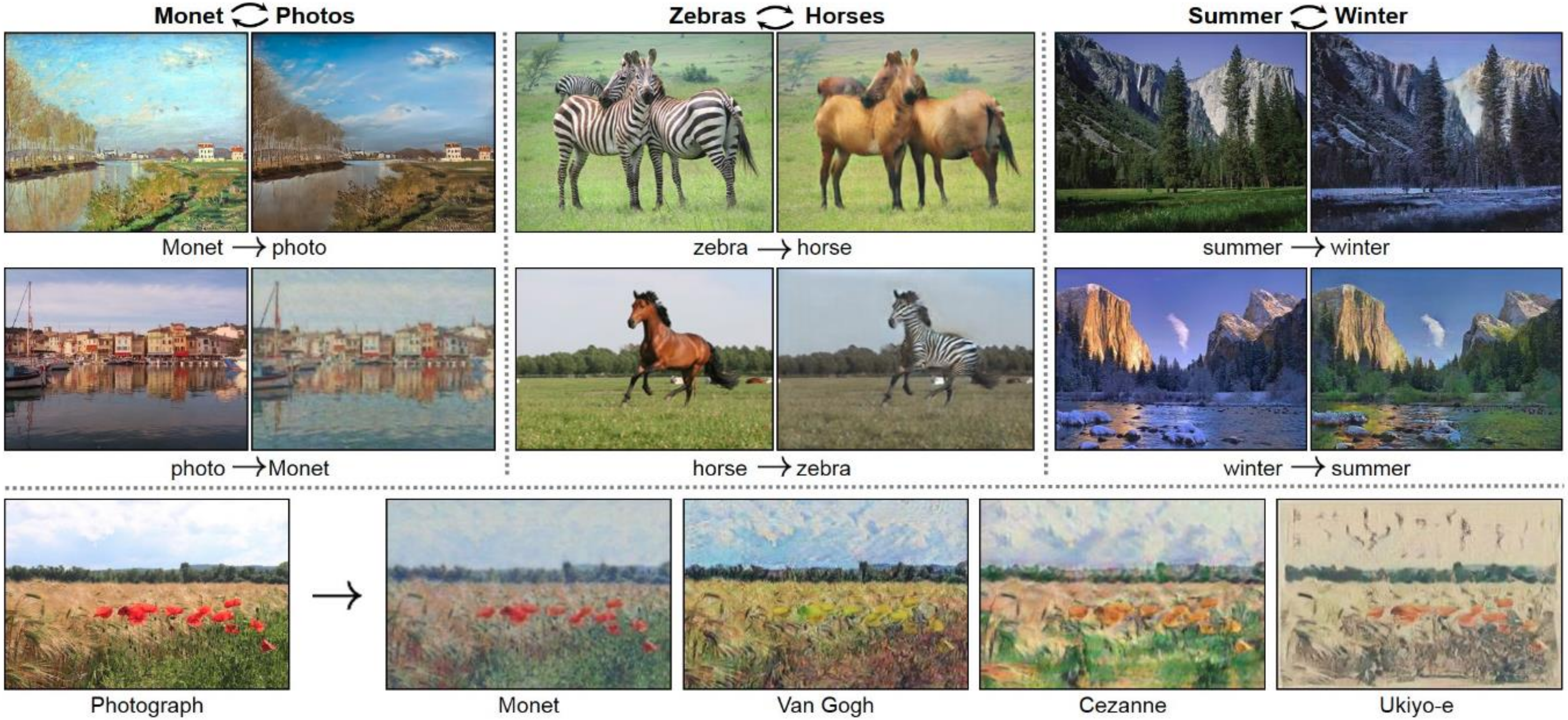
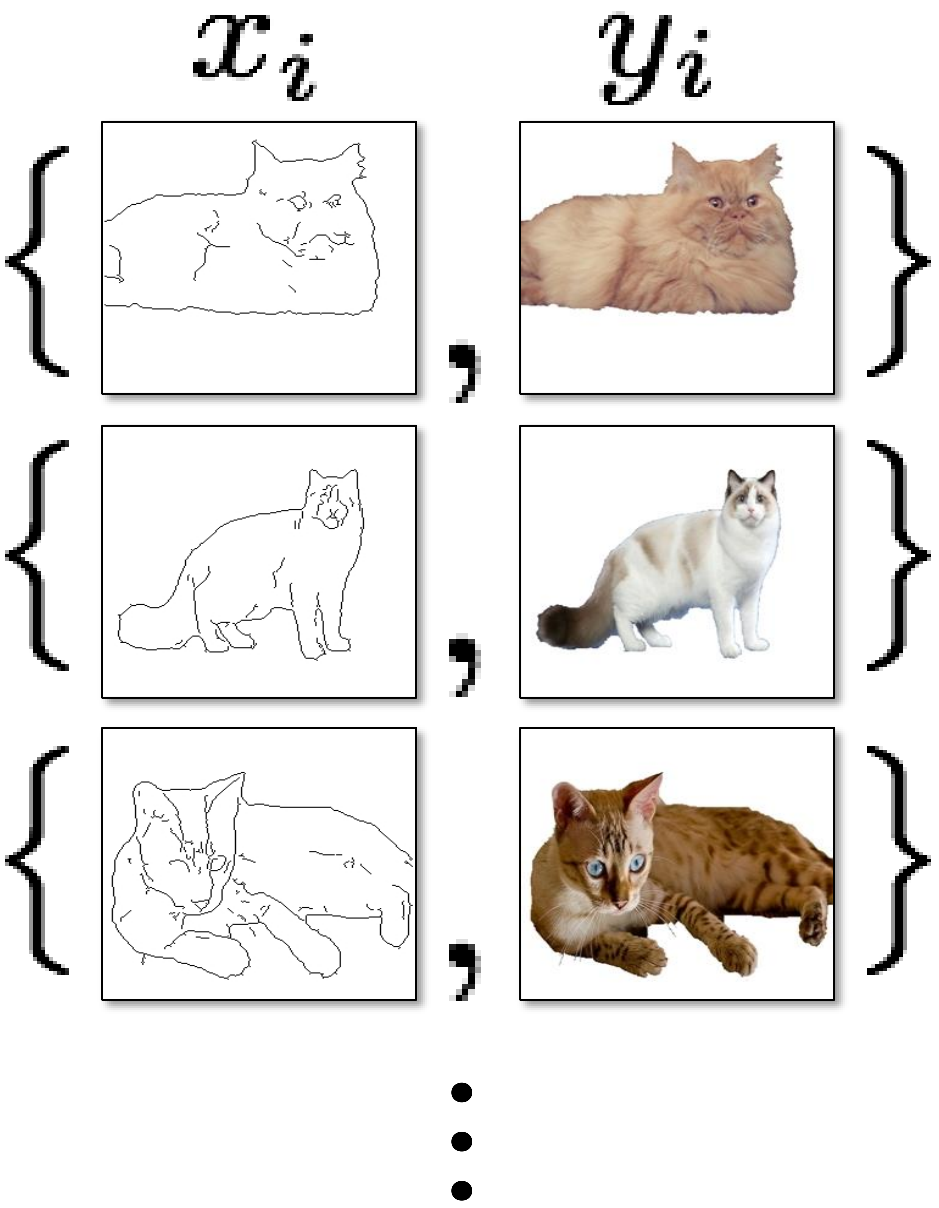
New Capabilities in Unsupervised Image to Image Translation

