Domain Adaptive Semantic Segmentation through Structure Enhancement

Fengmao Lv[1⋆], Qing Lian[1⋆], Guowu Yang[1], Guosheng Lin[2], Sinno Jialin Pan[2], and Lixin Duan[1]

[1] Big Data Research Center, University of Electronic Science and Technology of China

[2] School of Computer Science and Engineering, Nanyang Technological University

fengmaolv@126.com, {lianqinglalala,lxduan}@gmail.com

guowu@uestc.edu.cn, {gslin,sinnopan}@ntu.edu.sg

Abstract. Although fully convolutional networks have recently achieved great advances in semantic segmentation, the performance leaps heavily rely on supervision with pixel-level annotations which are extremely expensive and time-consuming to collect. Training models on synthetic data is a feasible way to relieve the annotation burden. However, the domain shift between synthetic and real images usually lead to poor generalization performance. In this work, we propose an effective method to adapt the segmentation network trained on synthetic images to real scenarios in an unsupervised fashion. To improve the adaptation performance for semantic segmentation, we enhance the structure information of the target images at both the feature level and the output level. Specifically, we enforce the segmentation network to learn a representation that encodes the target images' visual cues through image reconstruction, which is beneficial to the structured prediction of the target images. Further more, we implement adversarial training at the output space of the segmentation network to align the structured prediction of the source and target images based on the similar spatial structure they share. To validate the performance of our method, we conduct comprehensive experiments on the "GTA5 to Cityscapes" dataset which is a standard domain adaptation benchmark for semantic segmentation. The experimental results clearly demonstrate that our method can effectively bridge the synthetic and real image domains and obtain better adaptation performance compared with the existing state-of-the-art methods.

Keywords: Unsupervised domain adaptation, semantic segmentation, deep learning, transfer learning.

1 Introduction

Semantic segmentation is a critical and challenging task in computer vision, which aims at predicting the class label of each pixel in images. Over the past years, deep convolutional networks have achieved great advances in semantic

⋆ The first two authors contribute equally to this work.

segmentation [9,1,18]. However, the pixel-level annotation is an extremely heavy work. Specifically, we need more than 1 hour to annotate a single image in the Cityscapes dataset [3]. Training models on synthetic images can be a promising way to relieve the tedious annotation burden as their pixel-level labels can be automatically generated. Unfortunately, the domain shift between the synthetic images and real-world scenarios will degenerate the prediction results on real images. Therefore, domain adaptation should be considered to adapt the segmentation network trained on synthetic images to real images, given labeled source data and unlabeled target data. Although the recently proposed feature adaptation methods can bridge the source and target domains through learning domain-invariant features with adversarial mechanism [7,13,2], they cannot ensure that these features encode the structure information of the target images, since semantic segmentation is a highly structured prediction task.

In this paper, we propose to improve the domain adaptation performance of segmentation networks through enhancing the structure information of the target images at both the feature level and the output level. The main contribution of our work is two-fold: 1) enforcing an intermediate feature to reconstruct the training images; 2) adversarially aligning the structured output of the source and target images. Specifically, the reconstruction branch can enforce the encoding representation to preserve the visual cues of the target images, which are beneficial to their structured prediction. On the other hand, the output-level structure enhancement can directly regularize the target image's structured prediction since both domains should share similar spatial layout and local context. We conduct experiments on "GTA5 to Cityscapes" which is a standard domain adaptation benchmark for semantic segmentation to evaluate the performance of our method. The experimental results clearly demonstrate that our method can effectively bridge both domains and obtain better adaptation results than the existing state-of-the-art methods.

2 Related Work

Over the past years, domain adaptation in computer vision has been primarily explored for the classification task. Overall, the main idea is to learn a "deep" representation that is domain invariant [15,4,16,5,10]. Thus far, unsupervised domain adaptation for semantic segmentation has not been widely explored. In [7], Hoffman *et al.* first proposed to adapt segmentation networks through domain adversarial learning in the feature space. In [2], Chen *et al* . further proposed class-specific domain adversarial learning framework, which aimed at reducing the domain divergence in each class. In [11], Murez *et al.* proposed to learn domain adaptive segmentation networks through directly translating the source images to the target ones at the pixel level. In [14], Tsai *et al.* proposed to align both domains at the structured output space. In short, the previous works mainly focused on angling the source and target domains through implementing adversarial learning at different levels, ranging from intermediate features to final predictions. In our method, our main idea is to enhance the structure

