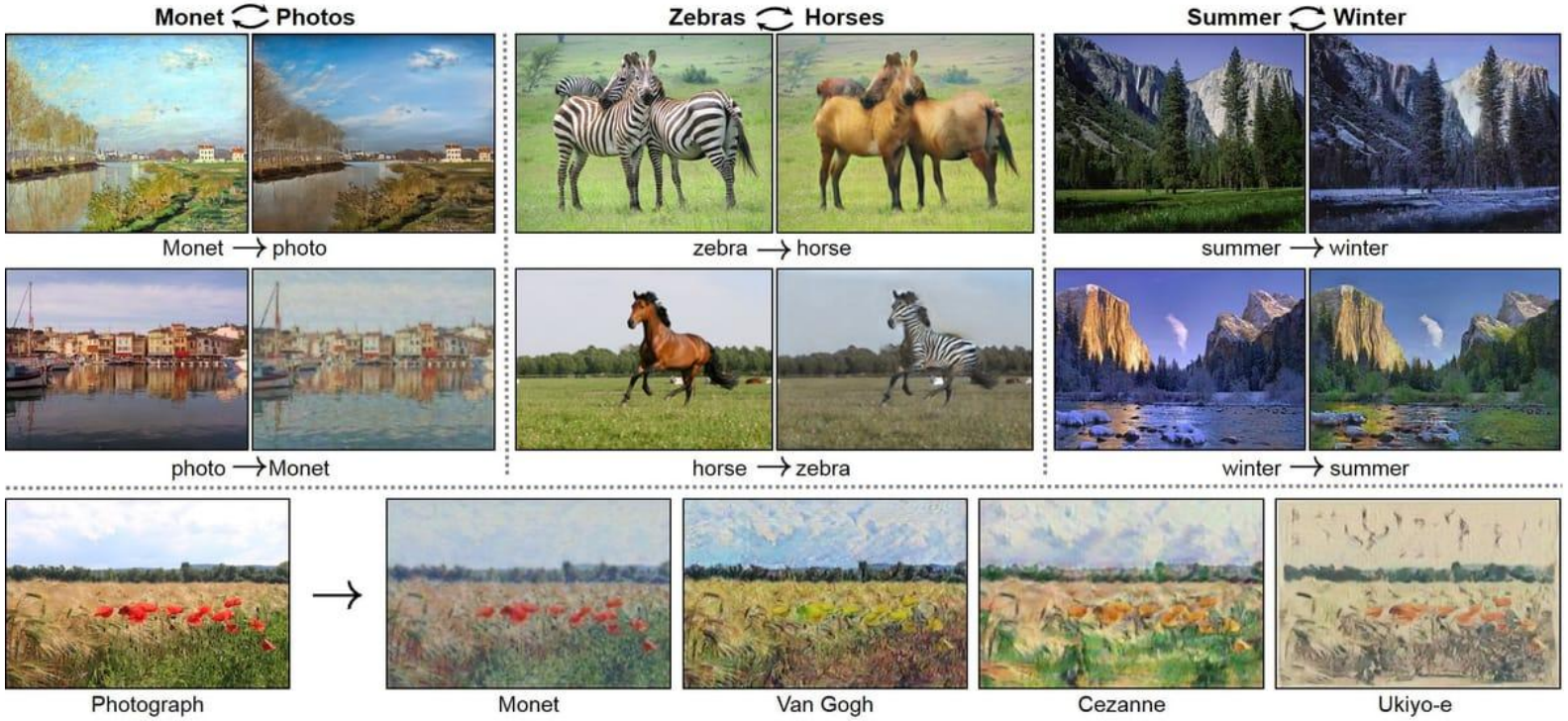
Practices in visual computing 2

Lab4: GANs

Simon Fraser University
Spring 2026

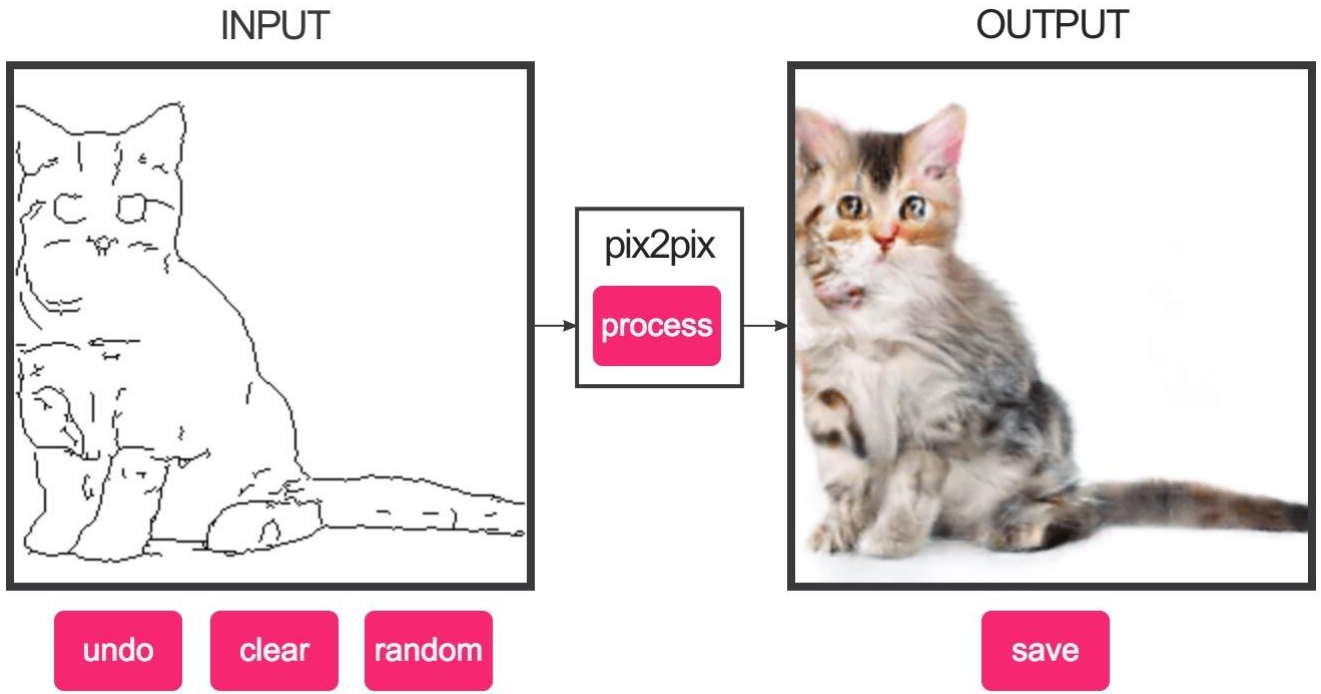
From Paired to Unpaired Translation

- Goal: learn a mapping between domains X and Y (e.g., photos ↔ paintings).
- Paired setting: we have aligned samples (x, y).
- Unpaired setting: we only have $x \sim p_x$ and $y \sim p_y$.
- Core challenge: realism is easy to enforce; structure preservation is the hard part.



Pix2Pix: Model Overview (Paired)

- Data: paired examples (x, y).
- Generator: $\hat{y} = G(x, z)$ (often z is dropped or injected via dropout).
- Discriminator: D(x, y) judges whether a pair is real or fake.
- Key idea: combine adversarial realism with a direct reconstruction term.



Pix2Pix: Loss Function

Conditional GAN loss

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))]$$

L_1 reconstruction loss

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1]$$

Final objective

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G)$$

CoGAN: Model Overview (Unpaired, Weight Sharing)

- Two generators: $G_1(z)$ for domain 1 and $G_2(z)$ for domain 2.
- Two discriminators: D_1 and D_2.
- Share high-level layers/weights between (G_1, G_2) and sometimes (D_1, D_2).
- Intuition: shared weights encourage a shared abstract representation even without pairs.

