On disentangled and few shot visual generation and understanding

Sagie Benaim
School of Computer Science, Tel Aviv University
Outline

• Part I: Disentangled representation and generation
  • Semi-supervised setting
  • Unsupervised setting

• Part II: Few shot generation
  • Image to Image translation
  • Video generation
What is a ‘disentangled representation’?

• A real world **high-dimensional** observation $x$ (image or video) can be represented by a **low-dimensional** latent variable $z$.

• $z$ corresponds **to semantically meaningful factors** of variation of $x$ such as: content, pose, style, etc.

• A change in a single factor of $z$ should correspond to a change in a single underlying factor of variation of $x$. 
What is a ‘disentangled representation’?

Why do we need it?

Style Transfer

Content Transfer

Pose Transfer

Shape Transfer

Video Prediction

Disentanglement: Supervision Level

• **Fully Supervised:** Each image in the dataset appears with or without each factor of variation.

• **Semi-Supervised (Set Level):** Each set of images (which may be different), appear with or without each factor of variation.

• **Unsupervised:** Strong assumptions about data-set which are incorporated into the model design.
Disentanglement: Supervision Level

• **Fully Supervised:** Each image in the dataset appears with or without each factor of variation.

• **Semi-Supervised (Set Level):** Each set of images (which may be different), appear with or without each factor of variation.

• **Unsupervised:** Strong assumptions about data-set which are incorporated into the model design.
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, $E^s_A(A)$. 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**, and **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

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

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

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

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

Ensures the separate encoder of B does not encode information about A

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

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

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

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

\[ G(E_c(b), E_A^s(a), 0) \]

Remove b’s smile
Add a’s glasses
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
Interpolations

Common Latent Space (Facial Features)

Separate A Latent Space (Smile)
Interpolations

Separate B Latent Space (Beard)

B₁ → B₂

Separate A Latent Space (Smile)

A₁

A₂
Losses “Necessary” and “Sufficient”

• 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
Attention-based Masked Generation

A + B’s mustache
Two Attributes
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>Press et al.</td>
<td>15.0 ± 0.6</td>
<td>167.7 ± 0.3</td>
<td>96.9%</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<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>
<td></td>
</tr>
<tr>
<td>Liu et al.</td>
<td>4.3 ± 0.3</td>
<td>129.0 ± 3</td>
<td>98.4%</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Fader</td>
<td>11.3 ± 0.7</td>
<td>155.6 ± 4.7</td>
<td>93.7%</td>
<td>0.89</td>
<td></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>Press et al.</td>
<td>16.6 ± 0.8</td>
<td>175.9 ± 1.4</td>
<td>100.0%</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<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>
<td></td>
</tr>
<tr>
<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>
<td></td>
</tr>
<tr>
<td>Fader</td>
<td>14.1 ± 0.6</td>
<td>162.6 ± 1.5</td>
<td>98.4%</td>
<td>0.76</td>
<td></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>Press et al.</td>
<td>15.3± 0.5</td>
<td>172.0± 4.7</td>
<td>100.0%</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<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>
<td></td>
</tr>
<tr>
<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>
<td></td>
</tr>
<tr>
<td>Fader</td>
<td>12.5± 0.3</td>
<td>137.7± 4.2</td>
<td>100.0%</td>
<td>0.76</td>
<td></td>
</tr>
</tbody>
</table>
Out of Domain Manipulation
Semi-Supervised Background-Foreground Segmentation Using Class Labels

<table>
<thead>
<tr>
<th>Input</th>
<th>GT</th>
<th>Ours</th>
<th>Press et al., Ahn et al., CAM</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>
</tr>
<tr>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
</tr>
<tr>
<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>
<tr>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
<td><img src="image16.png" alt="Image" /></td>
</tr>
</tbody>
</table>
Semi-Supervised Background-Foreground Segmentation Using Class Labels
Disentanglement: Supervision Level

- **Fully Supervised:** Each image in the dataset appears with or without each factor of variation.
- **Semi-Supervised (Set Level):** Each set of images (which may be different), appear with or without each factor of variation.
- **Unsupervised:** Strong assumptions about data-set which are incorporated into the model design.
Unsupervised learning of disentangled representations is fundamentally impossible without inductive biases both on the considered learning approaches and the data sets.”¹

¹ “Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations”. ICML 2019. Locatello et al.,
Content Style Disentanglement: StyleGAN
Content Style Disentanglement: StyleGAN
Instance Normalization