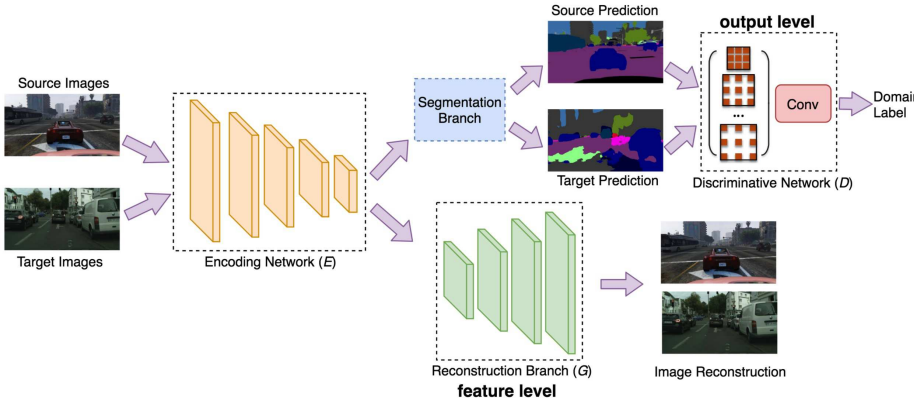

Fig. 1. The overall architecture of our model (best viewed in color).

information of the target images, which provides a reasonable regularization to their structured prediction.

3 Our method

In this paper, we focus on unsupervised domain adaptation for semantic segmentation. Our goal is to learn a segmentation network which can achieve good prediction results on the target domain, given source images I_S with pixel-level labels L_S and unlabeled target images I_T.

Overall, our adaptation method contains two major components, including reconstructing the training images and aligning the target images' structured prediction with adversarial training. Figure 1 shows the overview of our method. **Image Reconstruction**: our main idea aims at adapting the segmentation network trained on the source images through learning a representation that encodes the visual cues of the target images. This is achieved through enforcing an intermediate layer to reconstruct the training images. As displayed in Figure 1, the encoding network is shared by both the segmentation branch and the reconstruction branch. The reconstruction branch can regularize the encoding network to enhance the target images' structure information.

Throughout this paper, we denote the encoding network and the decoding network as E and G, respectively. The segmentation branch is represented as S. We define our image reconstruction loss as

$$\begin{aligned} &\min_{E,G,S} \mathcal{L}(E,G,S) \\ &s.t. \quad \mathcal{L}(E,G,S) = \lambda_{rec}\mathcal{L}_{rec} + \mathcal{L}_{seg} \\ &\qquad\qquad = \lambda_{rec}(L_1(G \circ E(I_S), I_S) + L_1(G \circ E(I_T), I_T)) \\ &\qquad\qquad + L_{sup}(S \circ E(I_S), L_S), \end{aligned} \tag{1}$$

where the former part is the reconstruction term for the training images and L_{sup} is the segmentation supervision term for the source images. In our method,

the image reconstruction is implemented with L_1 loss. Though ideally we only need to consider the reconstruction of the target images, the reconstruction of the source images can help the training of the decoding network.

Output Adaptation: Further more, we implement adversarial training at the output space of the segmentation network to align the structured prediction on the source and target images since both domains should share similar spatial layouts. As displayed in Figure 1, a discriminative network is invoked to discriminate whether a softmax prediction is from the source domain or the target domain. In contrast, the segmentation network $S \circ E(\cdot)$ will try to cheat the discriminator in order to make the target images' structured predictions resemble the source images' pixel maps. This can provide gradient updates to the segmentation network when the target images' predictions are not structured reasonably. As a whole, the segmentation network and the discriminative network play a minimax game.

To retain the spatial information, D is specified as a fully convolutional network, which discriminates the domain label of each spatial unit. Following [17], we adopt Atrous Spatial Pyramid Pooling (ASPP) in our discriminative network since this can help to align the structured output at multiple scales. The adversarial loss are formulated as

$$\begin{aligned}\max_{D} \min_{E,S} \mathcal{L}_{adv} =& \mathbb{E}_{I_T \sim \mathcal{X}_t}[\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\log(1 - D_{i,j}(S \circ E(I_T)))] \\ &+ \mathbb{E}_{I_S \sim \mathcal{X}_s}[\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\log(D_{i,j}(S \circ E(I_S)))].\end{aligned} \tag{2}$$

H and W are the height and width of the discriminator's output, respectively.

In conclusion, with the above sub-objectives, our finial objective function is defined as

$$\max_{D} \min_{E,S,G} \mathcal{L}_{seg} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv}. \tag{3}$$

In our defined minimax game, we alternately optimize each sub-network, while holding the other parts fixed. The parameters of the encoding network E is updated by averaging the gradients from each branch.