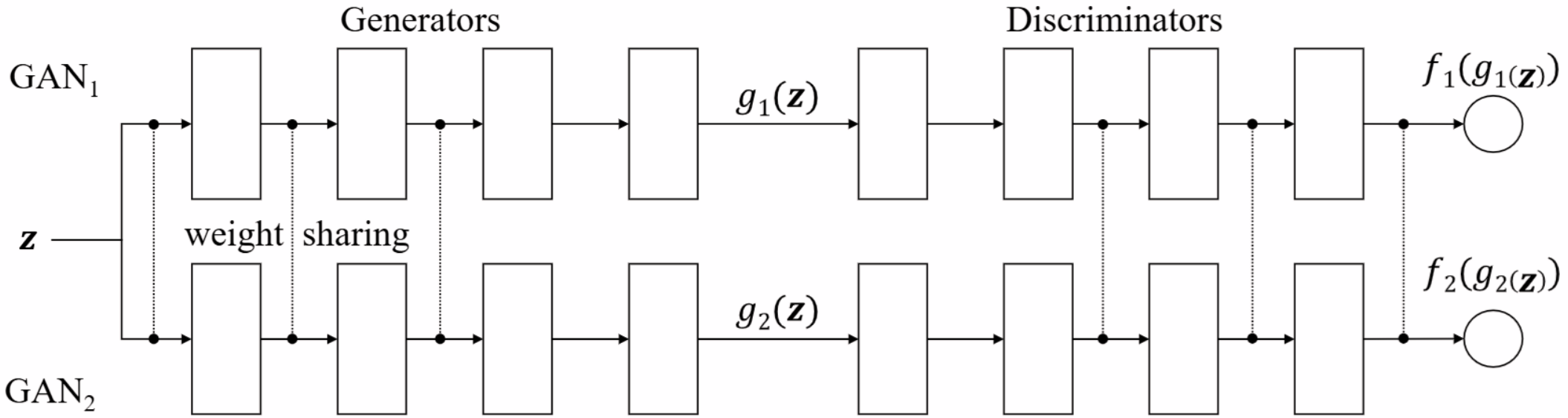
CoGAN: Loss Function

Independent GAN losses (per domain)

$$\mathcal{L}_{GAN}^{(i)}(G_i, D_i) = \mathbb{E}_{x_i \sim p_i}[\log D_i(x_i)] + \mathbb{E}_z[\log(1 - D_i(G_i(z)))], \quad i \in \{1, 2\}$$

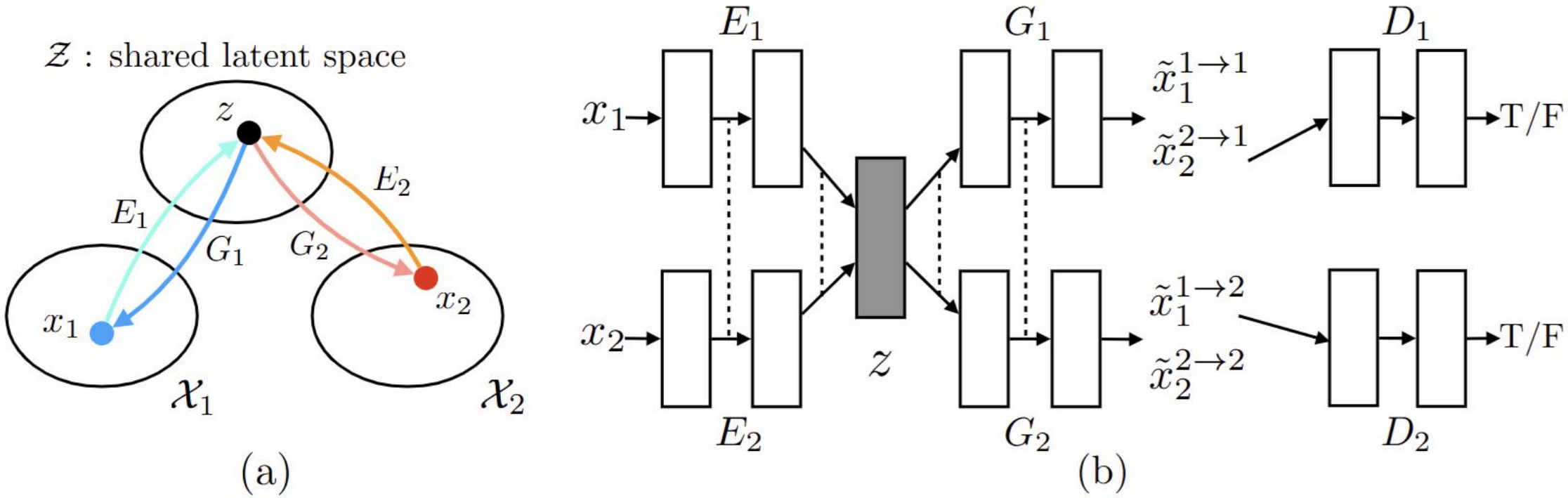
Total (conceptually)

