Towards Multi-Source Adaptive Semantic Segmentation

Paolo Russo[1,2][0000−0002−1886−3491], Tatiana Tommasi[3][0000−0001−8229−7159], and Barbara Caputo[1,3][0000−0001−7169−0158]

[1] Istituto Italiano di Tecnologia, Italy
[2] Sapienza Università di Roma, Italy
[3] Politecnico di Torino, Italy
paolo.russo@iit.it, {tatiana.tommasi, barbara.caputo}@polito.it

Abstract. When applying powerful deep learning approaches on real world tasks like pixel level annotation of urban scenes it becomes clear that even those strong learners may fail dramatically and are still not ready for deployment in the wild. For semantic segmentation, one of the main practical challenges consists in finding large annotated collection to feed the data hungry networks. Synthetic images in combination with adaptive learning models have shown to help with this issue, but in general, different synthetic sources are analyzed separately, not leveraging on the potential growth in data amount and sample variability that could result from their combination. With our work we investigate for the first time the multi-source adaptive semantic segmentation setting, proposing some best practice rule for the data and model integration. Moreover we show how to extend an existing semantic segmentation approach to deal with multiple sources obtaining promising results.

Keywords: Semantic Segmentation · Domain Adaptation.

1 Introduction

Semantic segmentation has recently become one of the most prominent task in computer vision. Indeed the ability to assign a label to each pixel of an input image is crucial whenever a very detailed description of the observed scene is needed, as in fine-grained object categorization [25] and autonomous driving [24, 21]. However, due to the complexity of manual labeling each image pixel, this task is plagued by the scarcity of large annotated datasets, which are instead essential to leverage the power of deep learning algorithms. Synthetic images appear a useful alternative, but they reduce only in part the described issue. In the case of urban scene scenarios for autonomous driving, computer games can be used to generate automatically images with their ground truth labels, but their level of realism is still low which induces the further need of domain adaptation methods. Thus, while solving the lack of data problem, other challenges come from the development of methods able to reduce the domain gap. Up today, those two aspects of the same problem has always been tackled separately. On

one side several research groups have focused on developing different simulators with an increasing set of visual details like urban layouts, buildings, vehicles and several weather conditions, with the aim of augmenting the realism of the produced images [4, 18]. On the other side, many recent works focus on integrating techniques to align the domains either at feature, pixel or output label space level, even considering combination of those levels with different adversarial losses [21, 2, 7]. Each of the proposed synthetic domains is generally used to train a model and test it on real images, but the different synthetic sources are always kept separated even if this choice limits again the amount and variance of annotated samples usable as source. The domain adaptation literature for object classification has shown that integrating multiple sources helps generalization [5, 26, 6]. With our work we import this strategy for the first time in the semantic segmentation framework, studying how the positive trend can be maintained by practically merging the two solutions described above. The path to this goal is not trivial due to the practical differences in class statistics across domains, as well as in texture, resolution and aspect ratio for which we propose best practice rules. Moreover, we go over the simple source sample combination, exploiting a multi-level strategy that adapts each single source to the target while cooperating with the adaptation of the joint data source. Besides the standard synthetic to real direction, we extend our analysis to the case of a synthetic dataset used as target when the source combines real images and a different synthetic collection. This setting allows to better understand the difference across various synthetic sources and paves the way to the simultaneous exploitation of both the synthetic-to-real and real-to-synthetic adaptive directions [19].

2 Related works

The deep learning revolution started within the context of object classification [9] but has rapidly extended to many other tasks. The first work to put semantic segmentation under the deep learning spotlight was [14] that showed how fully connected networks could be used to assign a label to each image pixel. Several following works have then extended the interest around this task proposing tailored architectures which involve multi-scale feature combinations [1, 23] or integrate context information [13, 27]. The main issue with deep semantic segmentation remains that of collecting a large amount of images with pixel-based expensive annotations. Some solutions in this sense have been proposed either developing methods able to deal with weak annotations [15, 8], or leveraging on other domain images, as the synthetic ones produced by 3D renderings of urban scenes [4, 17, 18]. To avoid the drop in performance due to the synthetic to real shift, domain adaptation techniques have been integrated with approaches involving different network levels. The most widely used solution consists in adding a domain classifier used adversarially to minimize the gap among different feature spaces [2]. In [21], adversarial learning is used both on the segmentation output and on inner network features. A third family of methods applies adaptation directly at the pixel level with GAN-based style transfer techniques [7].

