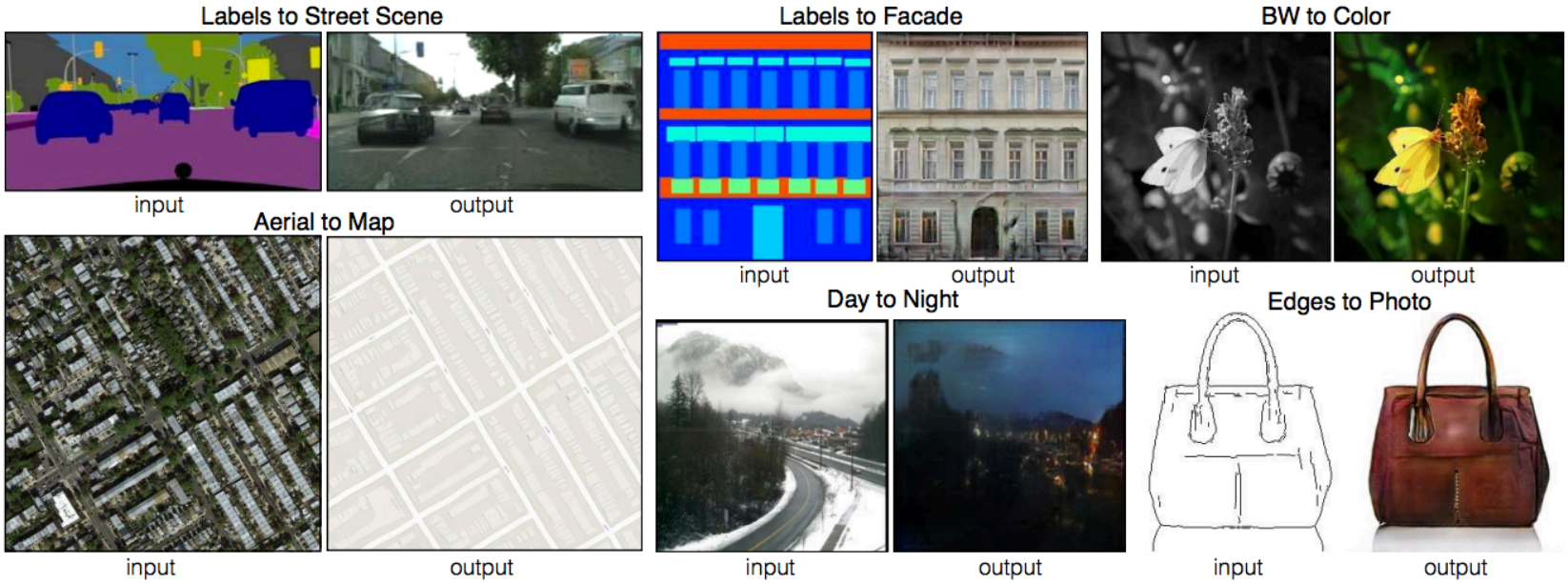
Multimodal Unsupervised Image-to-Image Translation

