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

Sagie Benaim
School of Computer Science, 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

Huang et al., ECCV 2018
Cannot transfer content

Press et al, ICLR 2019
Attribute Transfer

Liu et al, CVPR 2019
Add glasses

Transfer specific glasses

Target | Result
--- | ---

Source | Target | Result
--- | --- | ---

X

✓
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 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, $E^s_A(A)$. The **space of glasses**. $E^s_A(a)$ is the **glasses of a**.
3. "Separate" latent space for domain B, $E^s_B(B)$. The **space of smiles**. $E^s_B(b)$ is the **smile of 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:

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

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

\[ L_{\text{recon}}^A := \frac{1}{m_1} \sum_{i=1}^{m_1} \| G(E^c(a_i), E^s_A(a_i), 0) - a_i \|_1 \]

\[ L_{\text{recon}}^B := \frac{1}{m_2} \sum_{j=1}^{m_2} \| G(E^c(b_i), 0, E^s_B(b_j)) - b_j \|_1 \]
"Zero" Loss

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

\[
\mathcal{L}^A_{\text{zero}} := \frac{1}{m_2} \sum_{j=1}^{m_2} \| E^s_A(b_j) \|_1
\]

\[
\mathcal{L}^B_{\text{zero}} := \frac{1}{m_1} \sum_{i=1}^{m_1} \| E^s_B(a_i) \|_1
\]
Training:
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

Glasses to Smile

Glasses \cap 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)

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

Common Latent Space (Facial Features)

A<sub>1</sub> → A<sub>2</sub>

Separate B Latent Space (Beard)

B<sub>1</sub> → B<sub>2</sub>
Interpolations

**Common Latent Space (Facial Features)**

A<sub>1</sub>  A<sub>2</sub>

**Separate A Latent Space (Smile)**

A<sub>3</sub>  A<sub>4</sub>
Interpolations

Separate A Latent Space (Smile)

Separate B Latent Space (Beard)
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%
**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


- 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

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

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

Figure 6: Attr removal.
Out of Domain Manipulation

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


Many unmatched samples in domain B

+ One sample $x$ in domain A

$\rightarrow$ Analogue of $x$ in B
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
Phase II

1. Reconstruction Loss for A

2. Cycle Loss 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
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

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 analogous 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

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
<th>Input</th>
<th>Output</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pumpkins2Balls</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Footprints2SnowSteps</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ducks2Orcas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mountains2Pyramid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balls2Marbles</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plant2Logs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fig. 5. (a) Left: Input image $A$ (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

<table>
<thead>
<tr>
<th>Style (B)</th>
<th>Content (A)</th>
<th>Ours</th>
<th>Style (B)</th>
<th>Content (A)</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
<tr>
<td>Texture (B)</td>
<td>Content (A)</td>
<td>Ours</td>
<td>Texture (B)</td>
<td>Content (A)</td>
<td>Ours</td>
</tr>
<tr>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
</tr>
</tbody>
</table>

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

Style (B)  Content (A)  Ours

S  R  R  R
L  H  H  H
N  K  K  K
Videos
Thank You! Questions?
Numerical Results: Pretrained Classifier

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

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 $G(E^c(b), E^R_A(a), 0))$ in $A$. Similarly, in the reverse direction, we check the membership of $G(E^c(a), 0, E^R_B(b))$ 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?

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

Table 2. Given 20 randomly selected images $a \in A$ and $b \in B$, we consider the generated image $G(E^c(a), 0, E^g_B(b))$ 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 $h(x) = \sigma_d(Wx + b)$, 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.

<table>
<thead>
<tr>
<th>Augmentation</th>
<th>One-way cycle</th>
<th>Selective backprop</th>
<th>Accuracy (MNIST to SVHN)</th>
<th>Accuracy (SVHN to MNIST)</th>
</tr>
</thead>
<tbody>
<tr>
<td>False</td>
<td>False</td>
<td>False</td>
<td>0.07</td>
<td>0.10</td>
</tr>
<tr>
<td>True</td>
<td>False</td>
<td>False</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>False</td>
<td>True</td>
<td>False</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>True</td>
<td>True</td>
<td>False</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>True</td>
<td>True</td>
<td>True</td>
<td>0.19</td>
<td>0.20</td>
</tr>
<tr>
<td>True</td>
<td>False</td>
<td>True</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>False</td>
<td>True</td>
<td>False</td>
<td>0.22</td>
<td>0.23</td>
</tr>
<tr>
<td>True</td>
<td>True</td>
<td>No Phase II update of $E^S$ and $G^S$</td>
<td>0.16</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Additional settings:

| 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].

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Content</td>
<td>Summer2Winter</td>
<td>0.64</td>
<td>3.20</td>
<td>3.53</td>
<td>1.41</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>Winter2Summer</td>
<td>0.73</td>
<td>3.10</td>
<td>3.48</td>
<td>1.38</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>Monet2Photo</td>
<td>3.75</td>
<td>6.82</td>
<td>5.80</td>
<td>1.46</td>
<td>1.41</td>
</tr>
<tr>
<td></td>
<td>Photo2Monet</td>
<td>1.47</td>
<td>2.92</td>
<td>2.98</td>
<td>2.01</td>
<td>1.46</td>
</tr>
<tr>
<td>(ii) Style</td>
<td>Summer2Winter</td>
<td>1.64</td>
<td>6.51</td>
<td>1.62</td>
<td>1.69</td>
<td>1.69</td>
</tr>
<tr>
<td></td>
<td>Winter2Summer</td>
<td>1.58</td>
<td>6.80</td>
<td>1.31</td>
<td>1.69</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td>Monet2Photo</td>
<td>1.20</td>
<td>6.83</td>
<td>0.90</td>
<td>1.21</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td>Photo2Monet</td>
<td>1.95</td>
<td>7.53</td>
<td>1.91</td>
<td>2.12</td>
<td>1.88</td>
</tr>
</tbody>
</table>
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.

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