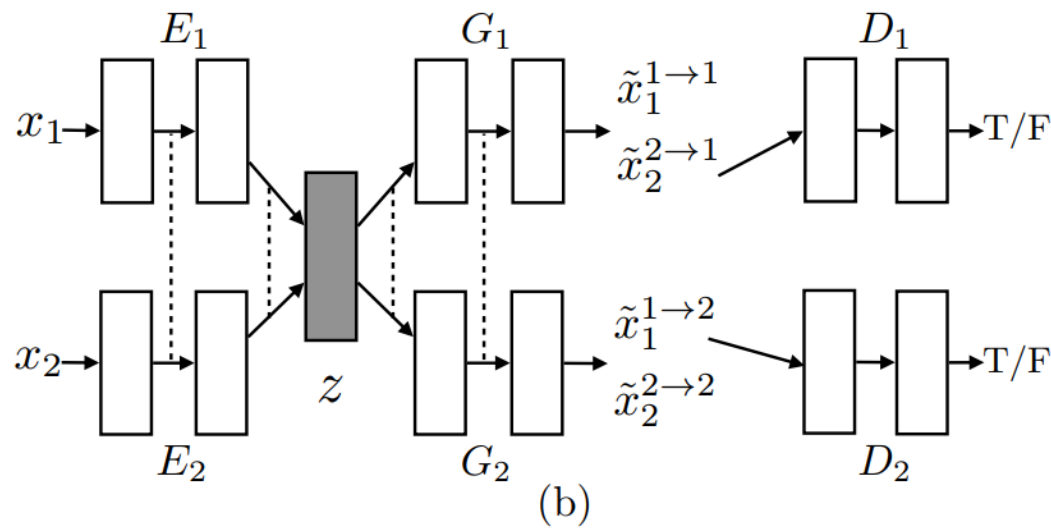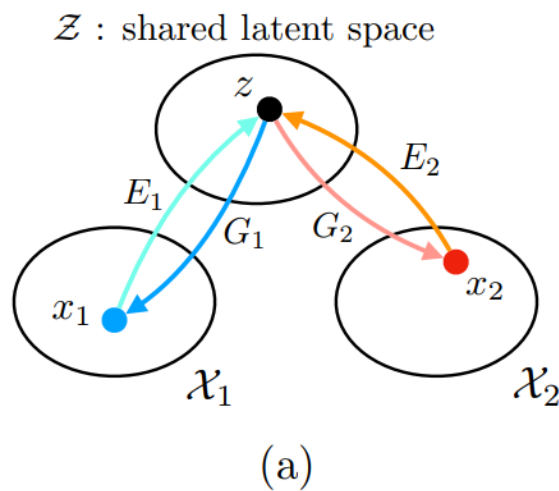

(a)   (b)

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

$$\mathcal{L}_{\text{VAE}_2}(E_2,G_2) + \boxed{\mathcal{L}_{\text{GAN}_2}(E_1,G_2,D_2)} + \mathcal{L}_{\text{CC}_2}(E_2,G_2,E_1,G_1)$$

$$\mathcal{L}_{\text{GAN}_1}(E_2,G_1,D_1) = \lambda_0 \mathbb{E}_{x_1 \sim P_{\mathcal{X}_1}}[\log D_1(x_1)] + \lambda_0 \mathbb{E}_{z_2 \sim q_2(z_2|x_2)}[\log(1 - D_1(G_1(z_2)))]$$

$$\mathcal{L}_{\text{GAN}_2}(E_1,G_2,D_2) = \lambda_0 \mathbb{E}_{x_2 \sim P_{\mathcal{X}_2}}[\log D_2(x_2)] + \lambda_0 \mathbb{E}_{z_1 \sim q_1(z_1|x_1)}[\log(1 - D_2(G_2(z_1)))]$$