Du Ang

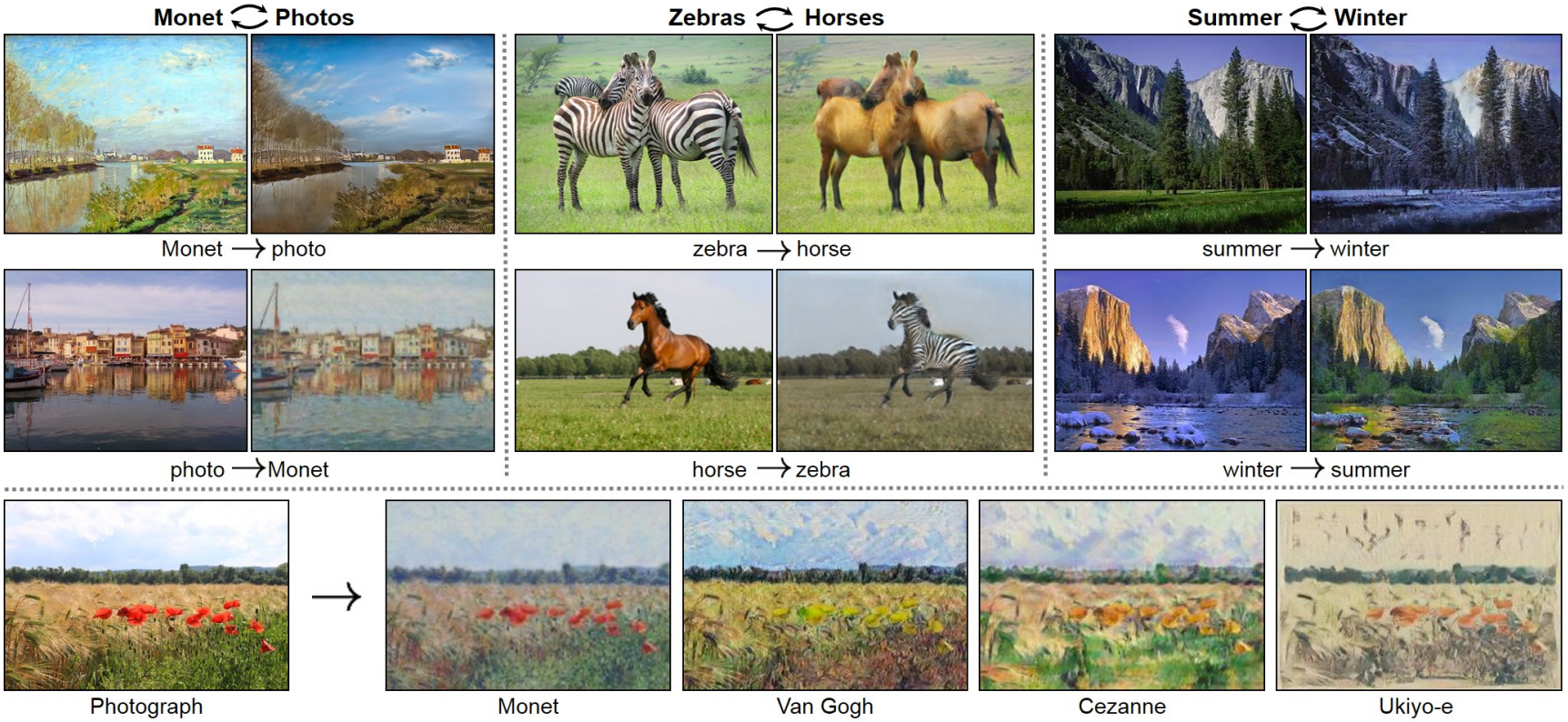
2018.04.18

pix2pix paired, unimodal



Isola et al., “Image-to-Image Translation with Conditional Adversarial Networks”. In CVPR, 2017.