Sagie Benaim and Lior Wolf

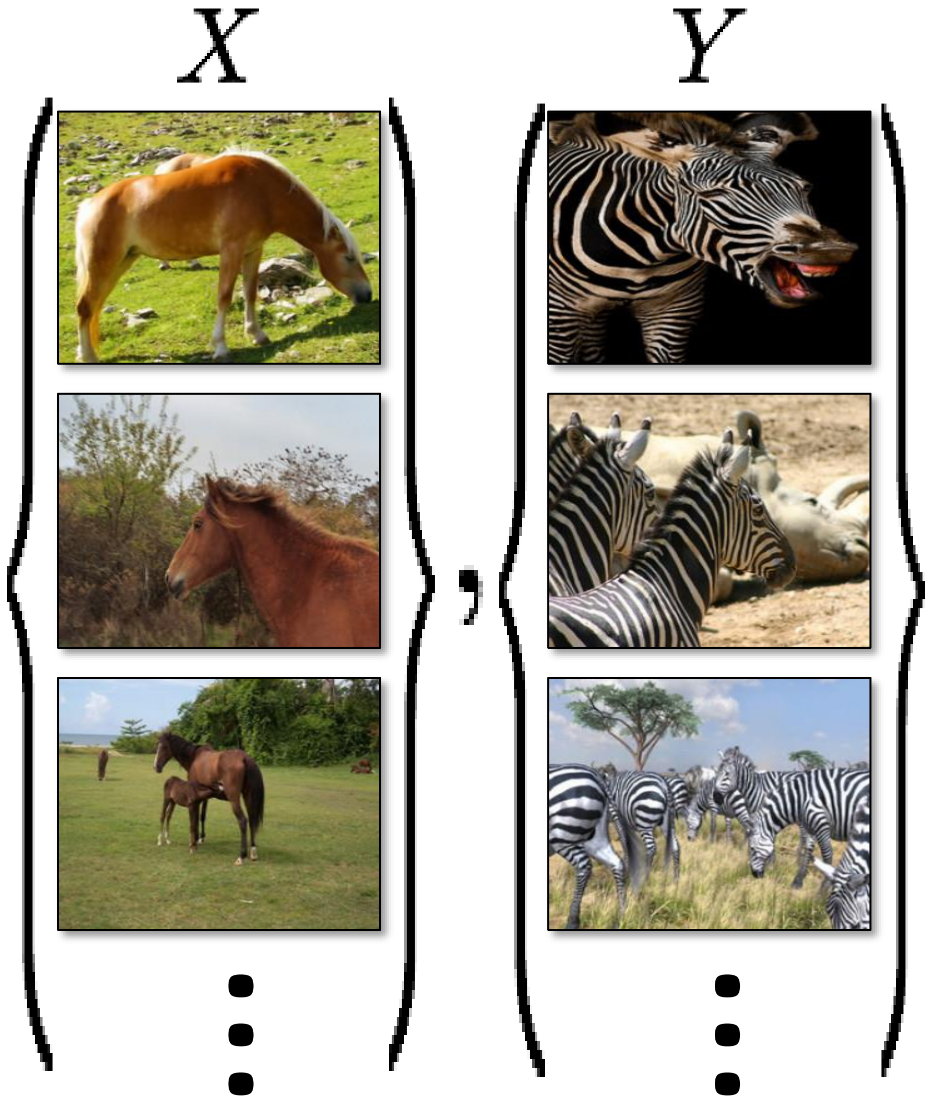
Image to Image Translation



Paired



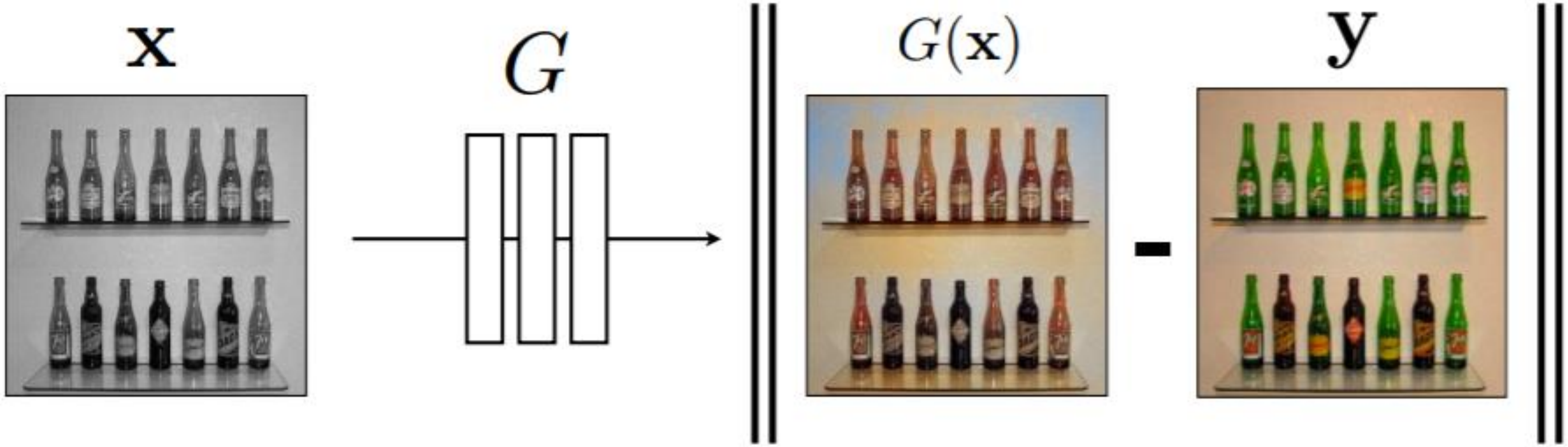
Unpaired



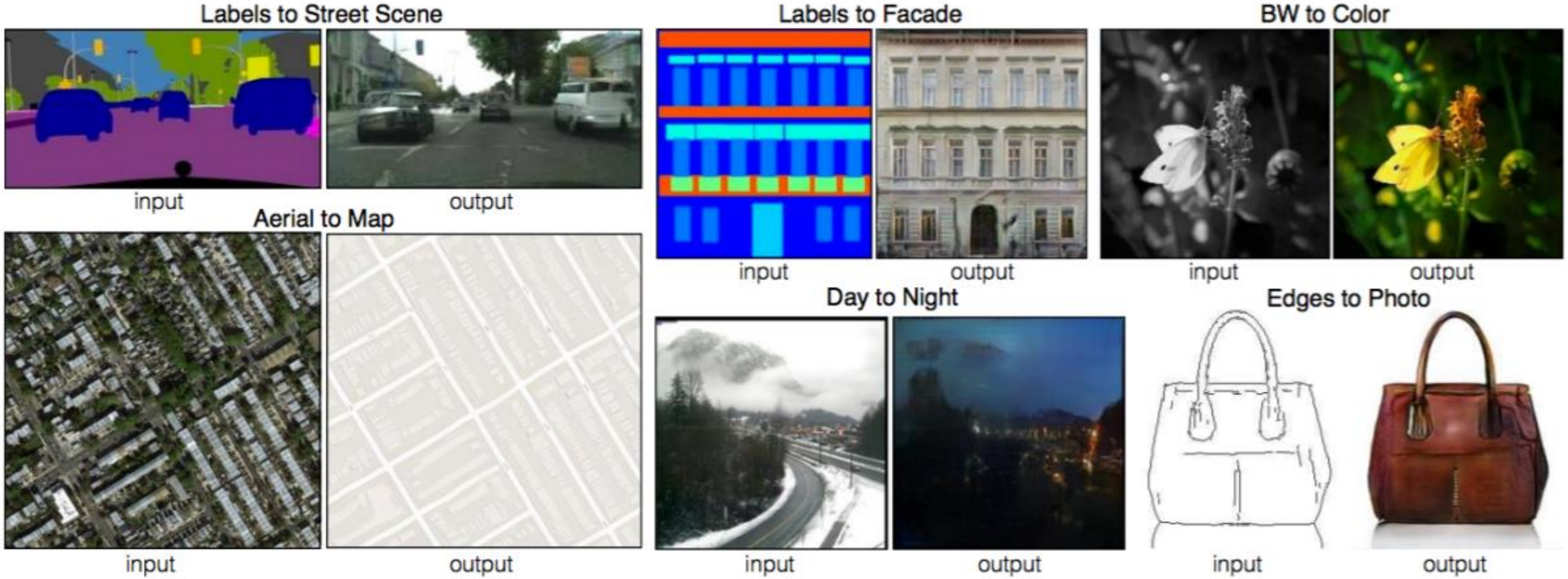
Fully Supervised: pix2pix

Conditional GAN

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



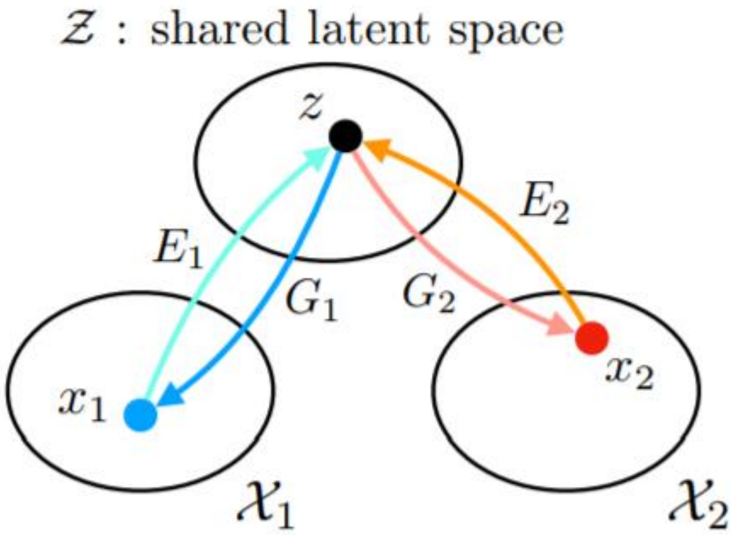
[Isola et al., CVPR 2017]



[Isola et al., CVPR 2017]

Unsupervised Alignment

- Highly related domains
  - “Unsupervised Image-to-Image Translation Networks” Liu et al.



Circular GANs

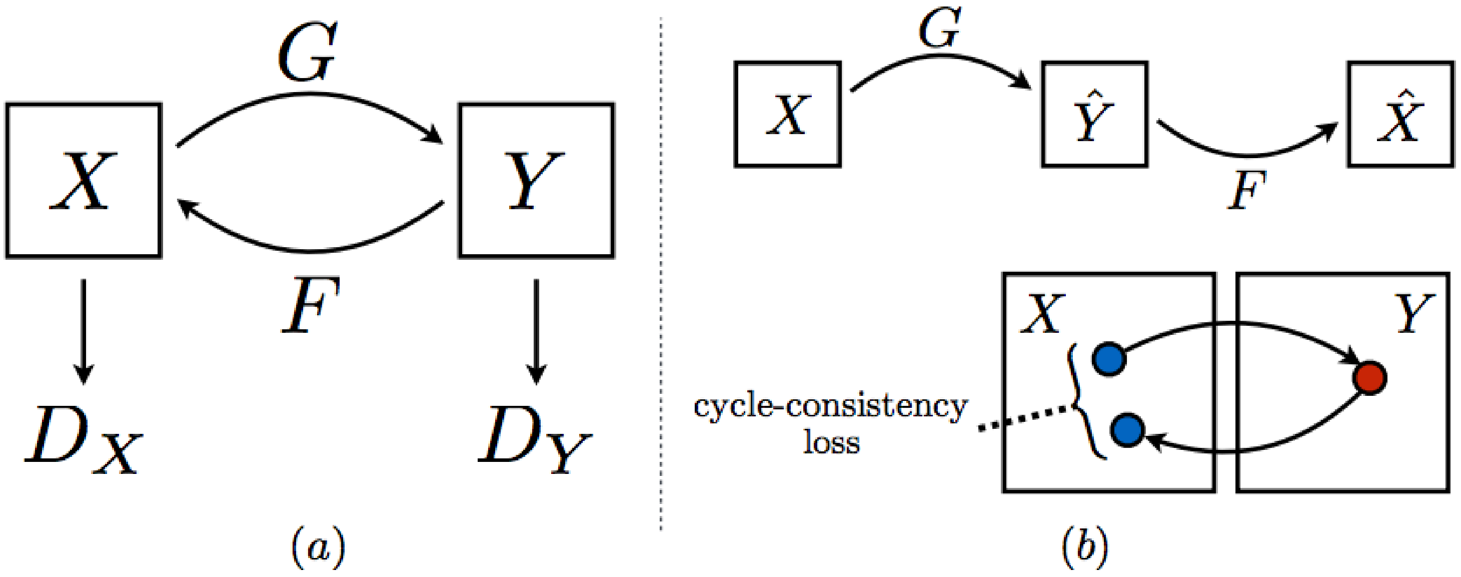
DiscoGAN: “Learning to Discover Cross-Domain Relations with Generative Adversarial Networks”. Kim et al. ICML’17.

CycleGAN: “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”. Zhu et al. arXiv:1703.10593, 2017.

DualGAN: “ Unsupervised Dual Learning for Image-to-Image Translation”. Zili et al. arXiv:1704.02510, 2017.

.

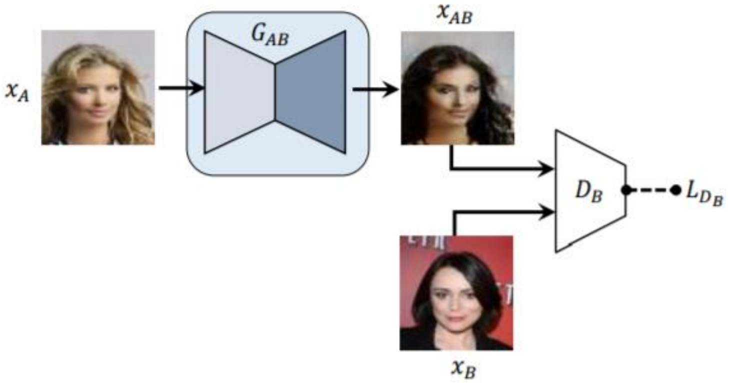
Circular GANs



Circular GANs (DiscoGAN, CycleGAN, DualGAN)

$$x \sim F(G(x))$$
$$y \sim G(F(y))$$

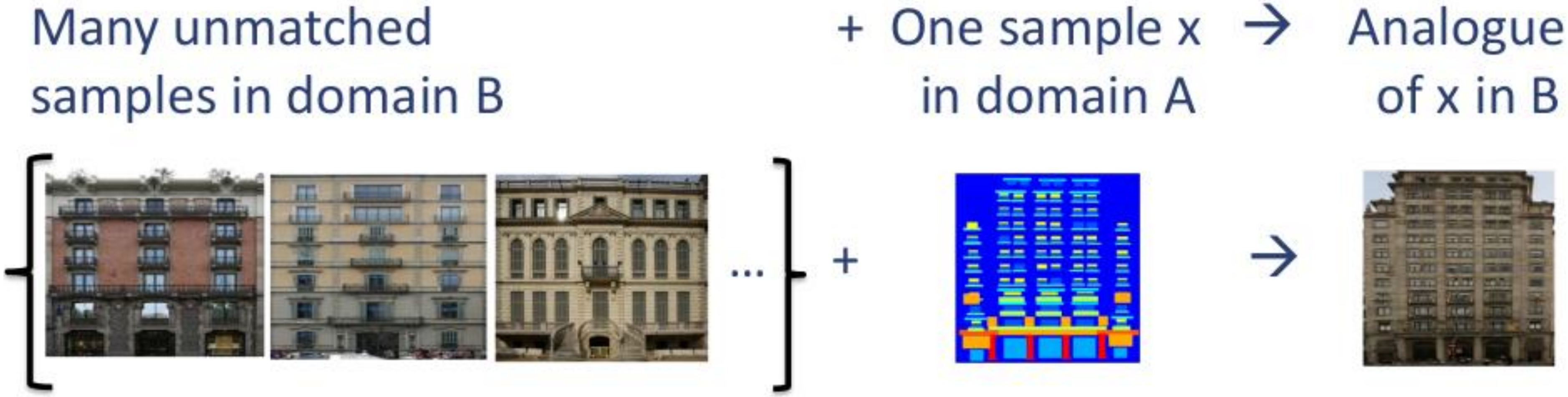
Building Block: Conditional GAN



$$\mathcal{L}_{\text{GAN}}(G_{AB}, D_B, \hat{p}_A, \hat{p}_B) = \mathbb{E}_{x_B \sim \hat{p}_B}[\log D_B(x_B)] + \mathbb{E}_{x_A \sim \hat{p}_A}[\log(1 - D_B(G_{AB}(x_A))]$$

