Cross-Domain Semantic Segmentation of Urban Scenes via Multi-Level Feature Alignment

Bin Zhang, Shengjie Zhao, Rongqing Zhang

Key Laboratory of Embedded System and Service Computing, Ministry of Education
School of Software Engineering,
Tongji University, Shanghai, 200092, China
Outline

1. Background
2. The Proposed Method
3. Experimental Results
4. Conclusion
In recent years, semantic segmentation has achieved great success due to the development of large-scale dataset.

### Advantage and Challenge of Semantic Segmentation

**Advantage**
- This technology has the potential to provide pixel-wise understanding of the scene.

**Challenge**
- The main challenge is the time-consuming process to collect and label the training dataset.

Therefore, how to fully utilize the synthetic dataset to improve the model performance in real-world scene is of great importance.
Domain adaptation is a representative method in transfer learning.

- **Advantage and Main Idea of Domain Adaptation**

**Advantage**
- Domain adaptation can reduce the retraining process of the model in the target domain.

**Main Idea**
- The main idea is to transfer the model trained on source domain dataset to target domain.
The existing domain adaptation techniques can be classified into supervised domain adaptation and unsupervised domain adaptation.

- **Supervised Domain Adaptation**
  Supervised domain adaptation assumes that both the source and target domain samples have the corresponding label.

- **Unsupervised Domain Adaptation**
  In the unsupervised domain adaptation setting, the source domain has a large number of annotated training examples while the target domain only have unlabeled data.
The Proposed Method

- Comparison between the segmentation results.
  (a) Directly applying DeepLab model trained on source images without modification.
  (b) Using our multi-level feature alignment method.
Cross-Domain Semantic Segmentation of Urban Scenes via Multi-Level Feature Alignment

The Proposed Method

Segmentation Module

Segmentation Map

Domain Classifier

Multi-Level Feature Alignment Module

Feature Reinforcement Unit

Multi-Level Feature Concatenation

ResNet-101

Stage 2
Stage 3
Stage 4
Stage 5

Source image

Translated source image

Target image

Translated target image

Domain Label
The Proposed Method

Image-to-Image Translation Network

- By mapping the source image into target domain, we enable our model to learn the segmentation task on labeled source data with target style.

- Two image translators learn to map samples across different domains while two adversarial discriminators try to discriminate them

$$\mathcal{L}_{adv}^{img} (I_{S\rightarrow T}, D_T) = \mathbb{E}_{x \sim \mathcal{X}_t} \left[ \log D_T(x) \right] + \mathbb{E}_{x \sim \mathcal{X}_s} \left[ \log (1 - D_T(I_{S\rightarrow T}(x))) \right]$$

$$\mathcal{L}_{adv}^{img} (I_{T\rightarrow S}, D_S) = \mathbb{E}_{x \sim \mathcal{X}_s} \left[ \log D_S(x) \right] + \mathbb{E}_{x \sim \mathcal{X}_t} \left[ \log (1 - D_S(I_{T\rightarrow S}(x))) \right]$$

$$\mathcal{L}_{cyc} (I_{S\rightarrow T}, I_{T\rightarrow S}) = \mathbb{E}_{x \sim \mathcal{X}_t} \left[ \| I_{T\rightarrow S}(I_{S\rightarrow T}(x)) - x \|_1 \right] + \mathbb{E}_{x \sim \mathcal{X}_s} \left[ \| I_{S\rightarrow T}(I_{T\rightarrow S}(x)) - x \|_1 \right]$$
The Proposed Method

Semantic Segmentation Network

• We utilize the DeepLab-v2 network with pretrained ResNet-101 backbone as our base model. We discard the last fully connected layer and modify the strides of the last two convolution layers to 1.

\[
\mathcal{L}_{seg} = \mathbb{E}_{(x,y) \sim \mathcal{X}^s} \left[ \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} y^{(h,w,c)} \log p^{(h,w,c)} \right]
\]

Multi-level Feature Alignment

• First, we concatenate the multi-stage feature and obtain the multi-level concatenated feature maps. Then a classifier tries to discriminate the multi-level feature representation.

\[
\mathcal{L}_d = \mathbb{E}_{x \sim \mathcal{X}^s} \left[ \log D \left( F^x_m \right) \right] + \mathbb{E}_{x \sim \mathcal{X}^t} \left[ \log (1 - D \left( F^x_m \right)) \right]
\]

\[
D \left( F^x_m \right) = \begin{cases} 0, & \text{if } x \sim \mathcal{X}^s \\ 1, & \text{otherwise.} \end{cases}
\]
The Proposed Method

- Detailed structure of the feature reinforcement unit. GConv stands for group convolution and GN represents group normalization.

**Algorithm 1** Training procedure of our network

**Input:** source dataset $\mathcal{X}^s$, target dataset $\mathcal{X}^t$, iteration number $\text{max\_iter}$, modulating factor $\lambda = 0.001$.

