Progressive Disentanglement for Instance-Invariant Domain Adaptive Object Detection

Aming Wu
amwu@xidian.edu.cn

June 9th, 2021
Applications of Object Detection
Progresses on Object Detection

- **Fast R-CNN**
- **SSD**
- **Mask R-CNN**
- **CornerNet**

Faster R-CNN - 2015

YOLO - 2016

SSD - 2017

CornerNet - 2018-


Domain-Shift Impact on Object Detection

- Training data and test data are from different domains
Adversarial Feature Learning

- Feature alignment through gradient reversal

Adversarial Feature Learning

● Domain adaptive Faster R-CNN

Adversarial Feature Learning

- **Strong-weak distribution alignment**

![Diagram of Adversarial Feature Learning]


Motivation

● The domain-specific information existing in the aligned features may affect the performance

● Domain-invariant features play a key role in transferring detection ability
Progressive Disentanglement
Progressive Disentanglement

- The goal of the base disentangled layer is to enhance the domain-invariant information in a middle-layer feature map

$$F_{di}^1 = E_{DIR}^1(F_b^1), \quad F_{ds}^1 = E_{DSR}^1(F_b^1), \quad F^1 = F_{di}^1 + F_b^1$$

- The progressive disentangled layer aims to extract instance-level domain-invariant features

$$F_b^2 = E_b^2(F^1), \quad F_{di}^2 = E_{DIR}^2(F_b^2), \quad F_{ds}^2 = E_{DSR}^2(F_b^2)$$
Three-Stage Optimization

- The goal of disentangled learning is to uncover a set of independent factors that give rise to the current observation.
- A detached optimization is devised to break the disentangled process into three sequential sub-processes.

(a) Feature Decomposition
(b) Feature Separation
(c) Feature Reconstruction
Three-Stage Optimization

Stage I: Feature Decomposition

Source

Target

\( E_{DSR}^1 \)  \( \rightarrow \)  \( E_{DIR}^1 \)  \( \rightarrow \)  \( E_b^1 \)  \( \rightarrow \)  \( E_b^2 \)  \( \rightarrow \)  \( E_{DSR}^2 \)

Domain Classifier

RPN

RoI Feat. Extractor

Classifier

Regressor

Domain Classifier
Three-Stage Optimization

- **Mutual Information Minimization**

  \[
  I(x, z) = \int \log \frac{dP_{xz}}{dP_x \otimes dP_z} dP_{xz}
  \]