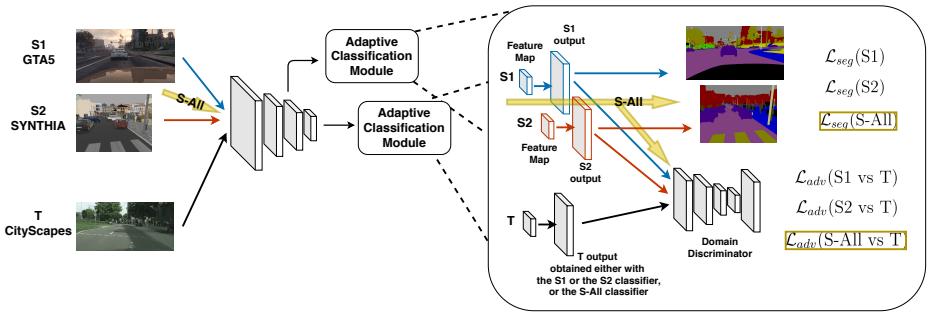

Fig. 1. Training Phase: our network has two *Adaptive Classification Modules* at different levels. In each module the source segmentation is predicted either with two separate source-specific branches or just using one overall *S-All* branch (we did not explicitly draw the *S-All* branch to avoid cluttering the image). The segmentation loss is computed based on the sources ground truth. Moreover a domain discriminator is used adversarially to reduce the domain shift comparing the target T either with each source-specific output, or with the output obtained by *S-All*.

Fig. 2. Test Phase: each classifier produces a semantic segmentation output (*S1*:blue, *S2*:red, *S-All*:yellow). For every pixel we apply a max-pooling operator over the three outputs. Finally the class assigned to the pixel is the one with the highest score over the C classes ($C = 19$ when testing on Cityscapes).

Other alternative strategies have focused on the introduction of critic networks to identify samples close to the classification boundary and exploit them to improve feature generalization [20], or defined a curriculum adaptation to focus first on easy and then on hard samples during the learning process [24], or even introduced tailored loss functions [28].

Our work is orthogonal to all those research efforts. Indeed up to our knowledge, none of the mentioned previous works have investigated the challenging case of multi-source adaptive semantic segmentation. We build over the multi-level approach presented in [21] and extend it to tackle two different sources and one target domain. Moreover, we investigate the effect of integrating a further pixel-level adaptive approach originally presented for unsupervised image style transfer [11] to further reduce the domain shift.

3 Method

An overall view of the proposed architecture can be seen in Figure 1. Our domain adaptation method starts with a segmentation network **G** which takes

the sources annotated images (I^s, Y^s) and the unlabeled target images (I^t) as input. The network ends with an *Adaptive Classification Module* that contain separate classification branches for each source as well as a domain discriminator **D**. Each source classification branch produces a segmentation softmax output $P^s = \mathbf{G}(I^s) \in \mathbb{R}^{H,W,C}$, where (H, W) are the height and width image dimensions and C is the number of categories. The used semantic segmentation loss is

$$\mathcal{L}^s_{seg}(I^s) = -\sum_{h,w} \sum_{c=1,ldots,C} Y^s_{h,w,c} \log(P^s_{h,w,c}), \tag{1}$$

where $s = 1, 2$ for the two sources.