• Let $a$ be a representation of images $I_a$

\[
IN(a) = \left( \frac{a - \mu(a)}{\sigma(a)} \right)
\]

• $\mu$ and $\sigma$ are computed along the spatial dimension of $a$.
• $\mu(a)$ and $\sigma(a)$ represent the **global statistics** of an image (such as brightness, contrast, lightning and global color changes)
Disentangle content from global statistics

• Let $a$ be a representation of images $I_a$

\[ a = \sigma(a) \left( \frac{a - \mu(a)}{\sigma(a)} \right) + \mu(a) \]

• $\mu(a)$ and $\sigma(a)$ represent the **global statistics** of an image (such as brightness, contrast, lightning and global color changes)

• **Content** represents information relating to shape and texture of objects.

• This gives unsupervised disentanglement of content and global statistics!
AdaIN – Adaptive Instance Normalization

• Let $a$, $b$ be a representation of images $I_a, I_b$

$$AdaIN(a, b) = \sigma(b) \left( \frac{a - \mu(a)}{\sigma(a)} \right) + \mu(b)$$

• Replace the global statistics of $a$ with that of $b$
Domain Adaptation

Supervised training on source domain and unsupervised on target domain

Source: GTAV

Target: Cityscapes
Unsupervised Domain Adaptation
Permuted AdaIN: Reducing the Bias Towards Global Statistics in Image Classification


\[ x^s_1, \ldots, x^s_n \]
\[ f^s_1, \ldots, f^s_n \]
\[ x^t_1, \ldots, x^t_n \]
\[ f^t_1, \ldots, f^t_n \]

Swap global statistics with probability \( p \)
Permuted AdaIN: Reducing the Bias Towards Global Statistics in Image Classification


Apply AdaIN with probability $p$
Unsupervised Domain Adaptation

Feature Extractor → Classifier

Source Domain

Target Domain

Apply pAdaIN between all layers

Fined Grained Domain Discriminator

$L_{seg}$
## Unsupervised Domain Adaptation

### GTVA to Cityscapes

<table>
<thead>
<tr>
<th>Method</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdaptSegNet [35]</td>
<td>86.5</td>
<td>36.0</td>
<td>79.9</td>
<td>23.4</td>
<td>23.3</td>
<td>93.5</td>
<td>35.2</td>
<td>14.8</td>
<td>83.4</td>
<td>33.3</td>
<td>75.6</td>
<td>58.5</td>
<td>27.6</td>
<td>73.7</td>
<td>32.5</td>
<td>35.4</td>
<td>3.9</td>
<td>30.1</td>
<td>28.1</td>
<td>42.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIBAN [28]</td>
<td>88.5</td>
<td>35.4</td>
<td>79.5</td>
<td>26.3</td>
<td>24.3</td>
<td>28.5</td>
<td>32.5</td>
<td>18.3</td>
<td>81.2</td>
<td>40.0</td>
<td>76.5</td>
<td>58.1</td>
<td>25.8</td>
<td>82.6</td>
<td>30.3</td>
<td>34.4</td>
<td>3.4</td>
<td>21.6</td>
<td>21.5</td>
<td>42.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLAN [29]</td>
<td>87.0</td>
<td>27.1</td>
<td>79.6</td>
<td>27.3</td>
<td>23.3</td>
<td>28.3</td>
<td>45.5</td>
<td>24.2</td>
<td>82.3</td>
<td>74.6</td>
<td>28.0</td>
<td>63.1</td>
<td>36.7</td>
<td>6.7</td>
<td>31.9</td>
<td>31.4</td>
<td>43.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AdaptPatch [36]</td>
<td>92.3</td>
<td>51.9</td>
<td>82.1</td>
<td>29.2</td>
<td>25.1</td>
<td>28.3</td>
<td>45.5</td>
<td>33.8</td>
<td>33.0</td>
<td>82.4</td>
<td>32.8</td>
<td>82.2</td>
<td>58.6</td>
<td>27.2</td>
<td>84.4</td>
<td>33.4</td>
<td>46.3</td>
<td>2.2</td>
<td>29.5</td>
<td>32.3</td>
<td>46.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADVENT [38]</td>
<td>89.4</td>
<td>33.1</td>
<td>81.0</td>
<td>66.6</td>
<td>29.2</td>
<td>18.3</td>
<td>79.9</td>
<td>47.7</td>
<td>78.8</td>
<td>30.5</td>
<td>84.8</td>
<td>38.5</td>
<td>44.5</td>
<td>1.7</td>
<td>31.6</td>
<td>32.4</td>
<td>45.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FADA [40]</td>
<td>92.5</td>
<td>47.5</td>
<td>85.1</td>
<td>37.6</td>
<td>32.8</td>
<td>33.4</td>
<td>33.8</td>
<td>18.4</td>
<td>85.3</td>
<td>37.7</td>
<td>83.5</td>
<td>63.2</td>
<td>39.7</td>
<td>87.5</td>
<td>32.9</td>
<td>47.8</td>
<td>1.6</td>
<td>34.9</td>
<td>39.5</td>
<td>49.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FADA [40] + pAdaIN</td>
<td><strong>93.3</strong></td>
<td><strong>55.7</strong></td>
<td><strong>85.6</strong></td>
<td><strong>38.3</strong></td>
<td><strong>29.6</strong></td>
<td><strong>31.2</strong></td>
<td><strong>34.2</strong></td>
<td><strong>17.8</strong></td>
<td><strong>86.2</strong></td>
<td><strong>41.0</strong></td>
<td><strong>88.8</strong></td>
<td><strong>65.1</strong></td>
<td><strong>37.1</strong></td>
<td><strong>87.6</strong></td>
<td><strong>45.9</strong></td>
<td><strong>55.1</strong></td>
<td><strong>15.1</strong></td>
<td><strong>39.4</strong></td>
<td><strong>31.1</strong></td>
<td><strong>51.5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Domain Adaptation