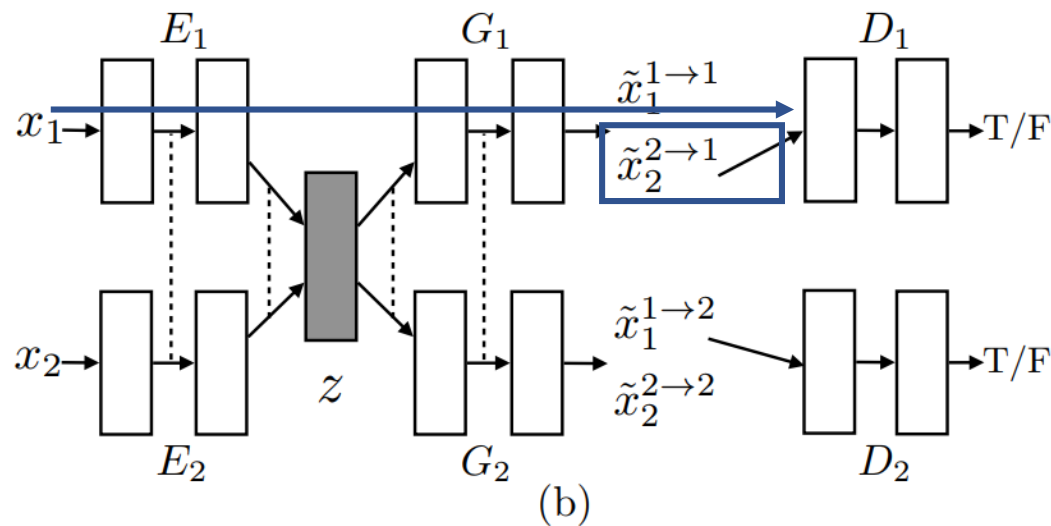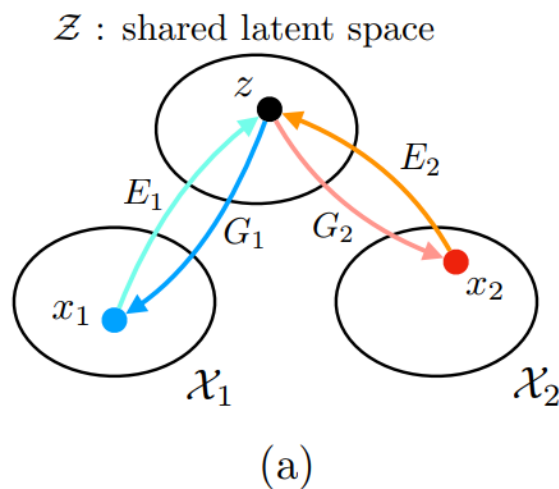Adversarial loss 에서는 domain-translated images 를 이용해 discriminator에서 판별함.

(a)

(b)

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

$$\mathcal{L}_{\text{VAE}_2}(E_2,G_2) + \boxed{\mathcal{L}_{\text{GAN}_2}(E_1,G_2,D_2)} + \mathcal{L}_{\text{CC}_2}(E_2,G_2,E_1,G_1)$$

$$\mathcal{L}_{\text{GAN}_1}(E_2,G_1,D_1) = \lambda_0 \mathbb{E}_{x_1 \sim P_{\mathcal{X}_1}}[\log D_1(x_1)] + \lambda_0 \mathbb{E}_{z_2 \sim q_2(z_2|x_2)}[\log(1 - D_1(G_1(z_2)))]$$

$$\mathcal{L}_{\text{GAN}_2}(E_1,G_2,D_2) = \lambda_0 \mathbb{E}_{x_2 \sim P_{\mathcal{X}_2}}[\log D_2(x_2)] + \lambda_0 \mathbb{E}_{z_1 \sim q_1(z_1|x_1)}[\log(1 - D_2(G_2(z_1)))]$$

