StarGAN v2: Diverse Image Synthesis for Multiple Domains

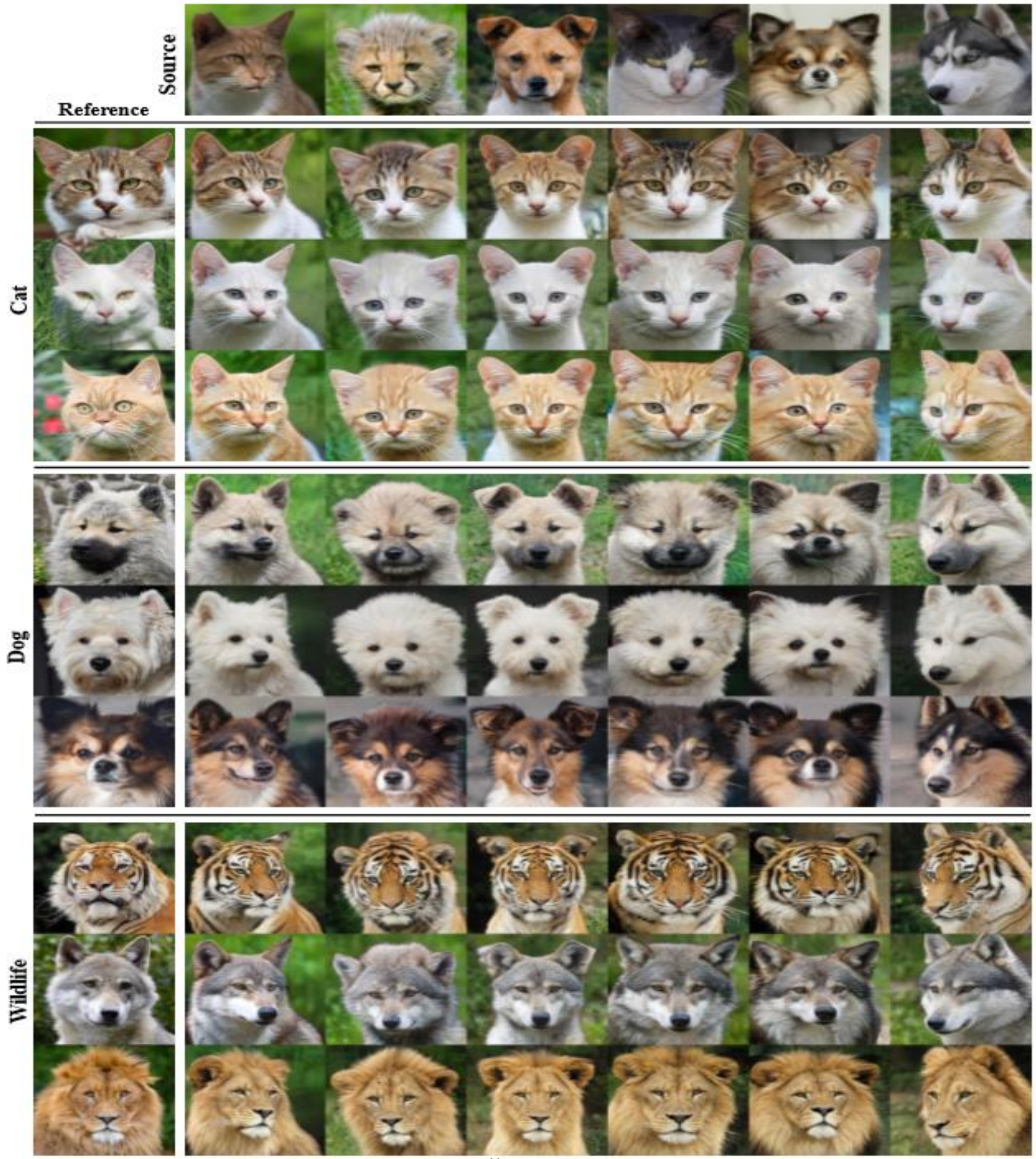
CVPR 2020
Naver Clova

Bochan Kim
2020. 05. 18 Mon

Results

Results

Results



GAN

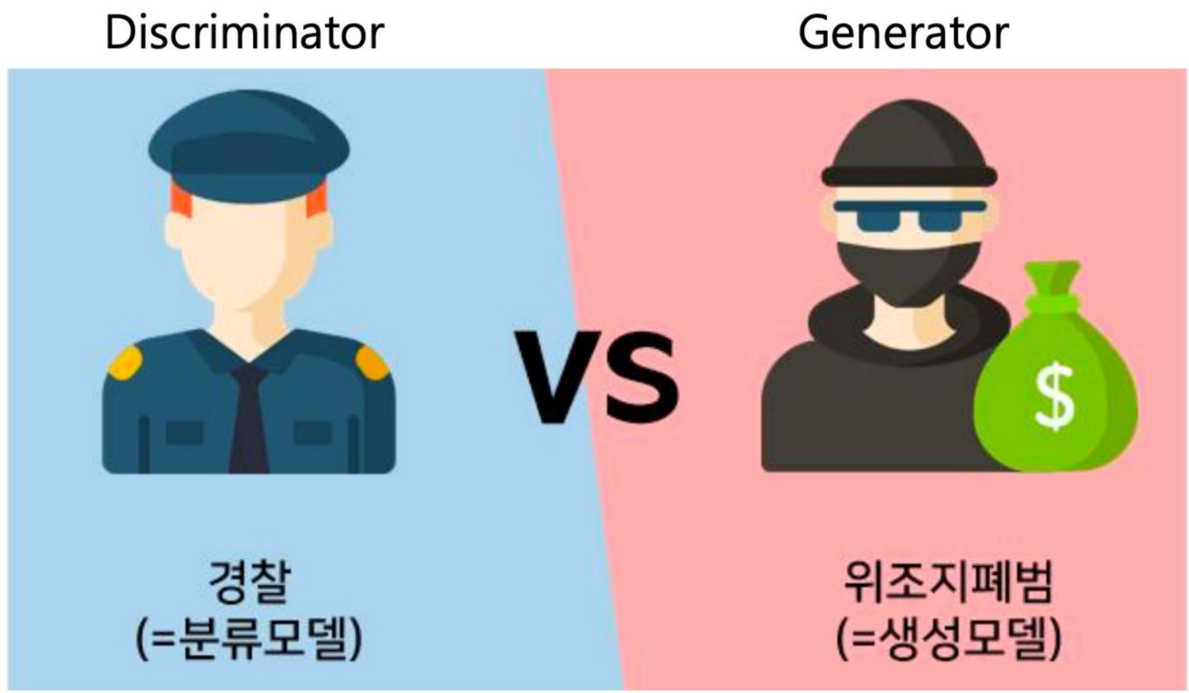
Image to Image Traslation

What is good I2I Translation model?