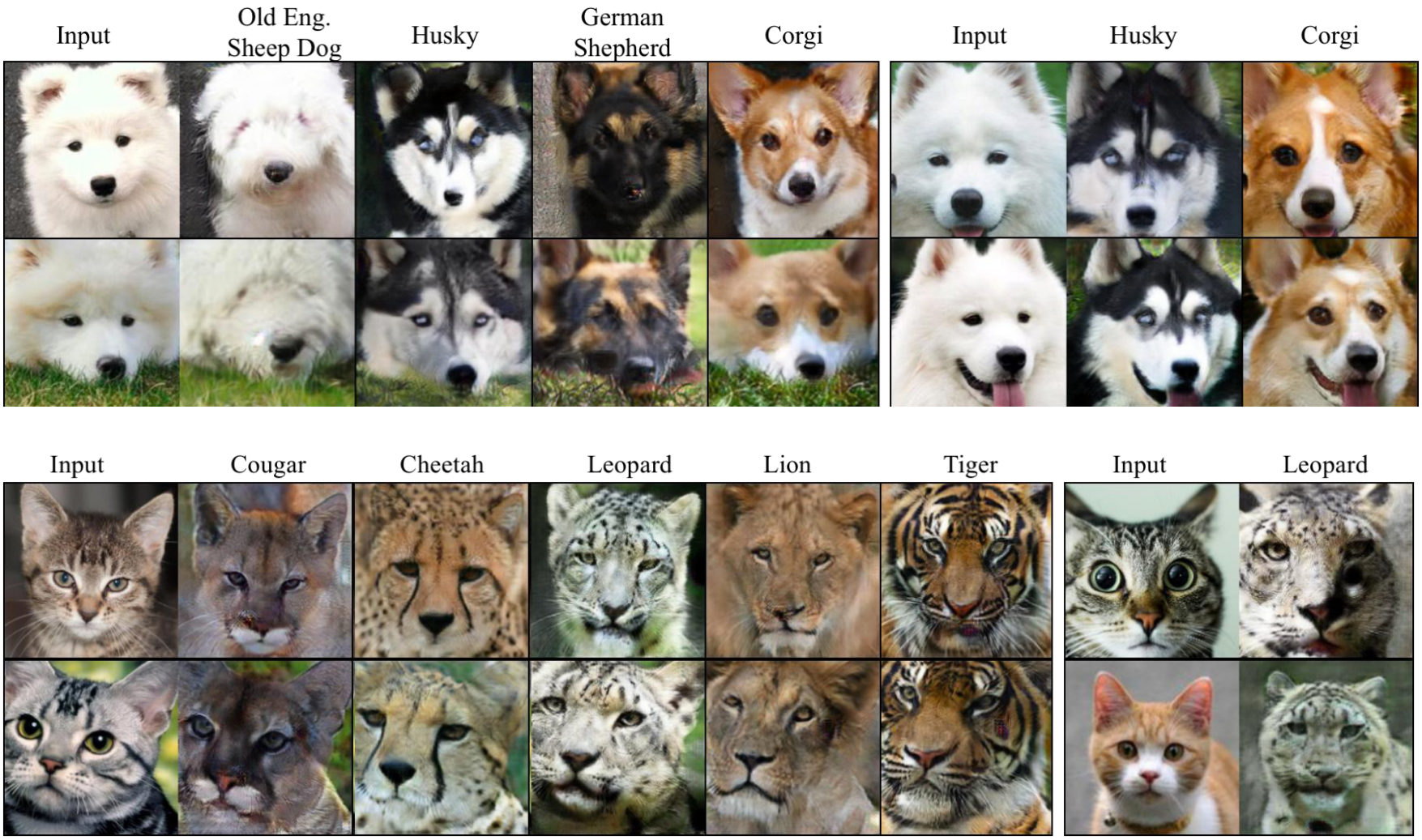
CycleGAN unpaired, unimodal



Zhu et al., “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”. In ICCV, 2017.

UNIT

unpaired, unimodal



BicycleGAN paired, multimodal

Zhu et al., “Toward Multimodal Image-to-Image Translation”. In NIPS, 2017.

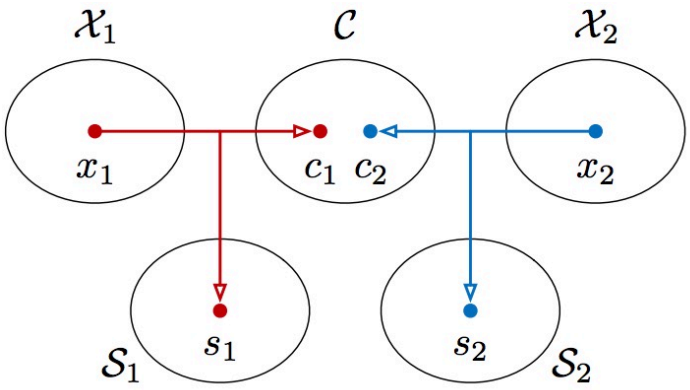
MUNIT unpaired, multimodal

(a) edges ↔ shoes

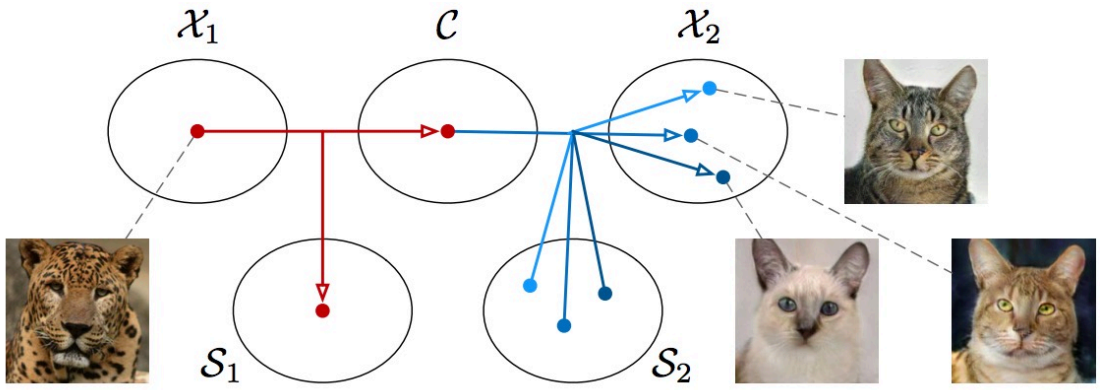
(b) edges ↔ handbags

Huang et al., “Multimodal UNsupervised Image-to-image Translation”. By NVIDIA Research, 2018.