The domain discriminator **D** takes as input the segmentation output of both the source and target data and is optimized through the binary loss

$$\mathcal{L}_d(P) = -\sum_{h,w} (1-z) \log(\mathbf{D}(P)_{h,w,0}) + (z) \log(\mathbf{D}(P)_{h,w,1}), \tag{2}$$

with z = 0 if the sample is drawn from the target domain, and z = 1 for the sample from the source domains. Finally the adversarial loss whose gradients backpropagates on the segmentation network to maximize the confusion between P^s and P^t is

$$\mathcal{L}_{adv}(I^t) = -\sum_{h,w} \log(\mathbf{D}(P^t)_{h,w,1}). \tag{3}$$

To further improve the adaptation effect involving inner-features, another adaptive classification module is also applied to a lower network level. Thus the overall loss is

$$\mathcal{L}(I_s, I_t) = \sum_{k=feature,output} \left\{ \sum_{s=1,2} \lambda^s_{seg} \mathcal{L}^s_{seg}(I^s) + \lambda^s_{adv} \mathcal{L}^s_{adv}(I^t) \right\}_k \tag{4}$$

and the network is optimized on the basis of the following criterion

$$\max_{\mathbf{D}} \min_{\mathbf{G}} \mathcal{L}(I_s, I_t). \tag{5}$$

We also repeated the whole training considering a single source branch that sees all the images together regardless of the domain identity: we indicate it as *S-All*, with its own $\mathcal{L}^{S-All}_{seg}$ loss. From the predictions of each available source and from *S-All*, we finally need a single segmentation target output. For this purpose we apply a max-pooling operator that runs on the prediction logits $\hat{Y}$ and selects the highest score per class, then followed by a second max-pooling over the classes:

$$\text{Assigned Label}(h, w) = \max_{c=1\ldots,C} \max_{s=\{1,2,S-All\}} (\hat{Y}^s_{h,w,c}). \tag{6}$$

Note that, by keeping only *S-All* we fall back to the single source original method in [21].

3.1 Adding Pixel-level Adaptation

As explained above the proposed adaptation process is applied both at the output and at the feature level. Inspired by the extensive GAN-based literature on style-transfer, we integrated in our method also a pixel-level adaptation process, directly modifying the input images. Specifically we used the Unsupervised Image-to-Image Translation (UNIT, [11]) method. It assumes that a pair of corresponding images in two different domains can be mapped to the same latent code in a shared space. By using a Coupled GANs [12] and imposing weight sharing constraints on the mapping functions, the method is able to change the style of an image so that it looks like coming from a different domain. We applied UNIT to produce target-like copies of the source images. After this (totally unsupervised) pre-processing step, the proposed architecture is used on the new stylized sources.

4 Experiments

4.1 Datasets and Setup

We used three publicly available datasets in our experiments as detailed in the following.

Cityscapes [3] is a real-world, vehicle-egocentric image dataset collected in 50 cities in Germany and nearby countries. It provides a training set made of 2,993 images as well as 503 images for validation purpose, having 2048×1024 resolution. All the training, validation, and test images are accurately annotated with per pixel category labels by human experts. We followed the VisDA Semantic Segmentation challenge protocol, focusing on 19 labeled classes.

GTA5 [17] is composed by 24,966 images with resolution 1914×1052, synthesized from the homonym video game and set in Los Angeles. Ground truth and annotations are compatible with the Cityscapes dataset [3] that contains 19 categories. Depending on the role of the dataset in the experiments we used either all the available images (as source) or a 500 sample subset (as target).

Synthia [18] is made of 9400 images at 1280×760 resolution compatible with the Cityscapes dataset, but covering only 16 object categories. Even if the virtual city used to generate the synthetic images does not correspond to any of the real cities covered by Cityscapes, Synthia shows almost photo-realistic frames with different light conditions and weather, multiple season, and a great variety of dynamic objects. With the same approach of GTA5, we used the full dataset for training and the first 500 images while testing.