\( x^s_1, \ldots, x^s_n \)

\( x^t_1, \ldots, x^t_n \)

\( f^s_1, \ldots, f^s_n \)

\( f^t_1, \ldots, f^t_n \)

Swap global statistics with probability p
Image Classification

\[ x_1, \ldots, x_n \]

\[ f_1, \ldots, f_n \]

\[ x_{\pi(1)}, \ldots, x_{\pi(n)} \]

\[ f_{\pi(1)}, \ldots, f_{\pi(n)} \]

Swap global statistics with probability p
# Image Classification

## ImageNet

<table>
<thead>
<tr>
<th>Method</th>
<th>Architecture</th>
<th>Top-1 Accuracy</th>
<th>Top-5 Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>ResNet50</td>
<td>77.1</td>
<td>93.63</td>
</tr>
<tr>
<td>pAdaIN</td>
<td>ResNet50</td>
<td>77.7</td>
<td>93.93</td>
</tr>
<tr>
<td>Baseline</td>
<td>ResNet101</td>
<td>78.13</td>
<td>93.71</td>
</tr>
<tr>
<td>pAdaIN</td>
<td>ResNet101</td>
<td><strong>78.8</strong></td>
<td><strong>94.35</strong></td>
</tr>
<tr>
<td>Baseline</td>
<td>ResNet152</td>
<td>78.31</td>
<td>94.06</td>
</tr>
<tr>
<td>pAdaIN</td>
<td>ResNet152</td>
<td><strong>79.13</strong></td>
<td><strong>94.64</strong></td>
</tr>
</tbody>
</table>

## Cifar100

<table>
<thead>
<tr>
<th>Method</th>
<th>Architecture</th>
<th>CIFAR 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>PyramidNet</td>
<td>83.49</td>
</tr>
<tr>
<td>pAdaIN</td>
<td>PyramidNet</td>
<td><strong>84.17</strong></td>
</tr>
<tr>
<td>Baseline</td>
<td>ResNet18</td>
<td>76.13</td>
</tr>
<tr>
<td>pAdaIN</td>
<td>ResNet18</td>
<td><strong>77.82</strong></td>
</tr>
<tr>
<td>Baseline</td>
<td>ResNet50</td>
<td>78.22</td>
</tr>
<tr>
<td>pAdaIN</td>
<td>ResNet50</td>
<td><strong>79.03</strong></td>
</tr>
</tbody>
</table>
Robustness Towards Corruption
# Robustness Towards Corruption

## CIFAR100-C

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DenseNet-BC</td>
<td>59.3</td>
<td>59.6</td>
<td>55.4</td>
<td>59.2</td>
<td>53.9</td>
<td>55.2</td>
<td>38.9</td>
<td><strong>37.5</strong></td>
</tr>
<tr>
<td>ResNext-29</td>
<td>53.4</td>
<td>54.6</td>
<td>51.4</td>
<td>54.1</td>
<td>51.3</td>
<td>54.4</td>
<td>34.4</td>
<td><strong>31.6</strong></td>
</tr>
</tbody>
</table>

