Visual Analogies: The role of disentanglement and learning from few examples

Sagie Benaim
School of Computer Science, Tel Aviv University

TEL AVIV UNIVERSITY

Visual Analogies

Domain **A**

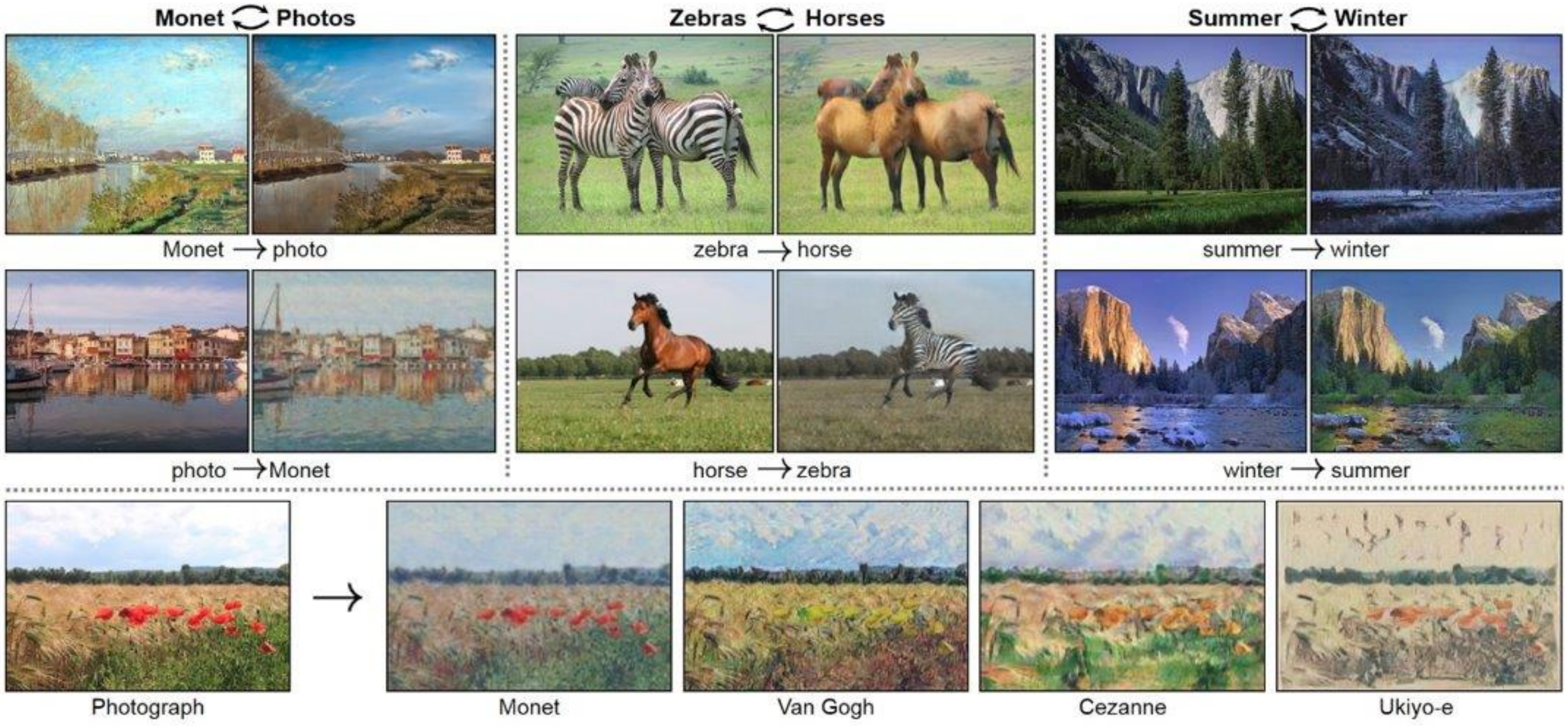
Domain **B**

DiscoGAN, Kim et al., ICML 2017

Visual Analogies

DiscoGAN, Kim et al., ICML 2017

Visual Analogies



CycleGAN, Zhu et al., ICCV 2017

Part I: The role of disentanglement in visual analogies

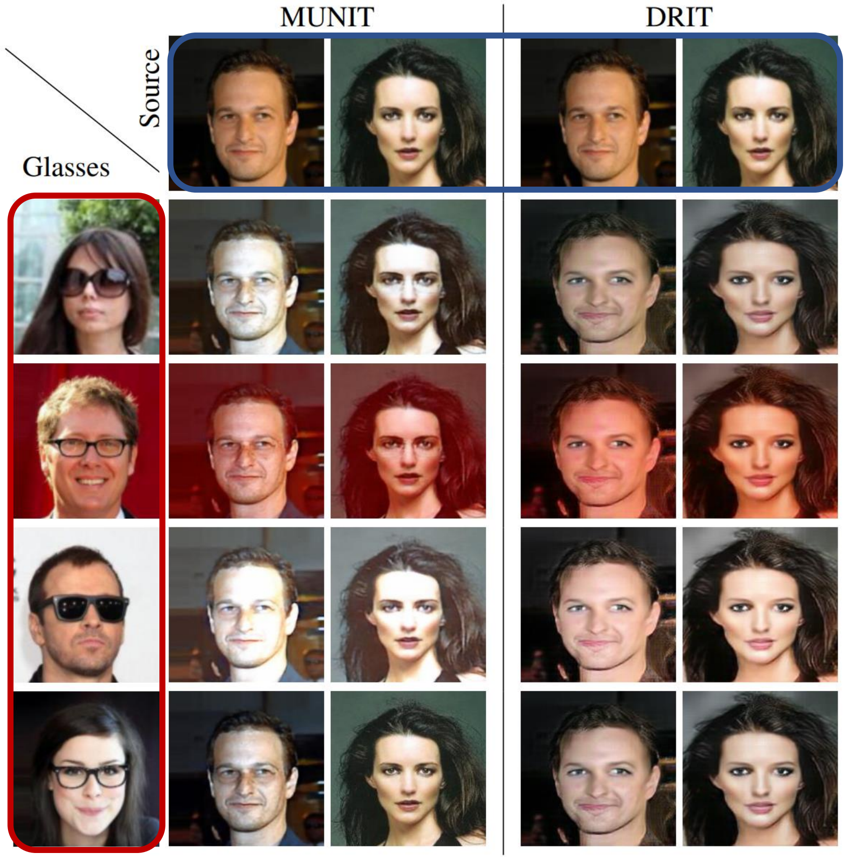
One to many problem

- CycleGAN and DiscoGAN produce a single output
- Many visual analogies exist
- MUNIT and DRIT: Style and texture variations

(b) edges ↔ handbags

Huang et al., ECCV 2018

Cannot transfer content



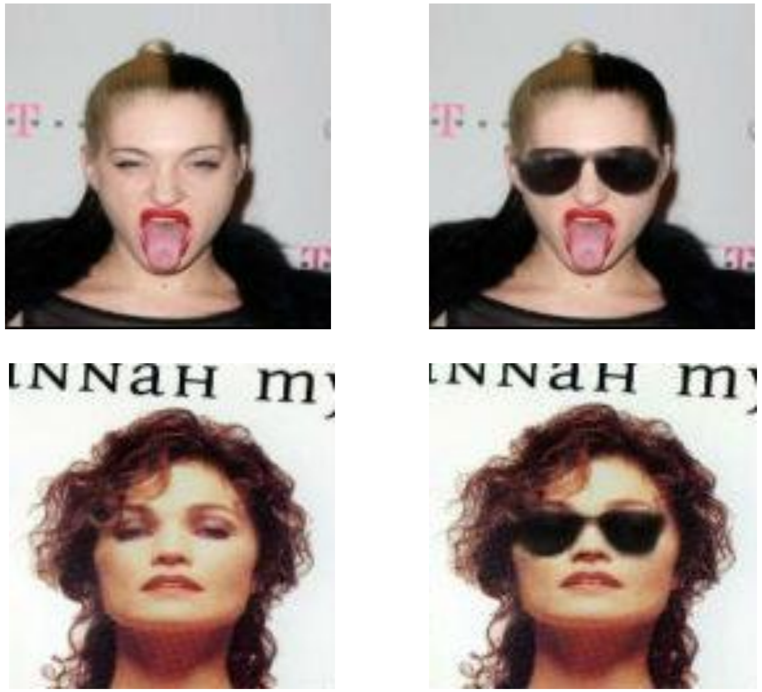
Press et al, ICLR 2019

Attribute Transfer

Liu et al, CVPR 2019

Add glasses

Target Result



Transfer specific glasses

Source Target Result

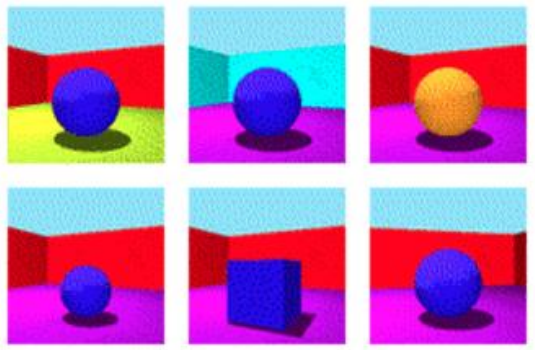
Domain Intersection and Domain Difference

S. Benaim, M. Khaitov, T. Galanti, L. Wolf. ICCV 2019.

Given two visual domains, disentangle the **separate (domain specific)** information and **common (domain invariant)** information.

Disentanglement in Literature

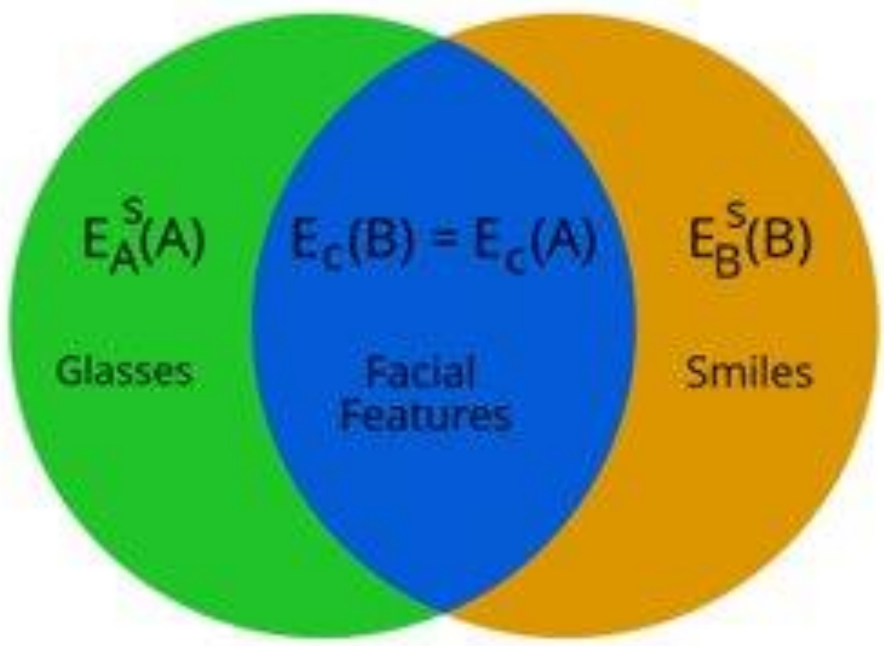
- BetaVAE, AnnealedVAE, FactorVAE and other works disentangle a particular 'pre-specified' property in a set of images, such as color, shape, size.



- We aim to disentangle the **separate (domain specific)** and **common (domain invariant).**

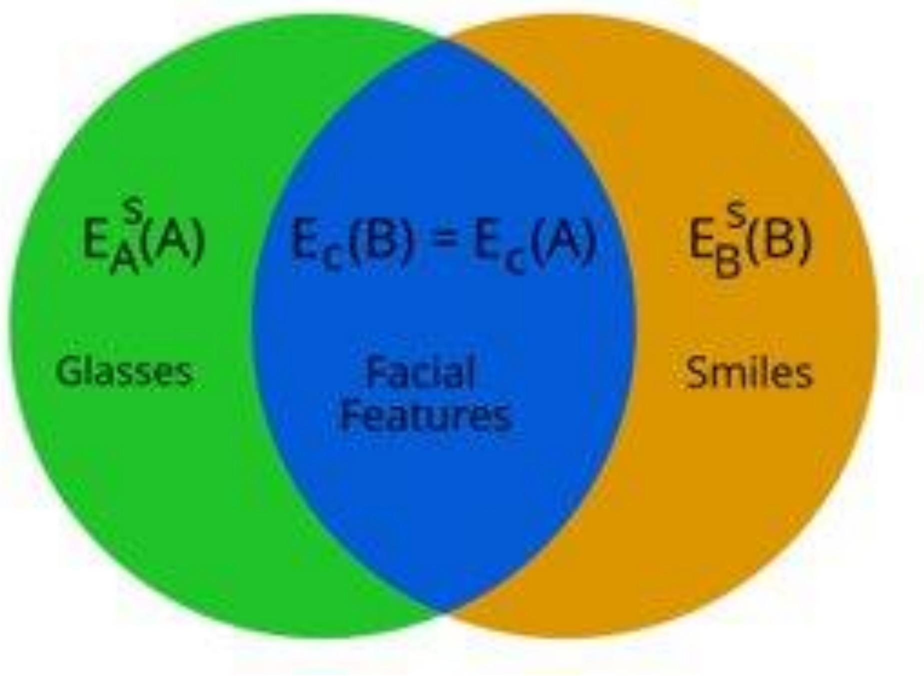
If A is **persons with glasses** and B is **smiling persons,** our method produces three latent spaces:

1. "Common" latent space, $E_c(A) = E_c(B)$. The space of **common facial features.** For $c \in A \cup B, E_c(c)$ is the **facial features of c**.



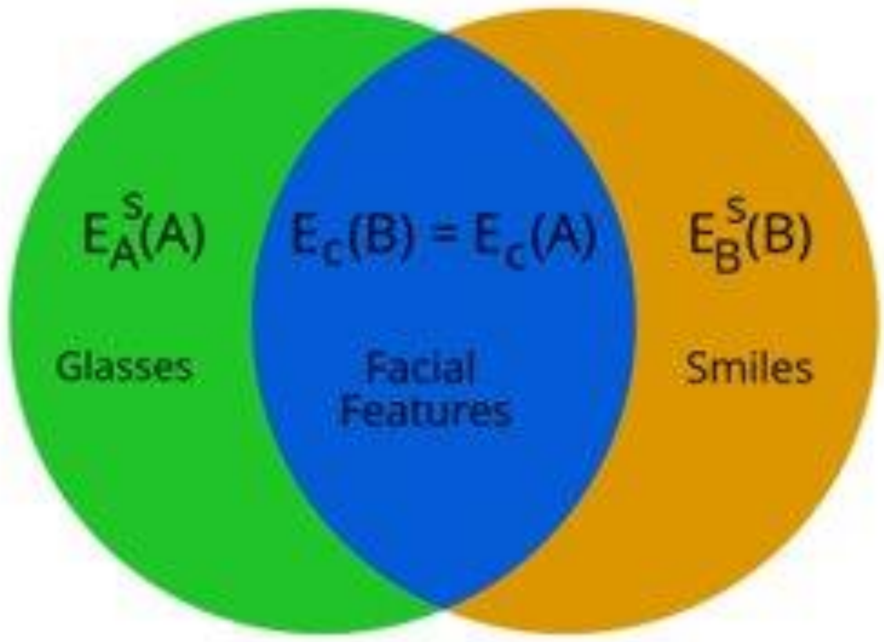
If A is **persons with glasses** and B is **smiling persons,** our method produces three latent spaces:

1. "Common" latent space, $E_c(A) = E_c(B)$. The space of **common facial features.** For $c \in A \cup B, E_c(c)$ is the **facial features of c**.
2. "Separate" latent space for domain A, $E_A^s(A)$. The **space of glasses.** $E_A^s(a)$ is the **glasses of** $\boldsymbol{a}$.