**Initialization:** Initialize the domain classifier $D$ and the two image translators from scratch. The segmentation network $G$ is pretrained on the ImageNet dataset.

1: train the image-to-image translation networks $I_{S \rightarrow T}$ and $I_{T \rightarrow S}$ by optimizing Eqn. (4)

2: repeat

3: $(x^s, y^s) \leftarrow$ sample source image from $\mathcal{X}^s$

4: $x^t \leftarrow$ sample target image from $\mathcal{X}^t$

5: obtain translated source image $\bar{x}^s = I_{S \rightarrow T}(x^s)$

6: generate feature maps for both $\bar{x}^s$ and $x^t$

7: $\mathcal{L}_{seg} \leftarrow$ compute the segmentation loss for source image by Eqn. (5)

8: $G \leftarrow \min \mathcal{L}_{seg} + \lambda \mathcal{L}_{adv}$

9: $D \leftarrow \min \mathcal{L}_d$

10: until $\text{max\_iter}$
Experimental Results

- **Datasets and Evaluation Metric**

  We verify the performance of our proposed approach on the GTA5 $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes domain adaptation tasks. Cityscapes is a large-scale dataset to evaluate the accuracy of semantic segmentation models, which covers the urban scenes of several European countries. It is split into a training set with 2,975 samples, a testing set with 1,525 samples, and a validation set with 500 samples. GTA5 dataset contains 24,966 high-definition images collected from a contemporary computer game called Grand Theft Auto V. The dataset is automatically annotated into 19 categories, which are consistent with the Cityscapes dataset. As for the evaluation metric, we choose the commonly adopted Intersection over Union (IoU) for fair comparison:

$$IoU = \frac{TP}{TP + FP + FN}$$
Experimental Results

**TABLE I**

The comparison results between baseline approaches and ours from GTA5 to Cityscapes.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Backbone</th>
<th>road</th>
<th>sidewalk</th>
<th>building</th>
<th>wall</th>
<th>fence</th>
<th>pole</th>
<th>t light</th>
<th>sign</th>
<th>tree</th>
<th>sky</th>
<th>person</th>
<th>rider</th>
<th>car</th>
<th>truck</th>
<th>bus</th>
<th>train</th>
<th>motorbike</th>
<th>bike</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN Wd [13]</td>
<td>VGG-16</td>
<td>70.4</td>
<td>32.4</td>
<td>62.1</td>
<td>14.9</td>
<td>5.4</td>
<td>10.9</td>
<td>14.2</td>
<td>2.7</td>
<td>79.2</td>
<td>21.3</td>
<td>64.6</td>
<td>44.4</td>
<td>4.2</td>
<td>70.4</td>
<td>8.0</td>
<td>7.3</td>
<td>0.0</td>
<td>35.5</td>
<td>27.1</td>
</tr>
<tr>
<td>TGC-F DA [20]</td>
<td>VGG-16</td>
<td>90.2</td>
<td>51.5</td>
<td>81.1</td>
<td>15.0</td>
<td>10.7</td>
<td>37.5</td>
<td>35.2</td>
<td>28.9</td>
<td>84.1</td>
<td>32.7</td>
<td>75.9</td>
<td>62.7</td>
<td>19.9</td>
<td>82.6</td>
<td>22.9</td>
<td>28.3</td>
<td>0.0</td>
<td>19.3</td>
<td>12.0</td>
</tr>
<tr>
<td>ROAD [21]</td>
<td>VGG-16</td>
<td>85.4</td>
<td>31.2</td>
<td>78.6</td>
<td>27.9</td>
<td>22.2</td>
<td>21.9</td>
<td>23.7</td>
<td>11.4</td>
<td>80.7</td>
<td>29.3</td>
<td>68.9</td>
<td>48.5</td>
<td>14.1</td>
<td>78.0</td>
<td>23.8</td>
<td>9.4</td>
<td>0.0</td>
<td>35.9</td>
<td></td>
</tr>
<tr>
<td>DLOW [22]</td>
<td>ResNet-101</td>
<td>87.1</td>
<td>33.5</td>
<td>80.5</td>
<td>24.5</td>
<td>13.2</td>
<td>29.8</td>
<td>29.5</td>
<td>26.6</td>
<td>82.6</td>
<td>26.7</td>
<td>81.8</td>
<td>55.9</td>
<td>25.3</td>
<td>78.0</td>
<td>33.5</td>
<td>38.7</td>
<td>0.0</td>
<td>22.9</td>
<td>34.5</td>
</tr>
<tr>
<td>CLAN [23]</td>
<td>ResNet-101</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>35.5</td>
<td>24.2</td>
<td>83.6</td>
<td>74.2</td>
<td>58.6</td>
<td>8.0</td>
<td>76.2</td>
<td>33.1</td>
<td>36.7</td>
<td>6.7</td>
<td>31.9</td>
<td>31.4</td>
<td>43.2</td>
</tr>
<tr>
<td>AdvEnt [24]</td>
<td>ResNet-101</td>
<td>89.4</td>
<td>33.1</td>
<td>81.0</td>
<td>26.6</td>
<td>26.8</td>
<td>27.2</td>
<td>33.5</td>
<td>24.7</td>
<td>83.9</td>
<td>36.7</td>
<td>78.8</td>
<td>57.8</td>
<td>30.5</td>
<td>84.8</td>
<td>38.5</td>
<td>44.3</td>
<td>1.7</td>
<td>31.6</td>
<td>32.4</td>
</tr>
<tr>
<td>Ours</td>
<td>ResNet-101</td>
<td>91.8</td>
<td>47.1</td>
<td>85.6</td>
<td>29.5</td>
<td>29.3</td>
<td>35.4</td>
<td>36.8</td>
<td>33.2</td>
<td>81.5</td>
<td>35.1</td>
<td>82.1</td>
<td>62.1</td>
<td>30.6</td>
<td>79.0</td>
<td>22.6</td>
<td>33.4</td>
<td>4.3</td>
<td>33.9</td>
<td>20.1</td>
</tr>
</tbody>
</table>