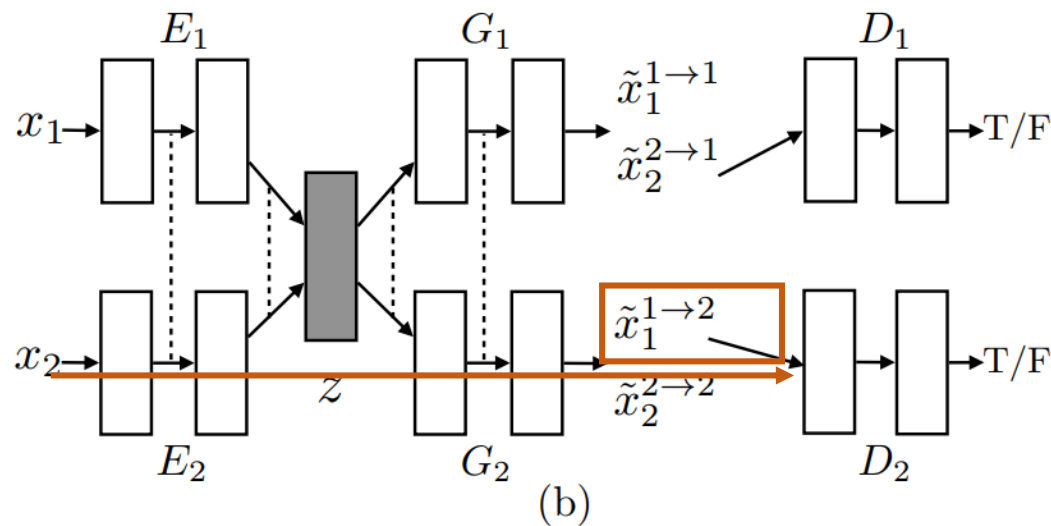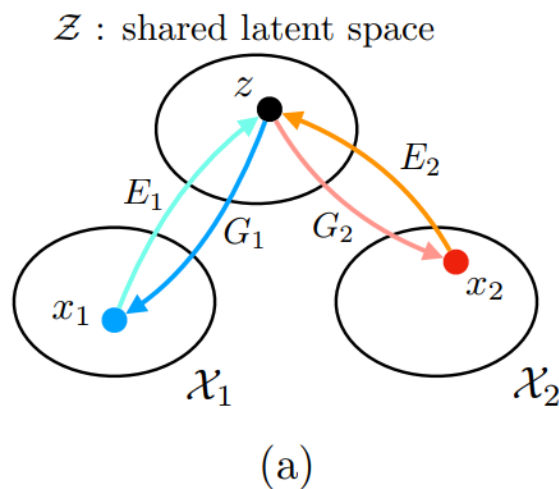Adversarial loss 에서는 domain-translated images 를 이용해 discriminator에서 판별함.

(a)

(b)

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

$$\mathcal{L}_{\text{VAE}_2}(E_2,G_2) + \boxed{\mathcal{L}_{\text{GAN}_2}(E_1,G_2,D_2)} + \mathcal{L}_{\text{CC}_2}(E_2,G_2,E_1,G_1)$$

$$\mathcal{L}_{\text{GAN}_1}(E_2,G_1,D_1) = \lambda_0 \mathbb{E}_{x_1 \sim P_{\mathcal{X}_1}}[\log D_1(x_1)] + \lambda_0 \mathbb{E}_{z_2 \sim q_2(z_2|x_2)}[\log(1 - D_1(G_1(z_2)))]$$

$$\mathcal{L}_{\text{GAN}_2}(E_1,G_2,D_2) = \lambda_0 \mathbb{E}_{x_2 \sim P_{\mathcal{X}_2}}[\log D_2(x_2)] + \lambda_0 \mathbb{E}_{z_1 \sim q_1(z_1|x_1)}[\log(1 - D_2(G_2(z_1)))]$$

Adversarial loss 에서는 domain-translated images 를 이용해 discriminator에서 판별함.

$\mathcal{Z}$: shared latent space

(a)

(b)