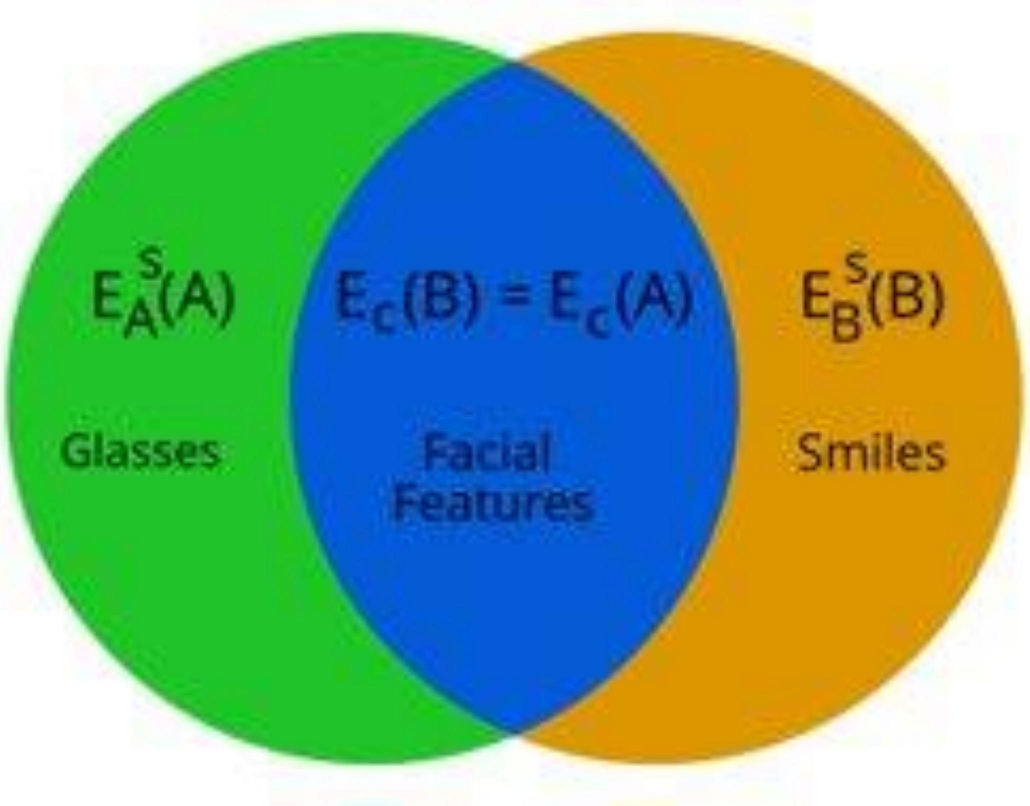


If A is **persons with glasses** and B is **smiling persons,** our method produces three latent spaces:

1. "Common" latent space, $E_c(A) = E_c(B)$. The space of **common facial features.** For $c \in A \cup B, E_c(c)$ is the **facial features of c**.
2. "Separate" latent space for domain A, $E_A^s(A)$. The **space of glasses.** $E_A^s(a)$ is the **glasses of $\boldsymbol{a}$**.
3. "Separate" latent space for domain B, $E_B^s(B)$. The **space of smiles.** $E_B^s(b)$ is the **smile of $\boldsymbol{b}$.**



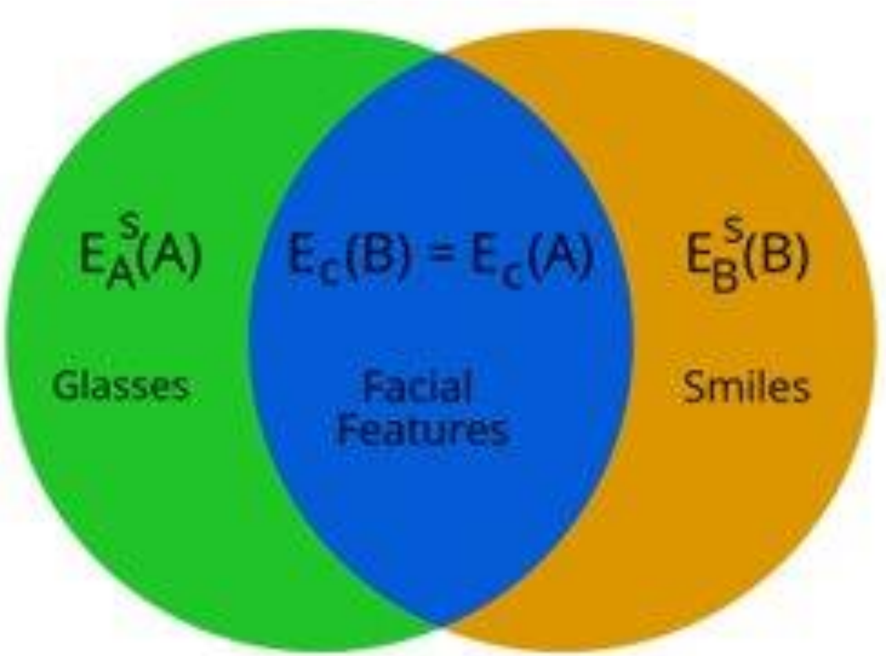
Given this disentangled representation, we generate a visual sample $G(E_c(c), E_A^s(a), E_B^s(b))$, having the **facial features of c, glasses of a, smile of b.**

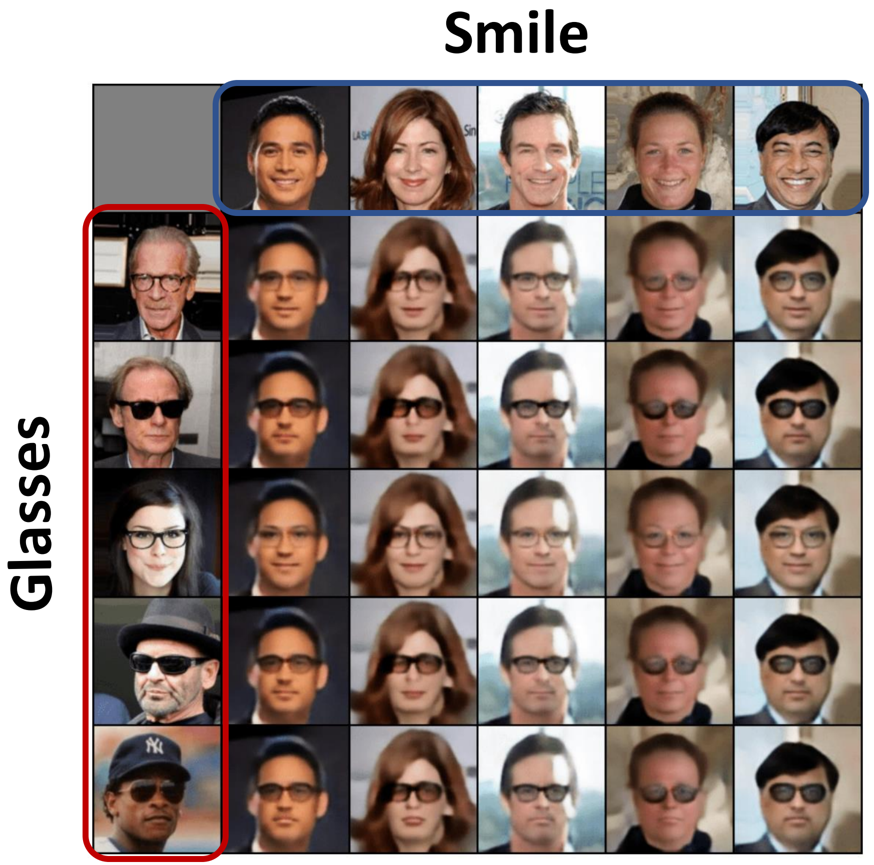


$$G(E_c(b), E_A^s(a), 0)$$

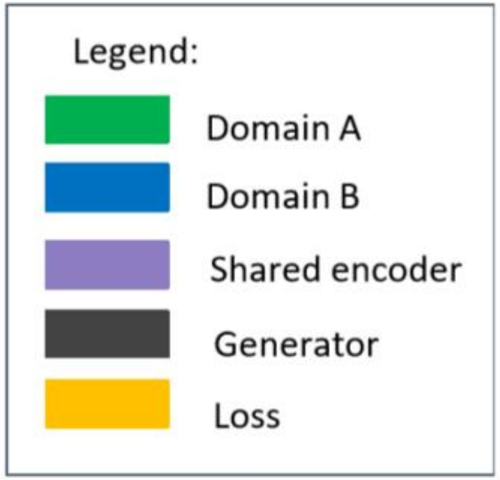
remove b’s smile

add a’s glasses





The "common" (or shared) Loss



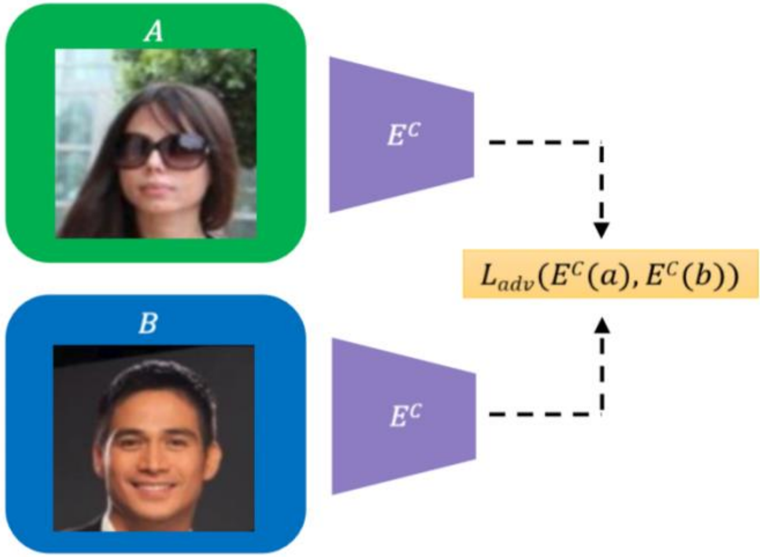
Ensures E_c encodes information common to both domains

Encoder E_c attempts to match distributions of $E_c(A)$ and $E_c(B)$:

$$\frac{1}{m_1}\sum_{i=1}^{m_1} l(d(E^c(a_i)),1) + \frac{1}{m_2}\sum_{j=1}^{m_2} l(d(E^c(b_j)),1)$$

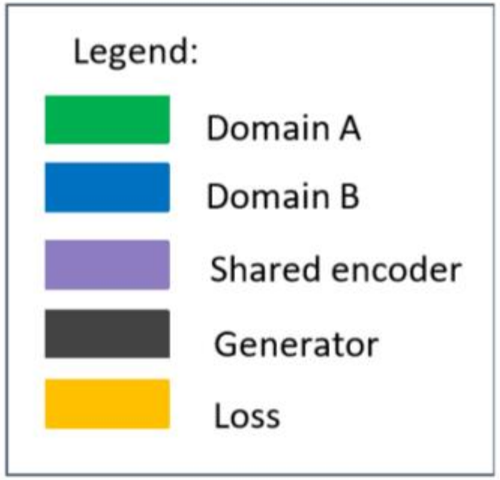
Discriminator d attempts to separate distributions:

$$\mathcal{L}_d := \frac{1}{m_1}\sum_{i=1}^{m_1} l(d(E^c(a_i)),0) + \frac{1}{m_2}\sum_{j=1}^{m_2} l(d(E^c(b_j)),1)$$



(c)

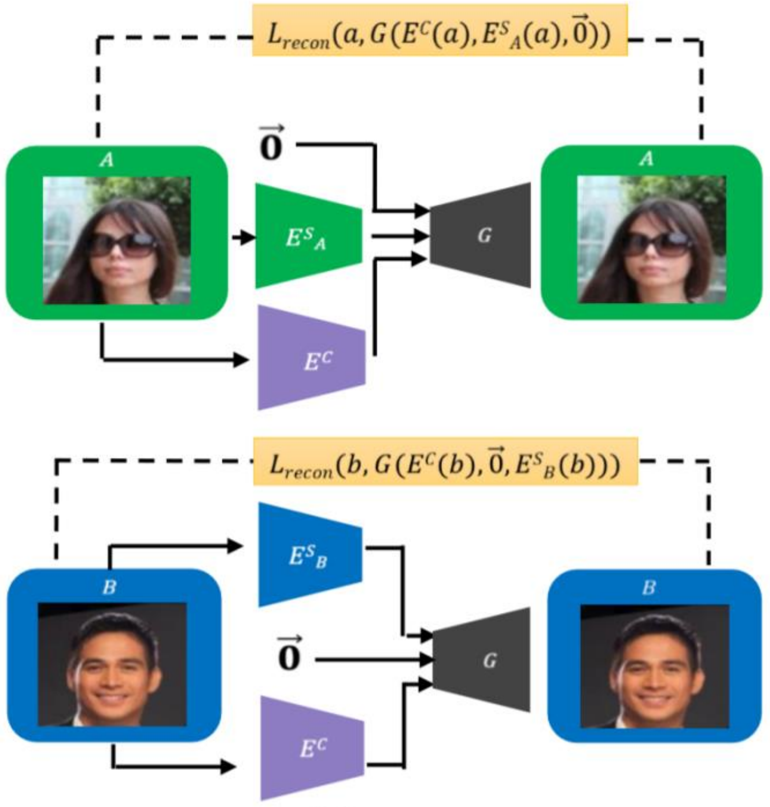
Reconstruction Losses



Ensures the "common" and "separate" encodings contain all the information in A or B

$$\mathcal{L}^{A}_{recon} := \frac{1}{m_1}\sum_{i=1}^{m_1} \|G(E^c(a_i), E^s_A(a_i), 0) - a_i\|_1$$

$$\mathcal{L}^{B}_{recon} := \frac{1}{m_2}\sum_{j=1}^{m_2} \|G(E^c(b_i), 0, E^s_B(b_j)) - b_j\|_1$$



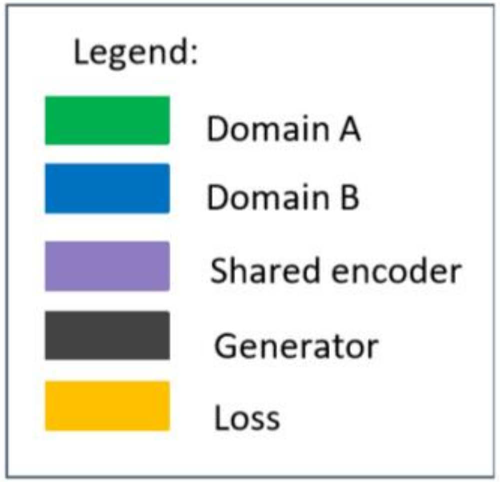
(b)