4 Experiments

4.1 Dataset

To evaluate the performance of our method, we conduct experiments on "GAT5 to Cityscapes", which is a standard benchmark of domain adaptation for semantic segmentation. Specifically, GAT5 is the dataset that contains 24,966 synthetic images with resolution of 1914×1052, rendered by the gaming engine Grand Theft Auto V. The pixel-level annotations of the GAT5 images are automatically generated. On the other hand, Cityscapes is a dataset that focuses on autonomous driving. The Cityscapes dataset consists of 2,975 images for training and 500 images in validation set. These images have a resolution of 2048×1024.

Table 1. Results of different methods on the "GTA5 to Cityscapes" dataset. Ablation studies are conducted for both the feature-level encoding and the output-level enhancement.

	road	sdwk	bldng	wall	fence	pole	light	sign	vgttn	trrn	sky	person	rider	car	truck	bus	train	mcycl	bcycl	mIoU
MCD[12]	**90.3**	31.0	78.5	19.7	17.3	28.6	30.9	16.1	**83.7**	30.0	69.1	58.5	19.6	**81.5**	23.8	30.0	5.7	25.7	14.3	39.7
CYCADA[6]	79.1	33.1	77.9	23.4	17.3	32.1	33.3	31.8	81.5	26.7	69.0	**62.8**	14.7	74.5	20.9	25.6	6.9	18.8	20.4	39.5
ROAD[7]	76.3	36.1	69.6	**28.6**	22.4	28.6	29.3	14.8	82.3	**35.3**	72.9	54.4	17.8	78.9	27.7	30.3	4.0	24.9	12.6	39.4
AdaptSeg[14]	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	**32.5**	35.4	3.9	30.1	28.1	42.4
RAN[17]	84.5	**36.9**	72.9	15.8	23.3	**39.4**	**41.8**	**36.8**	67.1	25.2	**89.1**	50.5	20.6	77.8	22.1	24.3	**22.8**	28.5	**37.9**	43.0
source only	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
feature-level	86.3	30.8	76.4	26.2	20.0	23.1	31.6	16.5	80.5	33.1	75.0	56.5	25.7	76.8	15.0	29.5	1.0	25.6	13.0	39.1
output-level	85.3	35.1	79.0	22.9	23.4	23.1	34.1	14.8	83.2	33.0	74.2	57.8	27.2	73.1	32.3	34.8	3.0	29.8	28.0	41.7
full model	88.9	31.3	**81.5**	28.3	**23.5**	28.7	37.1	30.2	82.2	33.1	76.8	59.7	**29.2**	80.8	28.9	**43.5**	4.2	**31.6**	32.3	**44.8**

We use 19 common semantic categories between GTA5 and Cityscapes as the labels. Following the existing state-of-the-art works [7,14], we train our domain adaptive segmentation network using the full GTA5 dataset and the Cityscapes training set with 2,975 images, and evaluate the performance on the Cityscapes validation set with 500 images.

4.2 Implementation Details

We adopt deeplabv2 as our baseline [1]. Specifically, the encoding network E is implemented with Resnet-101. The outputs of the *res5c* layer are fed into both the segmentation branch S and the reconstruction network G. G follows the identical architecture in [8], except that all the layers are shared by both domains. The discriminative network D contains 3 layer, including a ASPP layer with 4 dilated convolutional operators in parallel, and a convolutional layer followed sigmoid activations. The sampling rates in the ASPP layer are respectively set to 1, 2, 3 and 4. In our experiments, we use the PyTorch framework to implement our method. Overall, our experimental setting follows [14]. For E and S, we adopt stochastic gradient descent (SGD) with momentum of 0.9 as the optimizer. The parameters G and D are optimized by Adam with momentum of 0.99. In addition, we initialize the learning rate to 2.5×10^{-4} and decay it through the polynomial policy with power of 0.9. As the tradeoff parameters, λ_{rec} and λ_{adv} are set to 1.0×10^{-5} and 1.0×10^{-3}, respectively. The mIoU value is used as the metric of evaluation.