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

$$\mathcal{L}_{\text{VAE}_2}(E_2,G_2) + \mathcal{L}_{\text{GAN}_2}(E_1,G_2,D_2) + \boxed{\mathcal{L}_{\text{CC}_2}(E_2,G_2,E_1,G_1)}$$
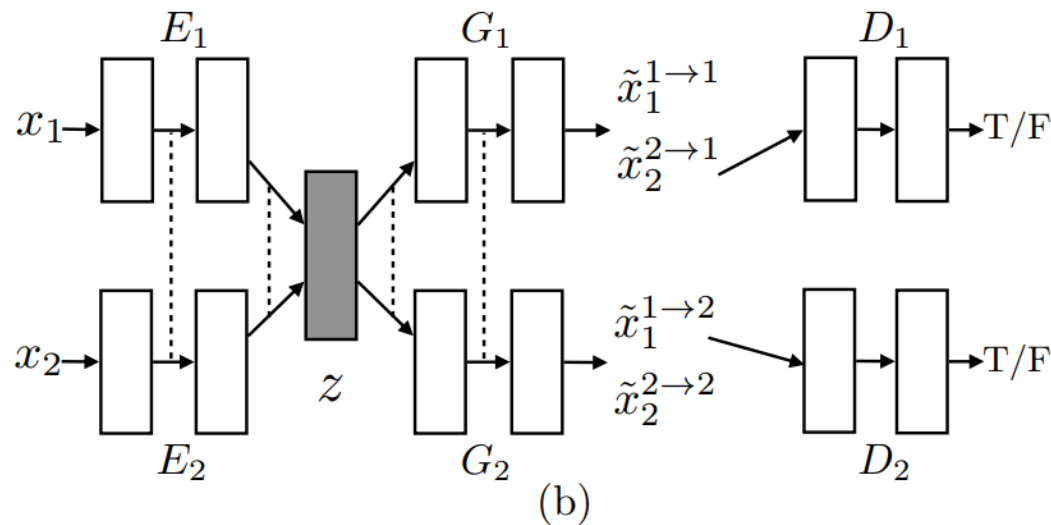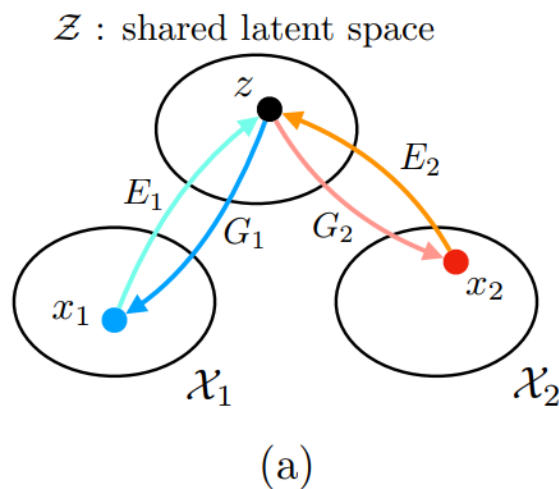
$$\mathcal{L}_{\text{CC}_1}(E_1,G_1,E_2,G_2) = \lambda_3 \text{KL}(q_1(z_1|x_1)||p_\eta(z)) + \lambda_3 \text{KL}(q_2(z_2|x_1^{1\to2}))||p_\eta(z)) -$$

$$\lambda_4 \mathbb{E}_{z_2 \sim q_2(z_2|x_1^{1\to2})}[\log p_{G_1}(x_1|z_2)]$$

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

$$\lambda_4 \mathbb{E}_{z_1 \sim q_1(z_1|x_2^{2\to1})}[\log p_{G_2}(x_2|z_1)].$$

Cycle-consistency loss
는 cycleGAN 과 동일한 개념.

$$x_1 = F_{2\to1}^*(F_{1\to2}^*(x_1))$$

$$x_2 = F_{1\to2}^*(F_{2\to1}^*(x_2))$$

(a)

(b)

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

$$\mathcal{L}_{\text{VAE}_2}(E_2,G_2) + \mathcal{L}_{\text{GAN}_2}(E_1,G_2,D_2) + \boxed{\mathcal{L}_{\text{CC}_2}(E_2,G_2,E_1,G_1)}$$

