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

Bin Zhang, Shengjie Zhao, Rongqing Zhang

Tongji University

Background

Semantic segmentation is an essential task in plenty of real-life applications such as virtual reality, video analysis, autonomous driving, etc. Recent advancements in fundamental vision-based tasks ranging from image classification to semantic segmentation have demonstrated deep learning-based models' high capability in learning complicated representation on large datasets. Nevertheless, manually labeling semantic segmentation dataset with pixel-level annotation is extremely labor-intensive. To address this problem, we propose a novel multi-level feature alignment framework for cross-domain semantic segmentation of urban scenes by exploiting generative adversarial networks. In the proposed multi-level feature alignment method, we first translate images from one domain to another one. Then the discriminative feature representations extracted by the deep neural network are concatenated, followed by domain adversarial learning to make the intermediate feature distribution of the target domain images close to those in the source domain. With these domain adaptation techniques, models trained with images in the source domain where the labels are easy to acquire can be deployed to the target domain where the labels are scarce. Experimental evaluations on various mainstream benchmarks confirm the effectiveness as well as robustness of our approach.

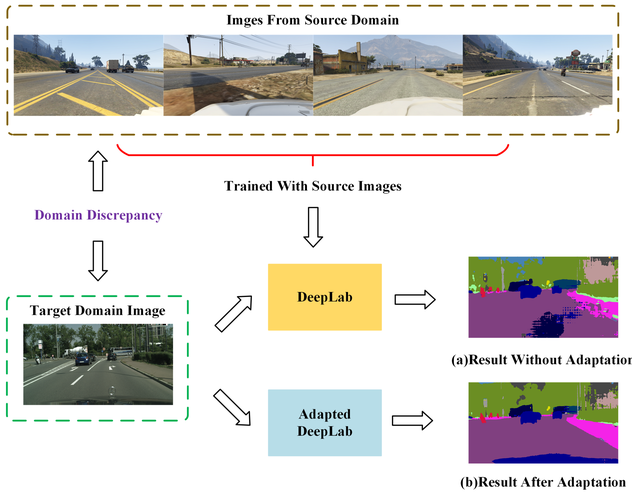

Fig. 1: Comparison between the segmentation results. (a) Directly applying DeepLab model trained on source images without modification. (b) Using our multi-level feature alignment method

Framework Overview

We start from explaining our unsupervised domain adaptation approach under the background of semantic segmentation. The overall architecture of our model is depicted in Fig.2, which consists of the following major parts: two image-to-image translation networks $I_{S\to T}$ and $I_{T\to S}$ that translate images from source domain to target domain and vice versa; feature extraction backbone (ResNet-101); semantic segmentation module; multi-level feature alignment module composed by feature reinforcement unit and domain classifier. We augment the conventional DeepLab-v2 model with our multi-level domain adaptation components, which greatly increase the models' domain adaptation ability.

Model

By mapping the source image into target domain, we enable our model to learn the segmentation task on labeled source data with target style. This is accomplished by the following adversarial loss function for $I_{S\to T}$ and D_T:

$$\begin{aligned}\mathcal{L}_{adv}^{img}\left(I_{S\to T}, D_T\right) = &\mathbb{E}_{x\sim\mathcal{X}^t}\left[\log D_T(x)\right] \\ &+ \mathbb{E}_{x\sim\mathcal{X}^s}\left[\log\left(1 - D_T\left(I_{S\to T}(x)\right)\right)\right],\end{aligned} \tag{1}$$

similarly for $I_{T\to S}$ and D_S:

$$\begin{aligned}\mathcal{L}_{adv}^{img}\left(I_{T\to S}, D_S\right) = &\mathbb{E}_{x\sim\mathcal{X}^s}\left[\log D_S(x)\right] \\ &+ \mathbb{E}_{x\sim\mathcal{X}^t}\left[\log\left(1 - D_S\left(I_{T\to S}(x)\right)\right)\right].\end{aligned} \tag{2}$$

Two image translators $I_{S\to T}$ and $I_{T\to S}$ learn to map samples across different domains while two adversarial discriminators D_S and D_T attempt to distinguish them.

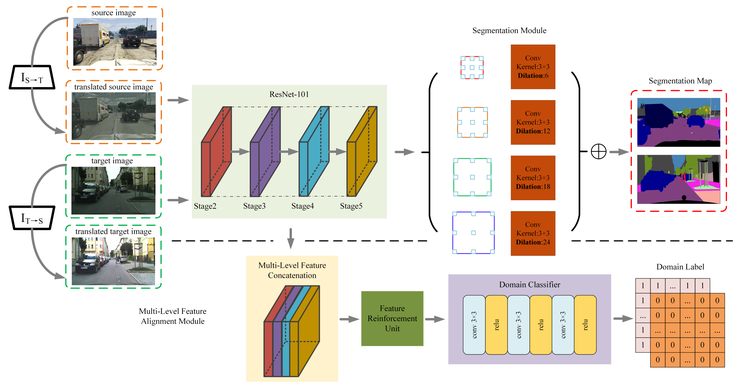

Fig. 2: Explanation of the pipeline of our framework for urban scene segmentation. The top part is ordinary semantic segmentation module (DeepLab-v2) and bottom part is our multi-level feature alignment module.