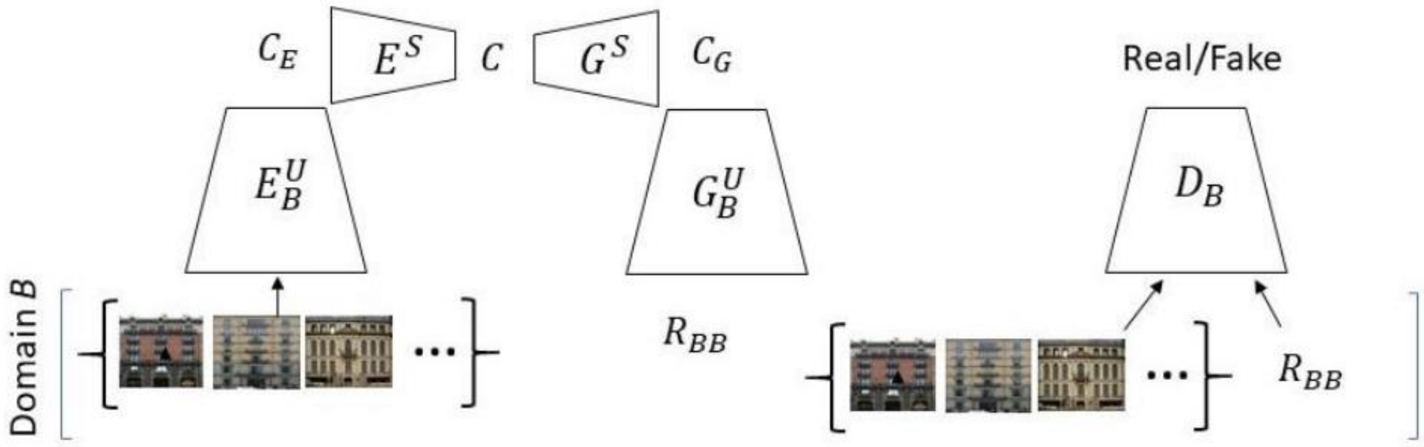
- Other GAN variants can be used: w-gan, improved w-gan, BEGAN, etc.

Less Supervision: Only a single image in domain A



One Shot Unsupervised Cross Domain Translation (NeurIPS 2018)

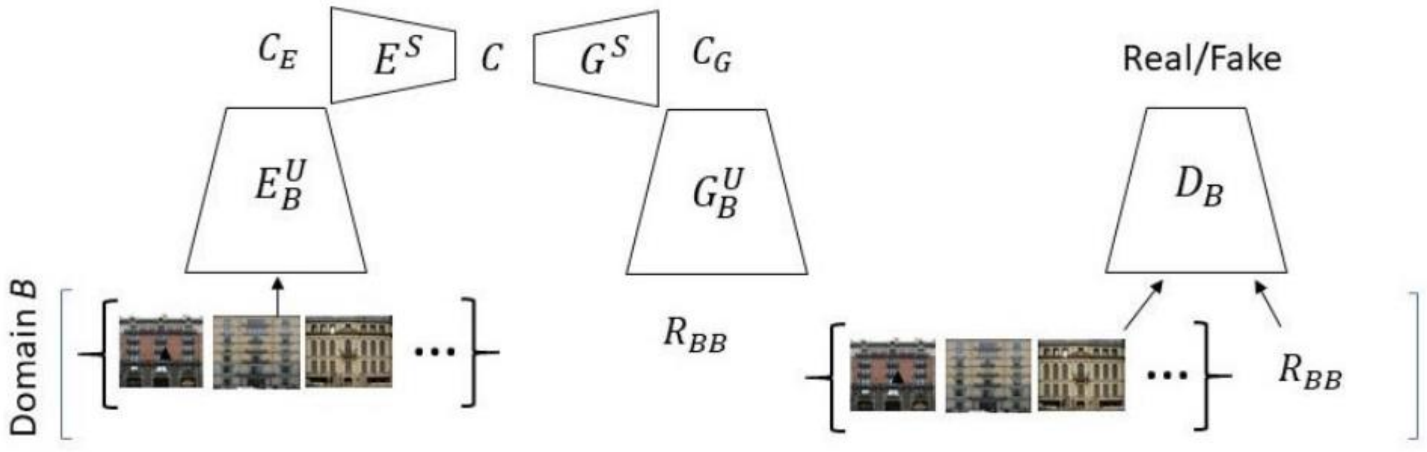
Phase I



$$\mathcal{L}_{REC_B} = \sum_{s \in P(\Lambda)} \|G_B(E_B(s)) - s\|_1$$

$$\mathcal{L}_{VAE_B} = \sum_{s \in P(\Lambda)} \mathrm{KL}(E_B \circ P(\Lambda) \| \mathcal{N}(0, I))$$

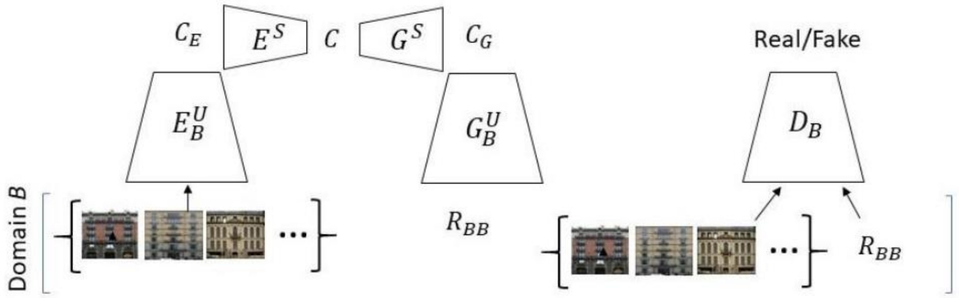
Phase I



$$\mathcal{L}_{\mathrm{GAN}_B} = \sum_{s \in P(\Lambda)} -\ell(\overline{D_B}(G_B(E_B(s))), 0)$$

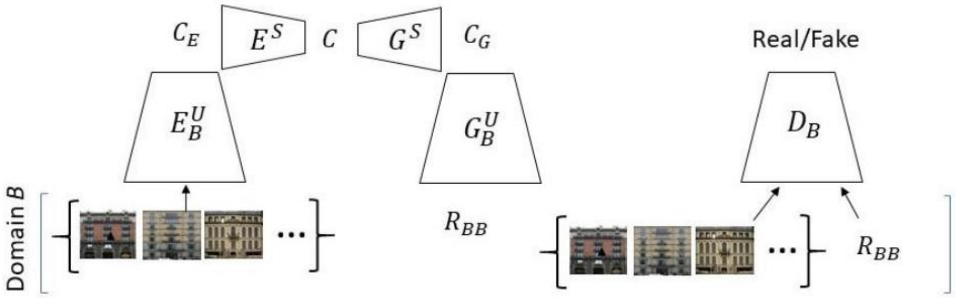
$$\mathcal{L}_{\mathrm{D}_B} = \sum_{s \in P(\Lambda)} +\ell(D_B(\overline{G_B}(\overline{E_B}(s))), 0) + \ell(D_B(s), 1)$$

Phase I



- Shared Latent Space assumption (UNIT Liu et al, CoGAN Liu et al, etc): Upper layers of the encoder and lower layers of the decoder should be shared to achieve successful translation.

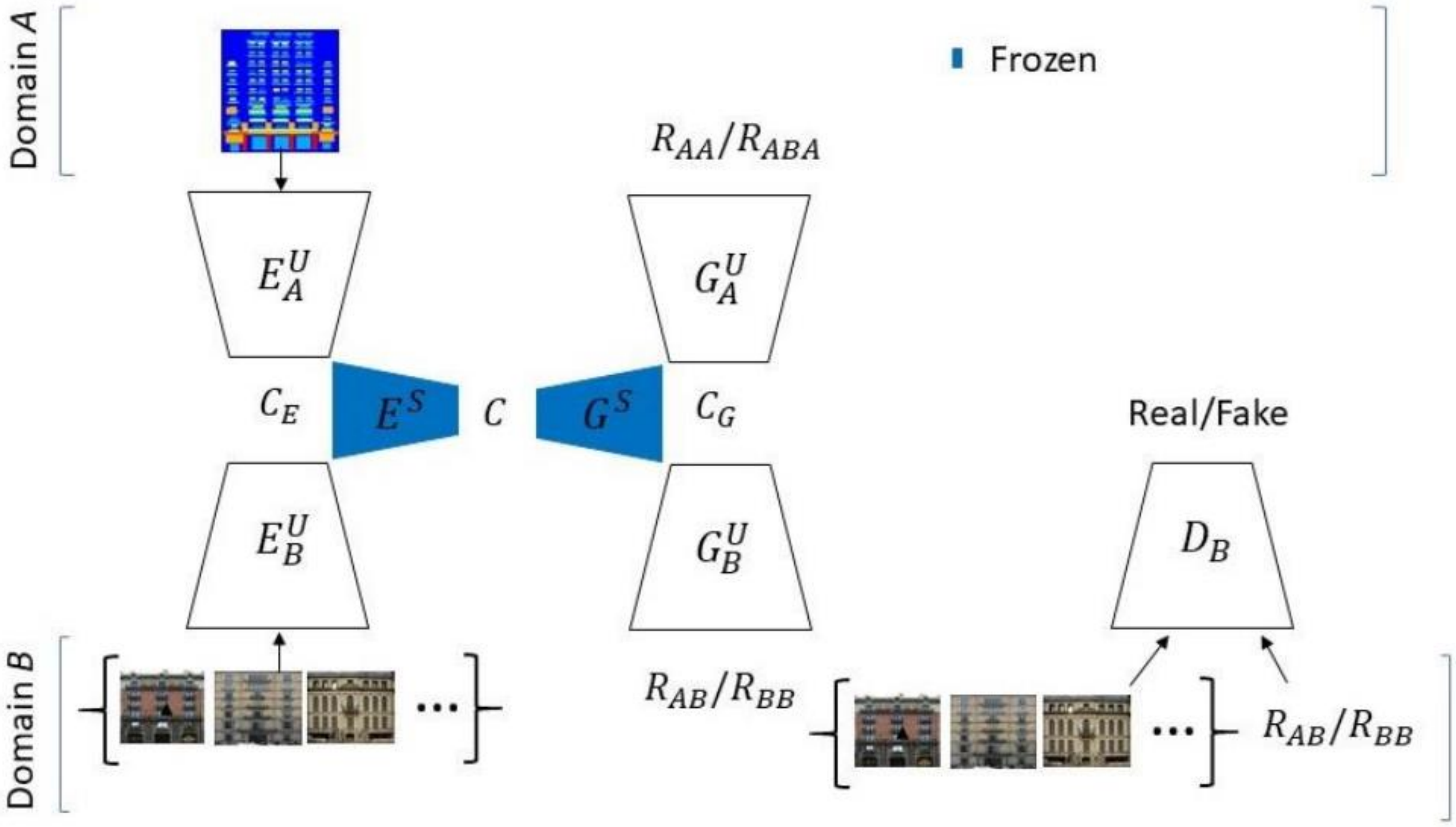
Phase I



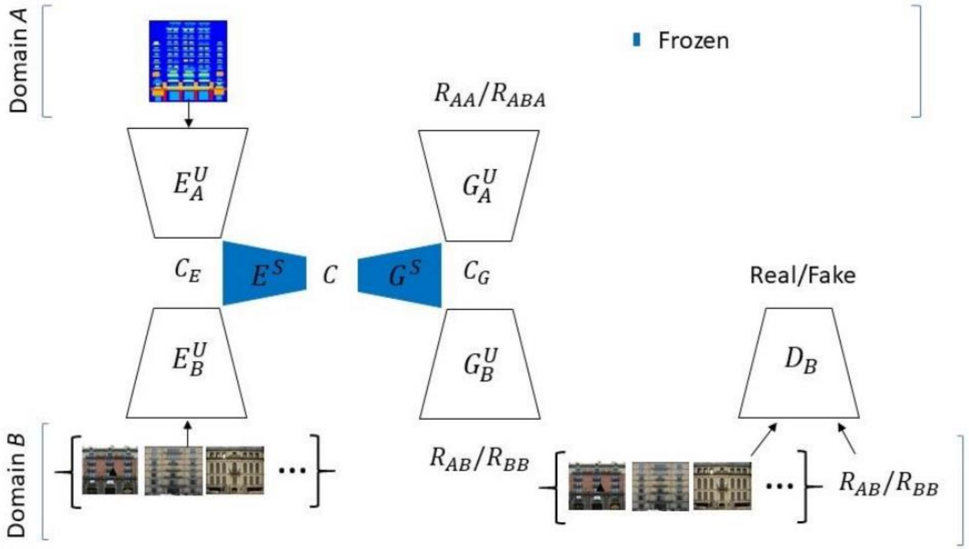
- Shared Latent Space assumption (UNIT Liu et al, CoGAN Liu et al, etc): Upper layers of the encoder and lower layers of the decoder should be shared to achieve successful translation.