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

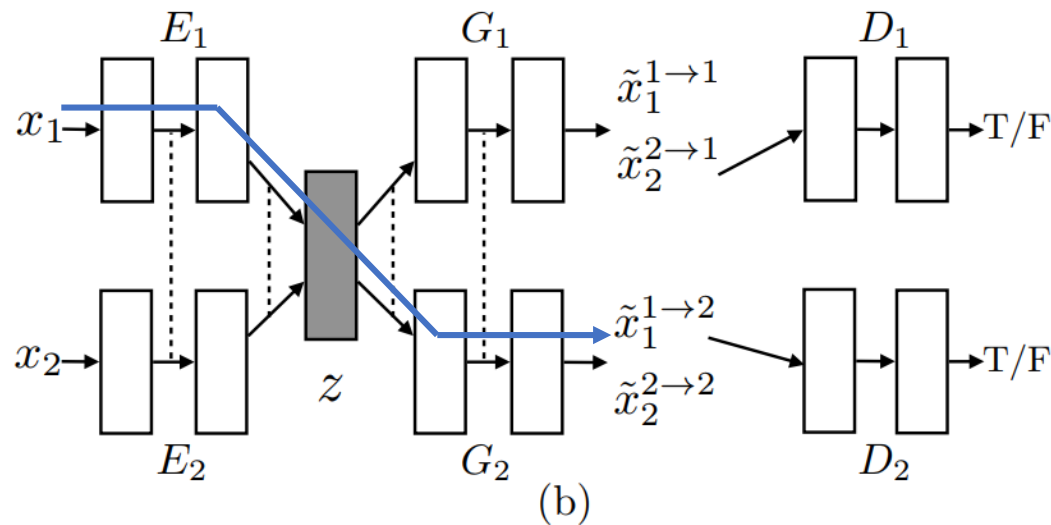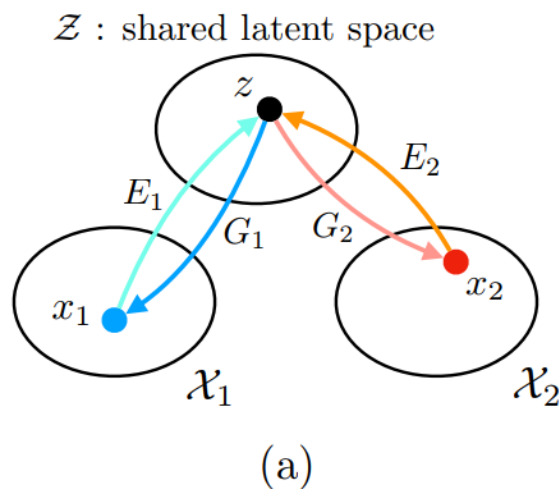$$\lambda_4 \mathbb{E}_{z_2 \sim q_2(z_2|x_1^{1\to2})}[\log p_{G_1}(x_1|z_2)]$$

Cycle-consistency loss
는 cycleGAN 과 동일한 개념.

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

$$\lambda_4 \mathbb{E}_{z_1 \sim q_1(z_1|x_2^{2\to1})}[\log p_{G_2}(x_2|z_1)].$$

$$x_1 = F_{2\to1}^*(F_{1\to2}^*(x_1))$$

$$x_2 = F_{1\to2}^*(F_{2\to1}^*(x_2))$$

$\mathcal{Z}$: shared latent space

(a)

(b)

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

$$\mathcal{L}_{\text{VAE}_2}(E_2,G_2) + \mathcal{L}_{\text{GAN}_2}(E_1,G_2,D_2) + \boxed{\mathcal{L}_{\text{CC}_2}(E_2,G_2,E_1,G_1)}$$
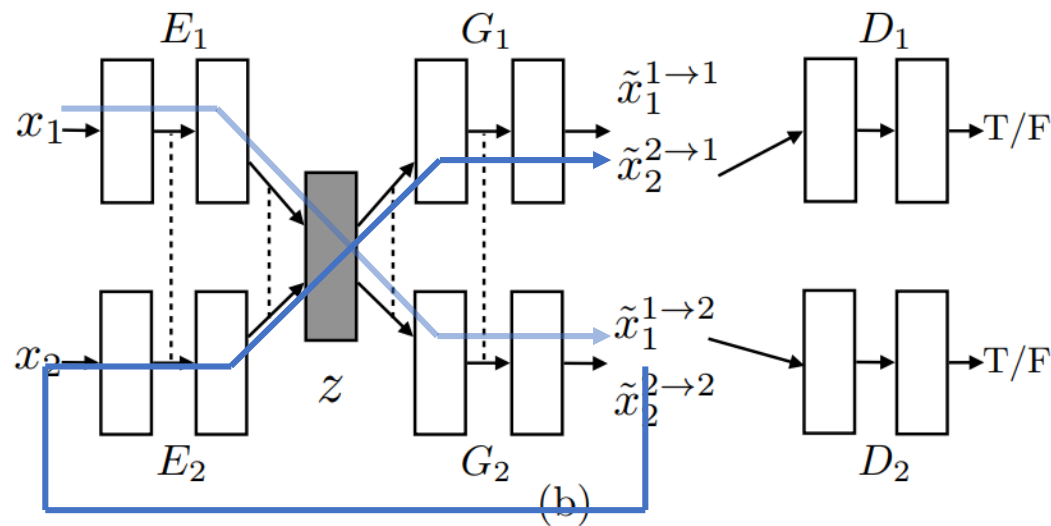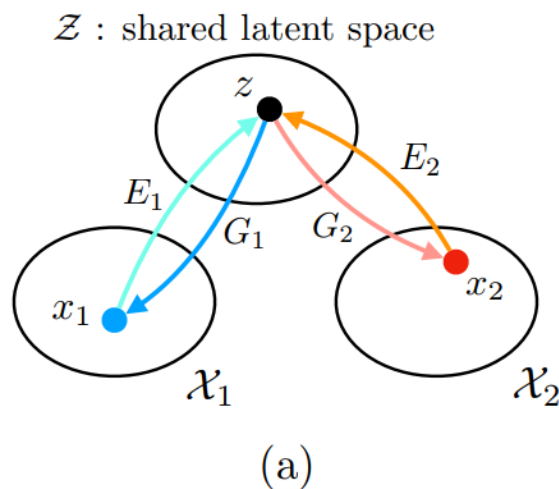
$$\mathcal{L}_{\text{CC}_1}(E_1,G_1,E_2,G_2) = \lambda_3 \text{KL}(q_1(z_1|x_1)||p_\eta(z)) + \lambda_3 \text{KL}(q_2(z_2|x_1^{1\to2}))||p_\eta(z)) -$$
$$\lambda_4 \mathbb{E}_{z_2 \sim q_2(z_2|x_1^{1\to2})}[\log p_{G_1}(x_1|z_2)]$$