4.3 Experimental Results

In Table 1 and Figure 2, we report our adaptation results both quantitatively and qualitatively. The results demonstrate that our adaptation method can effectively improve the structured predictions of the target images. From Figure 2, we can see that the structure information of the target images' predictions are significantly enhanced, which is consistent with our motivation. With our method, the target images' pixel-level predictions clearly delineate the real spatial layout. As displayed in Table 1, our method performs better than the existing

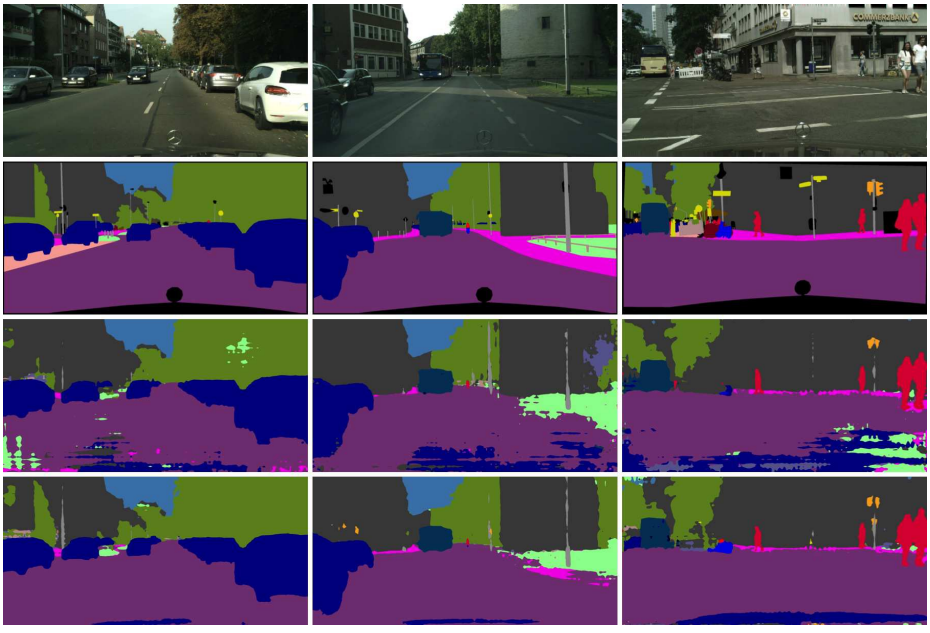

Fig. 2. The qualitative example results. The first row displays the target images, with their corresponding ground truth segmentation masks in the second row. The third and fourth rows display the results before adaptation and after adaptation with our adaptation method, respectively.

state-of-the-art methods. The ablation studies demonstrate that the feature-level encoding and the output-level enhancement can work complementarily to improve the adaptation performance. This can be ascribed to the fact that these two branches enhance the target images' structure information from complementary perspectives. Specifically, the reconstruction branch enforces the encoding representation to preserve the target images' visual cues such as the local contexts or spatial layouts, which are essential for the structured predictions. In contrast, the output-level enhancement can directly leverages the source images' pixel maps to regularize the target images' structured predictions.

5 Conclusion

In this paper, we propose an effective method to learn domain adaptive segmentation network in an unsupervised domain adaptation setting. Through enhancing the structure information of the target images at both the feature level and the output level, our method can effectively improve the domain adaptation performance of the segmentation networks. After adaptation using our method, the target images' pixel maps can clearly reveal their structure characteristics such as the spatial layout or the local context. The experimental results demonstrate that our method can effectively bridge the source and target domains.

References

1. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2018)
2. Chen, Y.H., Chen, W.Y., Chen, Y.T., Tsai, B.C., Wang, Y.C.F., Sun, M.: No more discrimination: Cross city adaptation of road scene segmenters. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2011–2020. IEEE (2017)
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
4. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495 (2014)
5. Haeusser, P., Frerix, T., Mordvintsev, A., Cremers, D.: Associative domain adaptation. In: International Conference on Computer Vision (ICCV). vol. 2, p. 6 (2017)
6. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. arXiv preprint arXiv:1711.03213 (2017)
7. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016)
8. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems. pp. 700–708 (2017)
9. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
10. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. arXiv preprint arXiv:1605.06636 (2016)
11. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. arXiv preprint arXiv:1712.00479 (2017)
12. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. arXiv preprint arXiv:1712.02560 **3** (2017)
13. Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S.N., Chellappa, R.: Unsupervised domain adaptation for semantic segmentation with gans. arXiv preprint arXiv:1711.06969 (2017)
14. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. arXiv preprint arXiv:1802.10349 (2018)
15. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
16. Zellinger, W., Grubinger, T., Lughofer, E., Natschläger, T., Saminger-Platz, S.: Central moment discrepancy (cmd) for domain-invariant representation learning. arXiv preprint arXiv:1702.08811 (2017)
17. Zhang, Y., Qiu, Z., Yao, T., Liu, D., Mei, T.: Fully convolutional adaptation networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6810–6818 (2018)
18. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 2881–2890 (2017)