- In fact, as we only have a single sample in A, these layers, represented by the shared encoder (Es) and shared decoder (Gs) can be trained with domain B samples **only**

Phase II

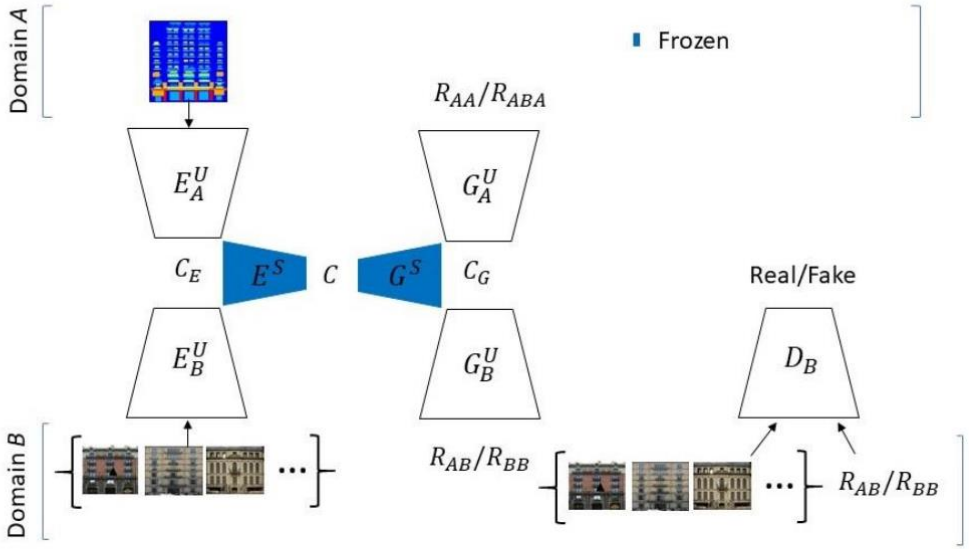

Domain A
■ Frozen
R_{AA}/R_{ABA}
E_A^U
G_A^U
C_E
E^S
C
G^S
C_G
Real/Fake
E_B^U
G_B^U
D_B
Domain B
...
R_{AB}/R_{BB}
...
R_{AB}/R_{BB}

Phase II



1. $\mathcal{L}_{REC_A} = \sum_{s \in P(x)} \|T_{AA}(s) - s\|_1$

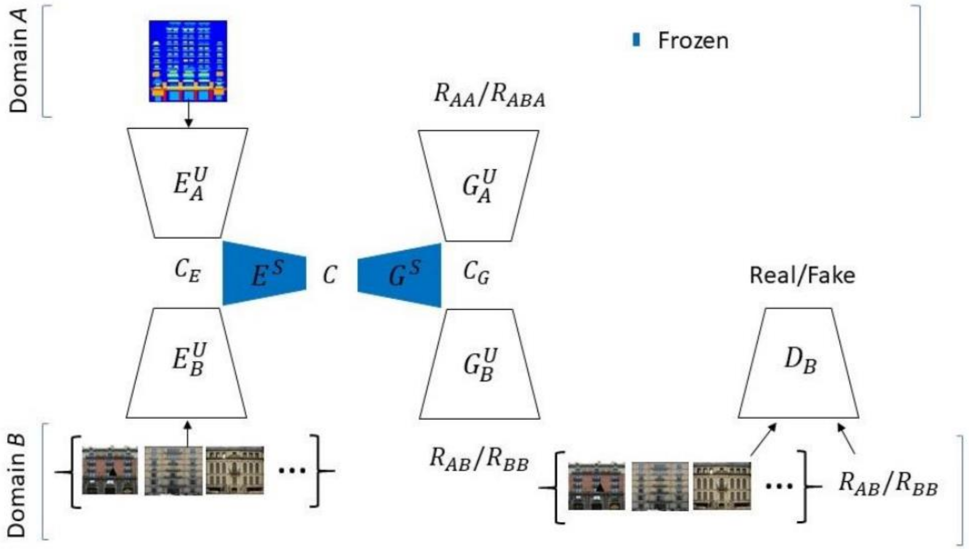
Phase II



1. $$\mathcal{L}_{REC_A} = \sum_{s \in P(x)} \|T_{AA}(s) - s\|_1$$

2. $$\mathcal{L}_{\text{cycle}} = \sum_{s \in P(x)} \|T_{BA}(T_{AB}(s)) - s\|_1$$

Phase II



1. $$\mathcal{L}_{REC_A} = \sum_{s\in P(x)} \|T_{AA}(s) - s\|_1$$

2. $$\mathcal{L}_{\text{cycle}} = \sum_{s\in P(x)} \|T_{BA}(T_{AB}(s)) - s\|_1$$

3. $$\mathcal{L}_{\text{GAN}_{AB}} = \sum_{s\in P(x)} -\ell(\overline{D_B}(T_{AB}(s)), 0)$$

$$\mathcal{L}_{\text{D}_{AB}} = \sum_{s\in P(x)} +\ell(D_B(\overline{T_{AB}}(s)), 0) + \ell(D_B(s), 1)$$

Selective Backpropagation

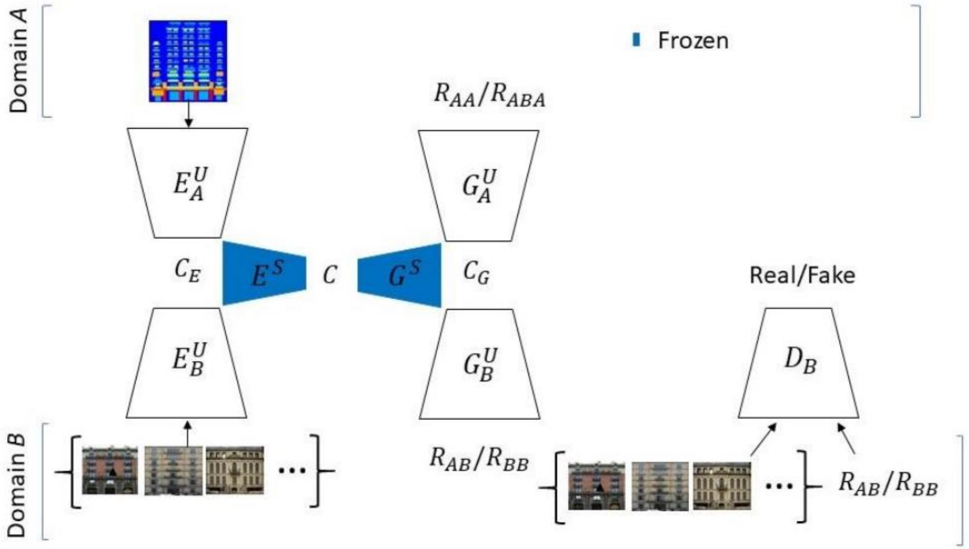
When training our network with x and its augmentations, backpropagation is applied **selectively** on the separate encoders and decoders only.

$$T_{BB} = G^U_B(\overline{G^S}(\overline{E^S}(E^U_B(x))))$$

$$T_{BA} = G^U_A(\overline{G^S}(\overline{E^S}(E^U_B(x))))$$

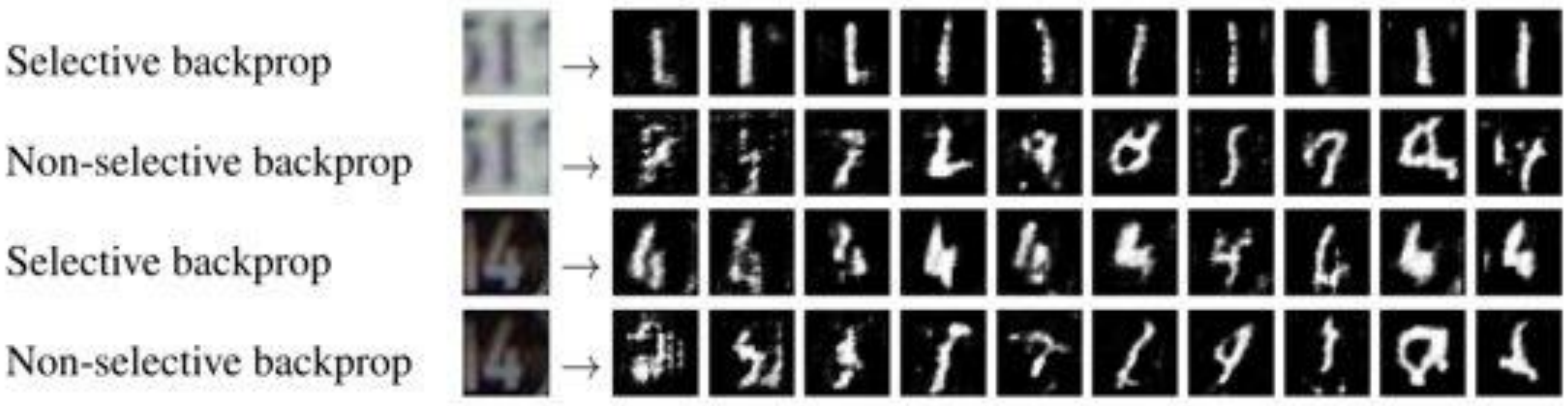
$$T_{AA} = G^U_A(\overline{G^S}(\overline{E^S}(E^U_A(x))))$$

$$T_{AB} = G^U_B(\overline{G^S}(\overline{E^S}(E^U_A(x))))$$



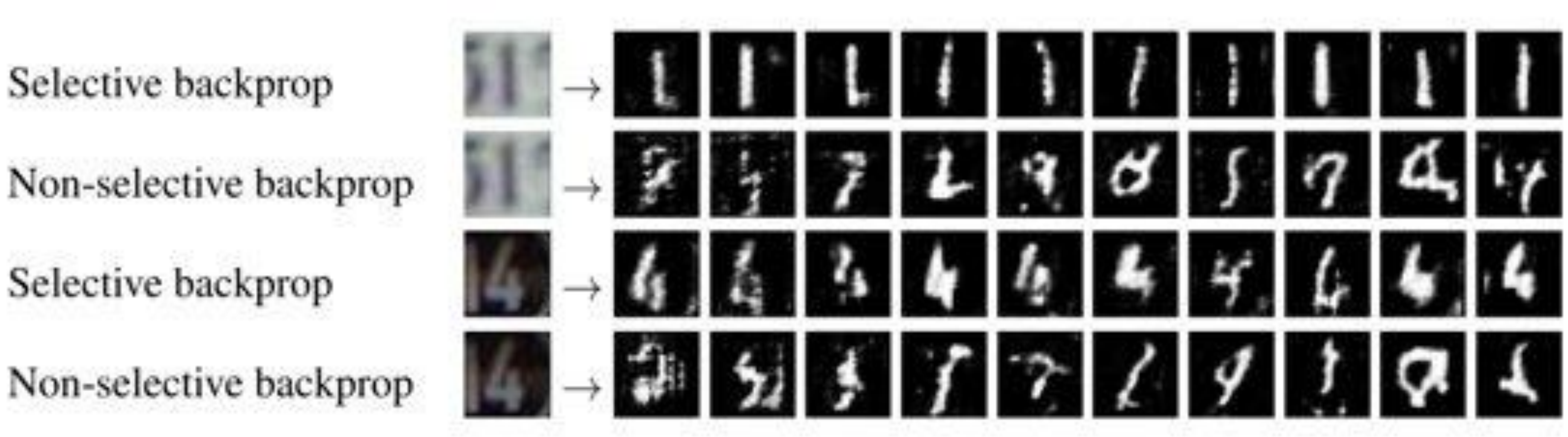
Selective Backpropagation

- Updating the shared encoder (Es) and decoder (Gs) with selective backpropagation turned off leads to **overfitting** on x, since for every shared representation, the unshared layers in domain A can still reconstruct this one sample.