- **Relation-consistency Loss**
Three-Stage Optimization

Stage II: Independent Feature Separation

- Source
  - Bird

- Target
  - Scene

- Estimator $\Phi^1$:
  - $E^1_{DSR}$
  - $E^1_{DIR}$

- Domain Classifier

- RPN

- Estimator $\Phi^2$:
  - $E^2_{DSR}$
  - $E^2_{DIR}$
  - $E^2_b$

- RoI Feat. Extractor

- Classifier

- Regressor

Three-Stage Optimization

Stage III: Feature Reconstruction

Comparison of Different Disentangled Methods

(a) Disentanglement on High-level Feature Vectors

(b) Disentanglement on Spatial Feature Maps

## Performance Analysis

### Table 1: Cityscapes ➔ FoggyCityscapes

<table>
<thead>
<tr>
<th>Method</th>
<th>backbone</th>
<th>person</th>
<th>rider</th>
<th>car</th>
<th>truck</th>
<th>bus</th>
<th>train</th>
<th>motorcycle</th>
<th>bicycle</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Only</td>
<td>VGG16</td>
<td>24.7</td>
<td>31.9</td>
<td>33.1</td>
<td>11.0</td>
<td>26.4</td>
<td>9.2</td>
<td>18.0</td>
<td>27.9</td>
<td>22.8</td>
</tr>
<tr>
<td>DAF [9]</td>
<td>VGG16</td>
<td>25.0</td>
<td>31.0</td>
<td>40.5</td>
<td>22.1</td>
<td>35.3</td>
<td>20.2</td>
<td>20.0</td>
<td>27.1</td>
<td>27.6</td>
</tr>
<tr>
<td>DT [30]</td>
<td>VGG16</td>
<td>25.4</td>
<td>39.3</td>
<td>42.4</td>
<td>24.9</td>
<td>40.4</td>
<td>23.1</td>
<td>25.9</td>
<td>30.4</td>
<td>31.5</td>
</tr>
<tr>
<td>SC-DA(Type3) [49]</td>
<td>VGG16</td>
<td>33.5</td>
<td>38.0</td>
<td>48.5</td>
<td>26.5</td>
<td>39.0</td>
<td>23.3</td>
<td>28.0</td>
<td>33.6</td>
<td>33.8</td>
</tr>
<tr>
<td>DMRL [14]</td>
<td>VGG16</td>
<td>30.8</td>
<td>40.5</td>
<td>44.3</td>
<td>27.2</td>
<td>33.4</td>
<td>20.7</td>
<td>34.5</td>
<td>28.4</td>
<td>32.2</td>
</tr>
<tr>
<td>MLDA [52]</td>
<td>VGG16</td>
<td>33.2</td>
<td>44.2</td>
<td>44.8</td>
<td>28.2</td>
<td>41.8</td>
<td>28.7</td>
<td>30.5</td>
<td>36.5</td>
<td>36.0</td>
</tr>
<tr>
<td>FSDA [53]</td>
<td>VGG16</td>
<td>29.1</td>
<td>39.7</td>
<td>42.9</td>
<td>20.8</td>
<td>37.4</td>
<td>24.1</td>
<td>26.5</td>
<td>29.9</td>
<td>31.3</td>
</tr>
<tr>
<td>MAF [11]</td>
<td>VGG16</td>
<td>28.2</td>
<td>39.5</td>
<td>43.9</td>
<td>23.8</td>
<td>39.9</td>
<td>33.3</td>
<td>29.2</td>
<td>33.9</td>
<td>34.0</td>
</tr>
<tr>
<td><strong>SW (B) [10]</strong></td>
<td>VGG16</td>
<td>29.9</td>
<td>42.3</td>
<td>43.5</td>
<td>24.5</td>
<td>36.2</td>
<td>32.6</td>
<td>30.0</td>
<td>35.3</td>
<td>34.3</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>VGG16</td>
<td>33.12</td>
<td>43.41</td>
<td>49.63</td>
<td>21.98</td>
<td>45.75</td>
<td>32.04</td>
<td>29.59</td>
<td>37.08</td>
<td><strong>36.57</strong></td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>ResNet101</td>
<td>32.82</td>
<td>44.37</td>
<td>49.57</td>
<td>33.02</td>
<td>46.10</td>
<td>37.97</td>
<td>29.90</td>
<td>35.26</td>
<td>38.63</td>
</tr>
</tbody>
</table>
**Performance Analysis**

### Table 2: Daytime $\rightarrow$ Nighttime

<table>
<thead>
<tr>
<th>Method</th>
<th>bus</th>
<th>bike</th>
<th>car</th>
<th>motor</th>
<th>person</th>
<th>rider</th>
<th>truck</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Only</td>
<td>21.1</td>
<td>19.6</td>
<td>45.8</td>
<td>12.9</td>
<td>22.4</td>
<td>22.3</td>
<td>31.6</td>
<td>25.1</td>
</tr>
<tr>
<td>DAF [9]</td>
<td>24.2</td>
<td>21.7</td>
<td>45.3</td>
<td>7.2</td>
<td>23.3</td>
<td>25.5</td>
<td>32.1</td>
<td>25.6</td>
</tr>
<tr>
<td>SW (B) [10]</td>
<td>26.0</td>
<td>18.3</td>
<td>44.1</td>
<td>10.4</td>
<td>23.0</td>
<td><strong>25.9</strong></td>
<td>33.4</td>
<td>25.9</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>33.1</strong></td>
<td>19.5</td>
<td>44.2</td>
<td><strong>13.2</strong></td>
<td>24.7</td>
<td>24.9</td>
<td><strong>35.5</strong></td>
<td><strong>27.9</strong></td>
</tr>
</tbody>
</table>

### Table 3: VOC $\rightarrow$ Watercolor

<table>
<thead>
<tr>
<th>Method</th>
<th>bike</th>
<th>bird</th>
<th>car</th>
<th>cat</th>
<th>dog</th>
<th>person</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Only</td>
<td>68.8</td>
<td>46.8</td>
<td>37.2</td>
<td>32.7</td>
<td>21.3</td>
<td>60.7</td>
<td>44.6</td>
</tr>
<tr>
<td>BDC-Faster [10]</td>
<td>68.6</td>
<td>48.3</td>
<td>47.2</td>
<td>26.5</td>
<td>21.7</td>
<td>60.5</td>
<td>45.5</td>
</tr>
<tr>
<td>DAF [9]</td>
<td>75.2</td>
<td>40.6</td>
<td>48.0</td>
<td>31.5</td>
<td>20.6</td>
<td>60.0</td>
<td>46.0</td>
</tr>
<tr>
<td>SW (B) [10]</td>
<td>82.3</td>
<td><strong>55.9</strong></td>
<td>46.5</td>
<td>32.7</td>
<td><strong>35.5</strong></td>
<td><strong>66.7</strong></td>
<td>53.3</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>95.8</strong></td>
<td>54.3</td>
<td><strong>48.3</strong></td>
<td>42.4</td>
<td>35.1</td>
<td>65.8</td>
<td><strong>56.9</strong></td>
</tr>
</tbody>
</table>