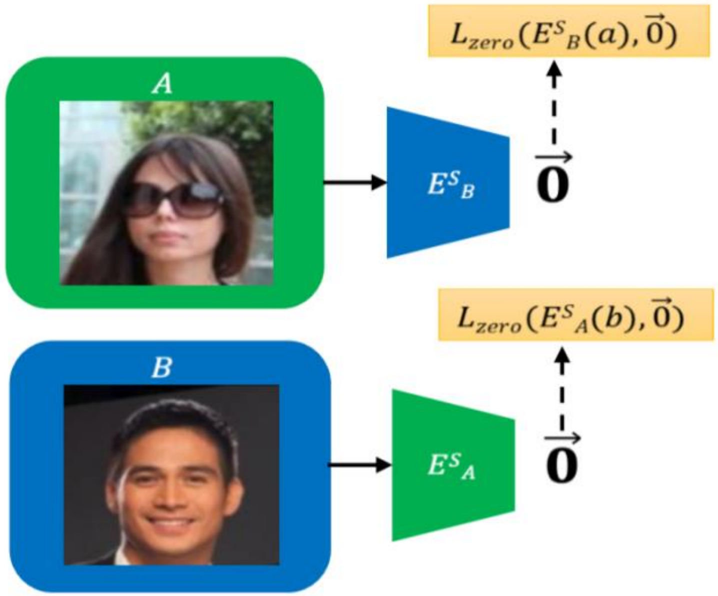
"Zero" Loss

Ensures the separate encoder of A (resp. B) does not encode information about B (resp. A)

$$\mathcal{L}^{A}_{zero} := \frac{1}{m_2}\sum_{j=1}^{m_2} \|E^{s}_{A}(b_j)\|_1$$

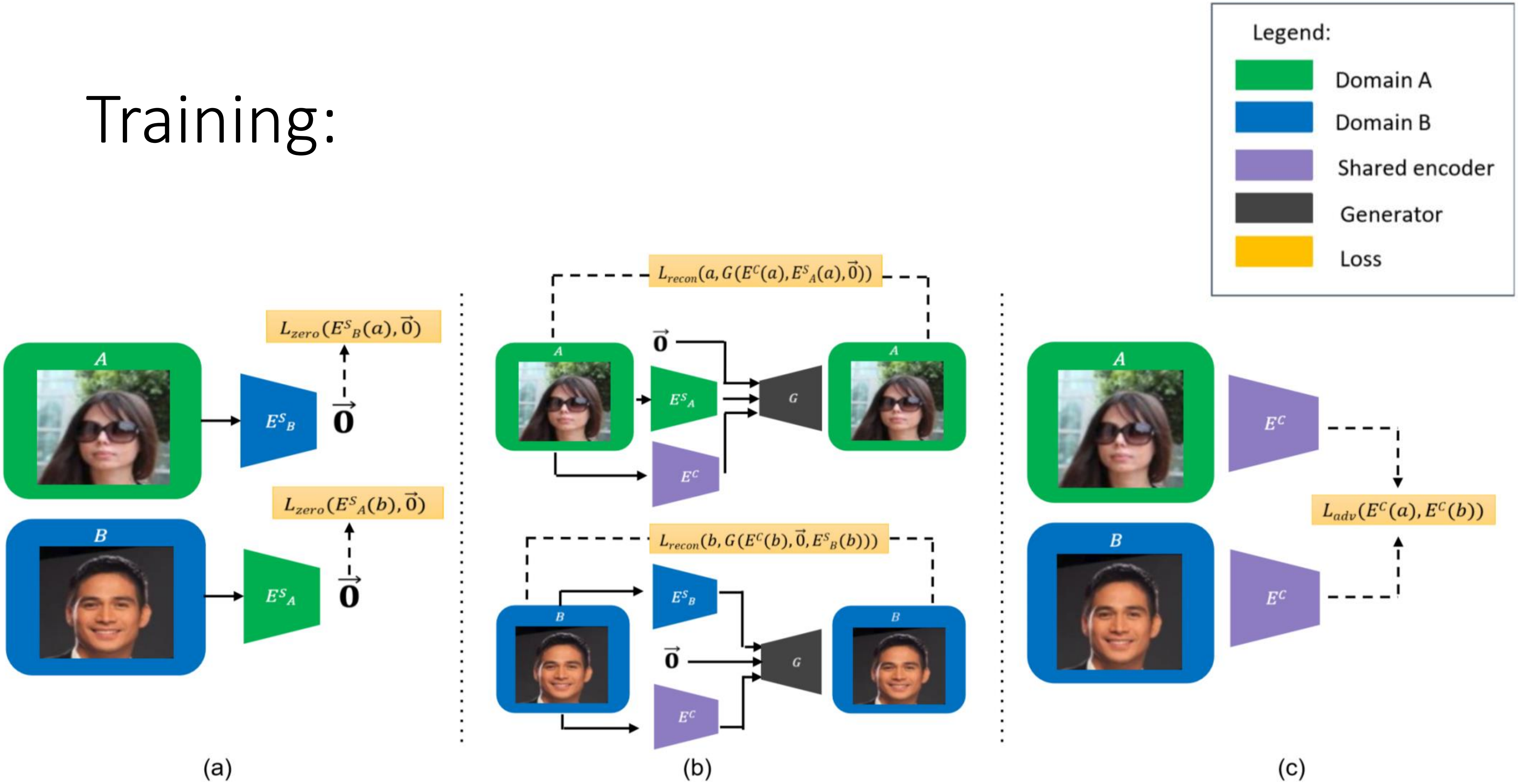
$$\mathcal{L}^{B}_{zero} := \frac{1}{m_1}\sum_{i=1}^{m_1} \|E^{s}_{B}(a_i)\|_1$$





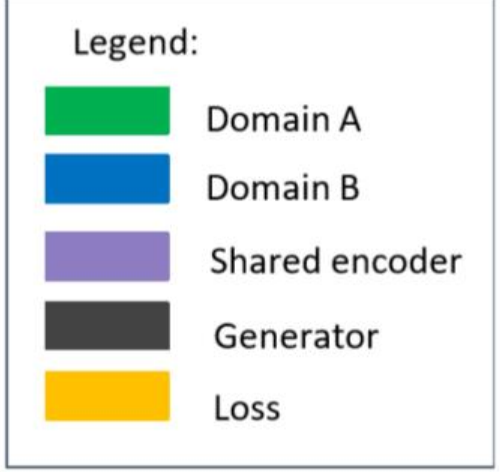
(a)

Training:



Legend:
Domain A
Domain B
Shared encoder
Generator
Loss
$L_{zero}(E^S{}_B(a),\vec{0})$
$L_{zero}(E^S{}_A(b),\vec{0})$
$L_{recon}(a,G(E^C(a),E^S{}_A(a),\vec{0}))$
$L_{recon}(b,G(E^C(b),\vec{0},E^S{}_B(b)))$
$L_{adv}(E^C(a),E^C(b))$
A
B
$E^S{}_B$
$E^S{}_A$
E^C
G
$\vec{0}$
(a)
(b)
(c)

Inference:

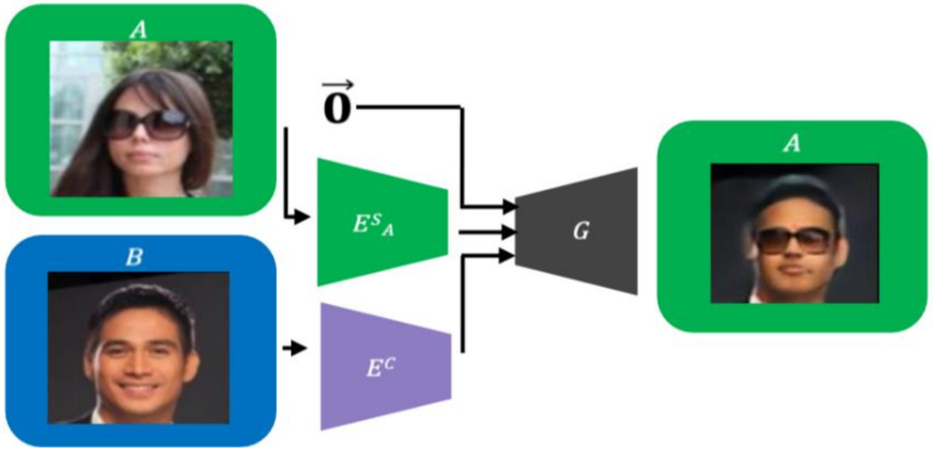


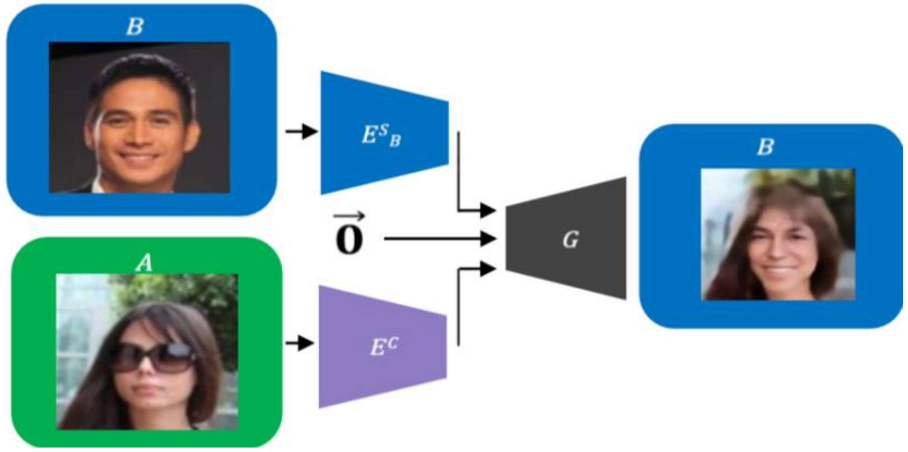
$$G(E_c(b), E_A^s(a), 0)$$

remove b's smile
add a's glasses

$$G(E_c(a), 0, E_A^s(b))$$

remove a's glasses
add b's smile





Results

Beard to Smile

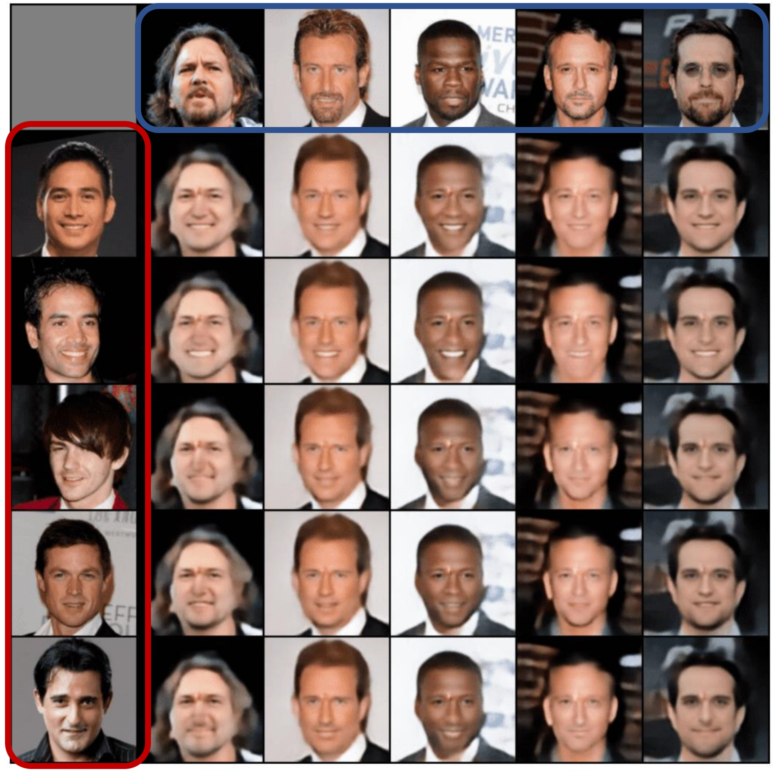


Figure 8. Translating from the domain of persons with facial hai to the domain of smiling persons.

Glasses to Smile

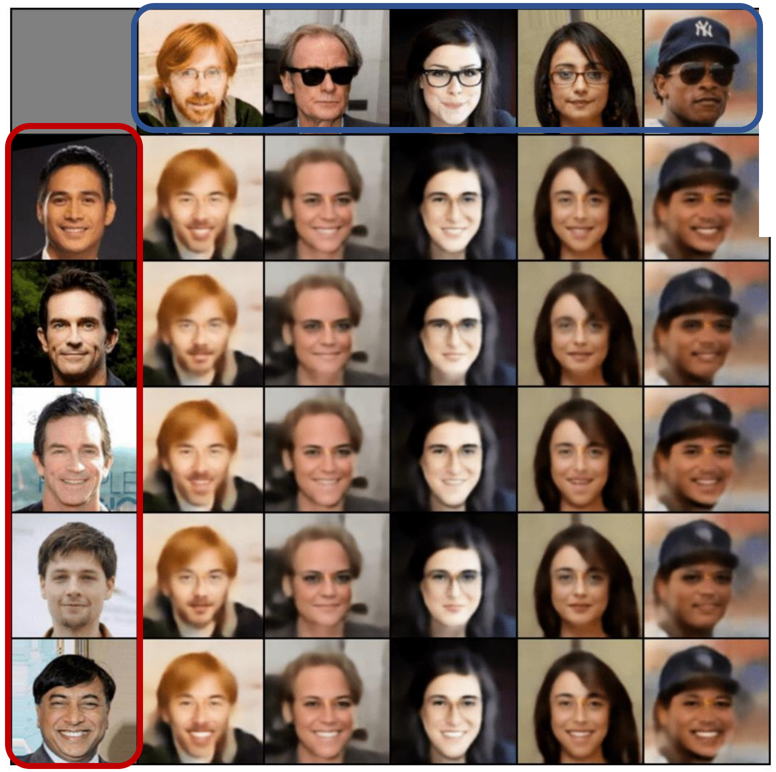
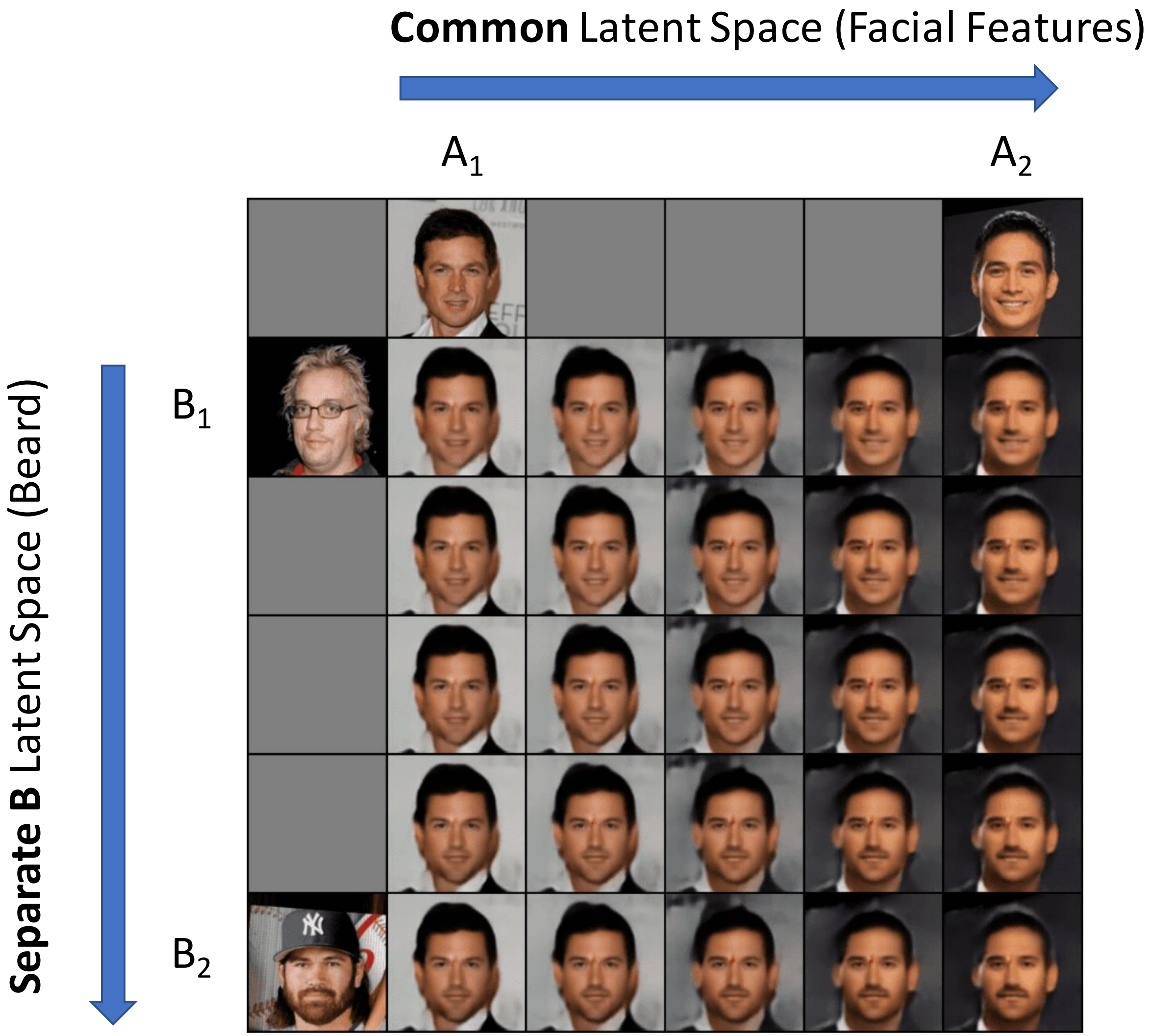