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

Cycle-consistency loss
는 cycleGAN 과 동일한 개념.

$$x_1 = F_{2\to1}^*(F_{1\to2}^*(x_1))$$
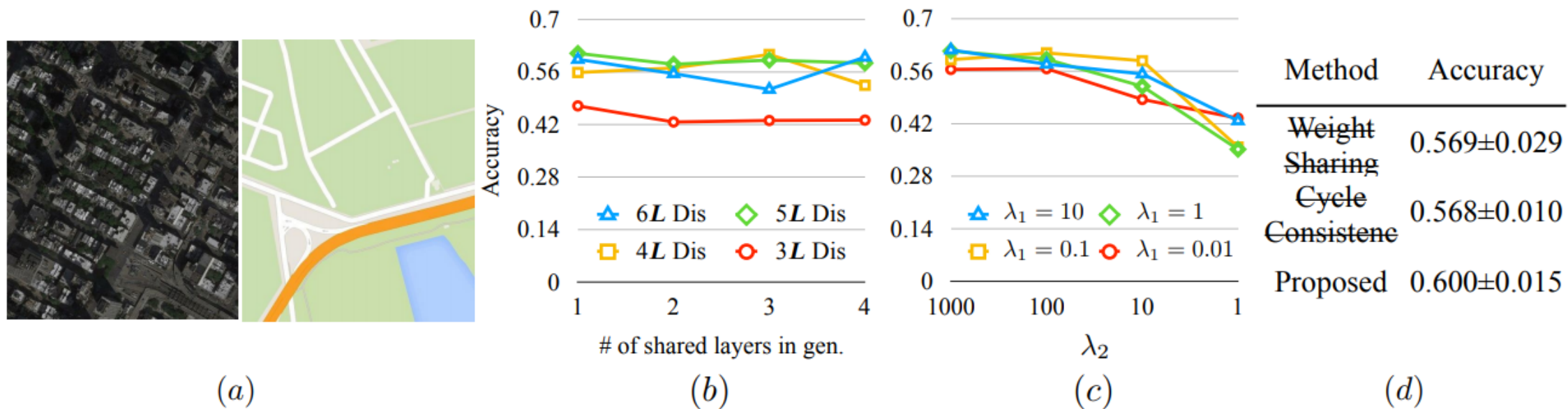
$$x_2 = F_{1\to2}^*(F_{2\to1}^*(x_2))$$

(a)

(b)