MUNIT method illustration

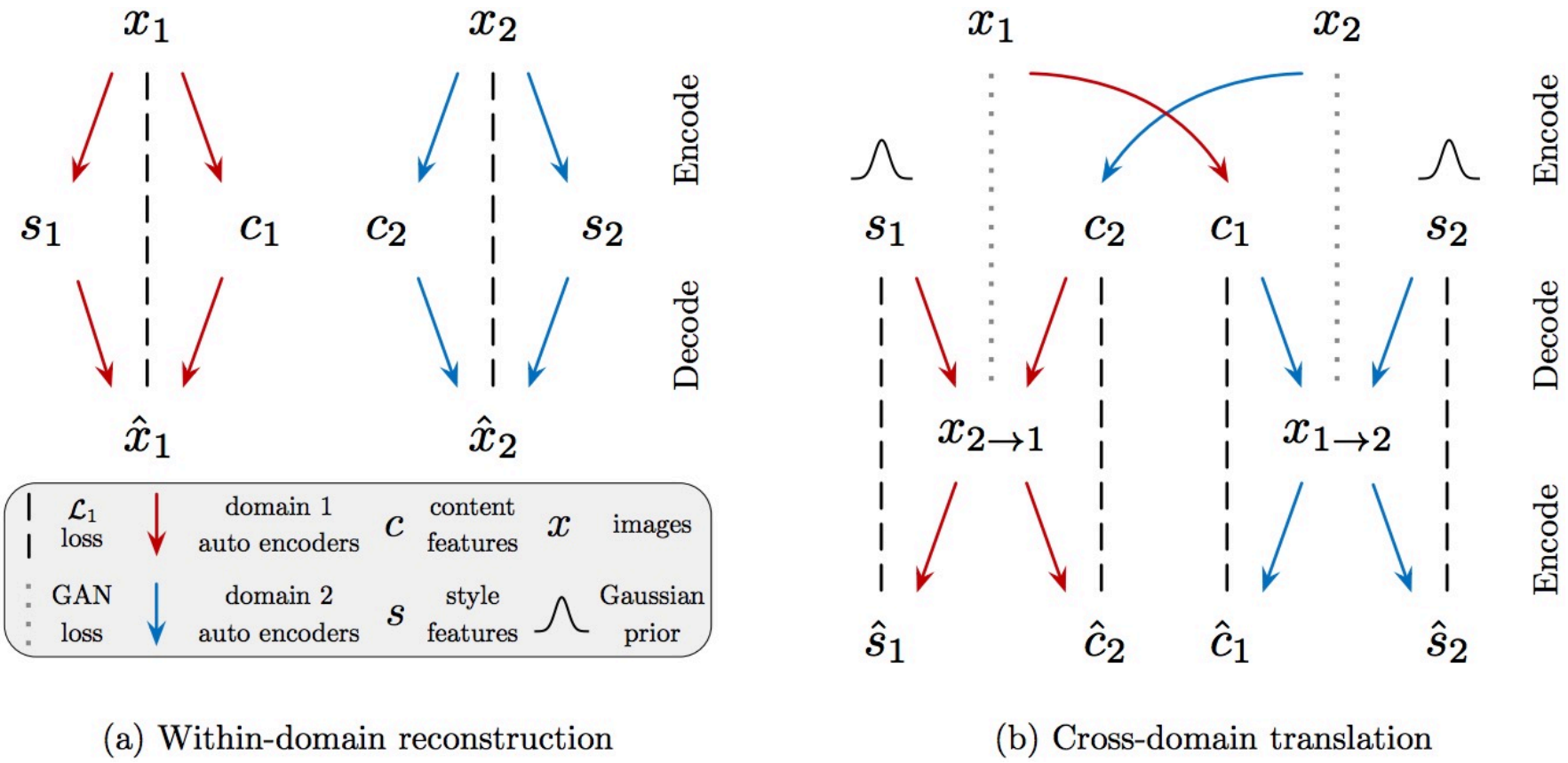


(a) Auto-encoding



(b) Translation

MUNIT model overview



(a) Within-domain reconstruction

(b) Cross-domain translation

MUNIT loss

Bidirectional reconstruction loss

- Image reconstruction $\text{image} \rightarrow \text{latent} \rightarrow \text{image}$