Trends / History

GAN

- Main Idea



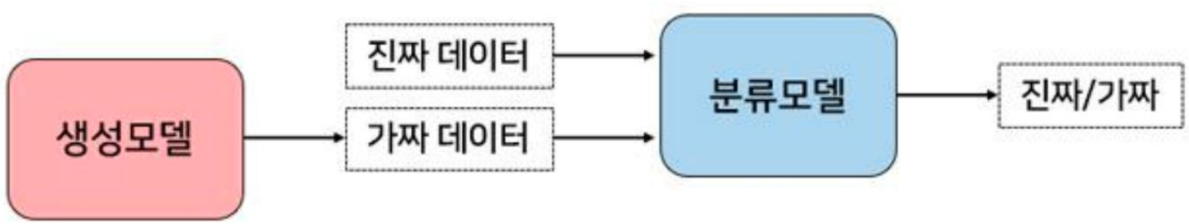
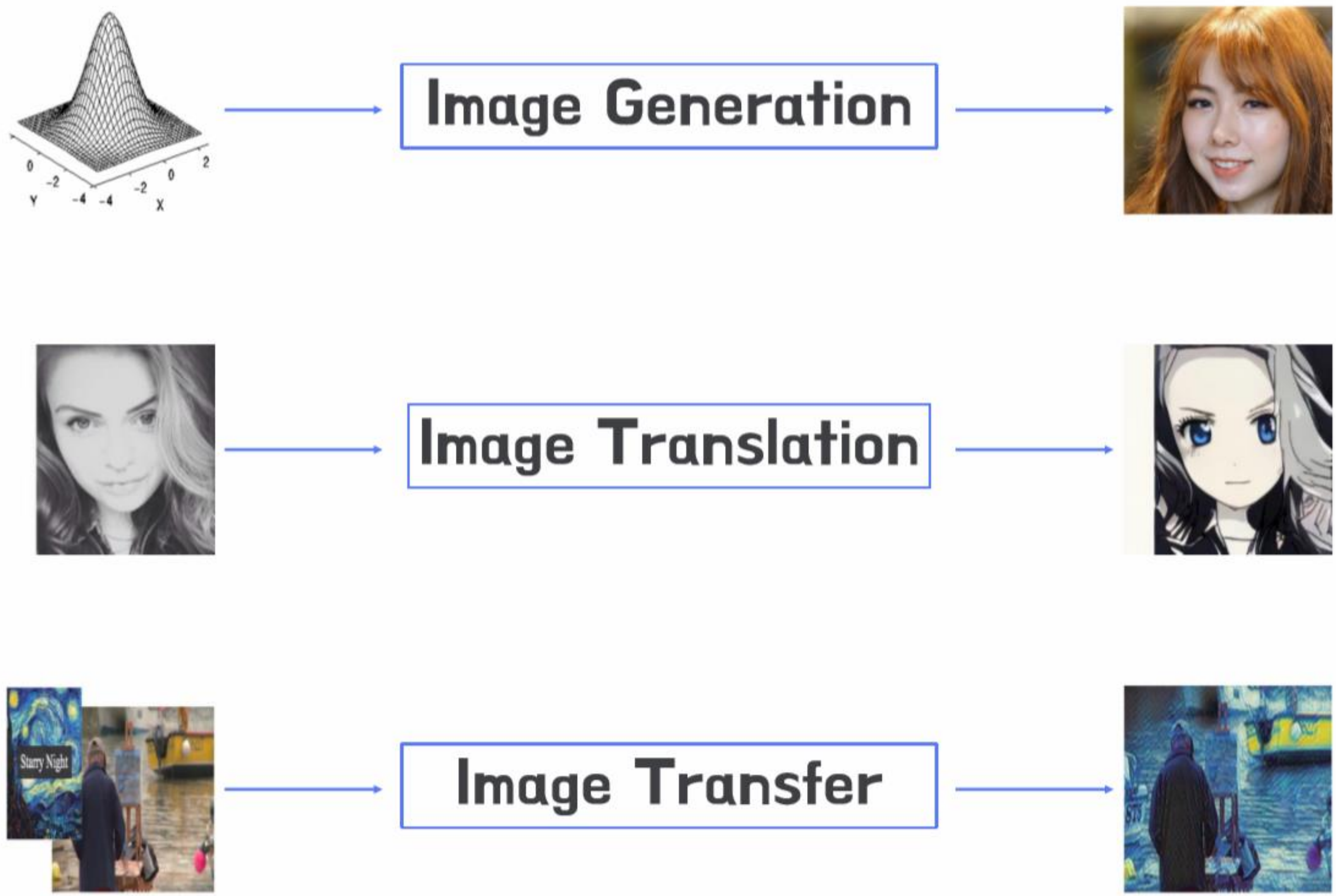


Image to Image Translation



Source : https://deview.kr/2019/schedule/279

What is good GAN model?

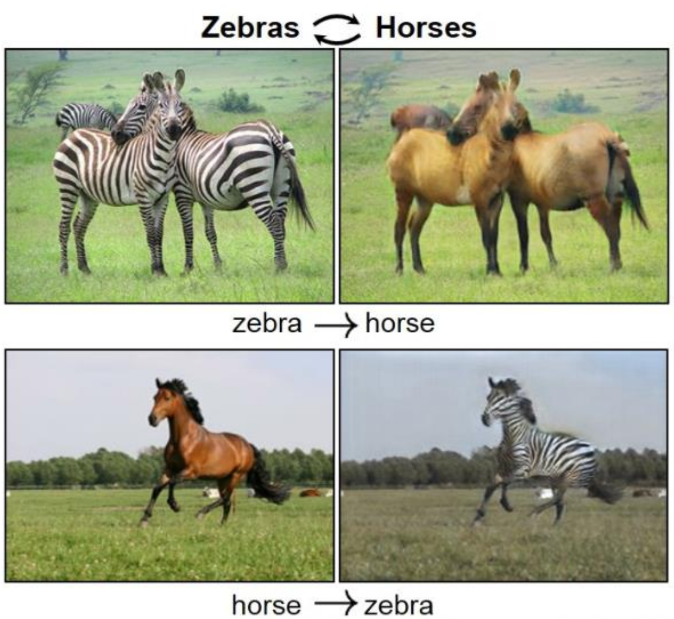
- Photo-realism

- Diversity
  - multi-modal is better than uni-modal
  - multi-domain is better than paired-domain

What is good GAN model?