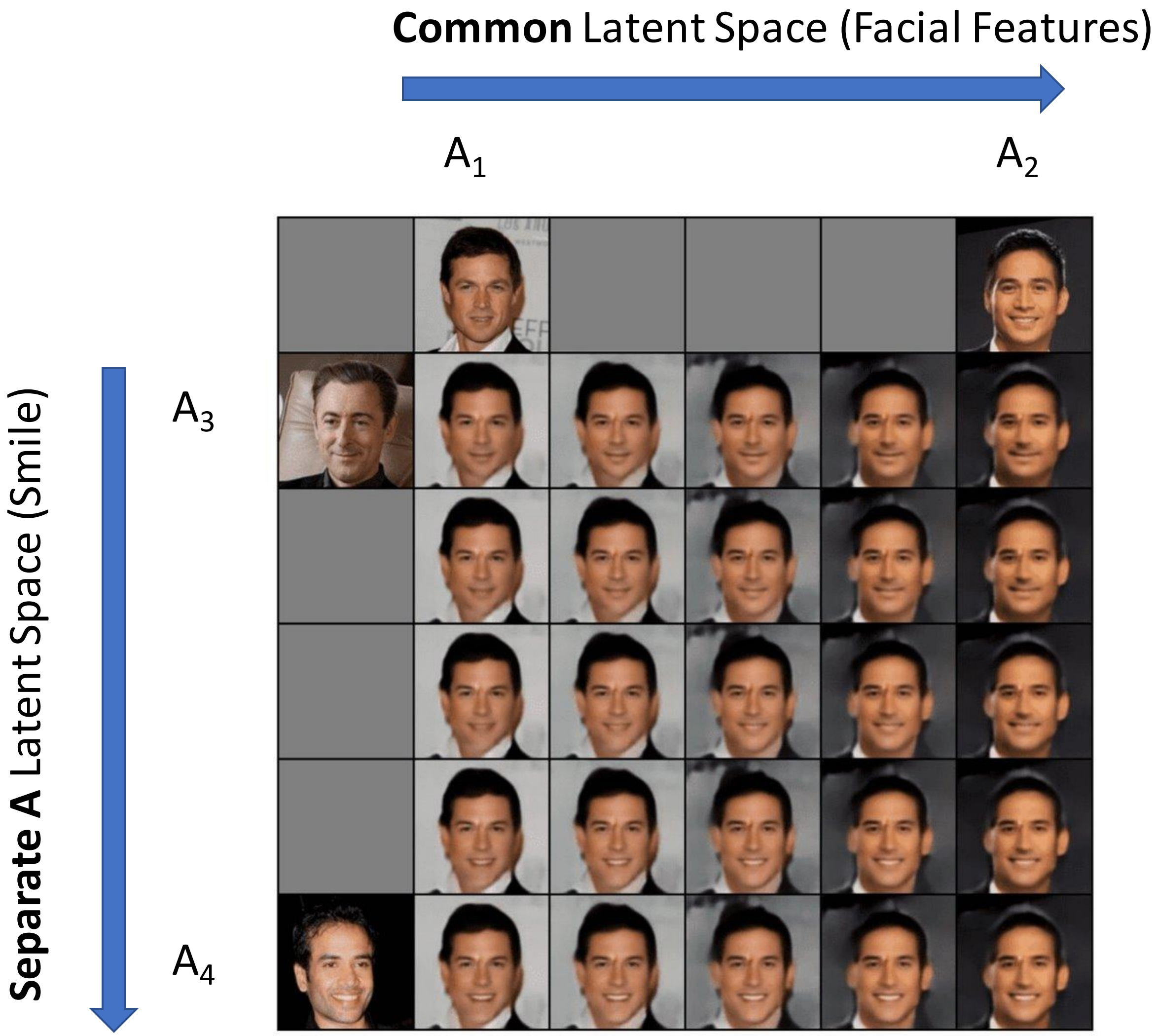
Figure 7. Translating from the domain of persons with glasses to the domain of smiling persons (reverse translation to Fig. 2 in main report)

Glasses ∩ Smile

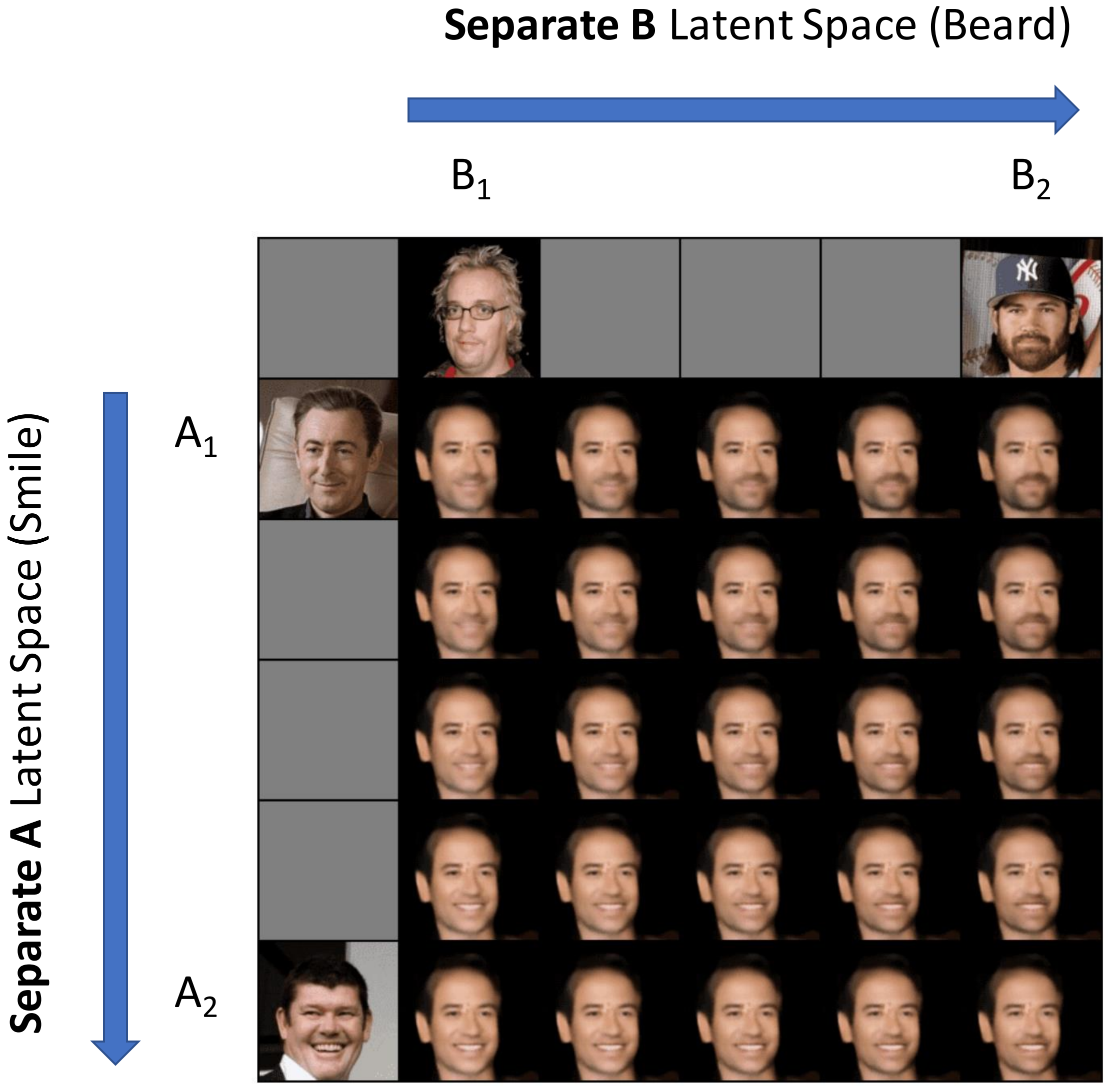
Interpolations



Interpolations



Interpolations



Domain Adaptation

- Our disentangled representation is useful for **Unsupervised** Domain Adaptation: **No labels at all.**
- A pretrained classifier is used to evaluate the percentage of images mapped to the same label in the target domain.
- Given an MNIST digit a, we randomly sample an SVHN digit b and consider the translation to SVHN as $G(E_c(a), 0, E_A^s(b))$.
- Achieve **SOTA:** MNIST to SVHN: 61.0%, Reverse: 41.0%

Theory

Definition 1 (Intersection). *We say that the two representations* $a = g(e^c(a), e^s_A(a), 0)$ *and* $b = g(e^c(b), 0, e^s_B(b))$ *form an intersection between* a *and* b*, if for any other representation* $a = \hat{g}(\hat{e}^c(a), \hat{e}^s_A(a), 0)$ *and* $b = \hat{g}(\hat{e}^c(b), 0, \hat{e}^s_B(b))$*, such that,* $\hat{g}$ *is invertible and* $\hat{e}^c(a) \sim \hat{e}^c(b)$*, we have:* $H(\hat{e}^c(a)) \leq H(e^c(a))$.

Theory

- Under mild assumptions (such as our losses being minimized):
  - $E^c(A)$ and $E^s_A(A)$ are independent (Similarly for B).
  - $E^c(A)$ captures the information underlying $e^c(A)$ (Similarly for B).
  - $E^s_A(A)$ holds the information underlying $e^s_A(A)$ (Similarly for B).
  - I.e. our losses are both **necessary and sufficient** for the desired **disentanglement**.

Masked Based Unsupervised Content Transfer

R. Mokady, S. Benaim, L. Wolf, A. Bermano. ICLR 2020.

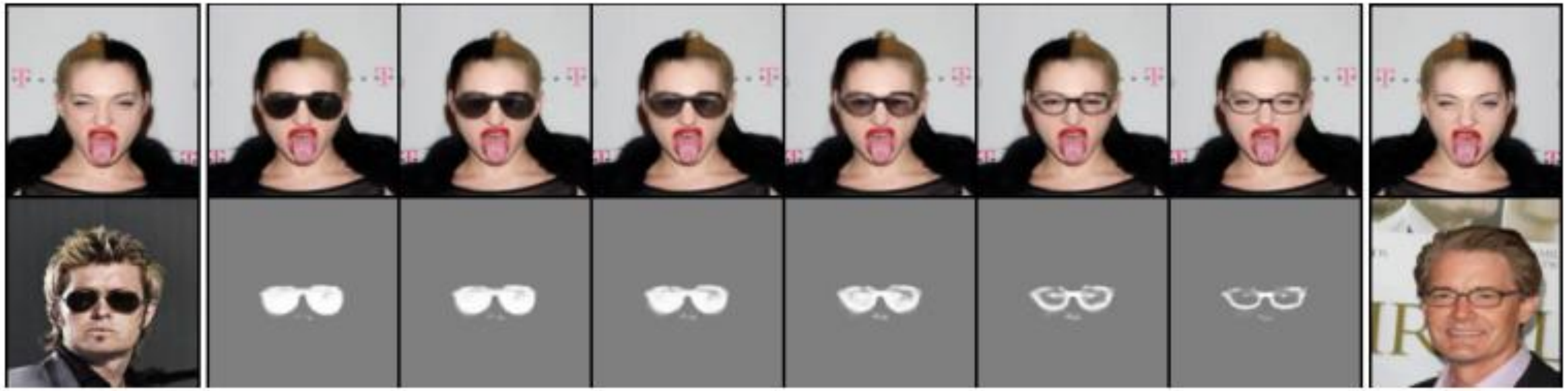
- Only a local change in the target is needed
- Learn a mask and adapt only the area in the masked area

Two Attributes

Smile to Glasses

Additional Content Transfer

Interpolation



Attribute Removal

Figure 6: Attr removal.

Table 6: Attribute removal for the task of Smile, Facial hair and Glasses.

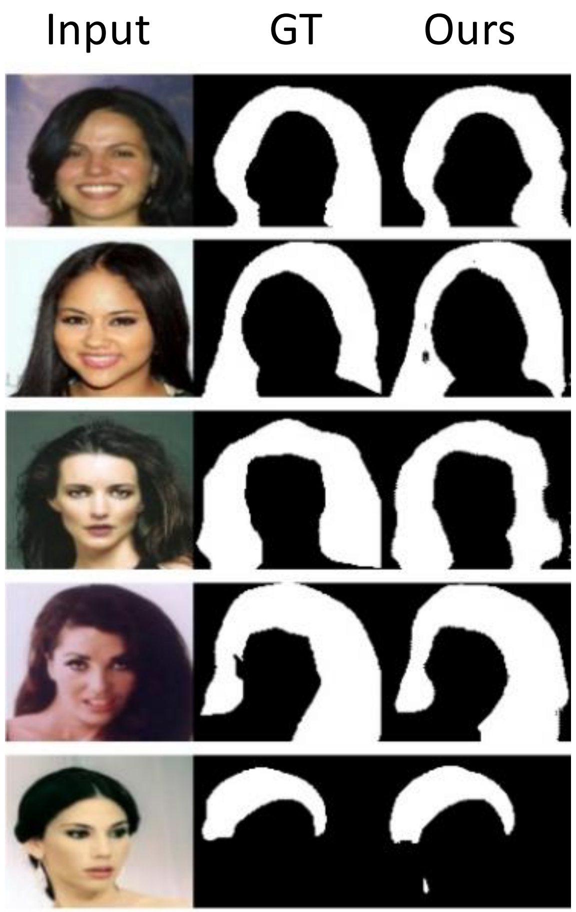
Task	Method	KID	FID	Class.	Sim.
Smile	Ours	2.6 ± 0.4	120.0 ± 2.6	96.9%	0.96
	Press et al.	15.0 ± 0.6	167.7 ± 0.3	96.9%	0.81
	He et al.	4.1 ± 0.4	127.7 ± 4.5	96.9%	0.95
	Liu et al.	4.3 ± 0.3	129.0 ± 3	98.4%	0.92
	Fader	11.3 ± 0.7	155.6 ± 4.7	93.7 %	0.89
Mustache	Ours	1.9 ± 0.5	119.0 ± 0.8	95.3 %	0.95
	Press et al.	16.6 ± 0.8	175.9 ± 1.4	100.0%	0.80
	He et al.	4.6 ± 0.5	130.0 ± 3.0	87.5%	0.96
	Liu et al.	14.0 ± 0.6	160.0 ± 3.3	87.5%	0.85
	Fader	14.1 ± 0.6	162.6 ± 1.5	98.4 %	0.76
Glasses	Ours	5.2 ± 0.5	136.5 ± 2.6	99.2%	0.87
	Press et al.	15.3 ± 0.5	172.0 ± 4.7	100.0%	0.73
	He et al.	8.3 ± 0.9	141.4 ± 6.8	100.0%	0.84
	Liu et al.	6.8 ± 0.3	141.8 ± 4.8	98.4%	0.86
	Fader	12.5 ± 0.3	137.7 ± 4.2	100.0%	0.76

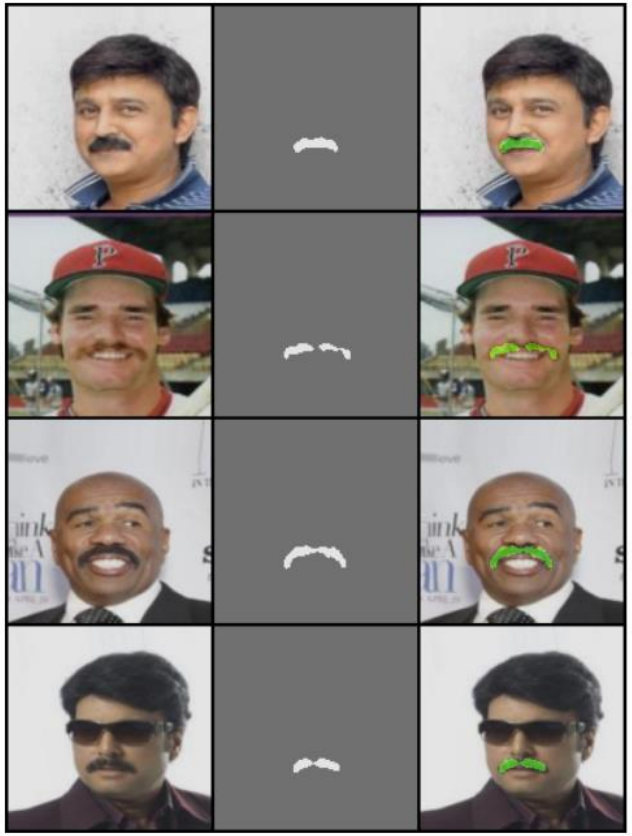
Out of Domain Manipulation

(a) (b)

Figure 23: Out of domain translation. (a) Results on extremely out of domain images. (b) Results obtained by manipulating LFW images.