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

- Latent reconstruction $\text{latent} \rightarrow \text{image} \rightarrow \text{latent}$

$$\mathcal{L}_{\text{recon}}^{c_1} = \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)}[||E_2^c(G_2(c_1, s_2)) - c_1||_1]$$
$$\mathcal{L}_{\text{recon}}^{s_2} = \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)}[||E_2^s(G_2(c_1, s_2)) - s_2||_1]$$

where $q(s_2)$ is the prior $\mathcal{N}(0, \mathbf{I})$, $p(c_1)$ is given by $c_1 = E_1^c(x_1)$ and $x_1 \sim p(x_1)$.

MUNIT loss

Adversarial loss

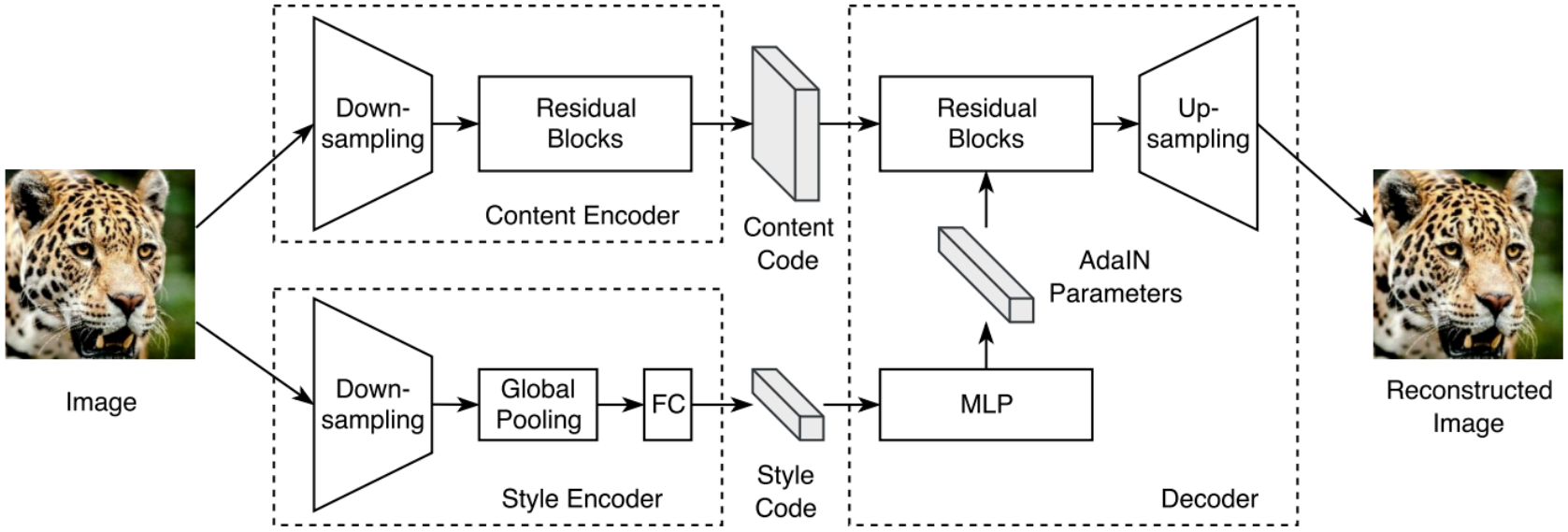
$$\mathcal{L}_{\text{GAN}}^{x_2} = \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)}[\log(1 - D_2(G_2(c_1, s_2)))] + \mathbb{E}_{x_2 \sim p(x_2)}[\log D_2(x_2)]$$

Total loss

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

where λ_x, λ_c, λ_s are weights that control the importance of reconstruction terms.