- Photo-realism

- Diversity
  - multi-modal is better than uni-modal
  - multi-domain is better than uni-domain



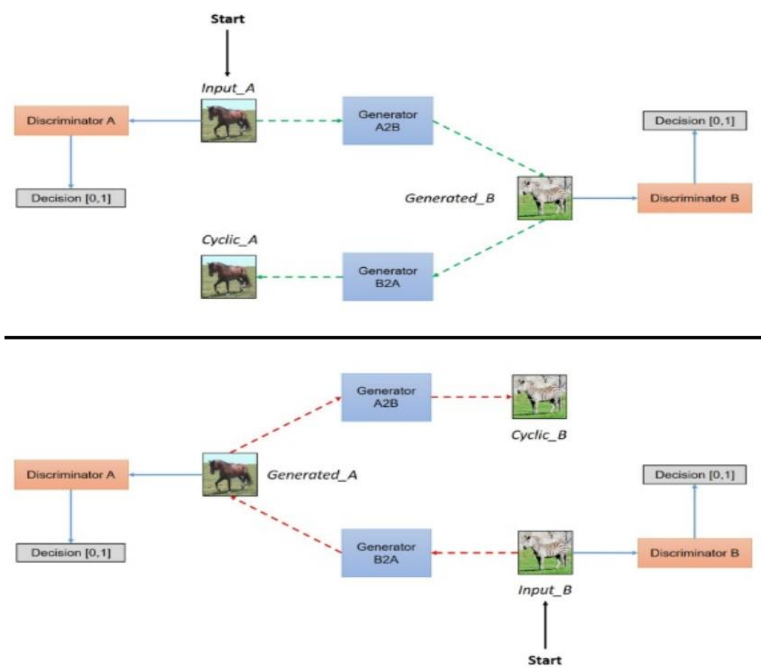
Uni-modal(cycleGAN)

Multi-modal(MUNIT)

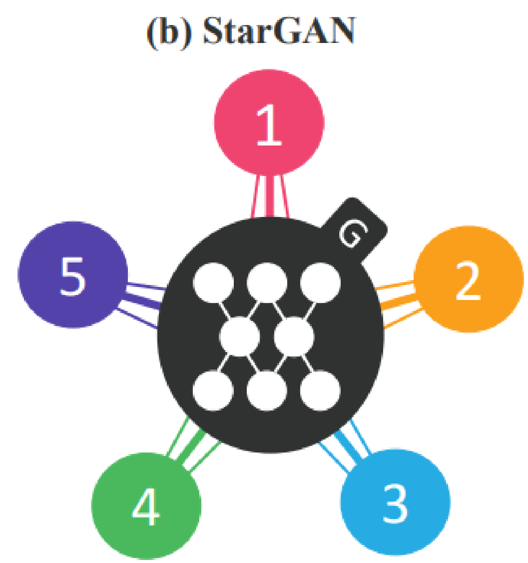
What is good GAN model?

- Photo-realism

- Diversity
  - multi-modal is better than uni-modal
  - multi-domain is better than uni-domain

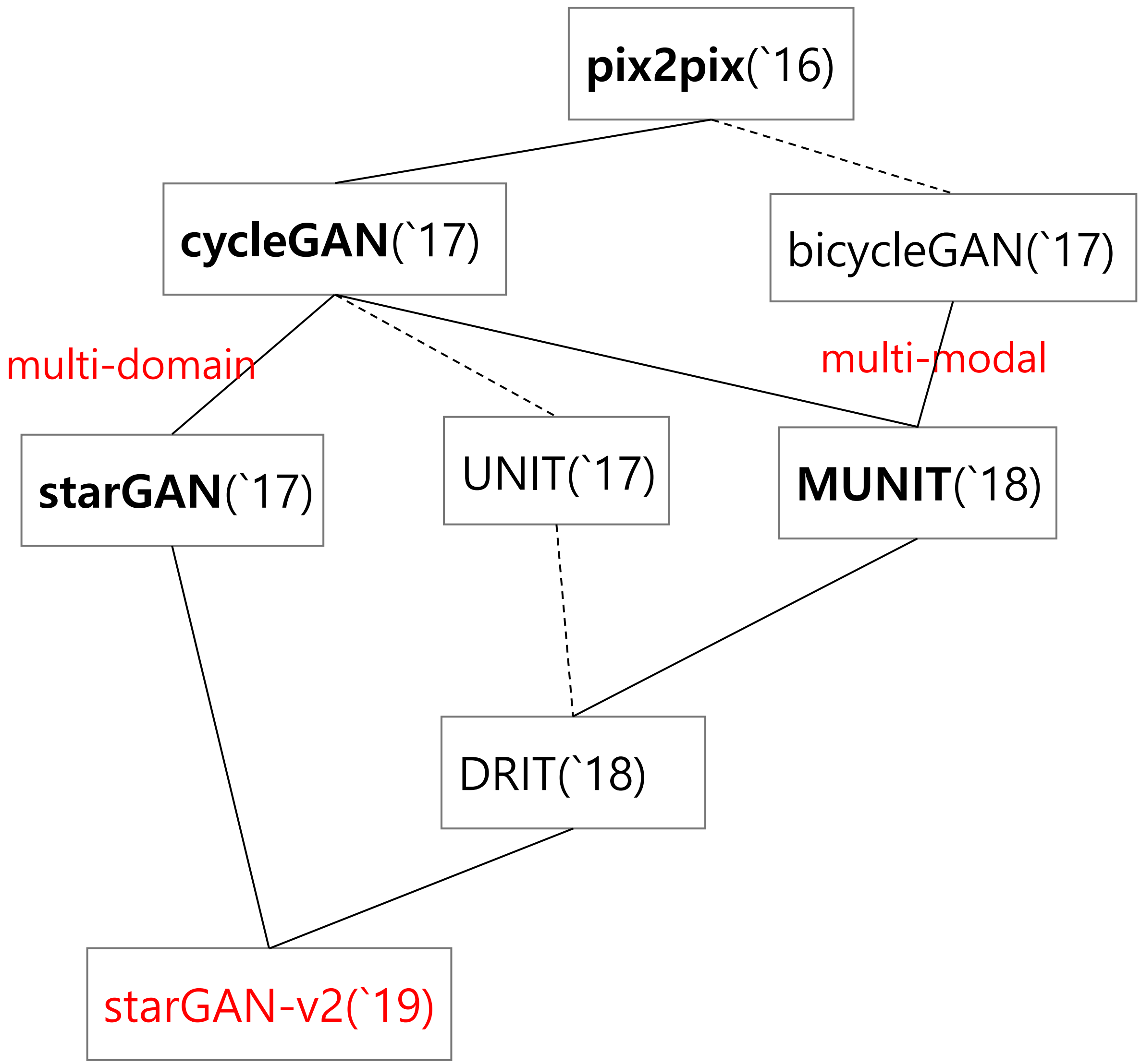


Uni-domain(cycleGAN)



Multi-domain (starGAN v1)