Figure 2: (a) Illustration of the Map dataset. Left: satellite image. Right: map. We translate holdout satellite images to maps and measure the accuracy achieved by various configurations of the proposed framework. (b) Translation accuracy versus different network architectures. (c) Translation accuracy versus different hyper-parameter values. (d) Impact of weight-sharing and cycle-consistency constraints on translation accuracy.
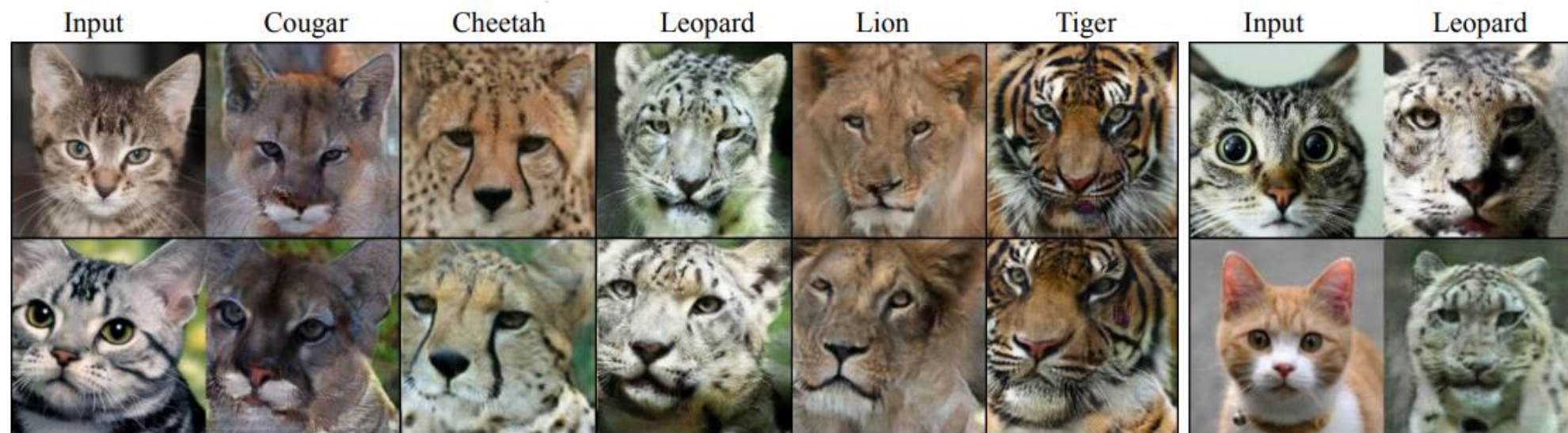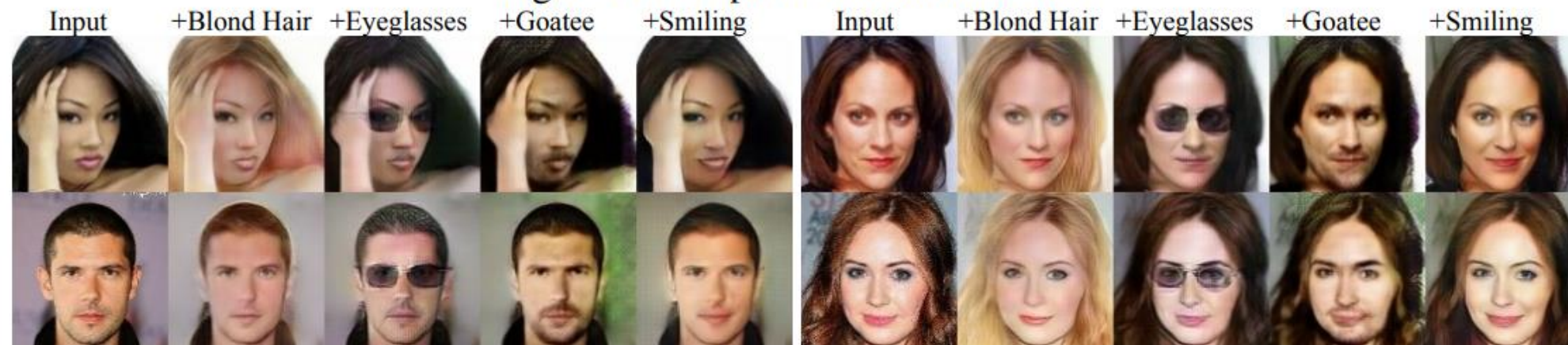
Figure 5: Cat species translation results.