Semi Supervised Segmentation Using Class Information

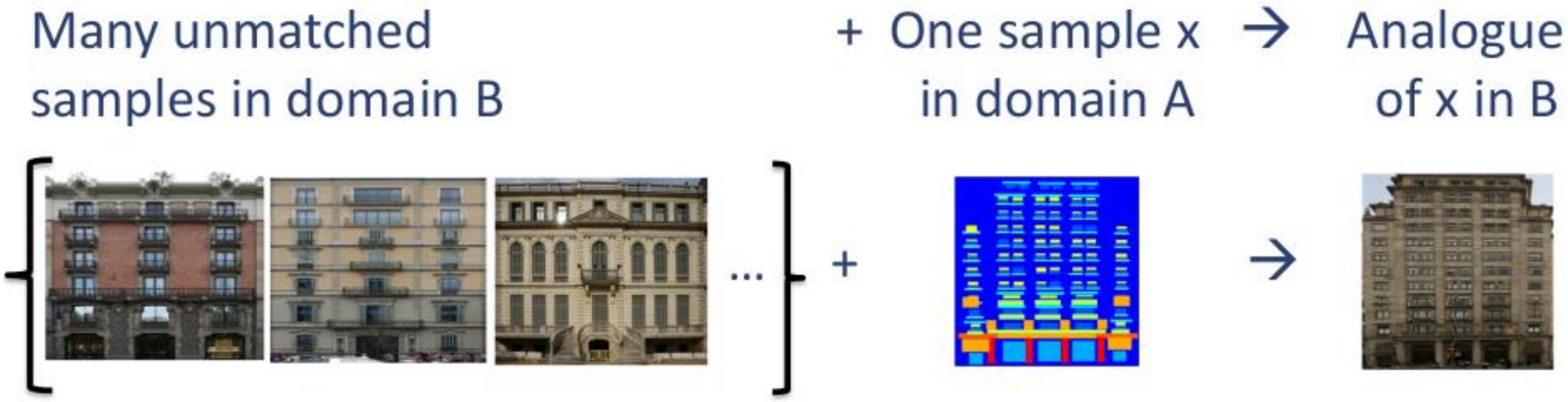




Part II: Generating analogies from few examples

One-shot unsupervised cross domain translation

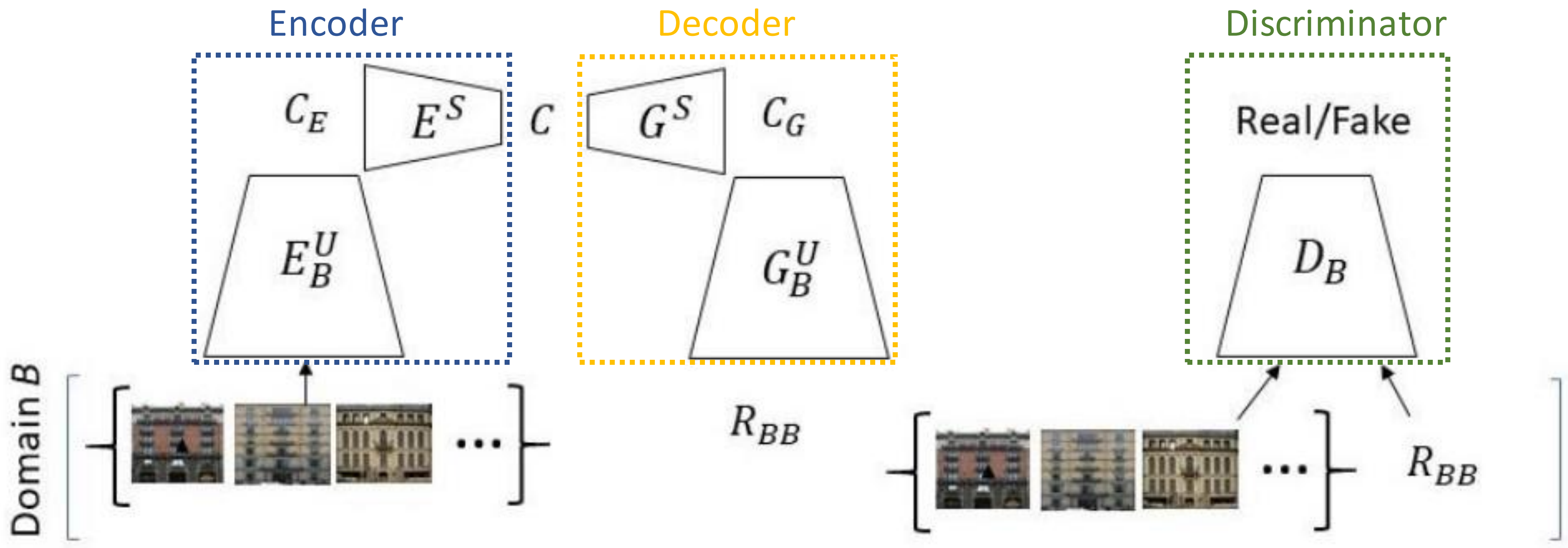
S. Benaim, L. Wolf. NeurIPS 2019.



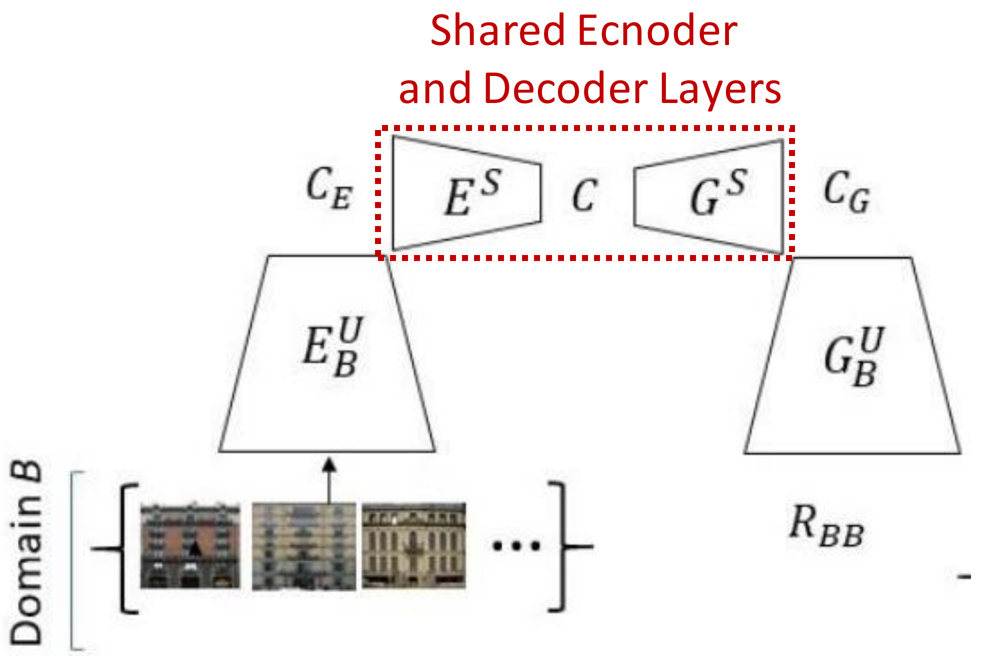
Phase I

For Domain B:

- Train a Variational Autoencoder
- Use a GAN loss to enhance visual quality



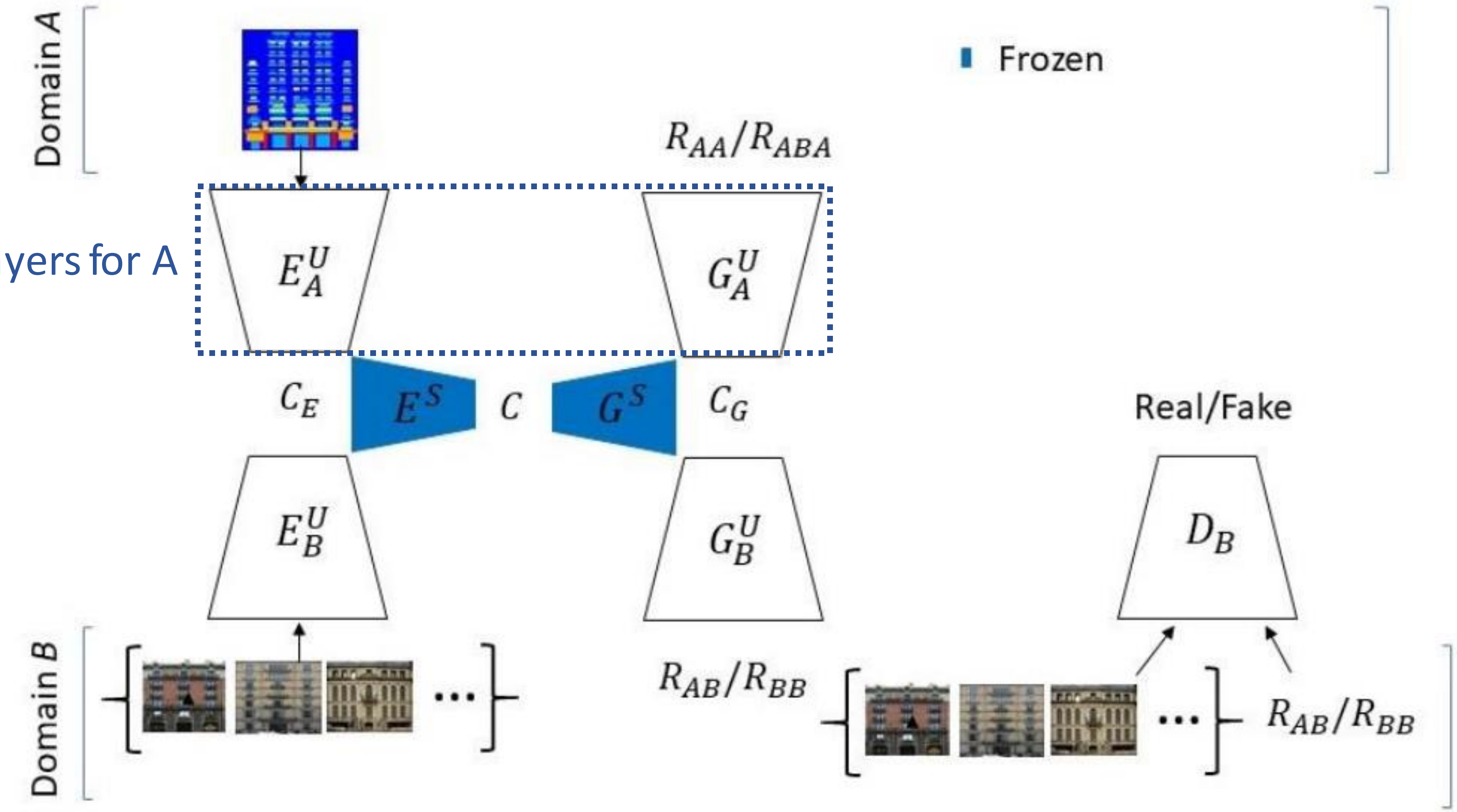
Phase I



- **Shared Latent Space** assumption (UNIT Liu et al, CoGAN Liu et al): Upper layers of the encoder and lower layers of the decoder should be shared to achieve successful translation.
- Shared encoder (Es) and shared decoder (Gs) can be trained with **domain B samples only**