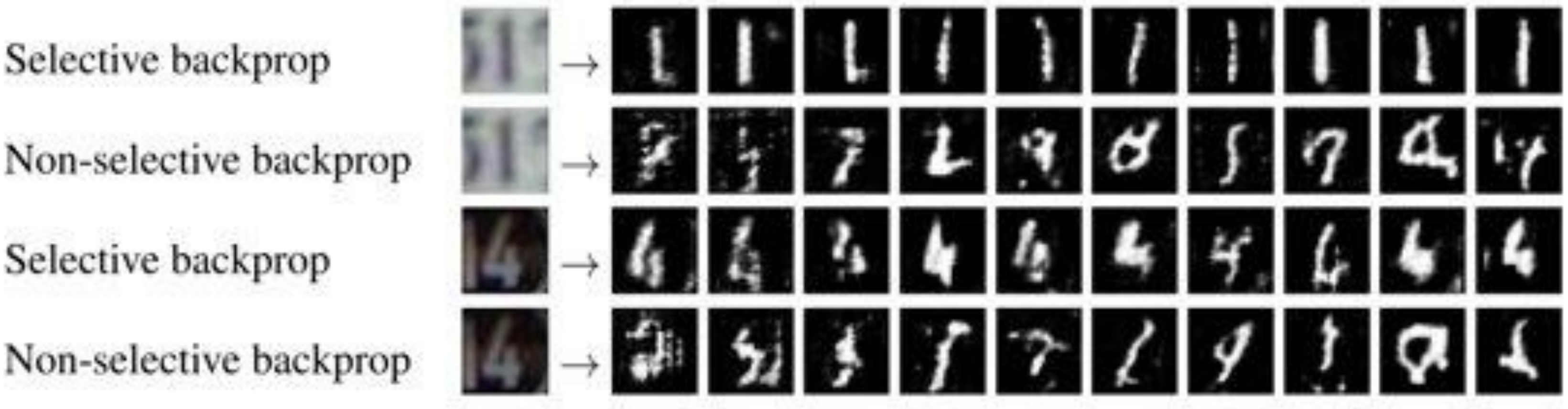
Selective Backpropagation

- Updating the shared encoder (Es) and decoder (Gs) with selective backpropagation turned off leads to **overfitting** on x, since for every shared representation, the unshared layers in domain A can still reconstruct this one sample.
- However, as the shared encoder (Es) and decoder (Gs) can be trained with domain B samples **only**, translation from domain A to B is still possible.

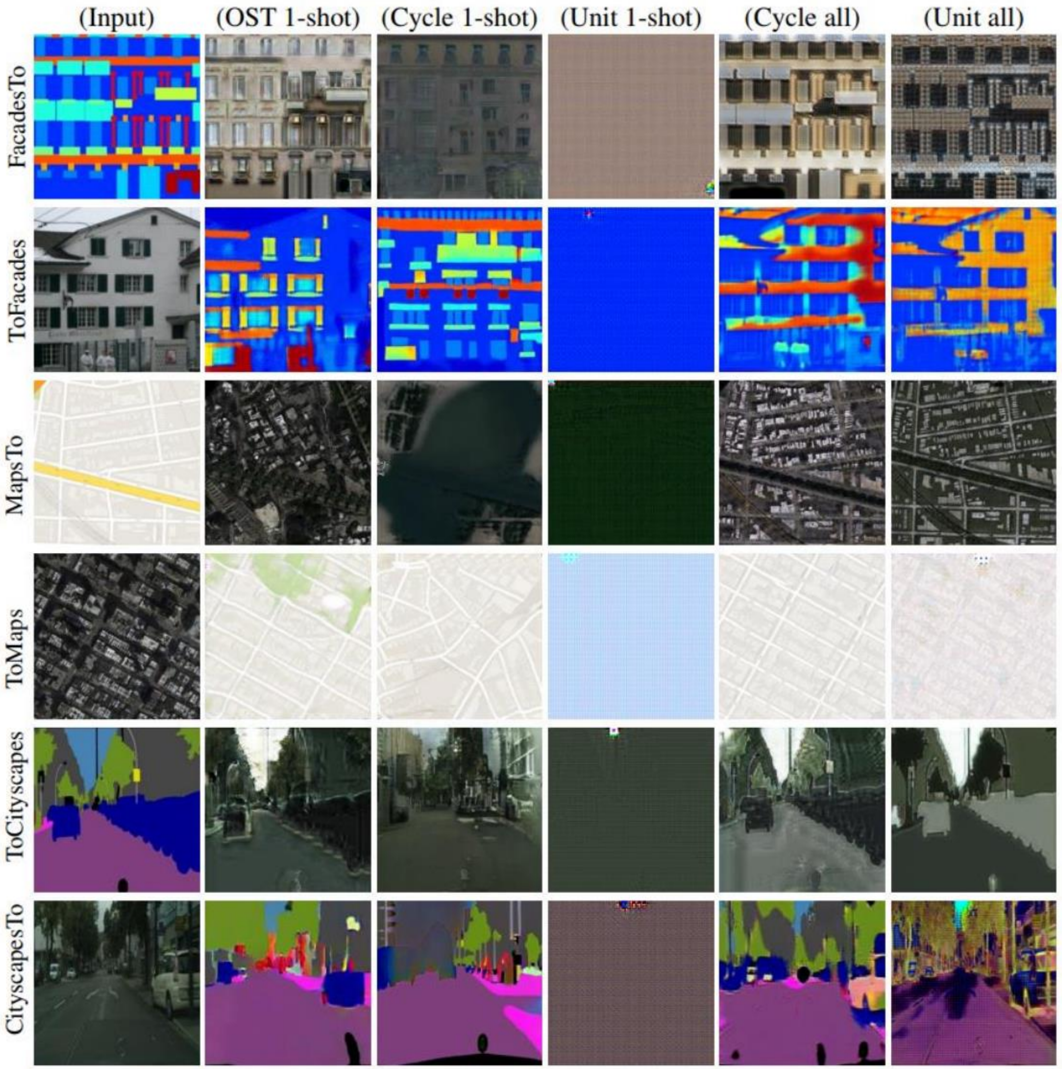


Selective Backpropagation

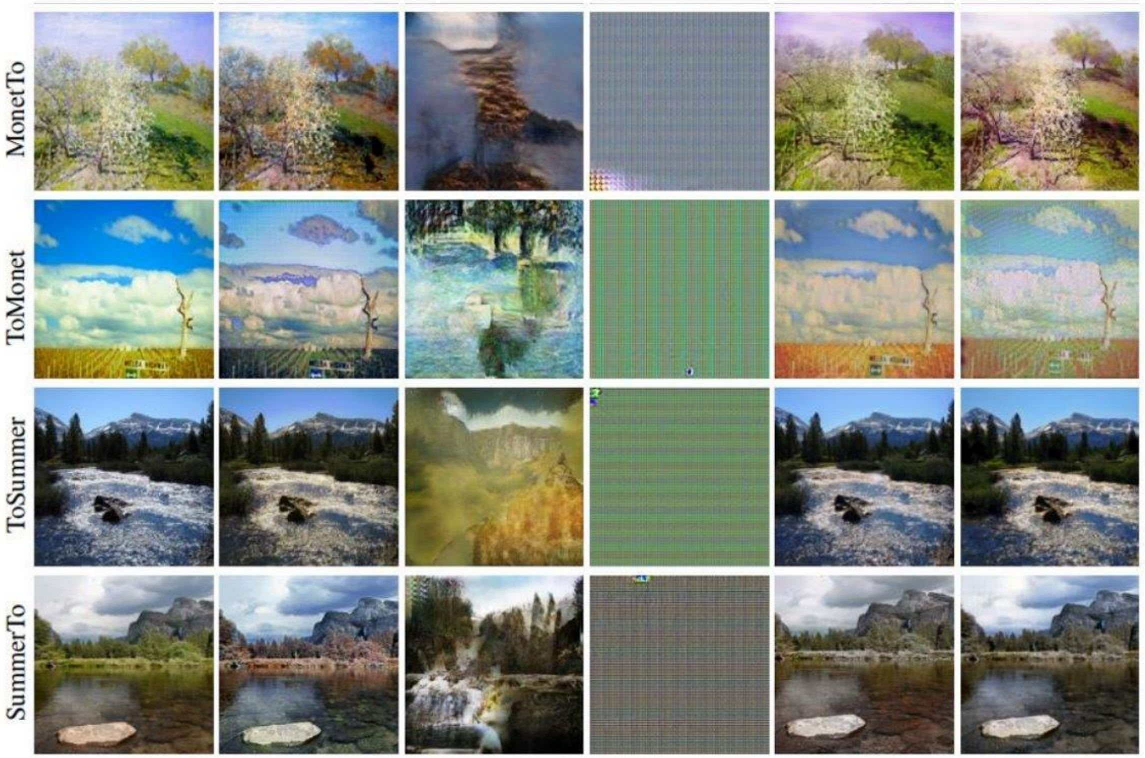
• Updating the shared encoder (Es) and decoder (Gs) with selective backpropagation turned off leads to **overfitting** on x, since for every shared representation, the unshared layers in domain A can still reconstruct this one sample.
• However, as the shared encoder (Es) and decoder (Gs) can be trained with domain B samples **only**, translation from domain A to B is still possible.
•Use of a Patch GAN as well as convolutional layers induces further prior on the network that allows for succesful translation given one input from domain A



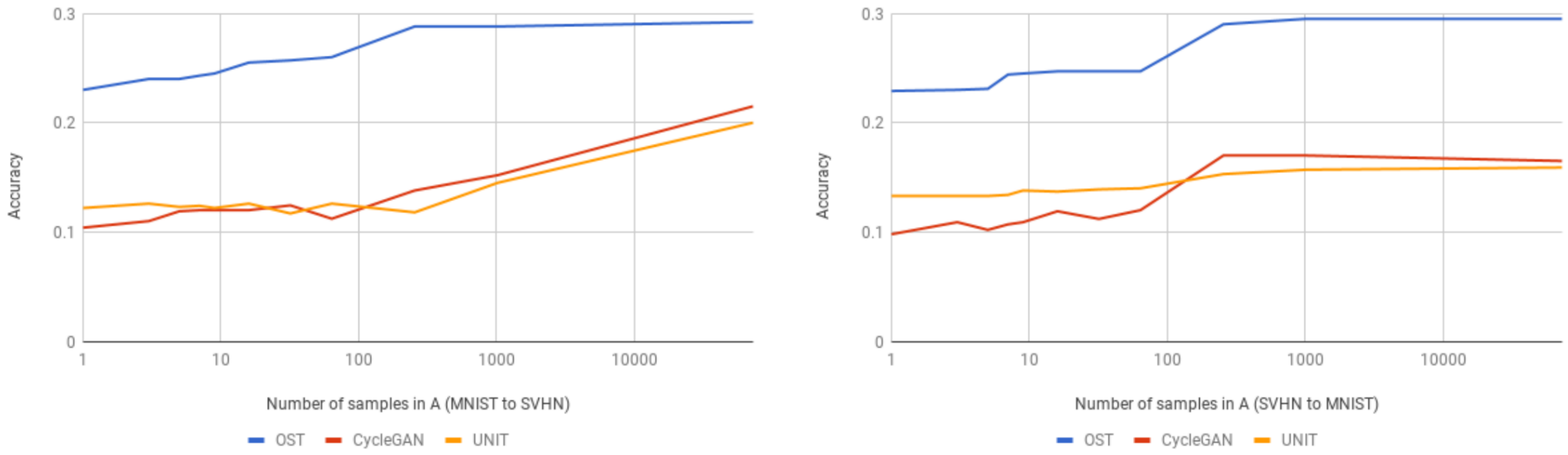
Qaulitative Results



Qaulitative Results



Quantitative Results



(a) (b)

Figure 3: (a) Translating MNIST images to SVHN images. x-axis is the number of samples in A (log-scale), y-axis is the accuracy of a pretrained classifier on the resulting translated images. The accuracy is averaged over 1000 independent runs for different samples. Blue: Our OST method. Yellow: UNIT [7]. Red: CycleGAN [2] . (b) The same graph in the reverse direction.

