UDA for Semantic Segmentation with Contrastive Learning

Zeynep Gerem
zeynep.gerem@tum.de

Anurag Singh
anurags.it@nsit.net.in

Abstract

Domain adaptation approaches focus on performing a given task like image classification, segmentation, etc. in a target domain with no (or very less) supervision, by utilising supervision in a related source domain. In this work, we aim to explore the contrastive learning for the task of unsupervised domain adaptation in semantic segmentation. Firstly, we aim to translate the source images to target and target images to source by replacing the low frequencies in the source domain with the target domain images using Fourier transform. This helps us achieve a make-shift translation of the source domain images into the target domain and vice-a-versa. We propose a pixel level contrastive loss using pseudo labels on each image and translated image pair to train our model in addition to the cross-entropy loss for the source image and its translated target image.

1. Introduction

In presence of large scale labeled data, deep neural networks have shown impressive performance on various computer vision tasks such as, image classification [7], segmentation [9], etc. However, these networks cannot achieve the same performance when they are tested on data from a different distribution. In domain adaptation, we aim to improve the network in a way that it can perform similarly in these cases (DA) [1, 16].

Unsupervised DA aims to adapt the model trained with labeled source data to operate on unlabeled target data. State-of-the-art UDA methods [18], [4] use adversarial training methods for the task. Instead, we use the method proposed in [21] which simply does a spectral transfer between source and target data for semantic segmentation.

In this work, we propose a method that uses pixel-wise contrastive loss together with a cross entropy loss for UDA task. Instead of doing the spectral transfer in one way as in [21], we apply it for both source and target images. in different settings. Since this transfer does not change semantic content, the transferred images should have the same semantic maps. The algorithm aims to minimize the difference between generated semantic maps for these images and achieve high performance on the target data as well.

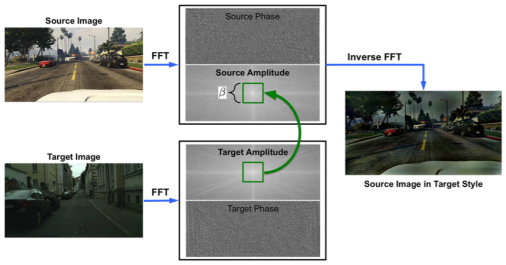

Figure 1. The source image is stylized into the target image by overwriting the low frequencies of the source image with target image in the spectral domain. This helps in style transfer in semantics preserving manner.

2. Related Work

Here, we discuss the related work in DA, semantic segmentation and also UDA in semantic segmentation.

Domain Adaptation: Domain adaptation is to build a model which performs nearly the same for the source and the target domain. To reduce discrepancy between datasets, studies have been done with adversarial training [18], [4]. However, stabilizing adversarial training is a hard task itself.

Semantic Segmentation: Semantic Segmentation has been explored for a long time and is still a demanding area. Since collecting data and labeling them manually for segmentation is hard, there have been recent studies using synthetic data [14]. In order to benefit from the synthetic data, UDA methods are explored for semantic segmentation task [21].

3. Spectral Transfer

Authors in [21] describe a method to map a source image to a target "style" without altering semantic content. A randomly sampled target image provides the style by swapping the low-frequency component of the spectrum of the source image with its own. The outcome "source image in target style" shows a smaller domain gap perceptually. It is also explained in form of a teaser in Fig. 1 taken from