Image to Image Translation

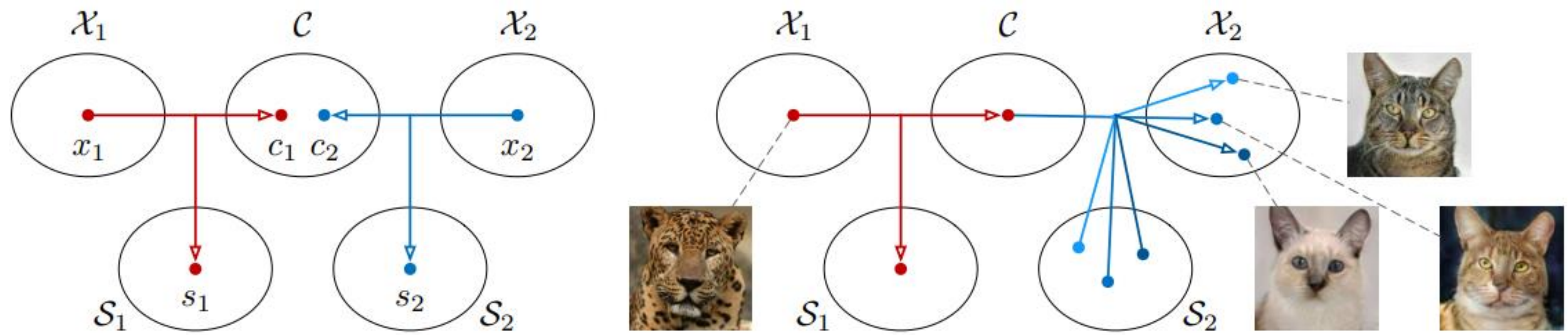


Source : https://deview.kr/2019/schedule/279

Trends of multi-modal and multi-domain (1)

- Image can be decomposed with content and style
  - content : domain invariant, e.g. pose
  - style: domain specific

- Encoder – Decoder
  - if style can be sampled from prior
    -> generate diverse style
    -> multi-modal



(a) Auto-encoding (b) Translation

Trends of multi-modal and multi-domain (2)

- (Decoding style using Adain)

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

- (Reconstruction / Cycle Consistency Loss)

- Unified(Single) Model for various domain
  -> multi-domain

Main Idea

- Both Multi-modal and Multi-domain
  - Diverse image of multiple domain within single framework

- How?
  - Multi-modal -> Encoder – Decoder

Main Idea

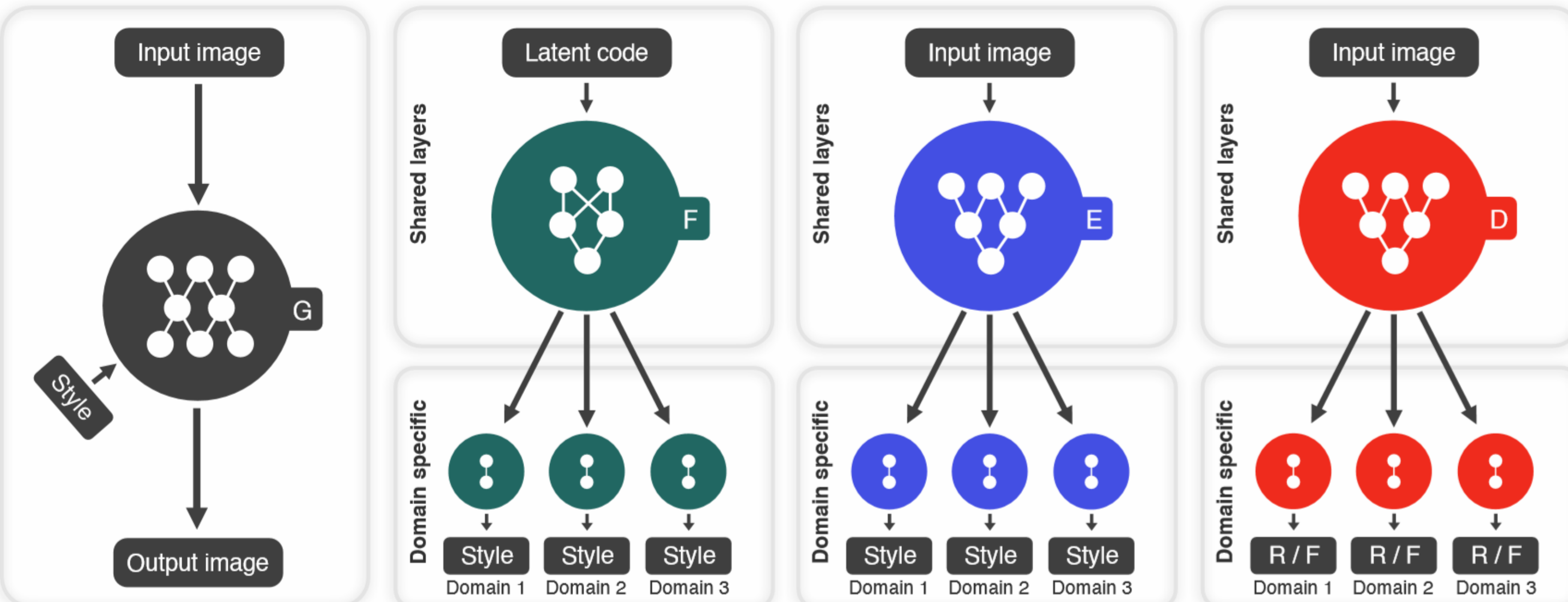
- Both Multi-modal and Multi-domain
  - Diverse image of multiple domain within single framework