Quantitative Results

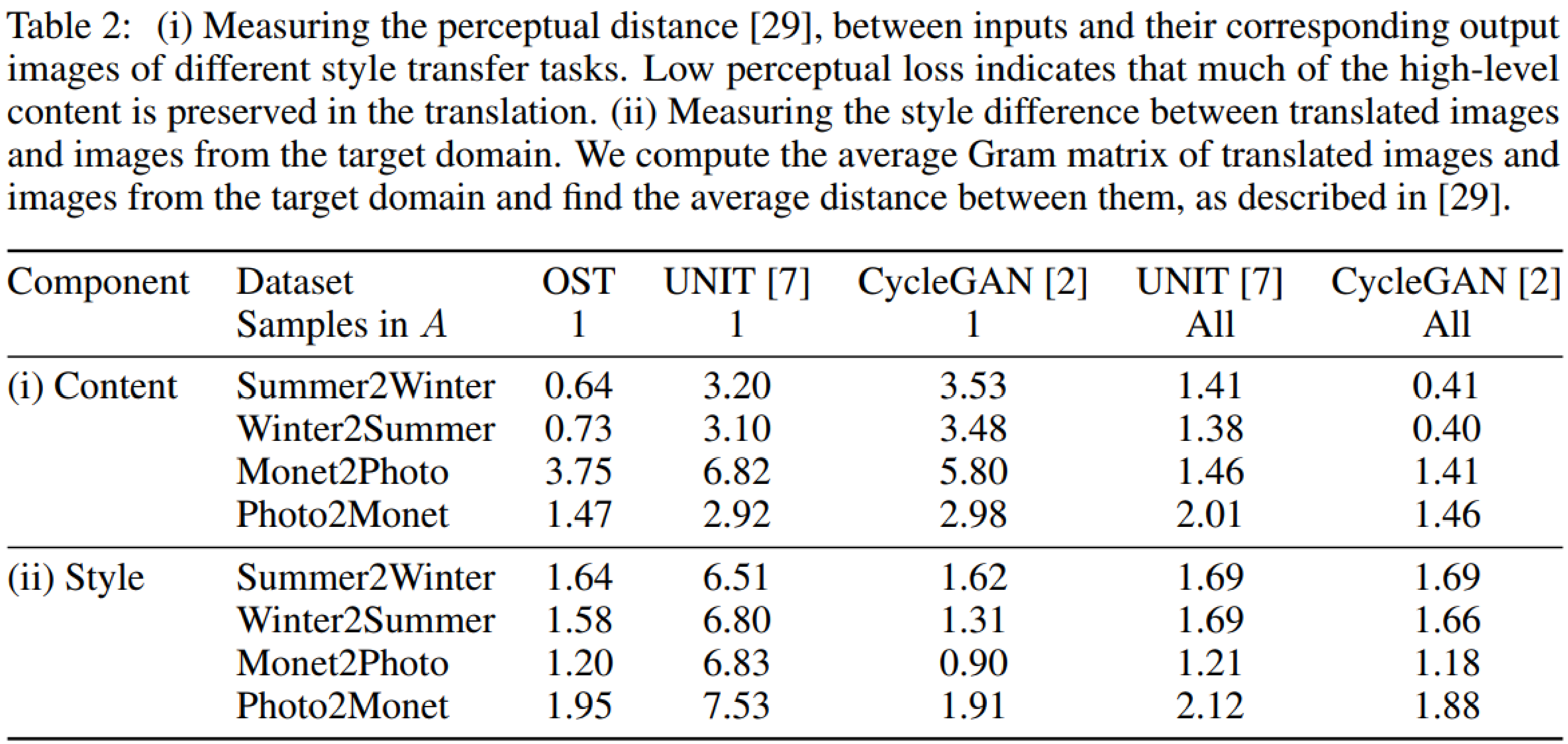
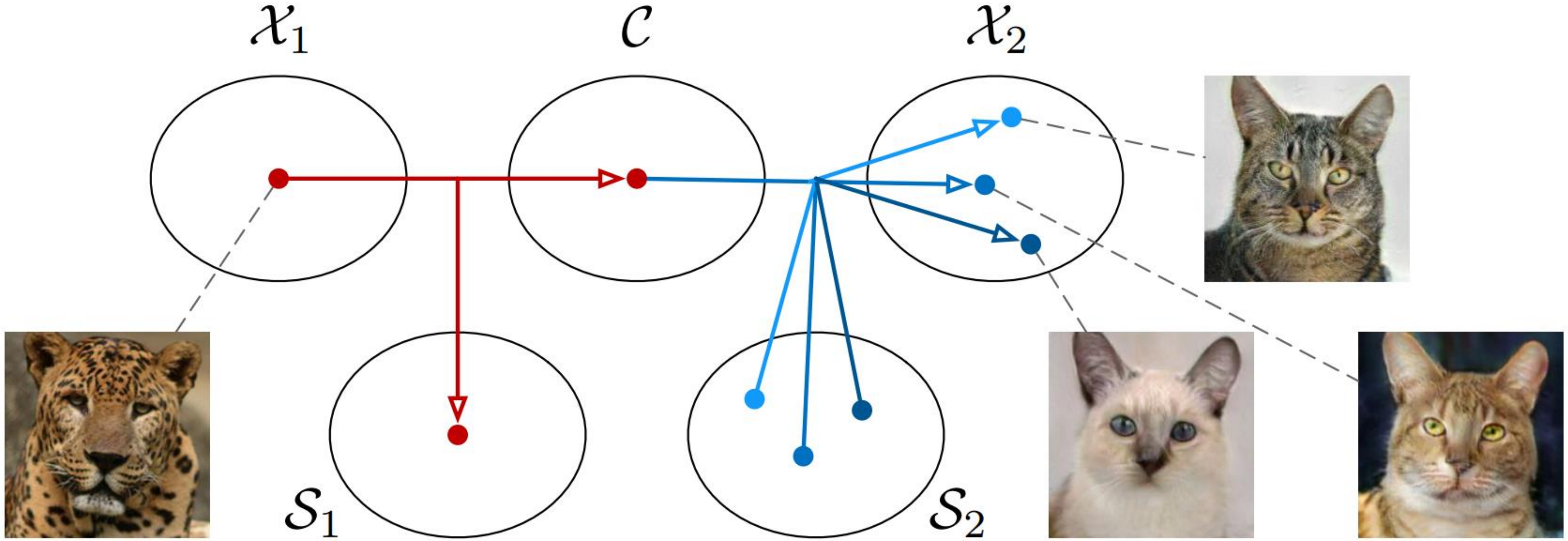


Table 2: (i) Measuring the perceptual distance [29], between inputs and their corresponding output images of different style transfer tasks. Low perceptual loss indicates that much of the high-level content is preserved in the translation. (ii) Measuring the style difference between translated images and images from the target domain. We compute the average Gram matrix of translated images and images from the target domain and find the average distance between them, as described in [29].

Component	Dataset	OST	UNIT [7]	CycleGAN [2]	UNIT [7]	CycleGAN [2]
	Samples in A	1	1	1	All	All
(i) Content	Summer2Winter	0.64	3.20	3.53	1.41	0.41
	Winter2Summer	0.73	3.10	3.48	1.38	0.40
	Monet2Photo	3.75	6.82	5.80	1.46	1.41
	Photo2Monet	1.47	2.92	2.98	2.01	1.46
(ii) Style	Summer2Winter	1.64	6.51	1.62	1.69	1.69
	Winter2Summer	1.58	6.80	1.31	1.69	1.66
	Monet2Photo	1.20	6.83	0.90	1.21	1.18
	Photo2Monet	1.95	7.53	1.91	2.12	1.88

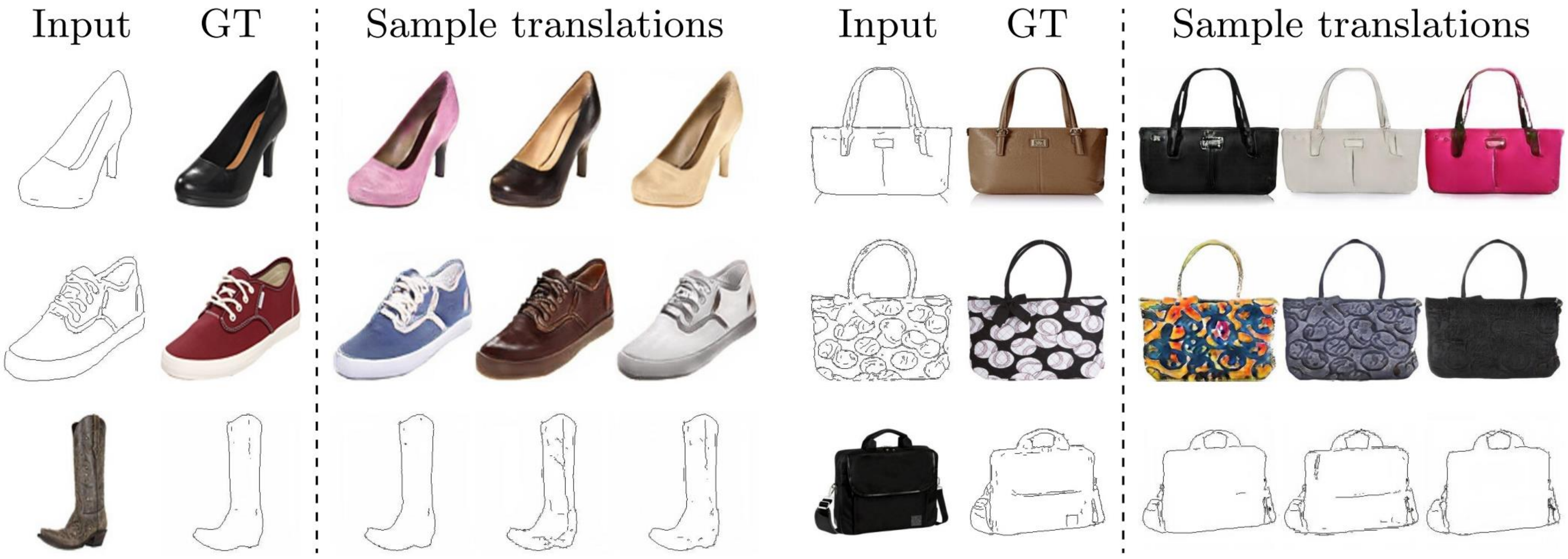
	Supervised	Unsupervised
Unimodal	Pix2pix, CRN, SRGAN	DistanceGAN, CycleGAN, DiscoGAN, DualGAN, UNIT, DTN, StarGAN, OST
Multimodal	pix2pixHD, BicycleGAN	MUNIT, Augmented CycleGAN

Emerging Disentanglement in Auto-Encoder Based Unsupervised Image Content Transfer (ICLR 2019)