$$\min_{G_1, G_2} \max_{D_1, D_2} \mathcal{L}_{GAN}^{(1)} + \mathcal{L}_{GAN}^{(2)} \quad \text{s. t. shared-weight constraints (architecture-level).}$$



UNIT: Model Overview (Shared Latent Space)

- Assumption: domains share a common latent space z.
- Encoders E_x, E_Y map images into z; decoders G_x, G_Y map z back.
- Combines VAE-style reconstruction + GAN realism + latent sharing (via weight sharing).
- Translation: $x \in X \rightarrow z = E_x(x) \rightarrow \hat{y} = G_Y(z)$.



UNIT: Loss Function (Typical Form)

VAE loss (per domain)

$$\mathcal{L}_{VAE}^{X} = \mathbb{E}_{x \sim p_X}[\|x - G_X(E_X(x))\|_1] + \beta \mathbb{E}_{x \sim p_X}[D_{KL}(q_X(z|x)\|\mathcal{N}(0, I))]$$

GAN loss (per domain)

$$\mathcal{L}_{GAN}^{X}(G_X, D_X), \quad \mathcal{L}_{GAN}^{Y}(G_Y, D_Y)$$

Latent / translation consistency

$$\mathcal{L}_z = \mathbb{E}_{x \sim p_X}[\|E_Y(G_Y(E_X(x))) - E_X(x)\|] + \mathbb{E}_{y \sim p_Y}[\|E_X(G_X(E_Y(y))) - E_Y(y)\|]$$

Total

$$\mathcal{L}_{UNIT} = \mathcal{L}_{VAE}^{X} + \mathcal{L}_{VAE}^{Y} + \mathcal{L}_{GAN}^{X} + \mathcal{L}_{GAN}^{Y} + \lambda_z \mathcal{L}_z$$

MUNIT: Model Overview (Content + Style)

- Decompose each image into content c and style s:

$$(c_X, s_X) = (E_X^c(x), E_X^s(x)), \quad (c_Y, s_Y) = (E_Y^c(y), E_Y^s(y))$$

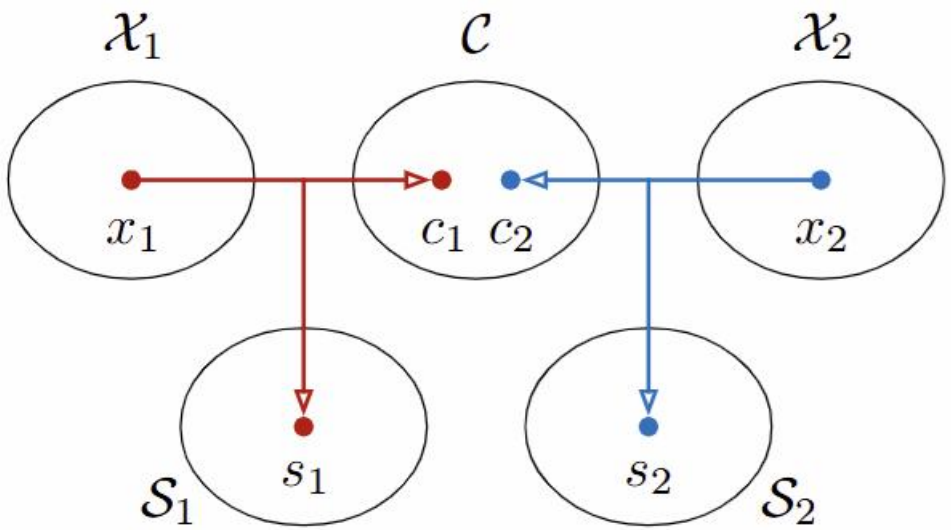
- Decoders/generators synthesize with both:

$$\hat{x} = G_X(c_X, s_X), \quad \hat{y} = G_Y(c_Y, s_Y)$$

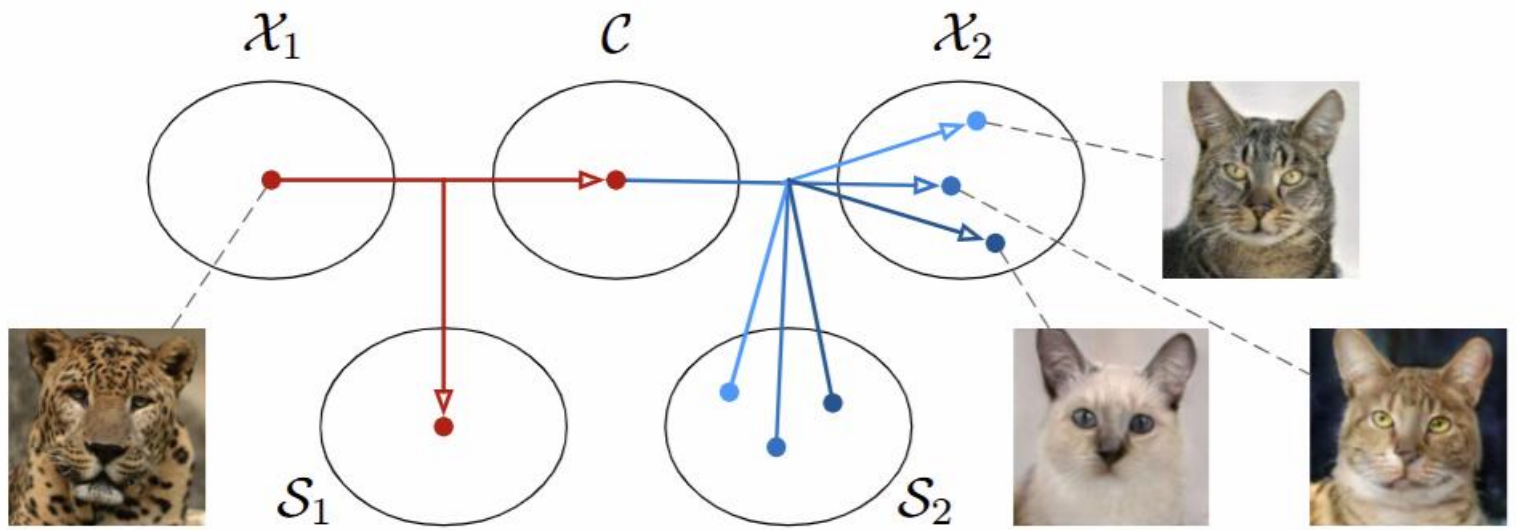
- Translation is multimodal by sampling style in the target domain:

$$\hat{y} = G_Y(c_X, s_Y), \quad s_Y \sim p(s_Y)$$

MUNIT: Model Overview (Content + Style)