## Category Wise Breakdown

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Network</th>
<th>Architecture</th>
<th>E</th>
<th>mCE</th>
<th>Noise</th>
<th>Blur</th>
<th>Weather</th>
<th>Digital</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>INet-C</td>
<td>ResNet50</td>
<td>22.9</td>
<td>76.7</td>
<td>80</td>
<td>82</td>
<td>83</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>INet-C</td>
<td>pAdaIN</td>
<td>22.3</td>
<td>73.8</td>
<td>78</td>
<td>79</td>
<td>81</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>C100-C</td>
<td>DenseNet-BC</td>
<td>24.2</td>
<td>38.9</td>
<td>60</td>
<td>51</td>
<td>41</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>C100-C</td>
<td>Augmix+pAdaIN</td>
<td>22.2</td>
<td>37.5</td>
<td>58</td>
<td>49</td>
<td>40</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>C100-C</td>
<td>ResNext-29</td>
<td>21.0</td>
<td>34.4</td>
<td>56</td>
<td>48</td>
<td>32</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>C100-C</td>
<td>Augmix+pAdaIN</td>
<td>17.3</td>
<td>31.6</td>
<td>58</td>
<td>48</td>
<td>24</td>
<td>20</td>
</tr>
</tbody>
</table>

Noise: Gauss. Shot Impulse Defocus Glass Motion Zoom Snow Frost Fog Bright Contrast Elastic Pixel JPEG
Part II: Few shot generation
Unconditional Few Shot Generation

SinGAN\(^1\)

StyleGAN2-ada\(^2\)

Image to Image Translation

CycleGAN, Zhu et al., ICCV 2017
Typical Training Setting

Paired

$\{x_i, y_i\}$

Unpaired

$\{X, Y\}$
One-Shot Unsupervised Cross Domain Translation


Domain B: Many unmatched samples

Domain A: A single sample \( x \)

Analogue of \( x \) in B
Phase I: Auto-Encoder for Domain B
Phase II: Shared Latent Space Assumption
Phase II: Adapt Outer Layers

$G_1: A \rightarrow B$

$G_2: B \rightarrow A$

Cycle Loss
Phase II: Adapt Outer Layers

\[ G_1 : A \rightarrow B \]

\[ G_2 : B \rightarrow A \]

Reconstruction Loss
Phase II: Adapt Outer Layers

$G_1: A \rightarrow B$

Adversarial PatchGAN Loss

$G_2: B \rightarrow A$

Patch Distribution of $x$ and its augmentations

Real\Fake

Real\Fake
Adapt All Layers: Overfitting

G_1: A -> B

G_2: B -> A
No Underfitting (Common Space Assumption)

$G_1: A \rightarrow B$

$G_2: B \rightarrow A$
Segmentation to Facade

One-Shot

Input

Ours  Cycle-1  Unit-1

Many Samples

Cycle-ALL  Unit-ALL
Facade to Segmentation

One-Shot

Many Samples

Input

Ours

Cycle-1

Unit-1

Cycle-ALL

Unit-ALL
Aerial View to Maps

One-Shot
- Ours
- Cycle-1
- Unit-1

Many Samples
- Cycle-ALL
- Unit-ALL
Maps to Aerial View

One-Shot

Many Samples

Input

Ours

Cycle-1

Unit-1

Cycle-ALL

Unit-ALL
Summer to Winter

One-Shot

Many Samples

Input

Ours

Cycle-1

Unit-1

Cycle-ALL

Unit-ALL
Winter to Summer

One-Shot

Many Samples

Input

Ours

Cycle-1

Unit-1

Cycle-ALL

Unit-ALL
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.
Structural Analogy
Structural Analogy

Target | Source | Output
Structural Analogy

Target  Source  Output
Structural Analogy
Cannot Change Object Shape
Structural Analogy

A

Structure

Analogy

Structure

B

Analogy

Structure
Motivation
Motivation
Motivation
Proposed Hierarchical Approach

Coarsest scale: Large Patches

\( \bar{a}^0 \) (Unconditional)
\( \bar{ab}^0 \) (Conditional)

LEVEL = 0

Finest scale: Small Patches

\( \bar{a}^N \) (Unconditional)
\( \bar{ab}^N \) (Conditional)

LEVEL = \( N \)
Unconditional Generation
Conditional Generation
Conditional Generation
Coarse and Mid Scales: Residual Training
Coarse and Mid Scales: \textbf{Residual} Training
Indirect Interaction Between Scales
Multiple Class Types