(b) Translation

Up to now: Style Only



(a) edges ↔ shoes

(b) edges ↔ handbags

Our Contribution: Full Content Disentanglement

- 2 Encoders:
  - Encoder e1 encodes common content from both domains
  - Encoder e2 encodes separate content from second domain
- One Decoder
  - Decoder g takes the concatenation of e1 and e2’s output and produces and image

Losses

$$\mathcal{L}_B = \frac{1}{m_2} \sum_{\boldsymbol{a} \in \mathbb{S}_B} \| g(e_1(\boldsymbol{b}), e_2(\boldsymbol{b})) - \boldsymbol{b} \|_1$$

Losses

$$\mathcal{L}_B = \frac{1}{m_2} \sum_{\boldsymbol{a} \in \mathbb{S}_B} \| g(e_1(\boldsymbol{b}), e_2(\boldsymbol{b})) - \boldsymbol{b} \|_1$$

$$\mathcal{L}_A = \frac{1}{m_1} \sum_{\boldsymbol{a} \in \mathbb{S}_A} \| g(e_1(\boldsymbol{a}), 0_{E_2}) - \boldsymbol{a} \|_1$$

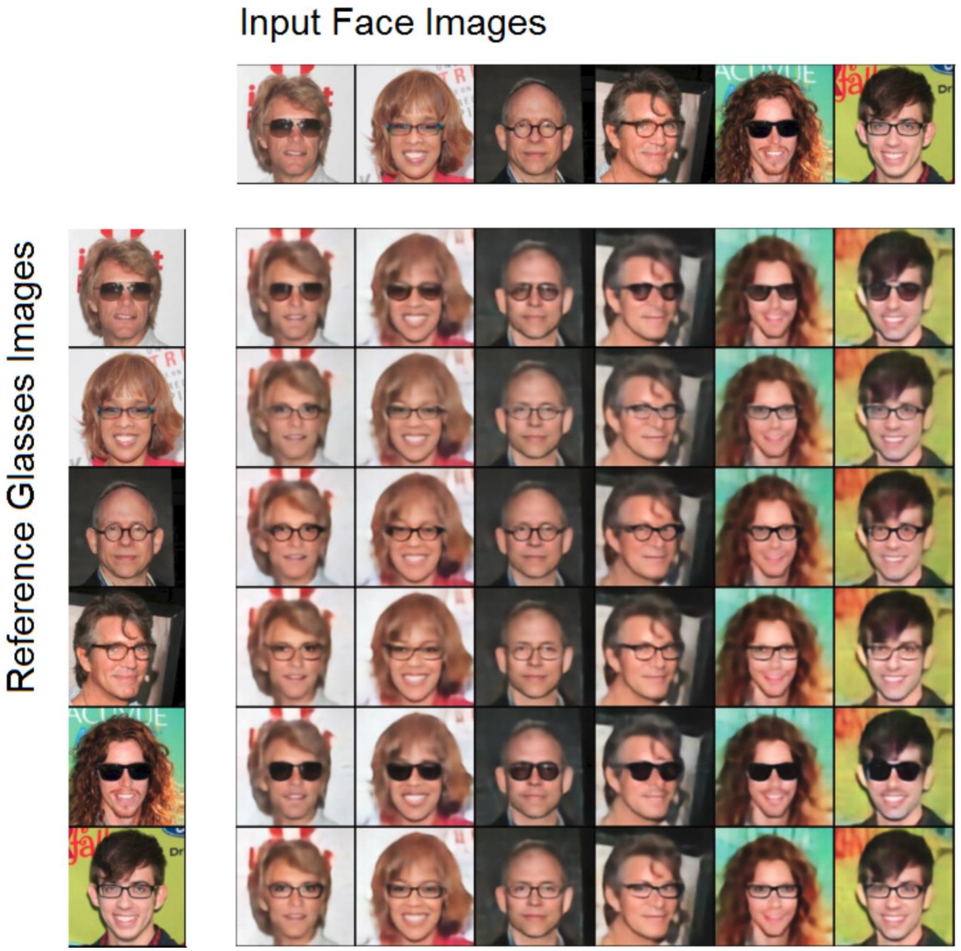
Losses

$$\mathcal{L}_B = \frac{1}{m_2} \sum_{\boldsymbol{a} \in \mathbb{S}_B} \| g(e_1(\boldsymbol{b}), e_2(\boldsymbol{b})) - \boldsymbol{b} \|_1$$

$$\mathcal{L}_A = \frac{1}{m_1} \sum_{\boldsymbol{a} \in \mathbb{S}_A} \| g(e_1(\boldsymbol{a}), 0_{E_2}) - \boldsymbol{a} \|_1$$

$$\mathcal{L}_D = \frac{1}{m_1} \sum_{\boldsymbol{a} \in \mathbb{S}_A} l(d(e_1(\boldsymbol{a})), 0) + \frac{1}{m_2} \sum_{\boldsymbol{b} \in \mathbb{S}_B} l(d(e_1(\boldsymbol{b})), 1)$$

Qualitative Results



Qualitative Results

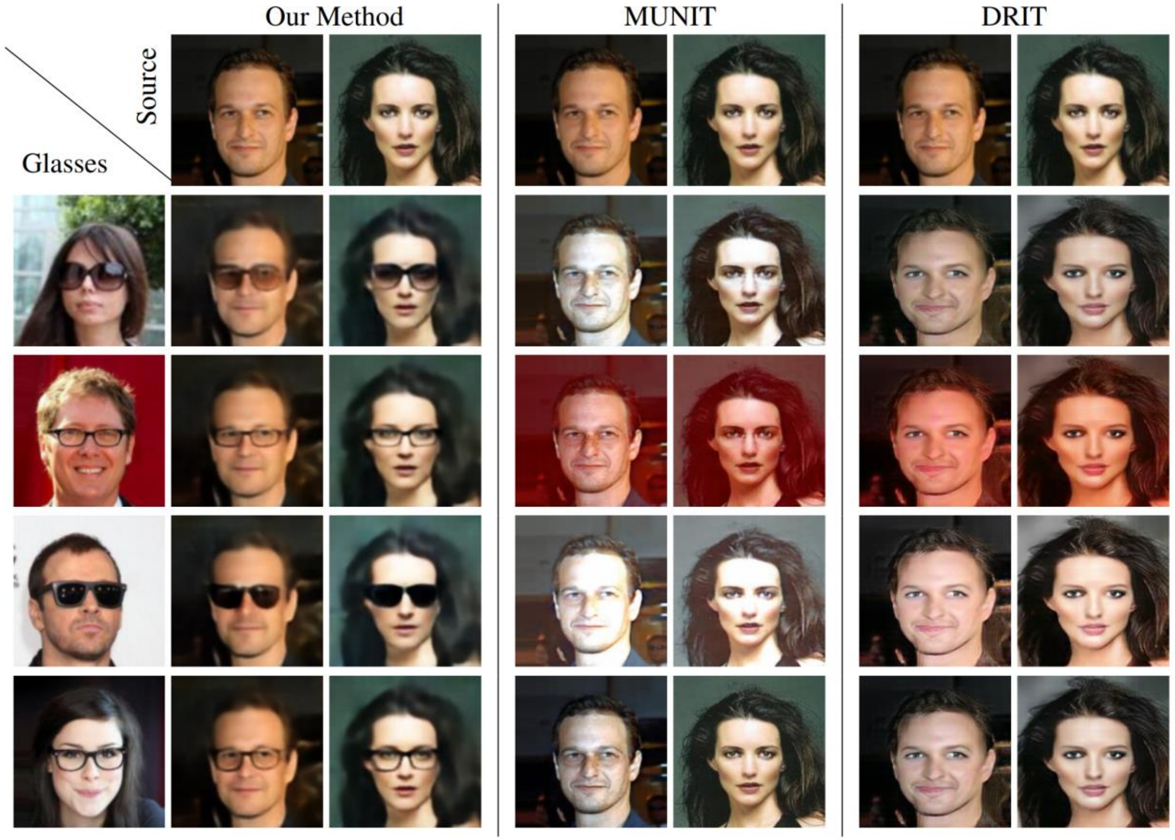


Figure 2: Glasses transfer. Our method vs literature baselines. Each image combines the domain A image in the top row, with the content of the guide image on the left column.

Qualitative Results

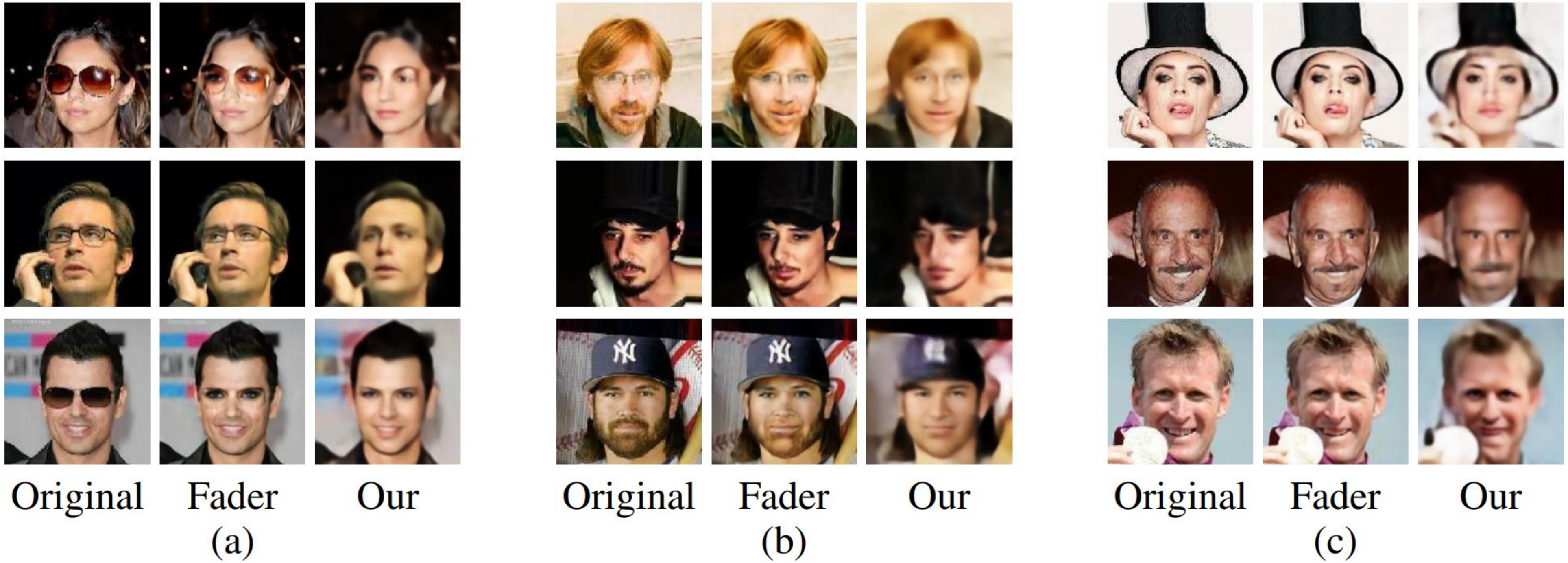


Figure 6: A comparison to the Fader networks of Lample et al. (2017) for the task of removing a feature. (a) Glasses. (b) Facial hair. (c) Mouth opening.

Quantitative Results

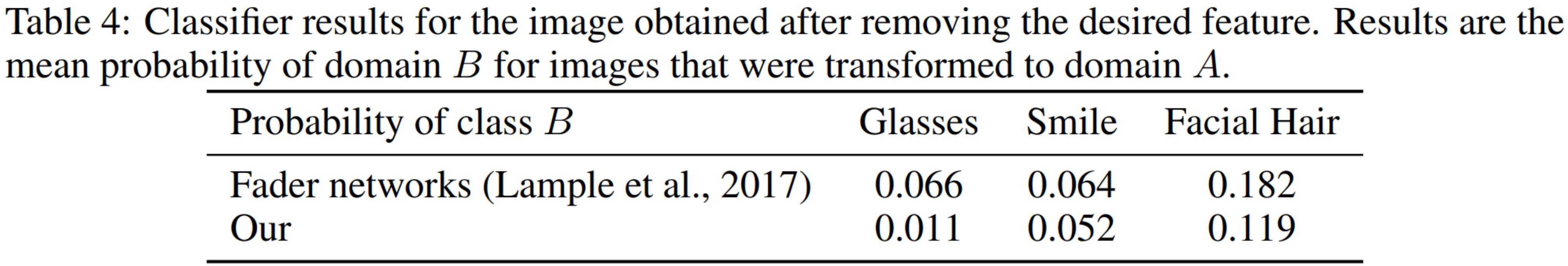


Table 4: Classifier results for the image obtained after removing the desired feature. Results are the mean probability of domain B for images that were transformed to domain A.