Phase II

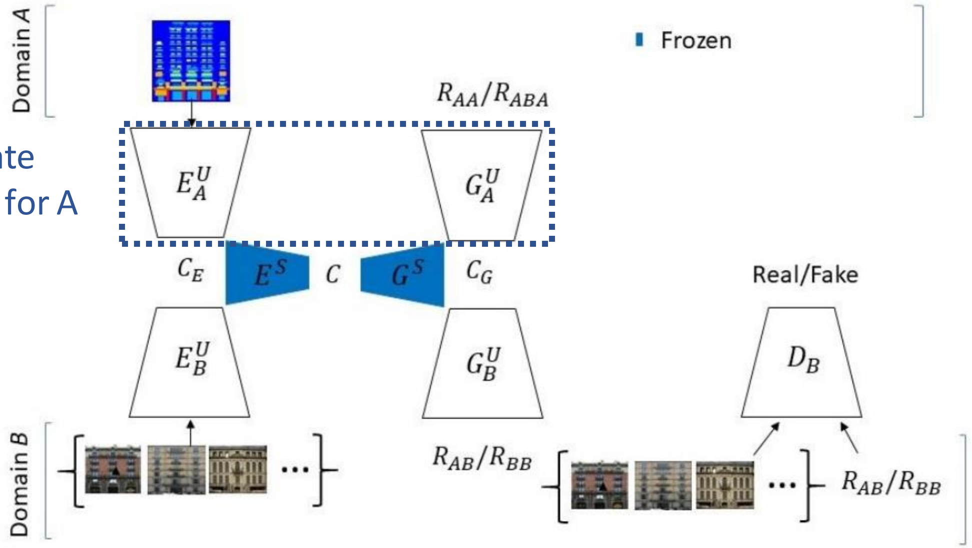
Separate Layers for A



Phase II

1. Reconstruction Loss for A

Separate Layers for A

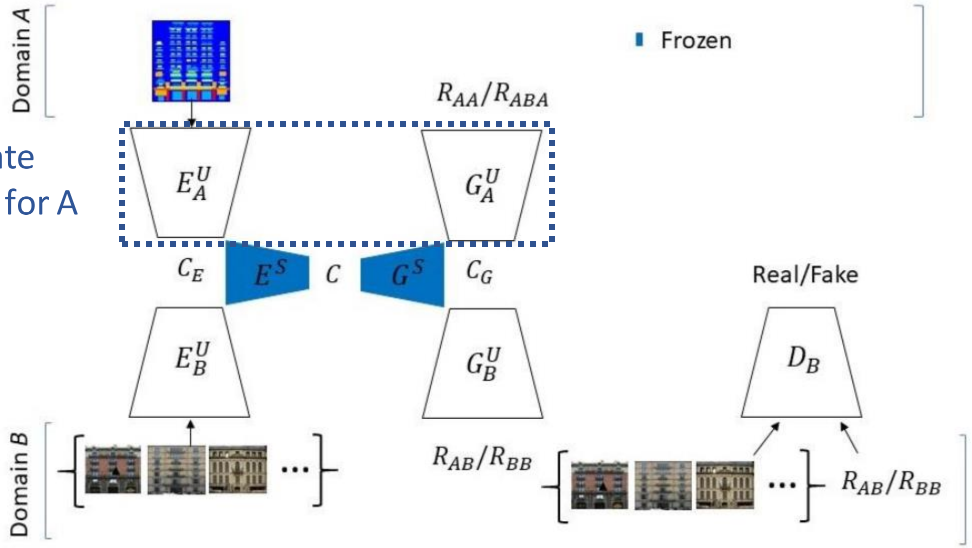


Phase II

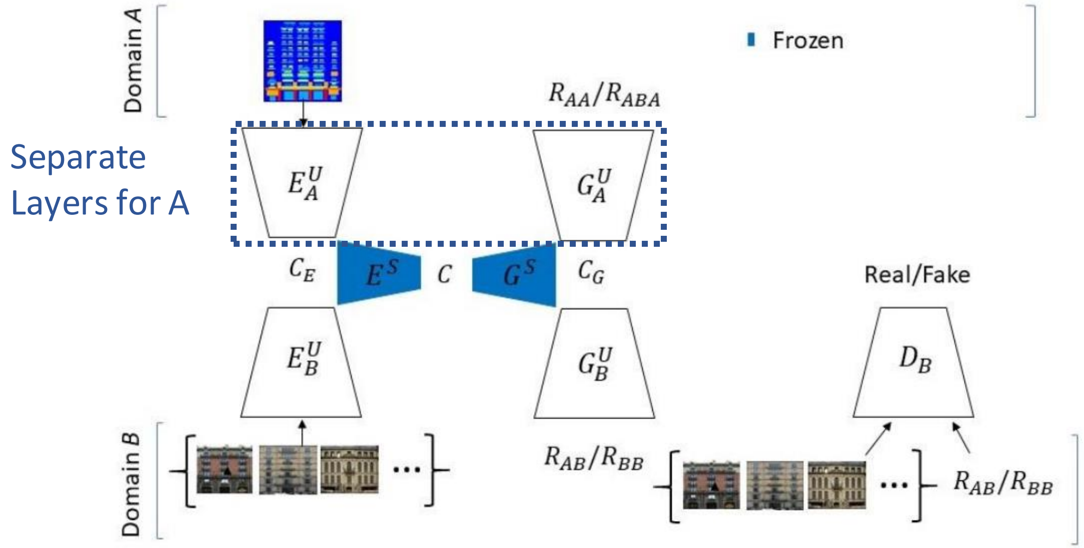
1. Reconstruction Loss for A

2. Cycle Loss for A

Separate Layers for A



Phase II



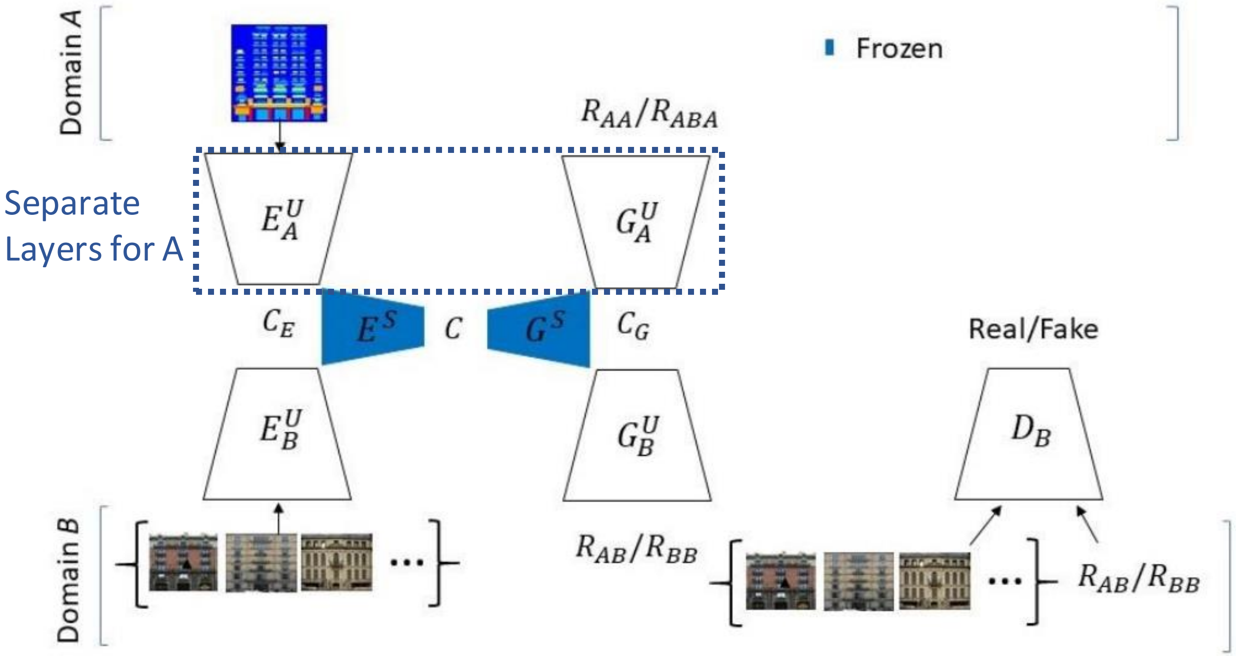
1. Reconstruction Loss for A

2. Cycle Loss for A

3. GAN loss on A --> B

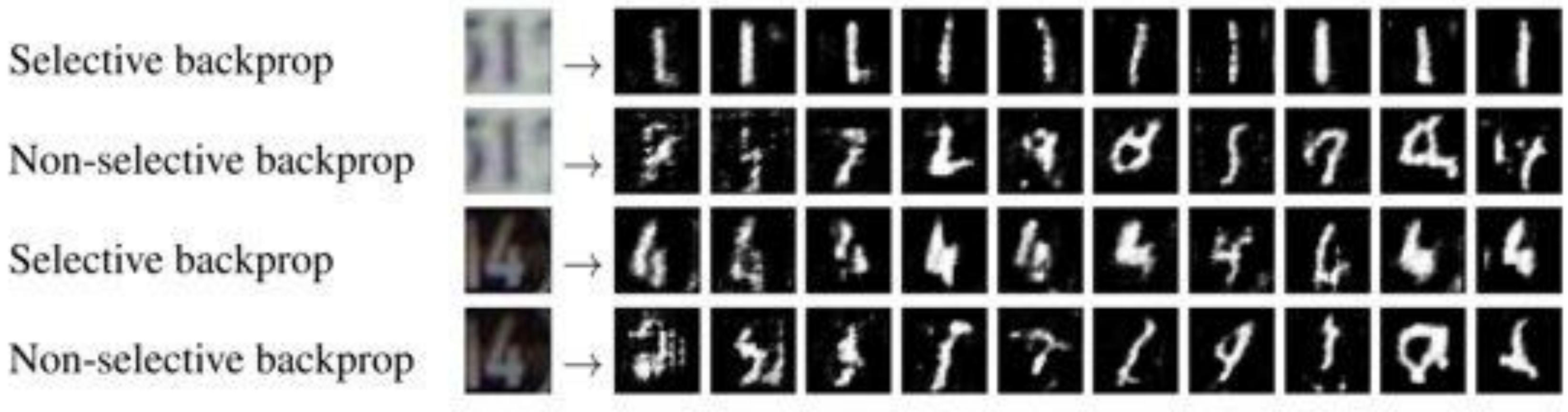
Selective Backpropagation

- Augmentations on A
- Patch discriminator
- Backpropagation is applied **selectively** on the separate encoders and decoders only.
- Similar to Transfer Learning - Finetuning on few layers



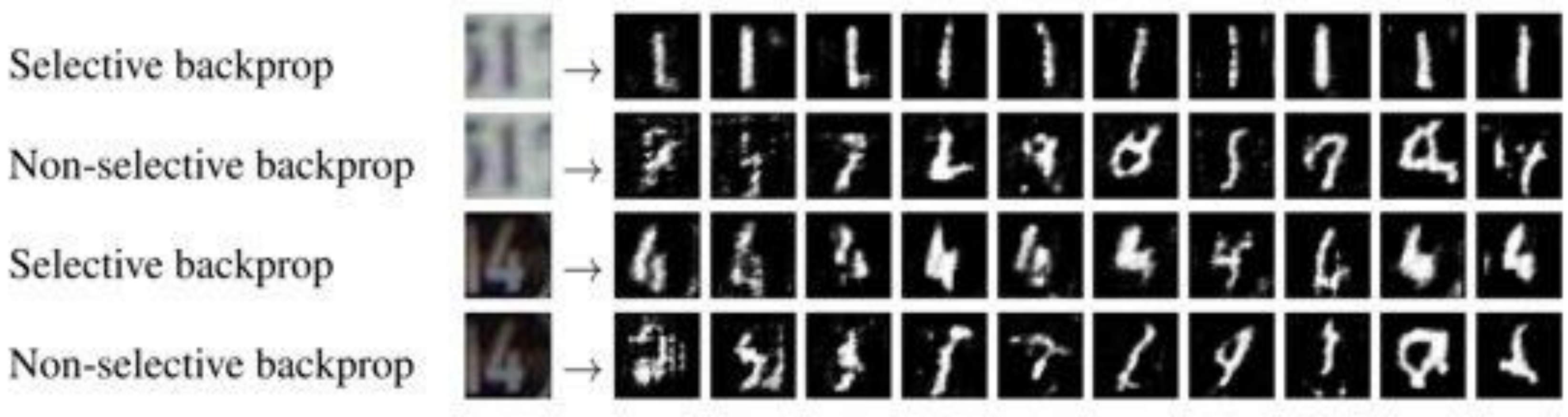
Selective Backpropagation

- Updating the shared encoder (Es) and decoder (Gs) with selective backpropagation turned off leads to **overfitting** on x



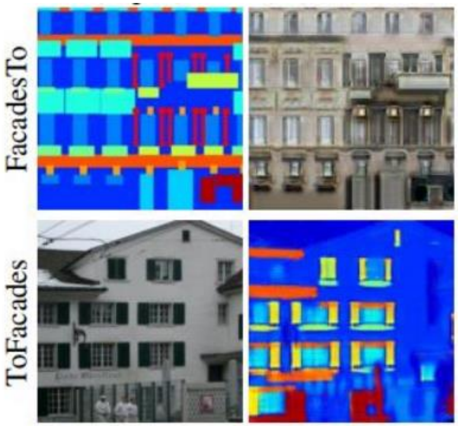
Selective Backpropagation

- Updating the shared encoder (Es) and decoder (Gs) with selective backpropagation turned off leads to **overfitting** on x
- 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.



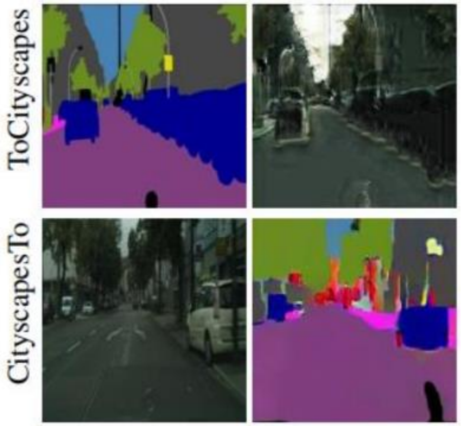
Qualitative Results

Input Output



Input Output

Input Output



Input Output

Input Output

Domain Adaptation

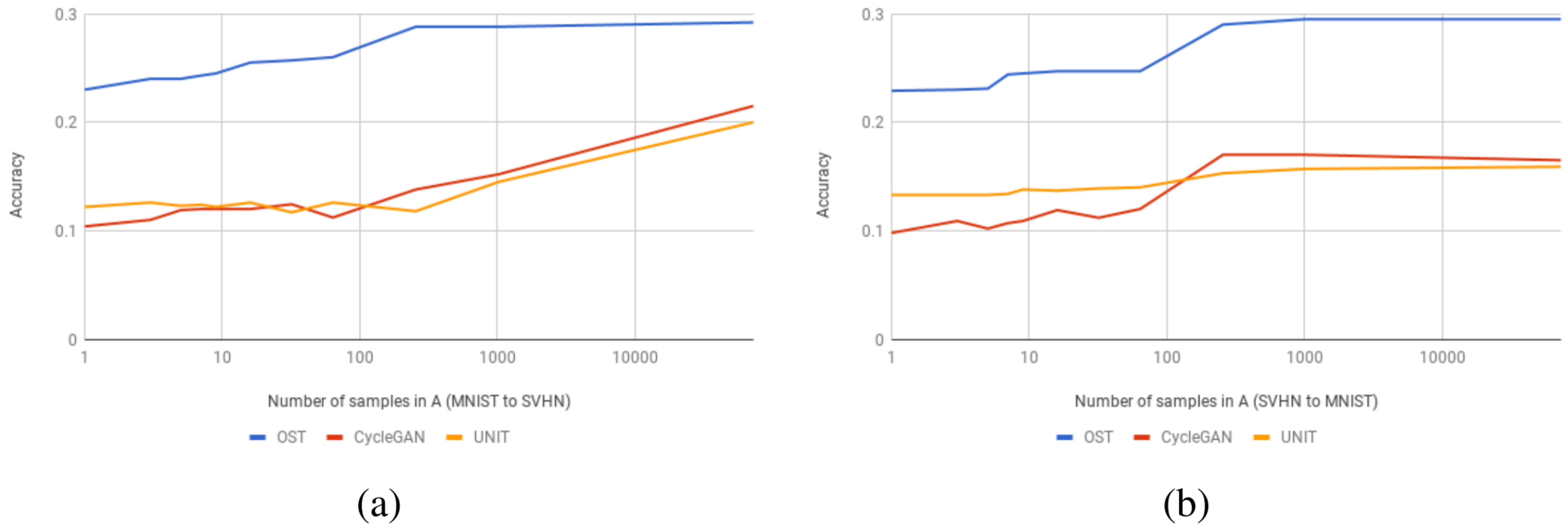


Figure 3: (a) Translating MNIST images to SVHN images. x-axis is the number of samples in A (log-scale), y-axis is the accuracy of a pretrained classifier on the resulting translated images. The accuracy is averaged over 1000 independent runs for different samples. Blue: Our OST method. Yellow: UNIT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

Structural-analogy from a Single Image Pair

S. Benaim, R. Mokady, A. Bermano, D Cohen-Or, L. Wolf. In Submission.

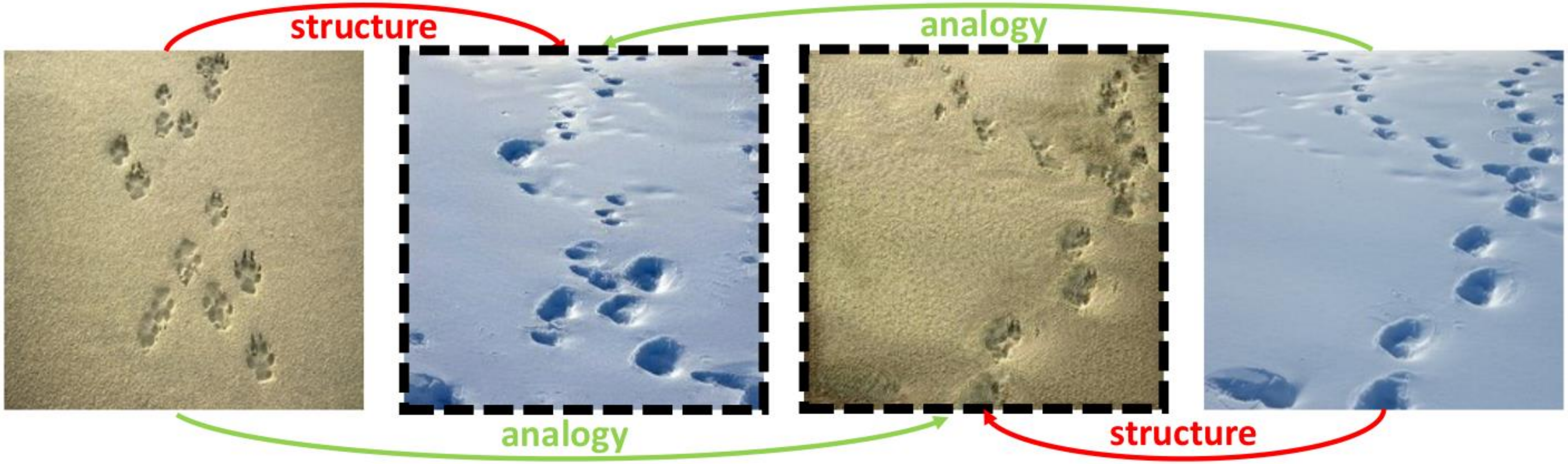


Fig. 1. Our method takes two images as input (left and right), and generates images that consist of features from one image, spatially structured analogically to the other.