(a) Auto-encoding



(b) Translation

MUNIT: Loss Function (Core Pieces)

Image reconstruction

$$\mathcal{L}^{x}_{recon} = \mathbb{E}_x[\|x - G_X(E^c_X(x), E^s_X(x))\|_1] + \mathbb{E}_y[\|y - G_Y(E^c_Y(y), E^s_Y(y))\|_1]$$

Content reconstruction (after translation)

$$\hat{y} = G_Y(c_X, s_Y),\ c_X = E^c_X(x),\ s_Y \sim p(s_Y), \quad \mathcal{L}^{c}_{recon} = \mathbb{E}_{x,\, s_Y}[\|E^c_Y(\hat{y}) - c_X\|_1] + (\text{symm.})$$

Style reconstruction

$$\mathcal{L}^{s}_{recon} = \mathbb{E}_{x,\, s_Y}[\|E^s_Y(\hat{y}) - s_Y\|_1] + (\text{symm.})$$

Style prior regularization (KL to Gaussian)

$$\mathcal{L}_{KL} = \mathbb{E}_x[D_{KL}(q_X(s|x)\|\mathcal{N}(0, I))] + \mathbb{E}_y[D_{KL}(q_Y(s|y)\|\mathcal{N}(0, I))]$$

Total

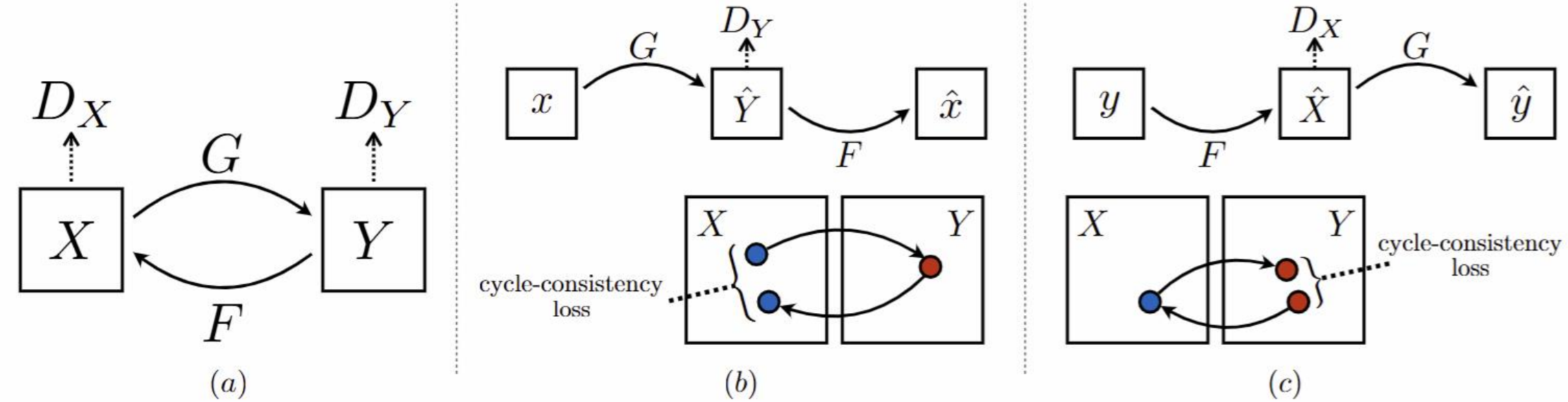
$$\mathcal{L}_{MUNIT} = \mathcal{L}_{GAN} + \lambda_x \mathcal{L}^{x}_{recon} + \lambda_c \mathcal{L}^{c}_{recon} + \lambda_s \mathcal{L}^{s}_{recon} + \lambda_{KL} \mathcal{L}_{KL}$$

CycleGAN: Model Overview (Unpaired, Pixel-Space Consistency)

- Two generators:

$$G : X \rightarrow Y, \qquad F : Y \rightarrow X$$

- Two discriminators: D_Y for domain Y and D_X for domain X.
- Idea: adversarial loss for realism + cycle consistency for structure.
- No paired data required.