- How?
  - Multi-modal -> Encoder – Decoder

  - Multi-domain -> Single Encoder and Decoder

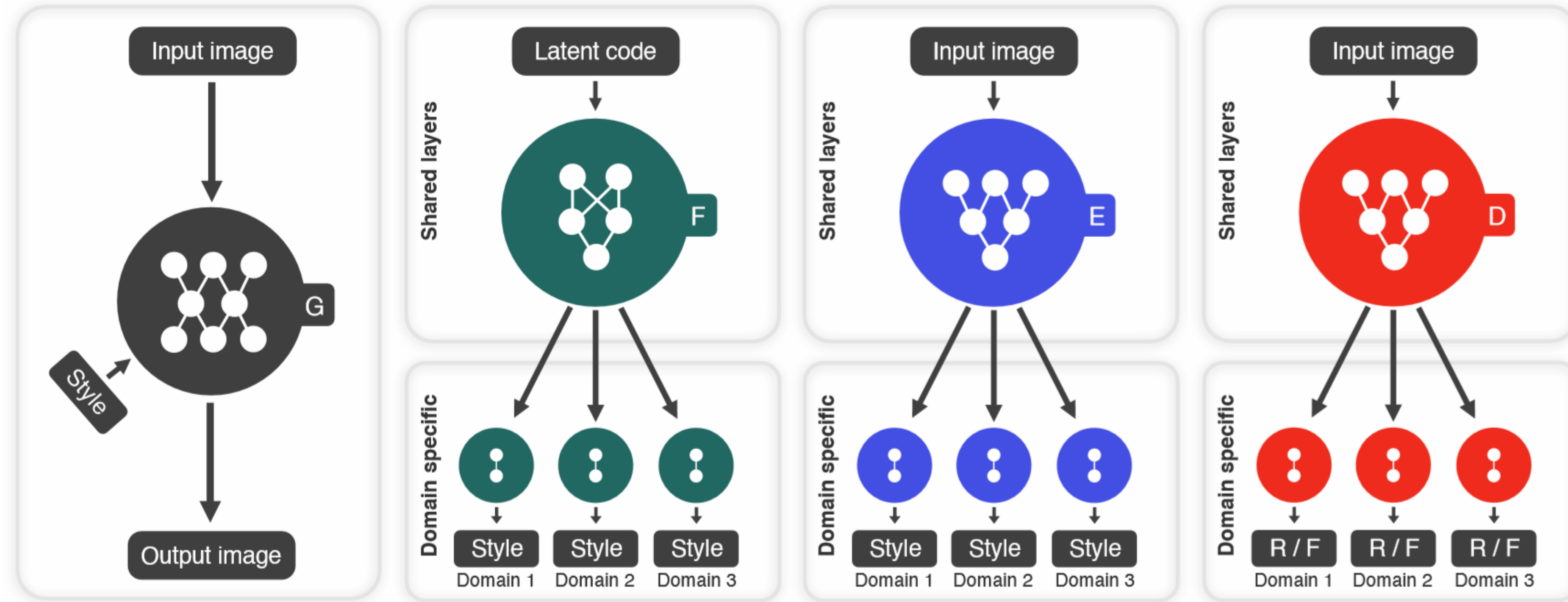
Model Architecture

- How?
  - Multi-modal -> Encoder – Decoder
  - Multi-domain -> Single Encoder and Decoder



(a) Generator (b) Mapping network (c) Style encoder (d) Discriminator

Model Architecture

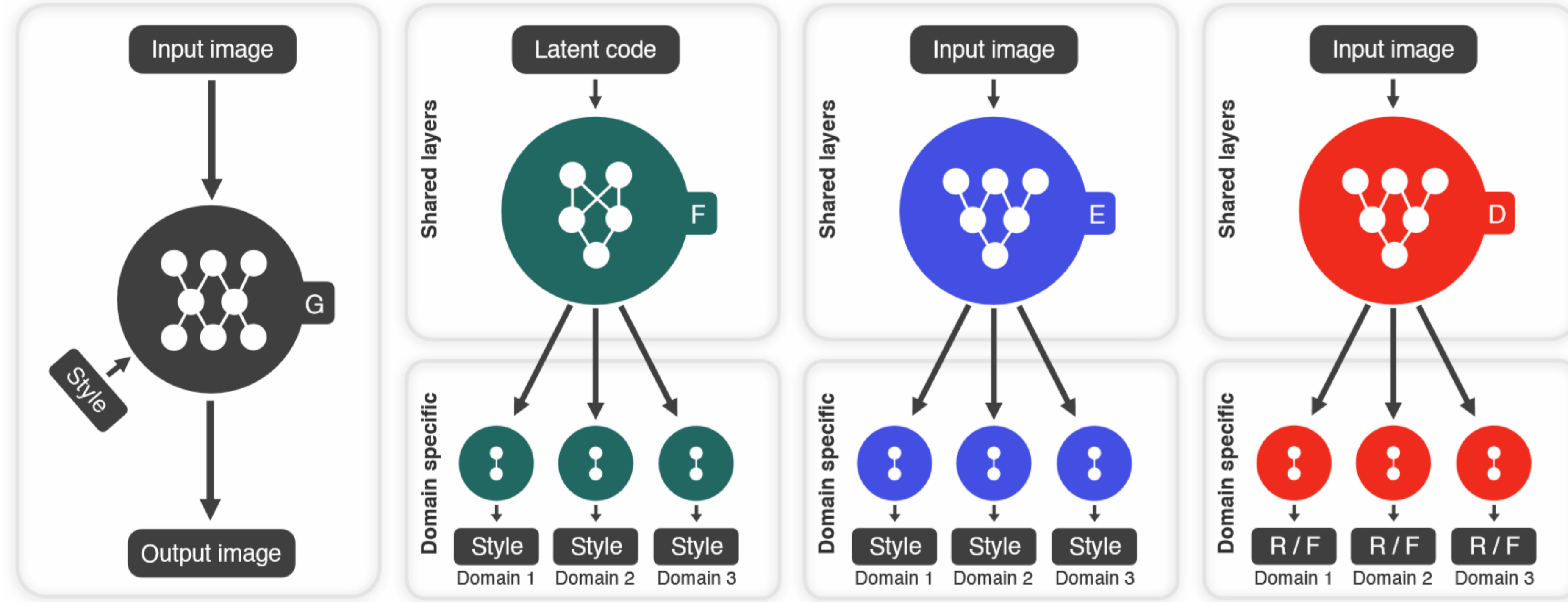


(a) Generator (b) Mapping network (c) Style encoder (d) Discriminator

How to Generate Image?

1. Latent-guided synthesis
   - random noise -> mapping network -> style code $\widetilde{\mathbf{s}} = F_{\widetilde{y}}(\mathbf{z})$
   - translate image using style code $G(\mathbf{x}, \widetilde{\mathbf{s}})$