Input

Output
Paired Generation

A

Unconditional

B

Unconditional
Paint to Image

Input  Sketch  Ours

Input  Sketch  Ours

Input  Sketch  Ours

Input  Sketch  Ours
Texture Transfer

Content

Texture

Ours
Style Transfer

Content

Style

Ours
Video Generation
Hierarchical Patch VAE-GAN: Generating Diverse Videos from a Single Sample

Extending 2D to 3D

Real

Ours

SinGAN [1] + 3D Convolution

ConSinGAN [2] + 3D Convolution

Proposed Approach: **Patch VAE**

Input video - $x^0$
Proposed Approach: **Patch VAE**

Encoder – $E(x^0)$
Proposed Approach: **Patch VAE**

Each feature $z_i, i = [1 \ldots K], K = T \times H \times W$, in the latent space is associated with a patch $\omega_i$
Proposed Approach: **Patch VAE**

Decoder - $\tilde{x}^0$
Proposed Approach: **Patch VAE**
Proposed Approach: Hierarchical Patch VAE

Coarsest scale: **Low** resolution and frame rate
- $x^0$ (Real)
- $\tilde{x}^0$ (Generated)
- LEVEL = 0

Finest scale: **High** resolution and frame rate
- $x^N$ (Real)
- $\tilde{x}^N$ (Generated)
- LEVEL = $N$
Proposed Approach: Hierarchical Patch VAE

LEVEL = 0
Proposed Approach: Hierarchical Patch VAE

Up-sampling block - $\tilde{x}^1$

LEVEL = 1
Proposed Approach: Hierarchical Patch VAE

Hierarchical up-sampling up to $\tilde{x}^M$

LEVEL $\leq M$
Proposed Approach: Hierarchical Patch VAE GAN

Up-sampling block $\hat{x}^{M+1}$

LEVEL = $M + 1$
Proposed Approach: Hierarchical Patch VAE GAN

Adversarial training

LEVEL = $M + 1$
Proposed Approach: Hierarchical Patch VAE GAN

Hierarchical up-sampling up to final resolution $\tilde{x}^N$

$M + 1 < \text{LEVEL} \leq N$
Effect of Number of patch-VAE levels

9 Levels Total

1 p-VAE – 8 p-GAN

8 p-VAE – 1 p-GAN

3 p-VAE – 6 p-GAN

Training Video
Effect of Number of patch-VAE levels

Total of 9 layers

Quality

Diversity

(Lower is Better)

(Higher is Better)
Conclusion (Disentanglement)

- Supervision level: supervised, semi-supervised and unsupervised.
- Semi-supervised generation -> good representation for downstream tasks.
- Unsupervised disentanglement of “global” statistics vs content using permuted AdaIN (applied on top of every convolutional layer) -> good for domain adaptation and many image classification tasks.
- Next: “semi-supervised” and “unsupervised” disentanglement for more complex tasks: e.g decompose illumination from a scene or decompose time-dependent from static factors in video.
Conclusion (Few shot generation)

- Image to Image Translation:
  - Weight sharing (shared latent space assumption)
  - Transformations (strong inductive bias)
  - Matching patches (dense similarity measure)
  - Next: Few shot image understanding: anomaly detection, retrieval?

- Video generation:
  - Patch VAE for coarse scales (large variety) and Patch GAN for fine scales (high fidelity)
  - Next: Temporal super resolution, temporal inpainting, etc
Papers (In order of appearance)

Thank You! Questions?
Unsupervised Domain Adaptation

### Generalization

<table>
<thead>
<tr>
<th>Method</th>
<th>Road</th>
<th>SW</th>
<th>Build</th>
<th>Wall</th>
<th>Fence</th>
<th>Pole</th>
<th>TL</th>
<th>TS</th>
<th>Veg.</th>
<th>Terrain</th>
<th>Sky</th>
<th>PR</th>
<th>Rider</th>
<th>Car</th>
<th>Truck</th>
<th>Bus</th>
<th>Train</th>
<th>Motor</th>
<th>Bike</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source only</td>
<td>57.9</td>
<td>17.4</td>
<td>71.5</td>
<td>19.3</td>
<td>18.3</td>
<td>25.39</td>
<td>32.5</td>
<td>16.8</td>
<td>82.3</td>
<td>28.2</td>
<td>78.0</td>
<td>55.3</td>
<td>31.3</td>
<td>71.6</td>
<td>19.1</td>
<td>26.8</td>
<td>9.2</td>
<td>26.3</td>
<td>13.7</td>
<td>37.0</td>
</tr>
<tr>
<td>Source only + pAdaIN</td>
<td>57.2</td>
<td>20.2</td>
<td>71.6</td>
<td>28.3</td>
<td>19.1</td>
<td>26.1</td>
<td>33.6</td>
<td>13.0</td>
<td>82.1</td>
<td>29.0</td>
<td>69.5</td>
<td>56.7</td>
<td>33.0</td>
<td>67.5</td>
<td>27.8</td>
<td>35.1</td>
<td>17.6</td>
<td>33.7</td>
<td>14.5</td>
<td>38.7</td>
</tr>
</tbody>
</table>