MUNIT auto-encoder architecture



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

MUNIT GAN architecture

Discriminator:

Least Squares GAN

Multi-scale discriminators

Domain-invariant perceptual loss

Evaluation metrics

Human Preference

Amazon Mechanical Turk (AMT)

LPIPS Distance

(Conditional) Inception Score

$$\mathrm{IS} = \mathbb{E}_{x_1 \sim p(x_1)}[\mathbb{E}_{x_{1\to 2} \sim p(x_{2\to 1}|x_1)}[\mathrm{KL}(p(y_2|x_{1\to 2})||p(y_2))]]$$

$$\mathrm{CIS} = \mathbb{E}_{x_1 \sim p(x_1)}[\mathbb{E}_{x_{1\to 2} \sim p(x_{2\to 1}|x_1)}[\mathrm{KL}(p(y_2|x_{1\to 2})||p(y_2|x_1))]]$$

Datasets

Edges ↔ shoes/handbags.

Animal image translation.

Street scene images.

Yosemite summer ↔ winter (HD).

Results

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.

Results

	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 $\mathcal{L}_{\text{recon}}^{x}$	6.0%	0.213	29.0%	0.191
MUNIT w/o $\mathcal{L}_{\text{recon}}^{c}$	20.7%	0.172	9.3%	0.185
MUNIT w/o $\mathcal{L}_{\text{recon}}^{s}$	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

† Trained with paired supervision.

Results

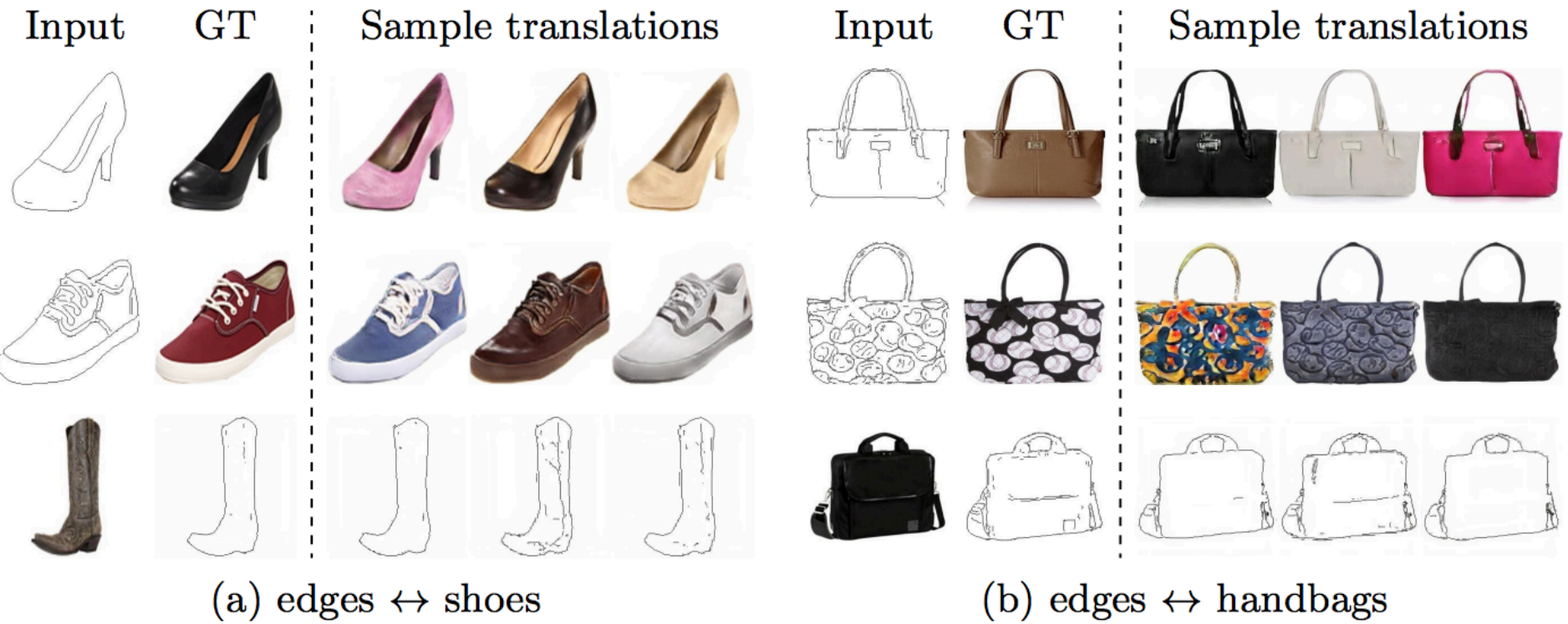


Fig. 5. Example results of (a) edges ↔ shoes and (b) edges ↔ handbags.