Probability of class B	Glasses	Smile	Facial Hair
Fader networks (Lample et al., 2017)	0.066	0.064	0.182
Our	0.011	0.052	0.119

Quantitative Results

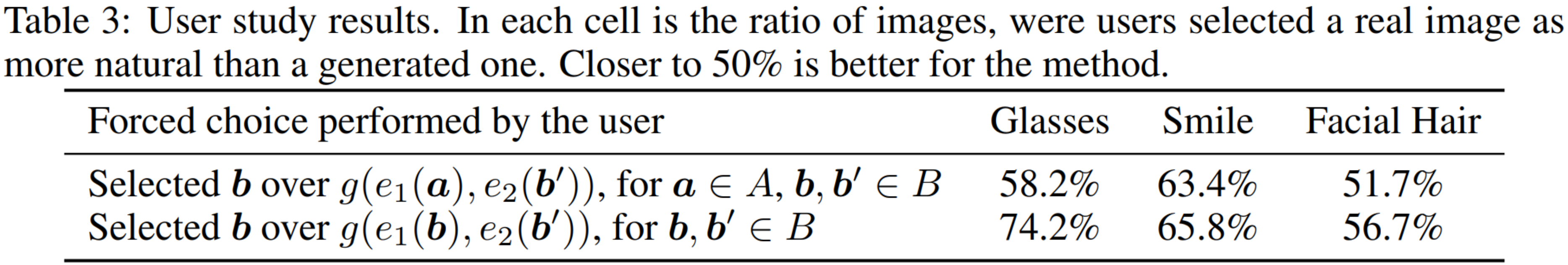


Table 3: User study results. In each cell is the ratio of images, were users selected a real image as more natural than a generated one. Closer to 50% is better for the method.

Forced choice performed by the user	Glasses	Smile	Facial Hair
Selected $\boldsymbol{b}$ over $g(e_1(\boldsymbol{a}), e_2(\mathbf{b}'))$, for $\boldsymbol{a} \in A$, $\mathbf{b}, \mathbf{b}' \in B$	58.2%	63.4%	51.7%
Selected $\boldsymbol{b}$ over $g(e_1(\mathbf{b}), e_2(\mathbf{b}'))$, for $\mathbf{b}, \mathbf{b}' \in B$	74.2%	65.8%	56.7%

Quantitative Results

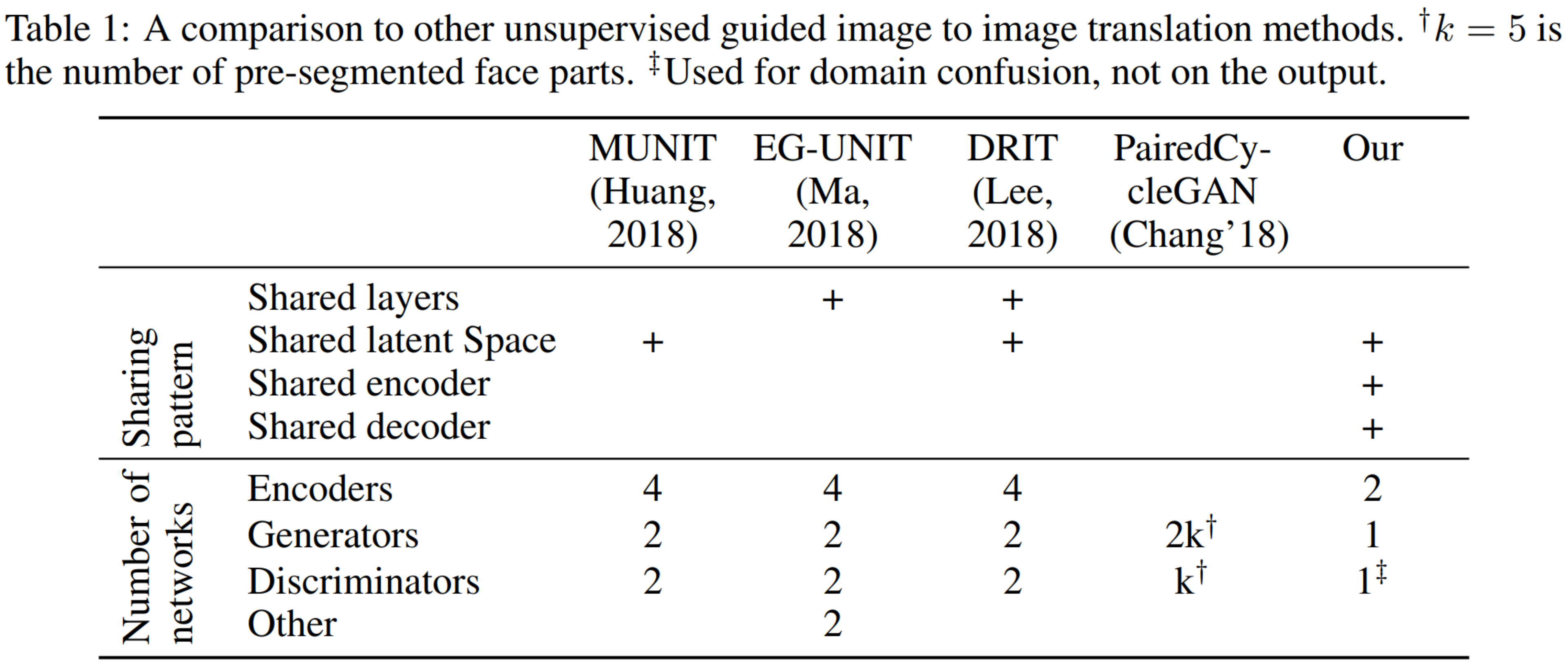


Table 1: A comparison to other unsupervised guided image to image translation methods. $^{\dagger}k = 5$ is the number of pre-segmented face parts. ‡Used for domain confusion, not on the output.

		MUNIT (Huang, 2018)	EG-UNIT (Ma, 2018)	DRIT (Lee, 2018)	PairedCycleGAN (Chang'18)	Our
Sharing pattern	Shared layers		+	+		
	Shared latent Space	+		+		+
	Shared encoder					+
	Shared decoder					+
Number of networks	Encoders	4	4	4		2
	Generators	2	2	2	$2k^{\dagger}$	1
	Discriminators	2	2	2	$k^{\dagger}$	$1^{\ddagger}$
	Other		2			

Other Domains?

- Audio Separation: Training data consists of a set of samples of mixed music and an unmatched set of instrumental music.
- Given a mixed sample, wish the separate the voice from the background instrumental music.

Key Elements

- After mapping the audio sample to a Spectrogram, can subtract the “background” from the “mixed” sample in “pixel space”, to get the “voice” only sample.

$$\begin{tikzcd}
 & GAN_C & R_2 & \\
a \arrow[r, "g"] & g(a) \arrow[u, dash, dashed] \arrow[ru, dash, dashed] \arrow[rr, "g"] \arrow[d, "a - g(a)"] & & g(g(a)) \arrow[lu, dash, dashed] \\
 & {} & &
\end{tikzcd}$$

$$\begin{tikzcd}
& {} \arrow[d, "a - g(a)"] & \\
& \overline{b} \in \overline{S}_B \arrow[r, dashed, no head, "R_4"] \arrow[d] & \overline{a} - g(\overline{a}) \\
c \arrow[r] \arrow[ddr, dashed, no head, "R_1"] \arrow[ddr, bend right, "g"'] \arrow[rr, dashed, no head, bend right=70, "R_3"'] & \overline{a} \in \overline{S}_B \times C \arrow[r, "g"] \arrow[d, dashed, no head, red] & g(\overline{a}) \arrow[u] \\
& GAN_A & \\
& g(c) &
\end{tikzcd}$$

Future reseach

- One Shot Domain Adaptation
- One Shot Image to Image translation in the reverse direction
- Other Domains: Audio, Video?
- Online Setting
- Finer Details of Disentanglement
- Other Domains where “Pixel Space” Subtraction is possible.

Thank You! Questions?

Minimality

- Potentially Infinitely many solutions preserving distance correlations

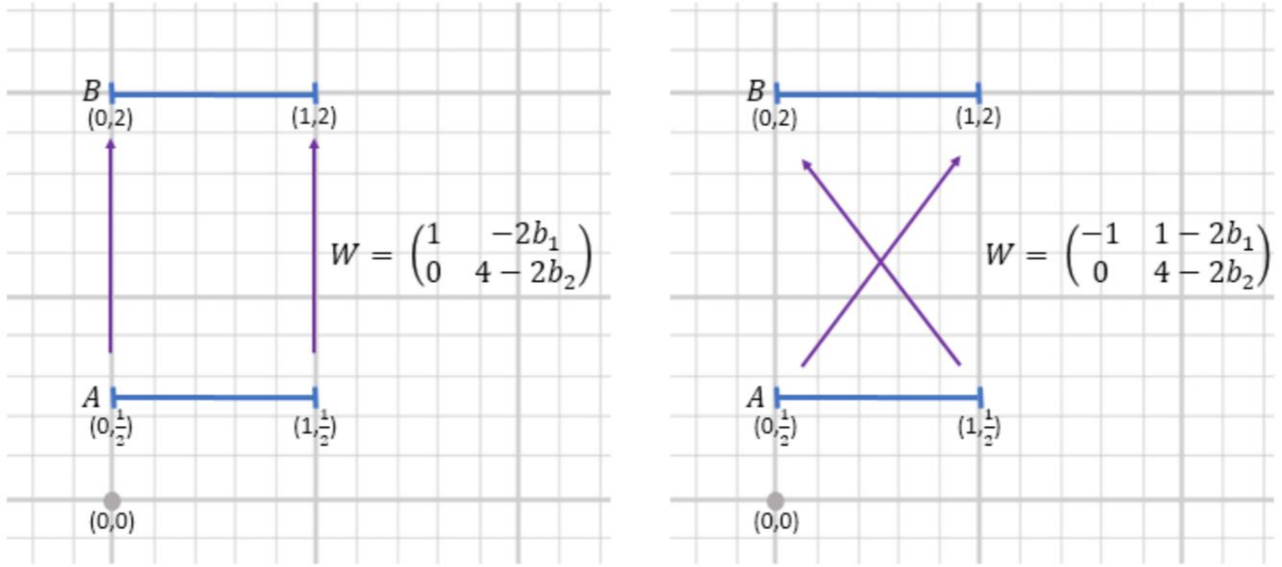

B
(0,2)
(1,2)
$W = \begin{pmatrix} 1 & -2b_1 \\ 0 & 4-2b_2 \end{pmatrix}$
A
$(0,\frac{1}{2})$
$(1,\frac{1}{2})$
(0,0)
B
(0,2)
(1,2)
$W = \begin{pmatrix} -1 & 1-2b_1 \\ 0 & 4-2b_2 \end{pmatrix}$
A
$(0,\frac{1}{2})$
$(1,\frac{1}{2})$
(0,0)

Figure 1: An illustrative example where the two domains are line segments in $\mathbb{R}^2$. There are infinitely many mappings that preserve the uniform distribution on the two segments. However, only two stand out as "semantic". These are exactly the two mappings that can be captured by a neural network with only two hidden neurons and Leaky ReLU activations, i.e., by a function $h(x) = \sigma_a(Wx + b)$, for a weight matrix W and the bias vector b.