### Table 4: VOC $\rightarrow$ Clipart

<table>
<thead>
<tr>
<th>Method</th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Only</td>
<td>35.6</td>
<td>52.5</td>
<td>24.3</td>
<td>23.0</td>
<td>20.0</td>
<td>43.9</td>
<td>32.8</td>
<td>10.7</td>
<td>30.6</td>
<td>11.7</td>
<td>13.8</td>
<td>6.0</td>
<td>36.8</td>
<td>45.9</td>
<td>48.7</td>
<td>41.9</td>
<td><strong>16.5</strong></td>
<td>7.3</td>
<td>22.9</td>
<td>32.0</td>
<td>27.8</td>
</tr>
<tr>
<td>BDC-Faster [10]</td>
<td>20.2</td>
<td>46.4</td>
<td>20.4</td>
<td>19.3</td>
<td>18.7</td>
<td>41.3</td>
<td>26.5</td>
<td>6.4</td>
<td>33.2</td>
<td>11.7</td>
<td>26.0</td>
<td>1.7</td>
<td>36.6</td>
<td>41.5</td>
<td>37.7</td>
<td>44.5</td>
<td>10.6</td>
<td>20.4</td>
<td>33.3</td>
<td>15.5</td>
<td>25.6</td>
</tr>
<tr>
<td>DAF [9]</td>
<td>15.0</td>
<td>34.6</td>
<td>12.4</td>
<td>11.9</td>
<td>19.8</td>
<td>21.1</td>
<td>23.2</td>
<td>3.1</td>
<td>22.1</td>
<td>26.3</td>
<td>10.0</td>
<td>10.0</td>
<td>19.6</td>
<td>39.4</td>
<td>34.6</td>
<td>29.3</td>
<td>1.0</td>
<td>17.1</td>
<td>19.7</td>
<td>24.8</td>
<td>19.8</td>
</tr>
<tr>
<td>SW (B) [10]</td>
<td>26.2</td>
<td>48.5</td>
<td>32.6</td>
<td>33.7</td>
<td>38.5</td>
<td>54.3</td>
<td>37.1</td>
<td><strong>18.6</strong></td>
<td>34.8</td>
<td><strong>58.3</strong></td>
<td>17.0</td>
<td>12.5</td>
<td>33.8</td>
<td>65.5</td>
<td>61.6</td>
<td><strong>52.0</strong></td>
<td>9.3</td>
<td>24.9</td>
<td>54.1</td>
<td><strong>49.1</strong></td>
<td>38.1</td>
</tr>
<tr>
<td>Ours</td>
<td>41.5</td>
<td><strong>52.7</strong></td>
<td>34.5</td>
<td>28.1</td>
<td>43.7</td>
<td><strong>58.5</strong></td>
<td>41.8</td>
<td>15.3</td>
<td><strong>40.1</strong></td>
<td>54.4</td>
<td><strong>26.7</strong></td>
<td>28.5</td>
<td><strong>37.7</strong></td>
<td><strong>75.4</strong></td>
<td>63.7</td>
<td>48.7</td>
<td><strong>16.5</strong></td>
<td><strong>30.8</strong></td>
<td><strong>54.5</strong></td>
<td><strong>48.7</strong></td>
<td><strong>42.1</strong></td>
</tr>
</tbody>
</table>
Experiments
Experiments
Experiments
## Experiments

### Ablation Analysis

<table>
<thead>
<tr>
<th>Method</th>
<th>OW</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>RC</th>
<th>C → F</th>
<th>V → W</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP layer</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>34.1%</td>
<td>52.9%</td>
</tr>
<tr>
<td>BP layer</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>33.6%</td>
<td>53.5%</td>
</tr>
<tr>
<td>BP layer</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>35.3%</td>
<td>55.3%</td>
</tr>
<tr>
<td>BP layer</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>35.5%</td>
<td>55.2%</td>
</tr>
<tr>
<td>BP layer</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td><strong>36.6%</strong></td>
<td><strong>56.9%</strong></td>
</tr>
<tr>
<td>SD layer</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>31.9%</td>
<td>50.3%</td>
</tr>
<tr>
<td>SD layer</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>32.1%</td>
<td>51.6%</td>
</tr>
<tr>
<td>SD layer</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>33.6%</td>
<td>53.1%</td>
</tr>
<tr>
<td>SD layer</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>33.5%</td>
<td>53.5%</td>
</tr>
<tr>
<td>SD layer</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td><strong>34.1%</strong></td>
<td><strong>54.6%</strong></td>
</tr>
</tbody>
</table>
Experiments

• Ablation Analysis

Detection Accuracy vs. IoU Threshold

(a) from Cityscapes to FoggyCityscapes

(b) from PASCAL VOC to Watercolor

(c) from PASCAL VOC to Clipart
Experiments

**Visualization Analysis**

Experiments:

- (a) GT
- (b) SD-Base
- (c) BP-Base
- (d) SD-DIR
- (e) BP-DIR
- (f) SD-DSR
- (g) BP-DSR
Experiments

- **Visualization Analysis**

(a) Foggy image
(b) P-Base
(c) P-DIR
(d) P-DSR
Experiments

Visualization Analysis
Thanks for your attention!