We ran each experiment by choosing two datasets as sources domains, and the third as target (unsupervised) domain. In previous works, the standard setting consists in evaluating the recognition performance only of the shared classes across domains, thus operating a subselection on Cityscapes when used against Synthia. We find it natural that different data collections may have only partially overlapping class sets and it should not be necessary to proceed every time to an ad-hoc class choice [22]. Thus, we decided to keep all the datasets with their

own original categories. Furthermore we investigate the effect of the resolution on the final segmentation accuracy considering a high and a low resolution case. In the first, all the images keep their own original size, while in the second they are all downscaled by halving the native image dimensions. Finally we remark that the three analyzed domains present remarkable differences on mean values. Since the adversarial approaches are very sensitive to non-zero mean data, we have chosen to work by removing from each dataset its own calculated image mean.

4.2 Implementation details

The main backbone of our segmentation network is the DeepLabv2 [1], which uses a ResNet-101 pretrained on ImageNet and COCO [10]. This architecture incorporates atrous convolution, which effectively enlarge the field of view of filters without increasing the number of parameters. Within the *Adaptive Classification Module* we have two separate network branches, one for each source, producing a 2D predictions followed by an interpolation function that rises the resolution to that of the original ground truth label (during training). At test time the same interpolation function was used to calculate accuracy using the target ground truth as reference. Following [21, 29], the module contains also a discriminator that classify the images on the basis of their source or target domain label. The discriminators model is the same of DCGAN [16], with convolutional layers interspersed by Leaky Relu non-linearities. Note that although there are two adaptive classification modules in the network, the classification output produced by the inner module has shown to be less reliable than the ending one which is actually the only used at test time.

The network is trained with the Adam solver and learning rate 0.0001, while for the architecture hyperparameters we kept the same values of [21]. The number of iterations was set to 50k, but we observed convergence already after 20k iterations.

4.3 Results

The main experiment results are reported in Table 1. The values reported are the mean Intersection Over Union (mIoU) which is the standard accuracy measurement used on semantic segmentation tasks.

The proposed method is able to improve the *S-All* results on almost all the performed experiments, even while the single source branch prove to reach lower accuracy w.r.t. the *S-All* result, getting a boost ranging from 0.4% to 1.2%, while w.r.t. the results without any adaptation at all (No Adapt column) the difference of performance are from 1.5% to 2.7%. Looking more into detail, the most difficult setting is the one with GTA5 as target domain, as the Synthia source domain fails to properly reach an acceptable accuracy, and this worsen the final performance in the full resolution case. The input data resolution has an impact on final accuracy ranging from 1.44% in the case of Cityscapes as target, to 4.41% in the case of Synthia target, showing that in order to obtain