Results

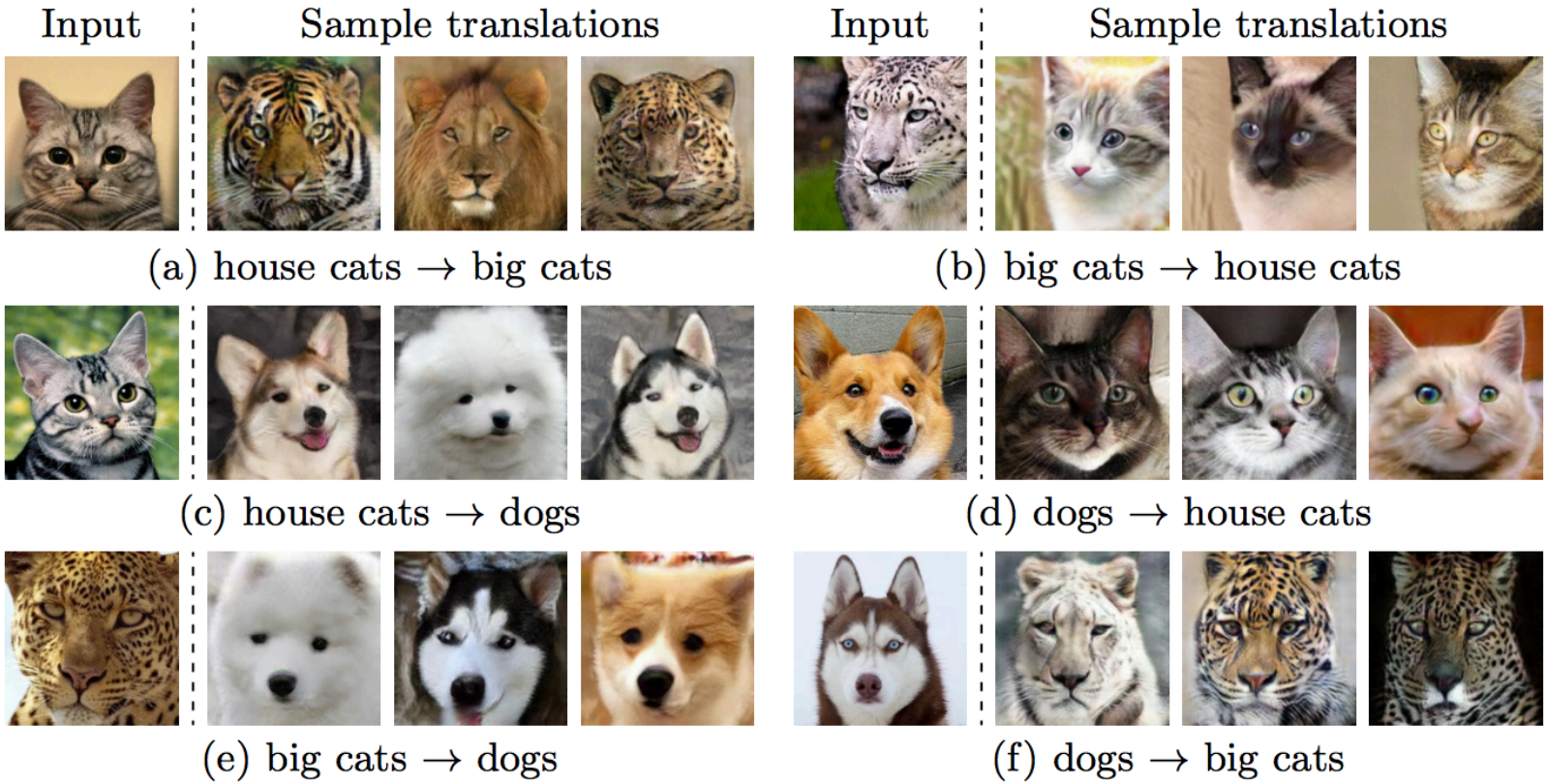


Fig. 6. Example results of animal image translation.

Results

	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

Results

Example-guided Image Translation

(b) edges → shoes

(b) big cats → house cats

Compare with style transfer methods