• Our method outperforms other methods by a large margin in segmenting “road”, “building”, “wall”, “fence”, “sky”, “rider”, and “motorbike” categories.

**TABLE II**

The comparison results between baseline approaches and ours from SYNTHIA to Cityscapes.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Backbone</th>
<th>road</th>
<th>sidewalk</th>
<th>building</th>
<th>wall</th>
<th>fence</th>
<th>pole</th>
<th>t light</th>
<th>sign</th>
<th>tree</th>
<th>sky</th>
<th>person</th>
<th>rider</th>
<th>car</th>
<th>truck</th>
<th>bus</th>
<th>train</th>
<th>motorbike</th>
<th>bike</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN Wd [13]</td>
<td>VGG-16</td>
<td>11.5</td>
<td>19.6</td>
<td>30.8</td>
<td>4.4</td>
<td>0.0</td>
<td>20.3</td>
<td>0.1</td>
<td>11.7</td>
<td>42.3</td>
<td>68.7</td>
<td>51.2</td>
<td>3.8</td>
<td>54.0</td>
<td>3.2</td>
<td>0.2</td>
<td>0.6</td>
<td>20.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TGC-F DA [20]</td>
<td>VGG-16</td>
<td>90.1</td>
<td>48.6</td>
<td>80.7</td>
<td>2.2</td>
<td>0.2</td>
<td>27.2</td>
<td>3.2</td>
<td>14.3</td>
<td>82.1</td>
<td>78.4</td>
<td>54.4</td>
<td>16.4</td>
<td>82.5</td>
<td>12.3</td>
<td>1.7</td>
<td>21.8</td>
<td>38.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROAD [21]</td>
<td>VGG-16</td>
<td>77.7</td>
<td>30.0</td>
<td>77.5</td>
<td>9.6</td>
<td>0.3</td>
<td>25.8</td>
<td>10.3</td>
<td>15.6</td>
<td>77.6</td>
<td>79.8</td>
<td>44.5</td>
<td>16.6</td>
<td>67.8</td>
<td>14.5</td>
<td>7.0</td>
<td>23.8</td>
<td>36.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLAN [23]</td>
<td>ResNet-101</td>
<td>81.3</td>
<td>37.0</td>
<td>80.1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>16.1</td>
<td>13.7</td>
<td>78.2</td>
<td>81.5</td>
<td>53.4</td>
<td>21.2</td>
<td>73.0</td>
<td>32.9</td>
<td>22.6</td>
<td>30.7</td>
<td>–</td>
</tr>
<tr>
<td>AdaptSegNet [14]</td>
<td>ResNet-101</td>
<td>79.2</td>
<td>37.2</td>
<td>78.8</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>9.9</td>
<td>10.5</td>
<td>78.2</td>
<td>80.5</td>
<td>53.5</td>
<td>19.6</td>
<td>67.0</td>
<td>29.5</td>
<td>21.6</td>
<td>31.3</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>ResNet-101</td>
<td>86.6</td>
<td>43.3</td>
<td>81.5</td>
<td>4.3</td>
<td>1.1</td>
<td>30.2</td>
<td>10.2</td>
<td>8.7</td>
<td>85.3</td>
<td>82.2</td>
<td>48.8</td>
<td>16.9</td>
<td>73.7</td>
<td>30.4</td>
<td>11.3</td>
<td>29.5</td>
<td>40.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Cross-Domain Semantic Segmentation of Urban Scenes via Multi-Level Feature Alignment

Experimental Results

- (a) Target Image
- (b) Source Only
- (c) Conventional Adapt
- (d) Ours
- (e) Ground Truth
Conclusions

- Contributions

1. We propose a novel multi-level feature alignment method for cross-domain semantic segmentation.

2. Our proposed MLFA provides a novel perspective of insight by incorporating the content and style alignment module.

3. The experimental results demonstrate that MLFA outperforms most current state-of-the-art unsupervised domain adaptation methods.
References