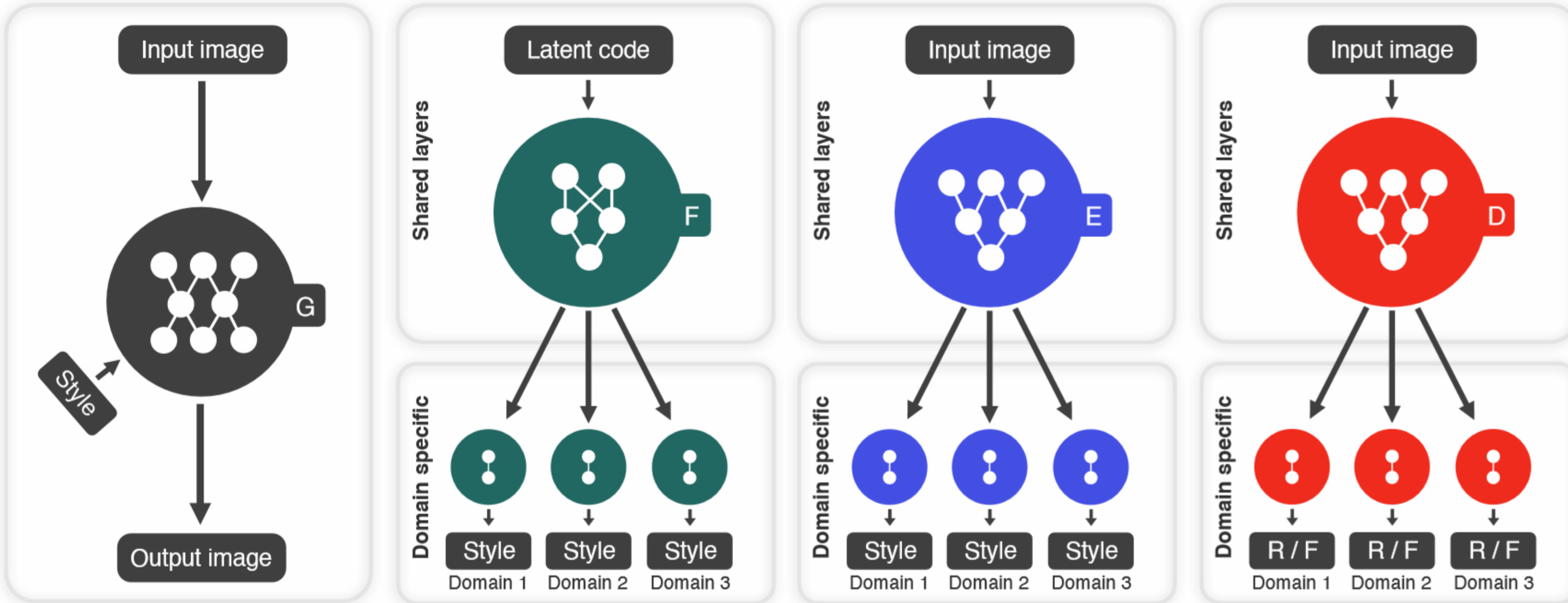
Model Architecture



(a) Generator (b) Mapping network (c) Style encoder (d) Discriminator

2. Reference-guided synthesis
   - extract style code from reference image $\tilde{s} = E_{\hat{y}}(\hat{x})$
   - translate image using style code $G(\mathbf{x}, \widetilde{\mathbf{s}})$