Domain Adaptation
SVHN to MNIST

Selective Adaptation
Adapt All Layers
Selective Adaptation
Adapt All Layers
Domain Adaptation

Domain A (Source)

Domain B (Target)

No Labels

With Labels
Unsupervised Domain Adaptation

Domain B (Target)

Domain A (Source)

No Labels

No Labels
Unsupervised Domain Adaptation

- Given an MNIST digit $a$, we randomly sample an SVHN digit $b$ and consider the translation to SVHN as $G(E_c(a), 0, E_A^S(b))$.
- Marginalize over samples in $b$.
- Achieve **SOTA**: MNIST to SVHN: 61.0%, Reverse: 41.0%
Training:

(a) $L_{zero}(E^s_B(a), \bar{o})$
(b) $L_{recon}(a, G(E^c(a), E^s_A(a), \bar{o}))$
(c) $L_{recon}(b, G(E^c(b), \bar{o}, E^s_B(b)))$

Legend:
- Green: Domain A
- Blue: Domain B
- Purple: Shared encoder
- Black: Generator
- Orange: Loss
Results

Beard to Smile

Glasses to Smile

Glasses $\cap$ Smile
Interpolations

**Common Latent Space (Facial Features)**

A1 A2

**Separate B Latent Space (Beard)**

B1 B2
Fully supervised (example)

Learning to Factorize and Relight a City. Liu et al., ECCV 2020
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 $G(E^c(b), E^g_A(a), 0))$ in $A$. Similarly, in the reverse direction, we check the membership of $G(E^c(a), 0, E^g_B(b))$ in $B$. 
Numerical Results: User Study

• Q1: Is the specific attribute of A (e.g. smile) removed?
• Q2: Is the guided image b’s specific attribute (e.g. glasses) added?
• Q3: Is the identity 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></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>Question (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>Question (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>Question (1) for [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>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 $G(E_c^c(a), 0, E_B^B(b))$ 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 \( b \).
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>False</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>True</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>
<tr>
<td>True</td>
<td>Two-way cycle</td>
<td>True</td>
<td>0.20</td>
<td>0.13</td>
</tr>
<tr>
<td>True</td>
<td>Two-way cycle</td>
<td>False</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>True</td>
<td>True</td>
<td>True</td>
<td><strong>0.23</strong></td>
<td><strong>0.23</strong></td>
</tr>
</tbody>
</table>
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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summer2Winter</td>
<td>0.64</td>
<td>3.20</td>
<td>3.53</td>
<td>1.41</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>Winter2Summer</td>
<td>0.73</td>
<td>3.10</td>
<td>3.48</td>
<td>1.38</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>Monet2Photo</td>
<td>3.75</td>
<td>6.82</td>
<td>5.80</td>
<td>1.46</td>
<td>1.41</td>
<td></td>
</tr>
<tr>
<td>Photo2Monet</td>
<td>1.47</td>
<td>2.92</td>
<td>2.98</td>
<td>2.01</td>
<td>1.46</td>
<td></td>
</tr>
<tr>
<td>(ii) Style</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summer2Winter</td>
<td>1.64</td>
<td>6.51</td>
<td>1.62</td>
<td>1.69</td>
<td>1.69</td>
<td></td>
</tr>
<tr>
<td>Winter2Summer</td>
<td>1.58</td>
<td>6.80</td>
<td>1.31</td>
<td>1.69</td>
<td>1.66</td>
<td></td>
</tr>
<tr>
<td>Monet2Photo</td>
<td>1.20</td>
<td>6.83</td>
<td>0.90</td>
<td>1.21</td>
<td>1.18</td>
<td></td>
</tr>
<tr>
<td>Photo2Monet</td>
<td>1.95</td>
<td>7.53</td>
<td>1.91</td>
<td>2.12</td>
<td>1.88</td>
<td></td>
</tr>
</tbody>
</table>
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 1</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 1</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 1</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>