Domain Intersection and Domain Difference

Sagie Benaim\textsuperscript{1}, Michael Khaitov\textsuperscript{1}, Tomer Galanti\textsuperscript{1}, Lior Wolf\textsuperscript{1,2}

\textsuperscript{1}Tel Aviv University
\textsuperscript{2}Facebook AI Research
Image to Image Translation
MUNIT: Style and Texture Changes

Sketch to Image Translation

(a) edges ↔ shoes

(b) edges ↔ handbags

Huang et al., ECCV 2018
DRIT, DRIT++: Similar Textural and Style Changes

Lee et al., ECCV 2018
Cannot Transfer Content!

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.
Attribute Transfer

Figure 6: Facial attribute editing results on the CelebA dataset. The rows from top to down are results of IcGAN [26], FaderNet [17], AttGAN [11], StarGAN [7] and STGAN.

Liu et al, CVPR 2019
Only a single Attribute!
For example, Fader Networks:

Figure 19. Translation from the domain of smiling persons to the domain of persons with glasses, using the Fader Networks method.
Given two visual domains, disentangle the separate (domain specific) information and common (domain invariant) information.
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, $E_A^s(A)$. The space of glasses. $E_A^s(a)$ is the glasses of a.

3. "Separate" latent space for domain B, $E_B^s(B)$. The space of smiles. $E_B^s(b)$ 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.
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

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

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

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

\[
L_{zero}^A := \frac{1}{m_2} \sum_{j=1}^{m_2} \|E_A^s(b_j)\|_1
\]

\[
L_{zero}^B := \frac{1}{m_1} \sum_{i=1}^{m_1} \|E_B^s(a_i)\|_1
\]
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

Separate B Latent Space (Beard)

Common Latent Space (Facial Features)
Interpolations

Common Latent Space (Facial Features)

Separate A Latent Space (Smile)
Interpolations

Separate A Latent Space (Smile)

Separate B Latent Space (Beard)
## 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 $G(E^c(b), E^c_{\lambda}(a), 0)$ in $A$. Similarly, in the reverse direction, we check the membership of $G(E^c(a), 0, E^c_{\beta}(b))$ 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 (1) ours</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>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>
<td></td>
</tr>
<tr>
<td>Question (2) ours</td>
<td>3.92 ±0.16</td>
<td>4.45 ±0.12</td>
<td>4.03 ±0.15</td>
<td>3.85 ±0.20</td>
<td>3.95 ±0.22</td>
<td></td>
</tr>
<tr>
<td>Question (3) ours</td>
<td>3.95 ±0.23</td>
<td>3.20 ±0.24</td>
<td>3.24 ±0.25</td>
<td>3.49 ±0.22</td>
<td>3.39 ±0.23</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question (1) for [20]</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>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>
<td></td>
</tr>
<tr>
<td>Question (2) for [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>Question (3) for [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 $G(E^c(a), 0, E_B^g(b))$ 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.
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

• 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” (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.

<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</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>
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

Figure 35: Additional Segmentation results for women’s hair. (a) original image, (b) ground truth segmentation, (c) our results, (d) the results of Press et al. (2019), (e) the results of Ahn & Kwak (2018), (f) results of CAM.
Code and paper available online:
https://github.com/sagiebenaim/DomainIntersection

Questions?