Model Architecture

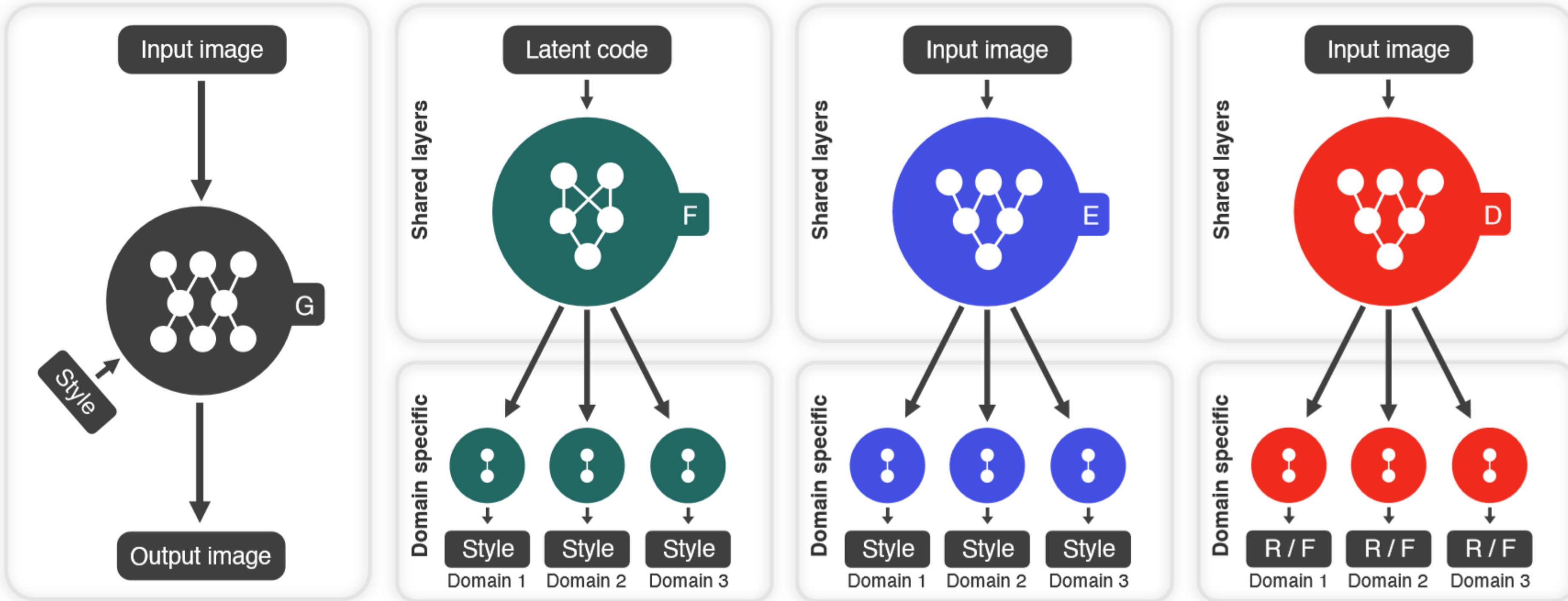


(a) Generator (b) Mapping network (c) Style encoder (d) Discriminator

- Adversarial Loss

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

Model Architecture



(a) Generator (b) Mapping network (c) Style encoder (d) Discriminator

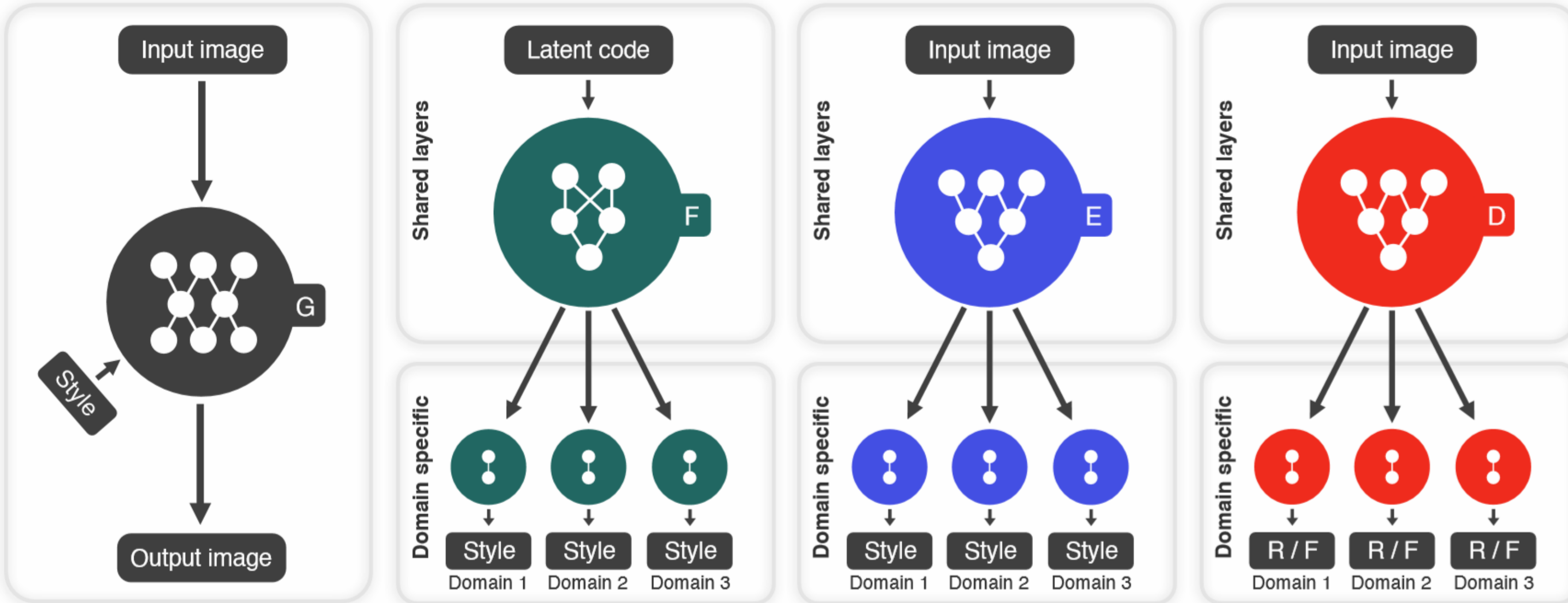
- Style Reconstruction Loss

$$\widetilde{\mathbf{s}} = F_{\widetilde{y}}(\mathbf{z})$$

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

$$\tilde{s} = E_{\hat{y}}\ (\hat{x})$$

Model Architecture



(a) Generator (b) Mapping network (c) Style encoder (d) Discriminator

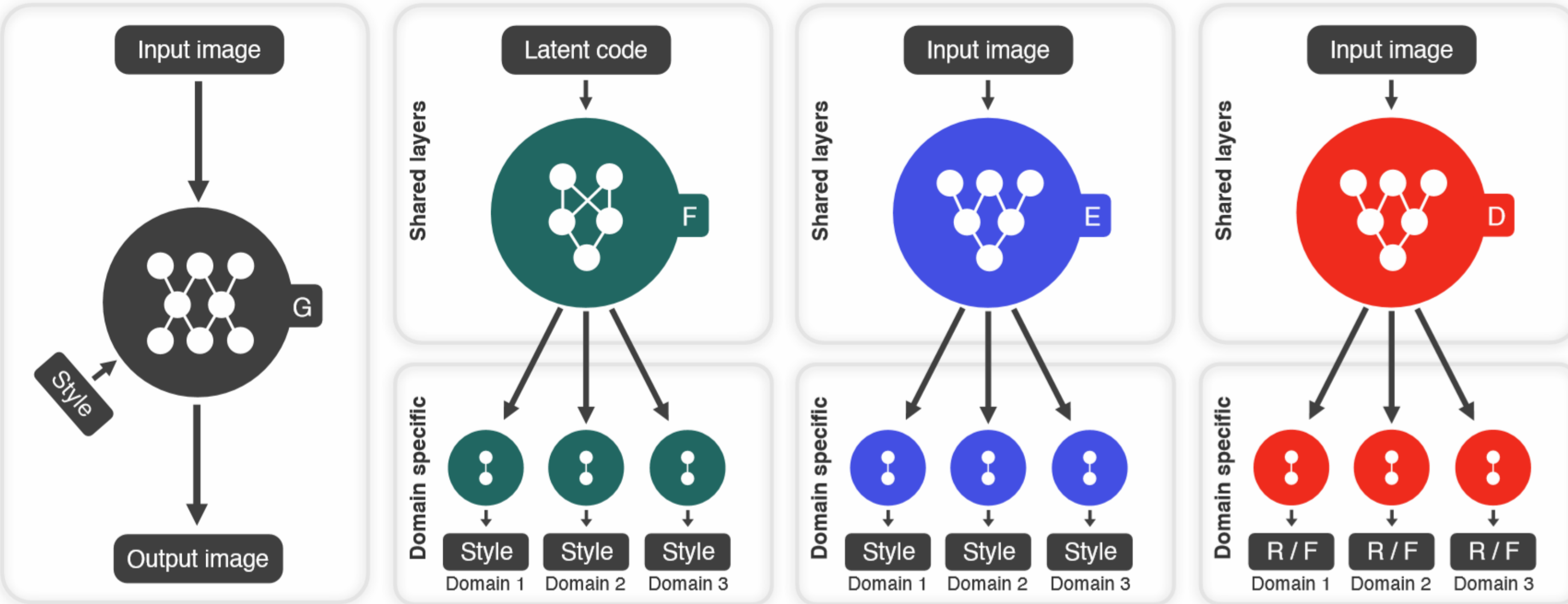
- Cycle Consistency Loss

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

$$\widetilde{\mathbf{s}} = F_{\widetilde{y}}(\mathbf{z})$$

$$\tilde{s} = E_{\hat{y}}\ (\hat{x})$$

Model Architecture



(a) Generator (b) Mapping network (c) Style encoder (d) Discriminator

- Style Diversification Loss

$$\widetilde{\mathbf{s}} = F_{\widetilde{y}}(\mathbf{z})$$

$$\mathcal{L}_{ds} = \mathbb{E}_{\mathbf{x},\widetilde{y},\mathbf{z}_1,\mathbf{z}_2}\left[\|G(\mathbf{x},\widetilde{\mathbf{s}}_1) - G(\mathbf{x},\widetilde{\mathbf{s}}_2)\|_1\right]$$

$$\tilde{s} = E_{\hat{y}}\ (\hat{x})$$

Objective

- Adversarial Loss

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

- Style Reconstruction Loss

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

- Cycle Consistency Loss

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

- Style Diversification Loss

$$\mathcal{L}_{ds} = \mathbb{E}_{\mathbf{x},\widetilde{y},\mathbf{z}_1,\mathbf{z}_2}\left[\|G(\mathbf{x},\widetilde{\mathbf{s}}_1) - G(\mathbf{x},\widetilde{\mathbf{s}}_2)\|_1\right]$$