First, we concatenate the multi-stage feature and obtain the multi-level concatenated feature maps f_m^x:

$$f_m^x = \left[f_2^x, f_3^x, f_4^x, f_5^x\right], \tag{3}$$

where [] denotes a concatenation operation. Second, f_m^x is fed to the feature reinforcement unit R, which is a modified Dense Block : we alter the concatenation rule for dense connections so that feature maps from the same stage can be located adjacent to each other. Besides, standard convolution is replaced with group convolution and group normalization layers take the place of batch normalization layers. The discriminator network takes the embeddings $F_m^x = R(f_m^x)$ as input and performs pixel-wise labeling task.

$$D\left(F_m^x\right) = \begin{cases} 0, \text{ if } x \sim \mathcal{X}^s. \\ 1, \text{ otherwise.} \end{cases} \tag{4}$$

The domain classifier helps to align the distributions between the feature maps of source images and target images by minimizing the following loss:

$$\begin{aligned}\mathcal{L}_d = &\mathbb{E}_{x\sim\mathcal{X}^s}\left[\log D\left(F_m^x\right)\right] \\ &+ \mathbb{E}_{x\sim\mathcal{X}^t}\left[\log\left(1 - D\left(F_m^x\right)\right)\right].\end{aligned} \tag{5}$$

An adversarial loss $\mathcal{L}_{adv}$ is also defined to maximize the discriminator loss:

$$\mathcal{L}_{adv} = \mathbb{E}_{x\sim\mathcal{X}^t}\left[\log D\left(F_m^x\right)\right]. \tag{6}$$

Experimental Results

GTA5 $\to$ Cityscapes: We present the segmentation performance in terms of mIoU in Table 1. To compare the segmentation results fairly, we adopt the identical evaluation protocol as specified by previous related works: 19 common classes between GTA5 and Cityscapes are chosen as valid category labels. Equipped with ResNet-101, our method outperforms other methods by a large margin in segmenting "road", "building", "wall", "fence", "sky", "rider", and "motorbike" categories. Those categories usually appear in urban scene images simultaneously and tend to be spatially connected to each other, some of which share highly similar texture appearance. Our proposed unsupervised domain adaptation strategy achieves the best result: mean IoU of 45.9%, which brings significant performance boost compared to both VGG-16 based and ResNet-101 based methods.

SYNTHIA $\to$ Cityscapes: The domain adaptation task from SYNTHIA to Cityscapes is more challenging since the computer-generated images are less realistic and SYNTHIA dataset has a larger domain variance with respect to real world street scenes compared to GTA5. The appearance of some objects in SYNTHIA is extremely divergent from those in Cityscapes and it is difficult for the network to learn transferable knowledge for those objects during the adaptation process. Although SYNTHIA dataset has a larger domain gap, our model is able to better discern confusing categories such as "road" and "sidewalk" compared to other baselines. Moreover, MLFA can segment the boundary of various objects more accurately. These results illustrate that even confronted with large domain difference, our method is still capable of maintaining better performance.

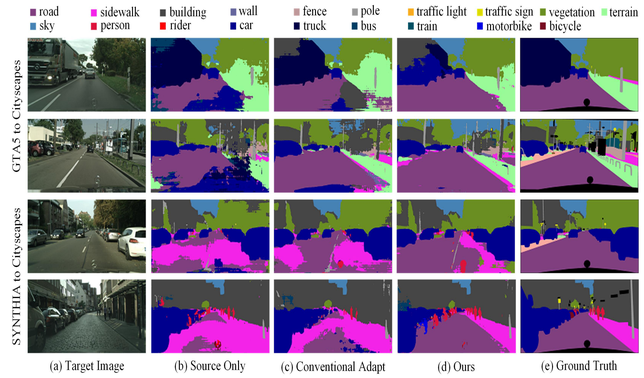

Fig. 3: Look, my method is better.

Conclusion

In this paper, we present an innovative framework to explore adaptive semantic segmentation. Our main contribution is the multi-level feature alignment (MLFA) module, which is formulated as a domain adversarial learning problem, aiming to learn both domain-invariant and domain-discriminative feature representation. By incorporating image-to-image translation network into our framework the domain variance can be further reduced. Extensive experiments on cross-domain segmentation tasks manifest that our implementation is superior to most current state-of-the-art methods.