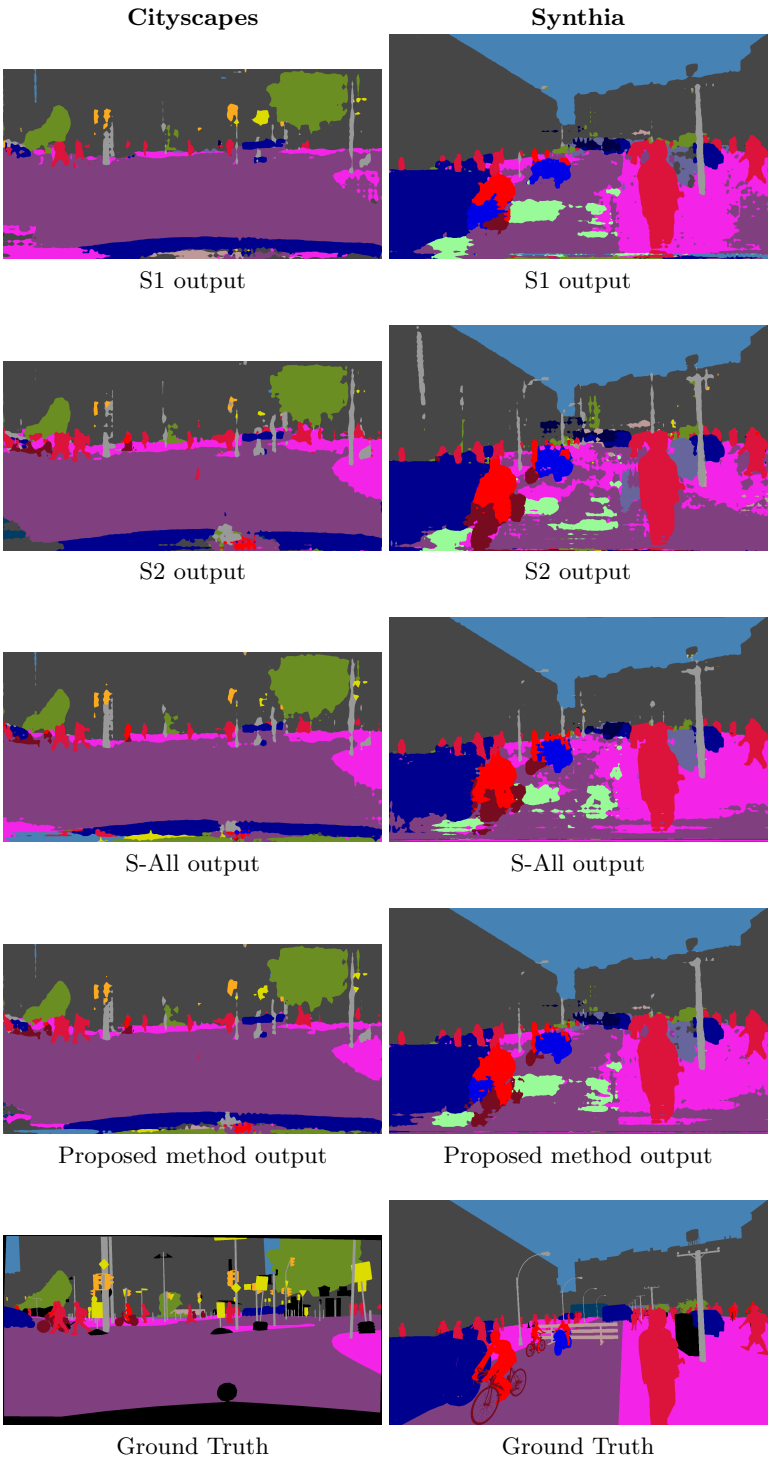

Fig. 3. Predicted labels in the case of Cityscapes and Synthia target datasets. The proposed method is able to better recognize some parts of the images like road pieces (dark violet) w.r.t. single branches or *S-All* approach.

Table 1. Performance values on the chosen experiments expressed with mIoU. The proposed method outperform the no adaptation results as well as single branches and *S-All* method on all the experiments but the one with GTA5 as target at high resolution, where it lags behind *S-All* result due to the poor performance of S2 branch.

Res	Sources	Target	No Adapt	*S1*	*S2*	*S-All*	Max Merge
Orig.	GTA5, Synthia	Cityscapes	39.98	39.55	34.51	41.81	**42.76**
	Cityscapes, GTA5	Synthia	35.55	35.25	34.07	36.37	**37.52**
	Cityscapes, Synthia	GTA5	37.97	41.17	23.60	**40.57**	39.49
Redu.	GTA5, Synthia	Cityscapes	-	39.44	33.36	40.89	**41.32**
	Cityscapes, GTA5	Synthia	-	30.52	30.02	32.87	**33.11**
	Cityscapes, Synthia	GTA5	-	44.93	23.28	41.87	**42.78**

Table 2. Intersection over Union for each experiment category. The experiments are performed on full resolution. Some particular categories (road, terrain, cars) seems to better exploit the power of the proposed method w.r.t the *S-All* one, and they contribute to the final accuracy increase due to their frequent presence on the scene.