Objective

- Full Objective

$$\min_{G,F,E} \max_{D} \quad \mathcal{L}_{adv} + \lambda_{sty}\,\mathcal{L}_{sty} - \lambda_{ds}\,\mathcal{L}_{ds} + \lambda_{cyc}\,\mathcal{L}_{cyc},$$

Results

- Latent-guided Systhesis

Method	CelebA-HQ		AFHQ	
	FID	LPIPS	FID	LPIPS
MUNIT [16]	31.4	0.363	41.5	0.511
DRIT [28]	52.1	0.178	95.6	0.326
MSGAN [34]	33.1	0.389	61.4	**0.517**
StarGAN v2	**13.7**	**0.452**	**16.2**	0.450
Real images	14.8	-	12.9	-

(b) Latent-guided synthesis on AFHQ

Results

- Reference-guided Systhesis

Method	CelebA-HQ		AFHQ	
	FID	LPIPS	FID	LPIPS
MUNIT [16]	107.1	0.176	223.9	0.199
DRIT [28]	53.3	0.311	114.8	0.156
MSGAN [34]	39.6	0.312	69.8	0.375
StarGAN v2	**23.8**	**0.388**	**19.8**	**0.432**
Real images	14.8	-	12.9	-

(b) Reference-guided synthesis on AFHQ

Contribution

- **Diverse** image of **multiple domain** within **single** framework
  - multi-modal + multi-domain model

- **Visual Quality**

- **AFHQ Dataset**
  - new high quality dataset of animal faces
  - with large inter-intra domain variation

Contribution

- **Diverse** image of **multiple domain** within **single** framework
  - well benchmarking other papers
  - (I think..) not perfect single encoder

- **Visual Quality**
  - (I guess..) that is not because of new proposed architecture
    but of highly optimized training methods (styleGAN?)

- **AFHQ Dataset**
  - thank you so much, I love you. sincerely

Contribution

LAYER	ACTVATION	NORM	OUTPUT SHAPE
Latent **z**	-	-	16
Linear	ReLU	-	512
Linear	ReLU	-	512
Linear	ReLU	-	512
Linear	ReLU	-	512
Linear	ReLU	-	512
Linear	ReLU	-	512
Linear $\star$ N	-	-	64 $\star$ N

Table 6. Mapping network architecture.

Init paper(2019.12)

TYPE	LAYER	ACTVATION	OUTPUT SHAPE
Shared	Latent **z**	-	16
Shared	Linear	ReLU	512
Shared	Linear	ReLU	512
Shared	Linear	ReLU	512
Shared	Linear	ReLU	512
Unshared	Linear	ReLU	512
Unshared	Linear	ReLU	512
Unshared	Linear	ReLU	512
Unshared	Linear	-	64

Table 6. Mapping network architecture.

Current

Contribution

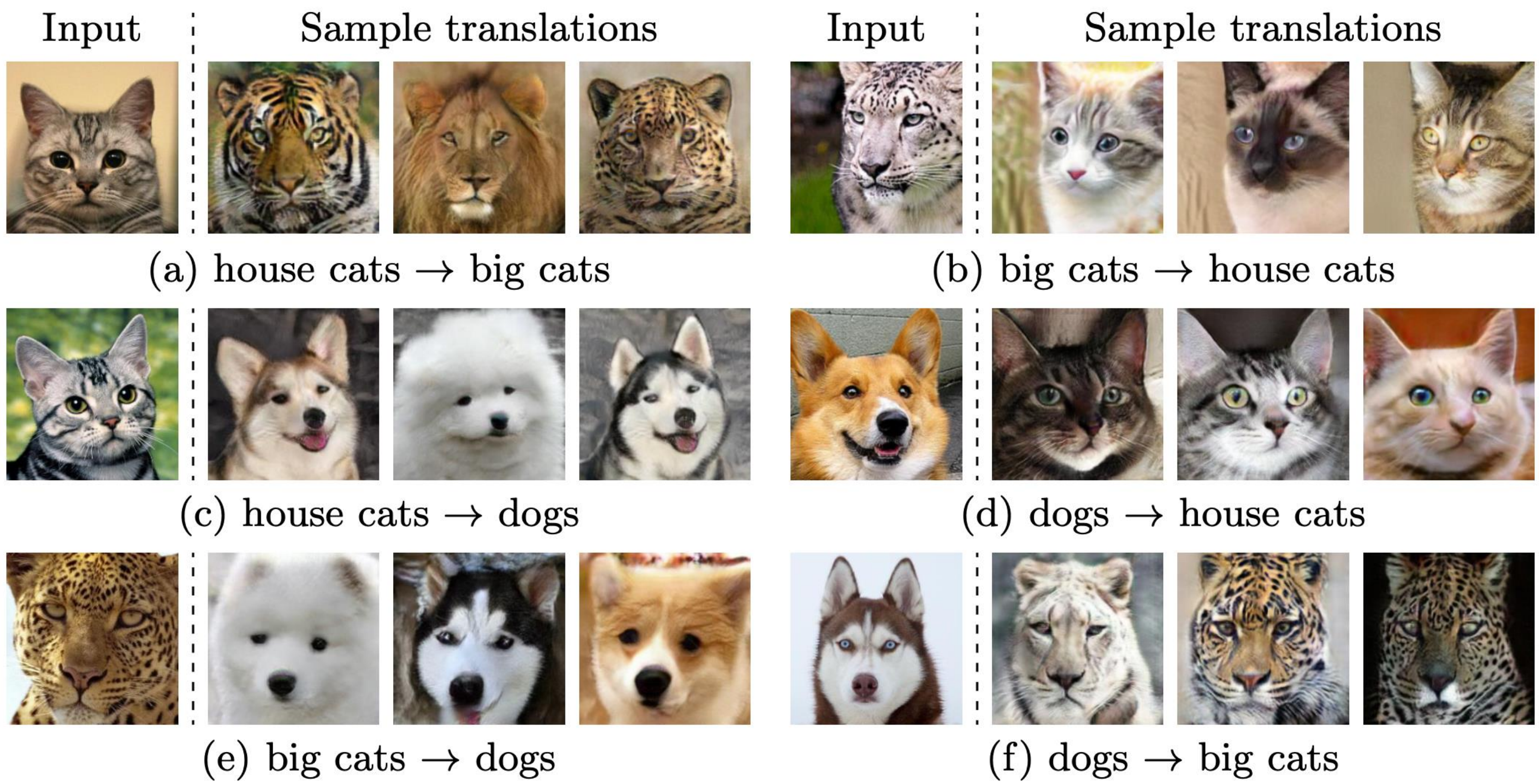


Fig. 6. Example results of animal image translation.

From MUNIT

Q & A

감사합니다.