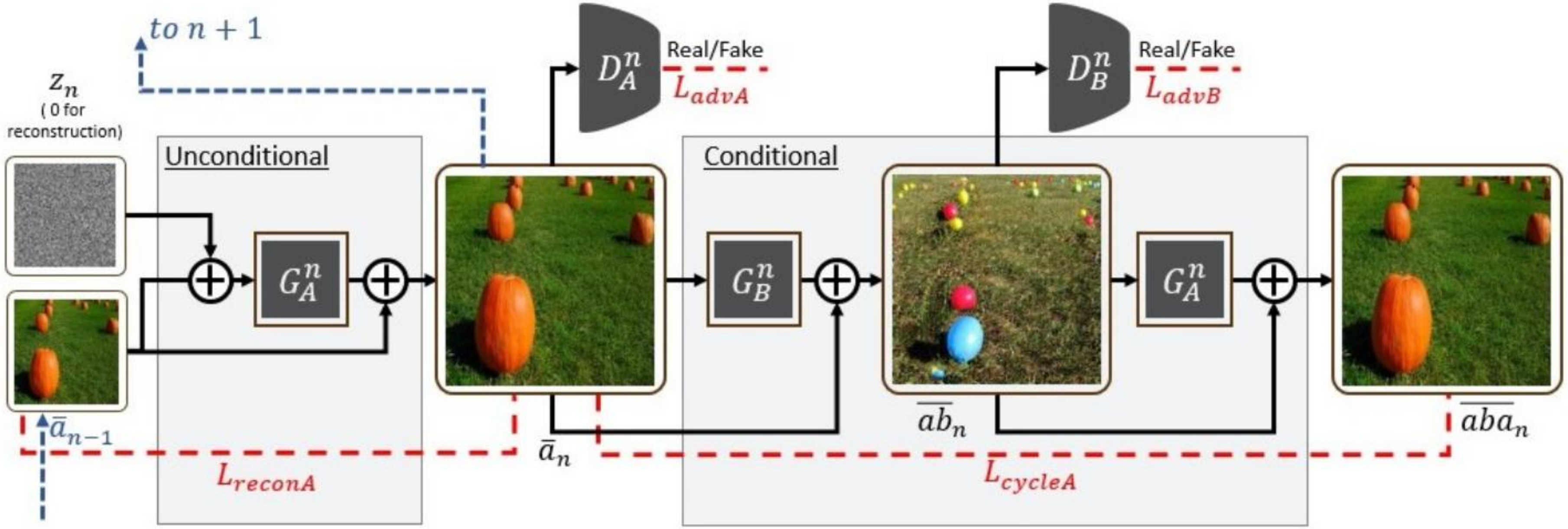
Main Idea

- In classical work (e.g Irani et al.), two visual signals are defined to be similar if **all patches** of one (at multiple scales) **are contained** in the other (completeness), **and vice versa** (coherence).
- Key idea: produce a mapping in which the **patch distribution** of a source image is mapped to its **corresponding patch distribution** of a target image and vice versa.
- When the multi-scale distributions match, in both directions, **completeness and coherence** are guaranteed.

Method

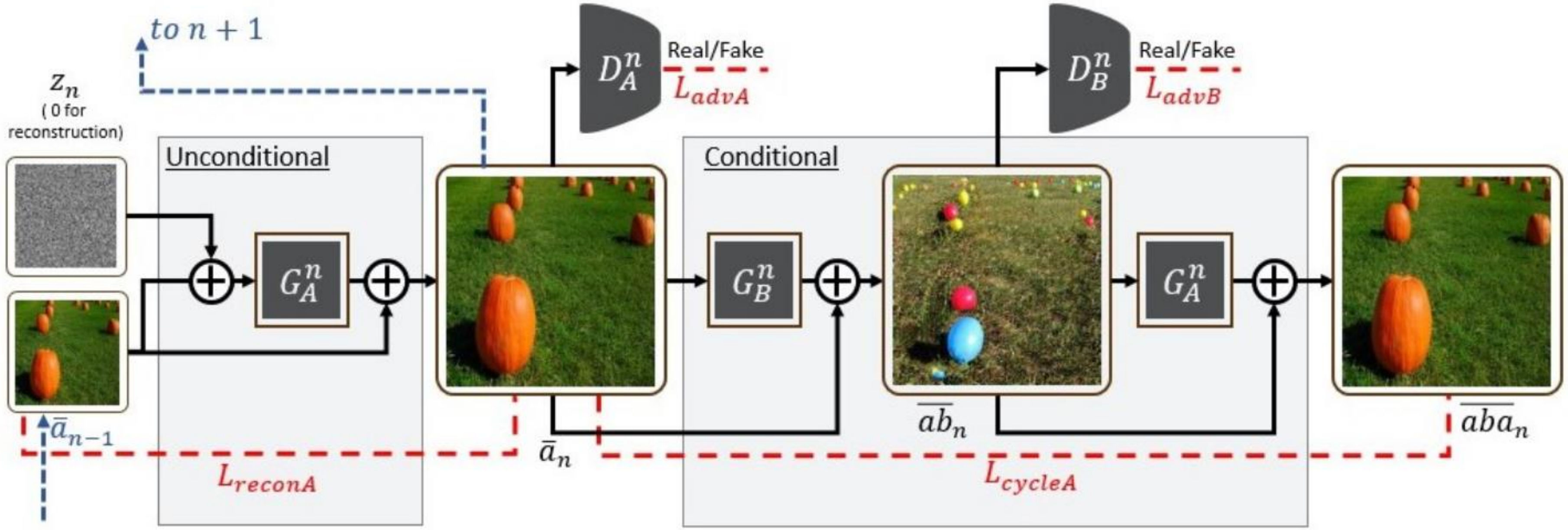
For each scale n:

- Unconditional Generation: Generate many samples of the same patch distribution
- Conditional Generation: Given a sample x, generate an analgeous sample using a conditional generator at scale n



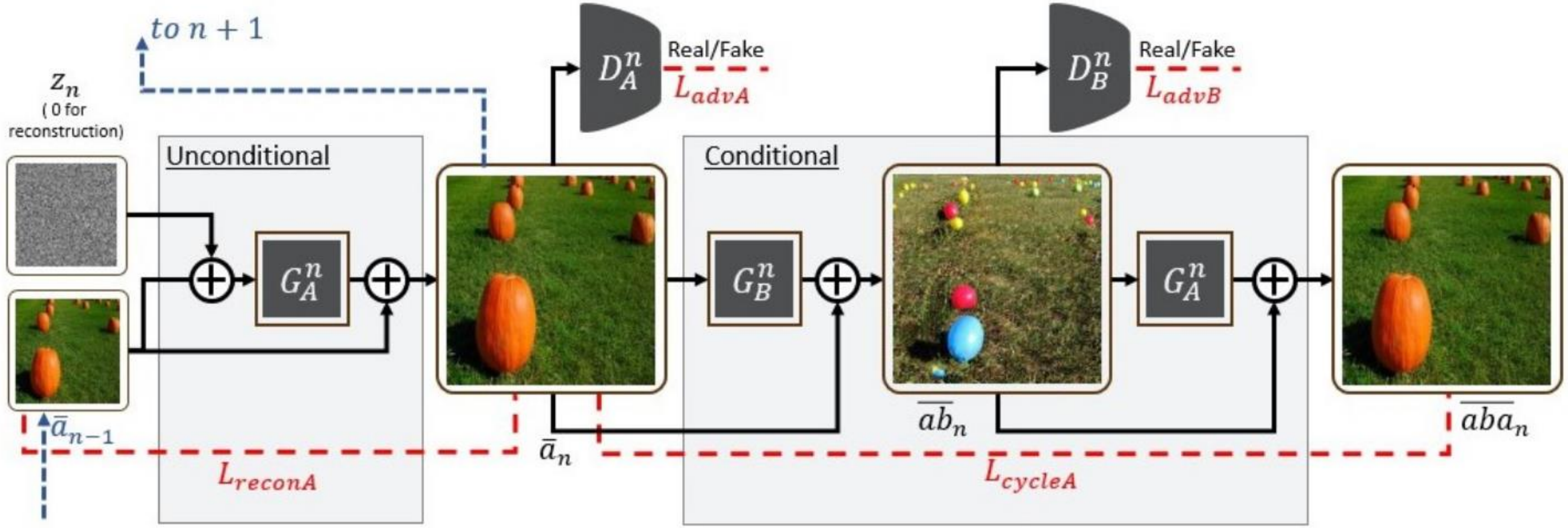
Method

- The same generator acts as both an unconditional generator and a conditional generator (same weights)
- The receptive field of the generator is fixed to 11x11 and the size of the image increases at each scale (level)
- Use of Patch-GAN or patch discriminator, to discriminate based on patches only

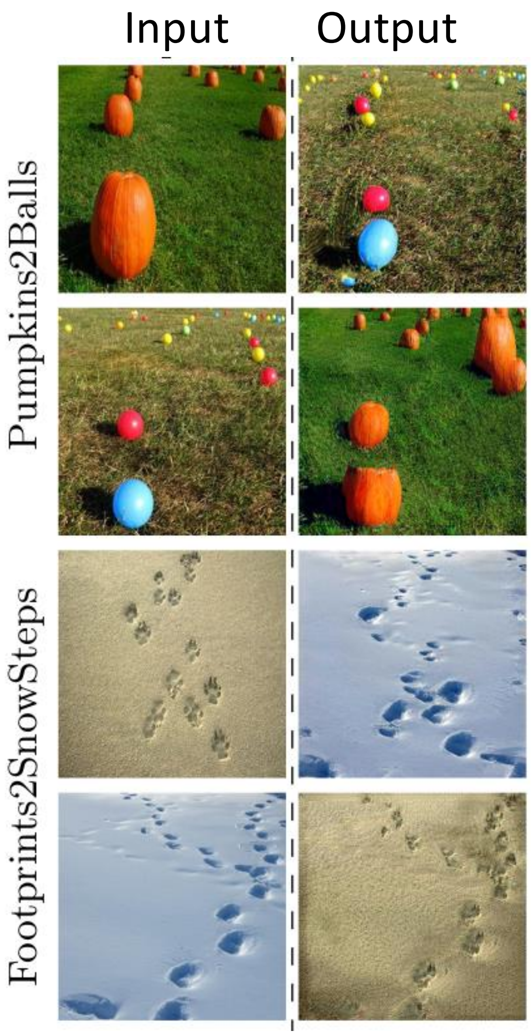


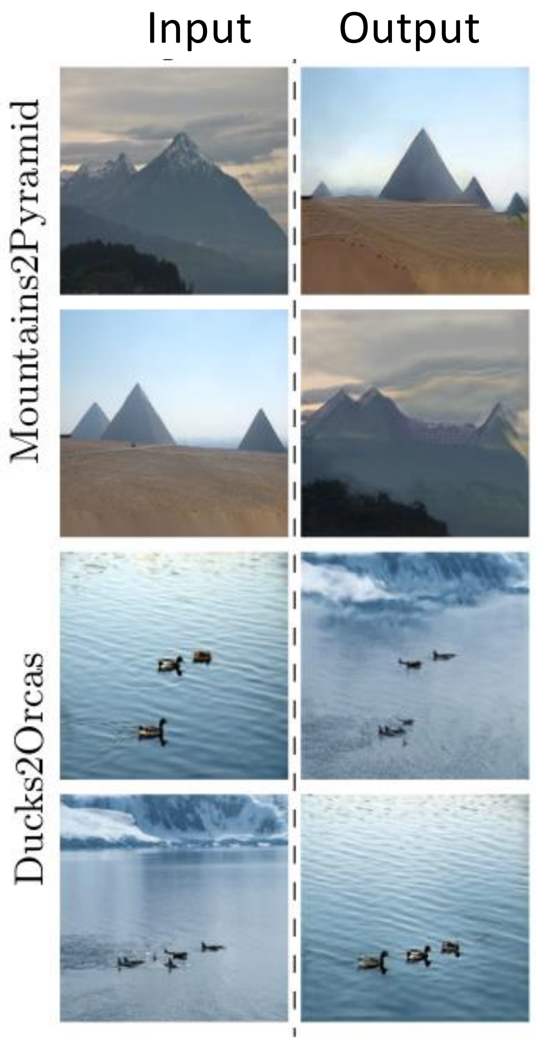
Losses

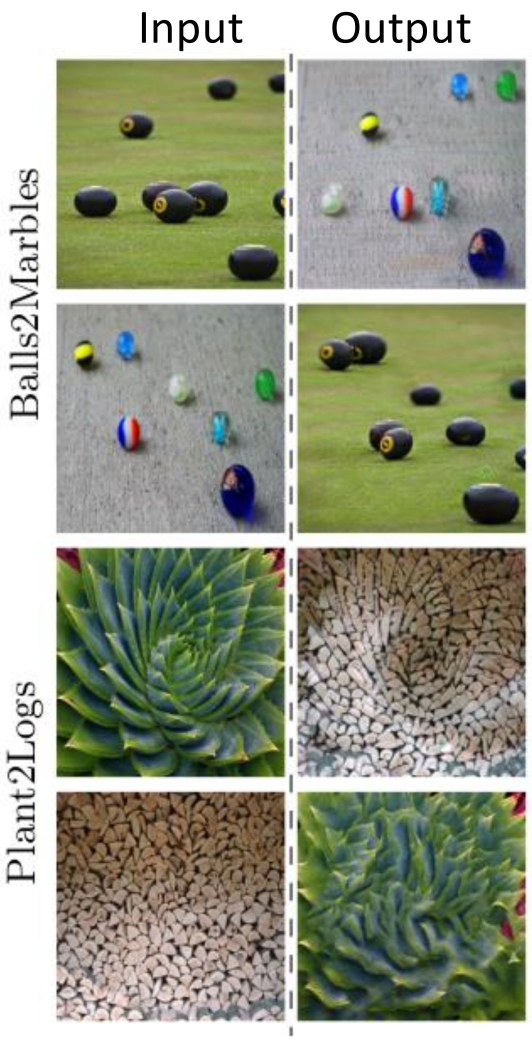
- Adversarial Patch-GAN Loss
- Cycle Loss (Conditional Generation)
- Reconstruction Loss (Unconditional Generation)