Setting	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motocycle	Bycycle
T: Cityscapes, *S-All*	85.0	36.3	79.9	21.5	18.0	29.5	25.5	19.3	81.4	23.5	78.0	57.6	23.9	75.4	35.0	40.7	2.6	31.7	29.9
T: Cityscapes, Max Merge	87.8	37.1	80.2	20.3	14.9	29.8	26.0	20.5	82.0	31.4	78.0	57.6	25.3	80.5	31.6	43.5	0.0	29.8	36.1
T: GTA5, Max Merge	82.4	29.4	56.5	41.6	6.7	31.1	26.4	19.3	64.4	7.4	88.1	42.8	50.3	74.2	36.8	31.4	0.0	32.5	29.3
T: Synthia, Max Merge	68.3	66.8	86.6	1.7	1.7	39.5	27.2	10.9	73.7	0.0	90.6	55.5	33.7	55.7	0.0	48.5	0.0	23.0	29.3

the best possible accuracy is preferable to keep resolution as high as possible, while at the same time demonstrates that in some cases a lower resolution can dramatically speed up the training phase (around 3x faster in our case) while losing a small amount of accuracy (target Cityscapes experiment).

Looking at per-class IoU measurements in table 2, we noticed how the overall increase of performance can be attributed to some specific classes IoU improvement; terrain, road, vegetation and car seem to be the classes which better take advantage of the proposed method. This effect can be noticed also in the produced images in Figure 3, where some parts of the road are better reproduced in our method w.r.t *S-All* output.

A final additional experiment have been performed by applying UNIT method to the GTA5 and Synthia datasets in order to convert their style to the Cityscapes one, after which the proposed architecture have been trained regularly with two stylized GTA5 and Synthia datasets as sources and Cityscapes as target. The measured accuracy obtained by merging the two branches *S1* and *S2* is 44.5%, which is very promising result, taking also into account that it can be further improved by exploiting *S-All* branch too. The UNIT architecture and our method have been trained separately because of the huge amount of GPU memory required in order to train them jointly.

5 Conclusions

We have presented a study on multi-sources domain adaptation on semantic segmentation tasks. The study revealed how simply putting all the sources together is a sub-optimal approach, and we proposed a simple method to leverage on individual sources as well as *S-All* method. The experiment performed show promising results, with a small but steady improvement on the majority of settings. Further investigation is required in order to better understand the effect of some parameters like the chosen data resolution and the datasets means, and the possibility of applying a style transfer method like UNIT jointly with the domain adaptation method into a fully integrated architecture.

References

1. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 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. Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *ICCV*, 2017.
3. Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
4. Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Conference on Robot Learning (CoRL)*, 2017.
5. Lixin Duan, Ivor W. Tsang, Dong Xu, and Tat-Seng Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *International Conference on Machine Learning (ICML)*, 2009.
6. Lixin Duan, Dong Xu, and Ivor W. Tsang. Domain adaptation from multiple sources: A domain-dependent regularization approach. *IEEE Transactions on Neural Networks and Learning Systems*, 23(3):504–518, March 2012.
7. Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018.
8. A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017.
9. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
10. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

11. Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.
12. Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016.
13. Wei Liu, Andrew Rabinovich, and Alexander C. Berg. Parsenet: Looking wider to see better. 2016.
14. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
15. Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015.
16. Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
17. Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016.
18. German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
19. Paolo Russo, Fabio Maria Carlucci, Tatiana Tommasi, and Barbara Caputo. From source to target and back: symmetric bi-directional adaptive gan. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.
20. Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. In *ICLR*, 2018.
21. Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
22. Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.
23. Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.
24. Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2020–2030, 2017.
25. Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, 14(2):119–135, Apr 2017.
26. Han Zhao, Shanghang Zhang, Guanhang Wu, Jo ao P. Costeira, José M. F. Moura, and Geoffrey J. Gordon. Multiple source domain adaptation with adversarial learning. In *Workshop of the International Conference on Learning Representations (ICLR-W)*, 2018.
27. Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.
28. Xinge Zhu, Hui Zhou, Ceyuan Yang, Jianping Shi, and Dahua Lin. Penalizing top performers: Conservative loss for semantic segmentation adaptation. In *ECCV*, 2018.

29. Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018.