CycleGAN: Loss Function (Assignment Core)

Adversarial losses

$$\mathcal{L}_{GAN}(G, D_Y) = \mathbb{E}_{y \sim p_Y}[\log D_Y(y)] + \mathbb{E}_{x \sim p_X}[\log(1 - D_Y(G(x)))]$$

$$\mathcal{L}_{GAN}(F, D_X) = \mathbb{E}_{x \sim p_X}[\log D_X(x)] + \mathbb{E}_{y \sim p_Y}[\log(1 - D_X(F(y)))]$$

Cycle consistency

$$\mathcal{L}_{cycle}(G, F) = \mathbb{E}_{x \sim p_X}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_Y}[\|G(F(y)) - y\|_1]$$

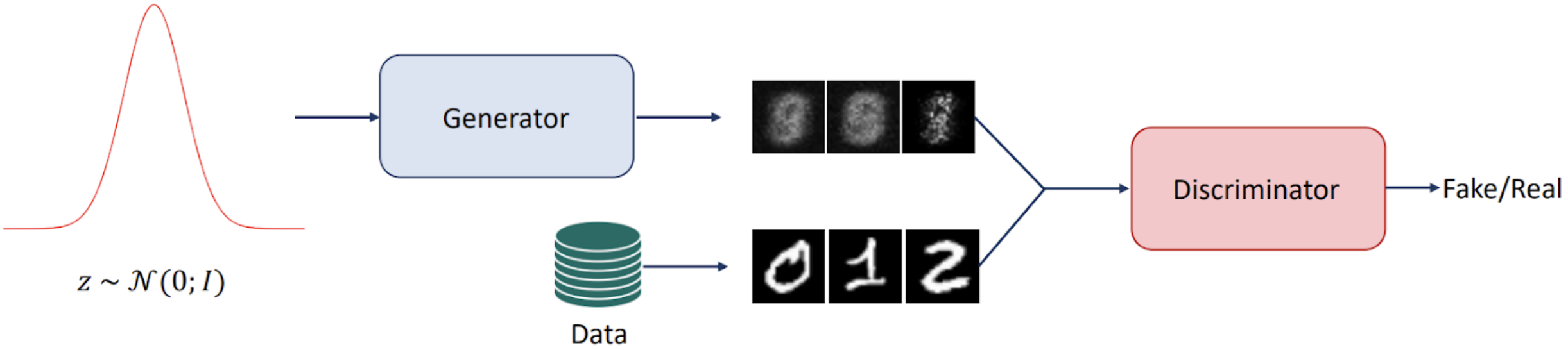
Total objective

$$\min_{G, F} \max_{D_X, D_Y} \mathcal{L}_{GAN}(G, D_Y) + \mathcal{L}_{GAN}(F, D_X) + \lambda \mathcal{L}_{cycle}(G, F)$$

Consistency Enforcement: Quick Comparison

- Pix2Pix: structure via paired supervision + L_1.
- CoGAN: weak coupling via shared weights.
- UNIT: consistency mainly in shared latent z (VAE + GAN).
- MUNIT: consistency mainly in content c (multimodal via style s).
- CycleGAN: consistency directly in pixel space via cycle loss.

GAN MNIST



Reference

- Pix2Pix: Isola et al., *Image-to-Image Translation with Conditional Adversarial Networks*, CVPR 2017.
- CoGAN: Isola et al., *Image-to-Image Translation with Conditional Adversarial Networks*, CVPR 2017.
- UNIT: Liu et al., *Unsupervised Image-to-Image Translation Networks*, NeurIPS 2017.
- MUNIT: Huang et al., *Multimodal Unsupervised Image-to-Image Translation*, ECCV 2018.
- CycleGAN: Zhu et al., *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*, ICCV 2017.