Visual Results





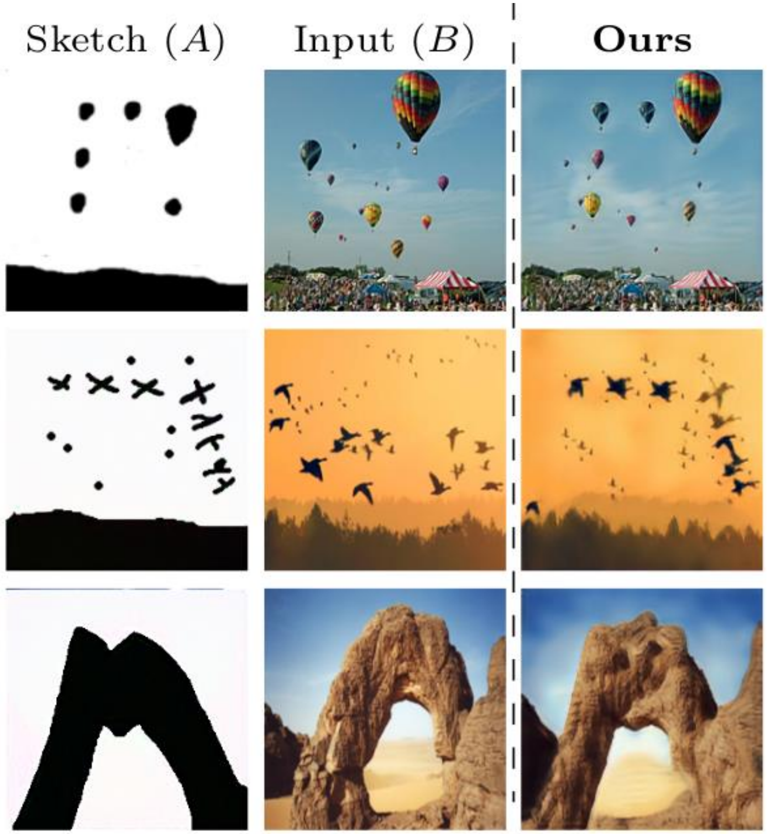


Random Generations

(a) (b)

Fig. 5. (a) Left: Input image $\underline{A}$ (hot air balloons). Right: Randomly generated samples $\overline{a}$ (top) and their translation $\overline{ab}$ (bottom). (b) As in (a) but for image B (birds).

Sketch to Image



Style and Texture

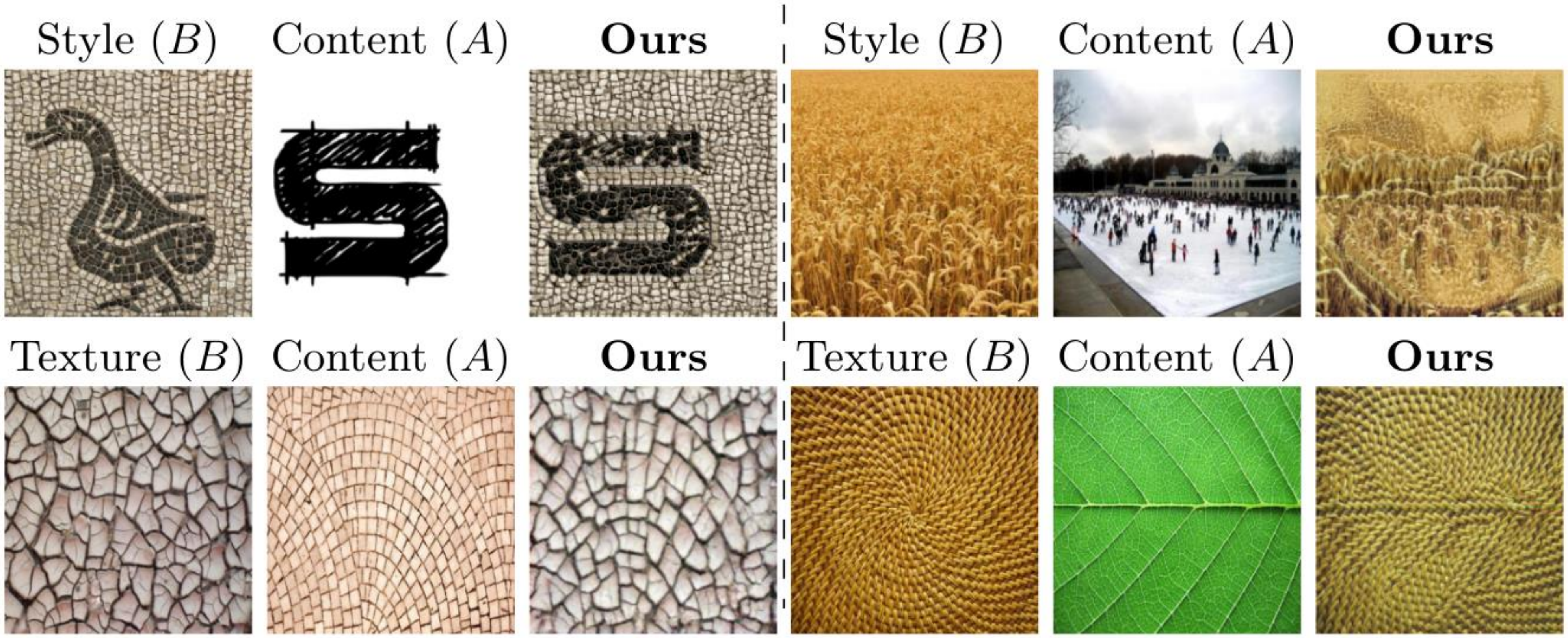
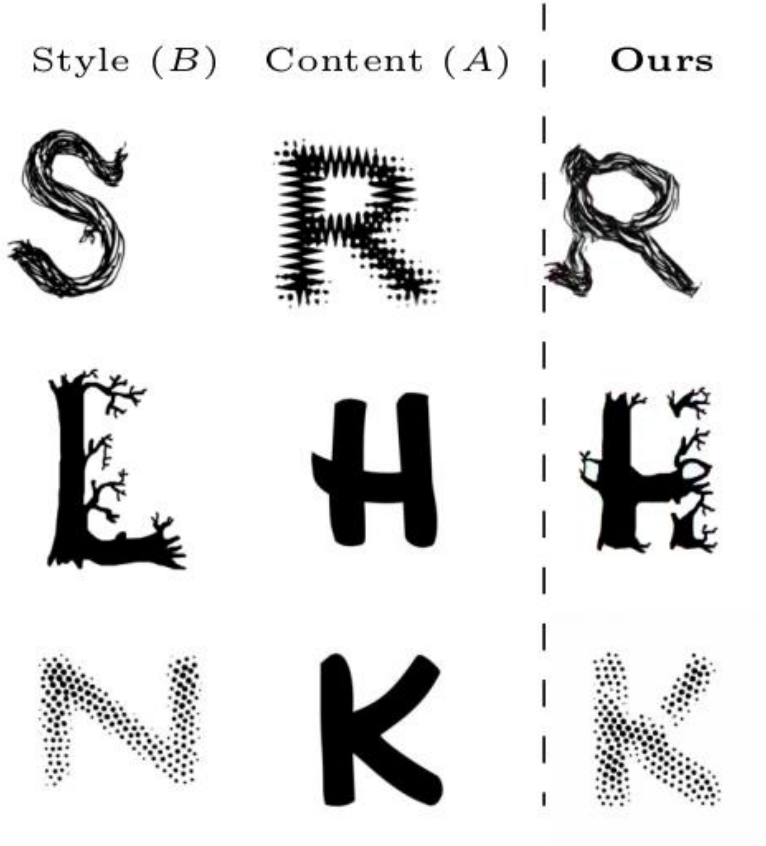


Fig. 7. An illustration of our method for the task of style and texture transfer.

Text Transfer



Videos

Thank You! Questions?

Numerical Results: Pretrained Classifier

	Smile To Glasses	Glasses To Smile	Facial Hair To Smile	Smile To Facial Hair	Facial Hair To Glasses	Glasses To Facial Hair
Fader networks [15]	76.8%	97.3%	95.4%	84.2%	77.8 %	85.2%
Guided content transfer [20]	45.8%	92.7%	85.6%	85.1%	38.6%	82.2%
MUNIT [12]	7.3%	9.2%	9.3%	8.4%	7.3%	8.5%
DRIT [16]	8.5%	6.3%	6.3%	10.3%	8.6%	10.1%
Ours	91.8%	99.3%	93.7%	87.1%	93.1%	97.2%

Table 1. We pretrain a classifier to distinguish between samples in A (e.g. images of persons with glasses) and samples in B (e.g. images of persons with smile). We then sample $a \in A$, $b \in B$ from the test samples and check the membership of the generated image $G(E^c(b), E^s_A(a), 0))$ in A. Similarly, in the reverse direction, we check the membership of $G(E^c(a), 0, E^s_B(b))$ in B.

Numerical Results: User Study

- Q1: Is the specific attribute of A (e.g smile) removed?
- Q2: Is the guided image b specific attribute (e.g glasses) added?
- Q3: Is the identify of a's image preserved?

	Smile To Glasses	Glasses To Smile	Facial Hair To Smile	Smile To Facial Hair	Facial Hair To Glasses	Glasses To Facial Hair
Question (1) ours	4.74 ±0.13	4.30 ±0.21	4.26 ±0.20	4.30 ±0.15	4.18 ±0.17	4.50 ±0.18
Question (2) ours	3.92 ±0.16	4.45 ±0.12	4.03 ±0.15	3.34 ±0.17	3.85 ±0.20	3.95 ±0.22
Question (3) ours	3.95 ±0.23	3.20 ±0.24	3.24 ±0.25	3.22 ±0.27	3.49 ±0.22	3.39 ±0.23
Question (1) for [20]	3.67 ±0.17	4.16 ±0.18	3.39 ±0.19	3.34 ±0.13	4.24 ±0.12	3.15 ±0.15
Question (2) for [20]	1.87 ±0.35	4.42 ±0.22	3.00 ±0.32	2.67 ±0.33	2.20 ±0.42	3.30 ±0.22
Question (3) for [20]	3.95 ±0.15	2.93 ±0.22	3.37 ±0.25	3.40 ±0.27	3.43 ±0.28	3.75 ±0.20

Table 2. Given 20 randomly selected images $a \in A$ and $b \in B$, we consider the generated image $G(E^c(a), 0, E^s_B(b)))$ and ask if (1) a's separate part is removed (2) b's separate part is added (3) a's common part is preserved (similarly in the reverse direction). Mean opinion scores in the range of 1 to 5 are reported, where higher is better.

Minimality

- Potentially Infinitely many solutions preserving distance correlations

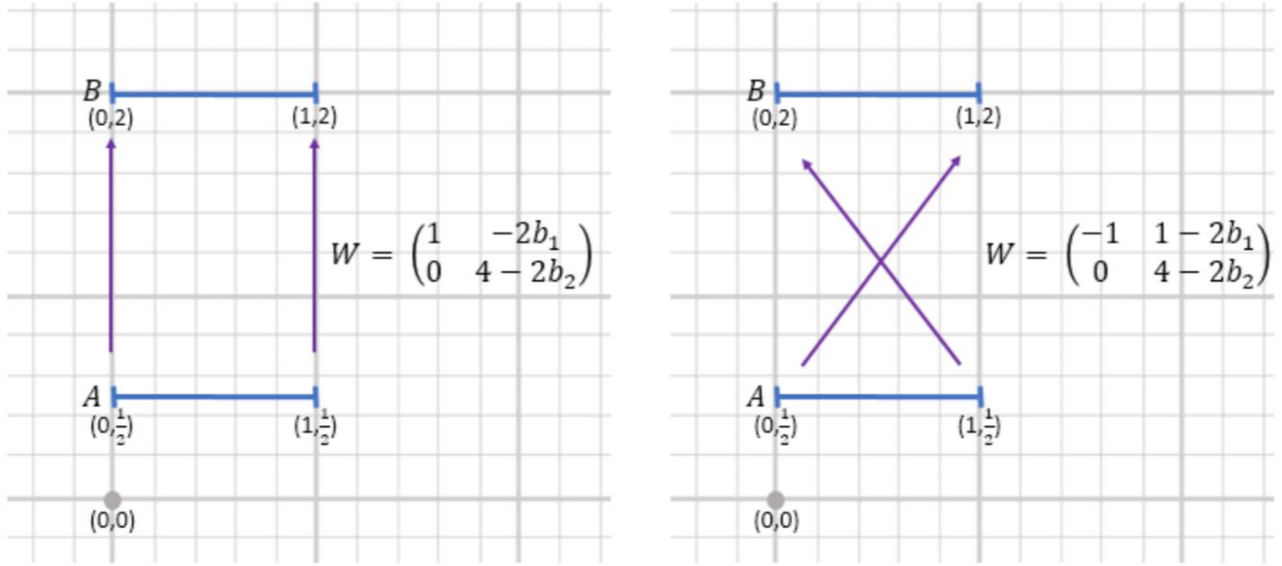


Figure 1: An illustrative example where the two domains are line segments in $\mathbb{R}^2$. There are infinitely many mappings that preserve the uniform distribution on the two segments. However, only two stand out as "semantic". These are exactly the two mappings that can be captured by a neural network 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.

Quantitative Results

Table 1: Ablation study for the MNIST to SVHN translation (and vice versa). We consider the contribution of various parts of our method on the accuracy. Translation is done for one sample.

Augment-ation	One-way cycle	Selective backprop	Accuracy (MNIST to SVHN)	Accuracy (SVHN to MNIST)
False	False	False	0.07	0.10
True	False	False	0.11	0.11
False	True	False	0.13	0.13
True	True	False	0.14	0.14
False	False	True	0.19	0.20
True	False	True	0.20	0.20
False	True	True	0.22	0.23
True	True	No Phase II update of E^S and G^S	0.16	0.15
True	Two-way cycle	True	0.20	0.13
True	Two-way cycle	False	0.11	0.12
True	True	True	**0.23**	**0.23**

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 Samples in A	OST 1	UNIT [7] 1	CycleGAN [2] 1	UNIT [7] All	CycleGAN [2] 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

Quantitative Results

Table 3: (i) Perceptual distance [29] between the inputs and corresponding output images, for various drawing tasks. (ii) Style difference between translated images and images from the target domain. (iii) Correctness of translation as evaluated by a user study.

	Method	Images to Facades	Facades to Images	Images To Maps	Maps to Images	Labels to Cityscapes	Cityscapes to Labels
(i)	OST 1	4.76	5.05	2.49	2.36	3.34	2.39
	UNIT [7] All	3.85	4.80	2.42	2.30	2.61	2.18
	CycleGAN [2] All	3.79	4.49	2.49	2.11	2.73	2.28
(ii)	OST 1	3.57	7.88	2.24	1.50	0.67	1.13
	UNIT [7] All	3.92	7.42	2.56	1.59	0.69	1.21
	CycleGAN [2] All	3.81	7.03	2.33	1.30	0.77	1.22
(iii)	OST 1	91%	90%	83%	67%	66%	56%
	UNIT [7] ALL	86%	83%	81%	75%	63%	37%
	CycleGAN [2] ALL	93%	84%	97%	81%	72%	45%