Figure 6: Attribute-based face translation results.

Figure 3: Street scene image translation results. For each pair, left is input and right is the translated image.
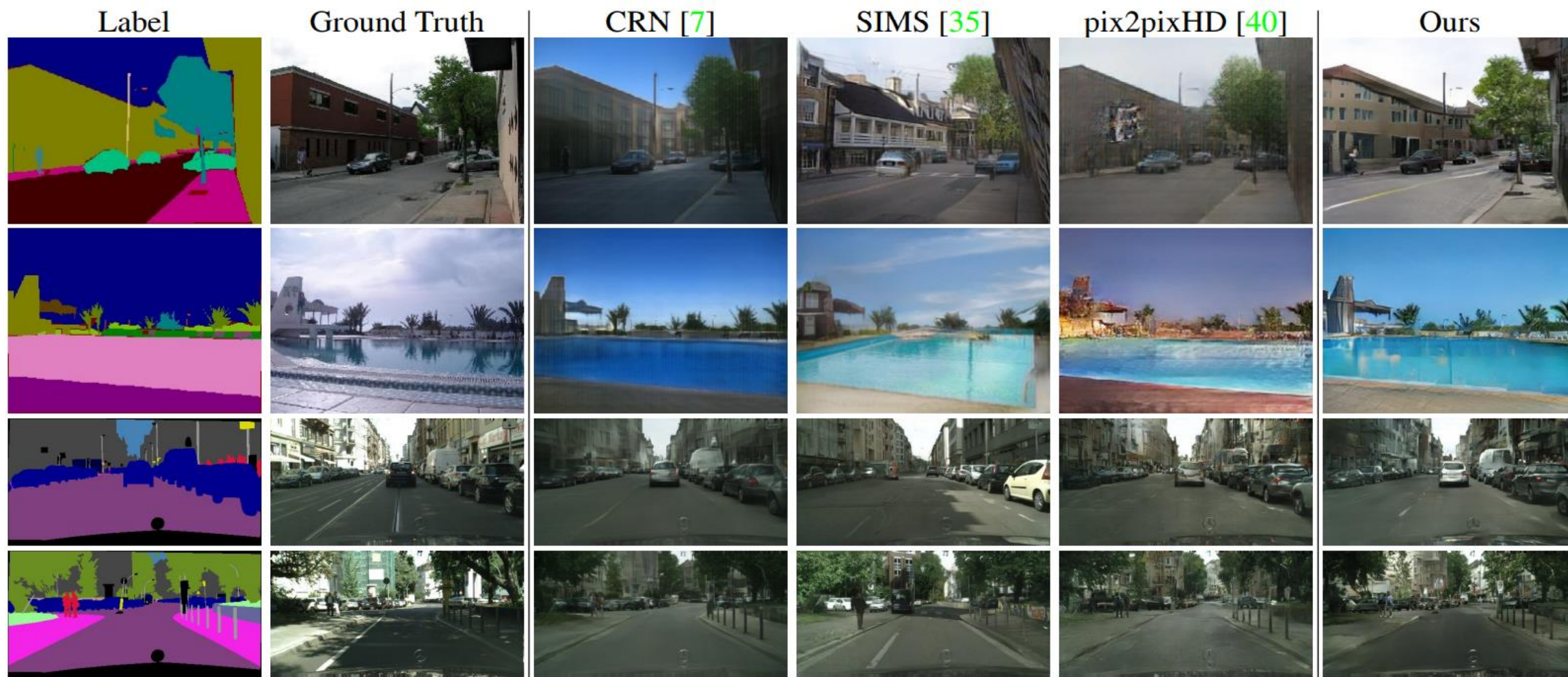
Figure 6: Visual comparison of semantic image synthesis results on the ADE20K outdoor and Cityscapes datasets. Our method produces realistic images while respecting the spatial semantic layout at the same time.

우리가 제안한 모델은 두 도메인 간 translation 을 매우 잘 한다.

Future work로 해결해야 할 과제는 다음과 같다.

      1. 현재 모델은 gaussian space assumption 때문에 unimodal. ➜ MUNIT 논문

      2. Saddle point 때문에 unstable training.