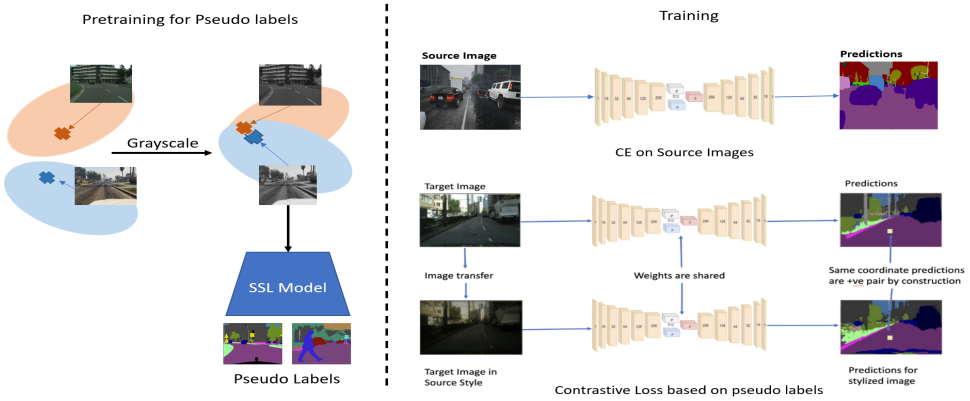

Figure 2. The above block diagram depicts the core idea of our approach and also the losses involved in training of our model

their work. Let $\mathcal{F}_A$ and $\mathcal{F}_P; \mathbb{R}^{HxWx3} \rightarrow \mathbb{R}^{HxWx3}$ be the amplitude and phase part of the fourier transform $\mathcal{F}$. Then $\mathcal{F}(I)$ represents Fourier transform of image I in the spectral domain. A mask M is introduced with a hyper parameter $\beta \in (0, 1)$ as $M_\beta = \mathbf{1}_{(h,w)\in[-\beta H:\beta H,-\beta W:\beta W]}$. where H, W are height and width of the image. Now the style transferred images can be obtained as follows:

$$I_{s\rightarrow t} = \mathcal{F}^{-1}[\mathcal{F}(I_t) \circ M_\beta + (1 - M_\beta) \circ \mathcal{F}(I_s)] \quad (1)$$

Where I_s and I_t are images from source and target domain respectively and $I_{s\rightarrow t}$ is source image in target style.

4. The proposed method

In our method, we train two models separately. The first model is trained to get pseudo labels and second model is trained with these pseudo labels in addition to the ground truth labels we have for source domain dataset.

4.1. The first model

In order to reduce the domain gap between source and target datasets, we train the first model with grayscale images from both domains. To train on target data, we perform entropy minimization using the loss

$$\mathcal{L}_{ent} = \sum_i \rho(- < \hat{I}_i^t, log(\hat{I}_i^t) >) \quad (2)$$

Where $\hat{I}^t$ is output of the model for I_t and $\rho(x) = (x^2 + 0.001^2)^\eta$ with hyperparameter η. In addition to this, we use cross entropy loss on source images. Together with scaling factor λ_{ent}, the overall loss function is

$$\mathcal{L} = \lambda_{ent}\mathcal{L}_{ent} + \mathcal{L}_{CE}^s \quad (3)$$

After training, we pass the target images through the model to obtain pseudo labels. Although we use grayscale images, there is still a domain gap between source and target sets. Comparison of mean intersection over union (mIoU) scores obtained from original and translated target images is given in the first section of Table 2. Also, qualitative comparison of pseudo labels with ground truth can be seen in Fig. 4.

4.2. The second model

To train the second model, we use contrastive loss in addition to cross entropy loss on source images in target style and entropy minimization on target images. The idea behind the contrastive learning is that for two images, the pixels from the same class are positives and the rest are negative. With contrastive learning, we aim to make positive pairs closer in embedding space and make negative pairs apart.

Let $I_{t\rightarrow s}$ denote a target image in source style obtained using spectral transfer described in the previous section. Taking inspiration from [22] let y_p^I denote the class label of pixel p in image I. Let N_c^I denote the number of pixel in I with pseudo label class c, and N^I denote the total number of pixels in I. Let f_p^I be a d-dimensional, unit-normalized feature extracted from I at pixel p. Let $\mathbf{1}_{pk}^{AB} = \mathbf{1}[y_p^A = y_k^B]$ and $e_{pq}^{AB} = exp(f_p^A.f_q^B/\tau)$ where τ is temperature hyperparameter. The contrastive loss between target image I_t and the translated image $I_{t\rightarrow s}$ is given by:

$$\mathcal{L}_{t,t\rightarrow s} = -\frac{1}{N^I}\sum_{p=1}^{N^I}\frac{1}{N_{y_p^I}^{I_{t\rightarrow s}}}\sum_{q=1}^{N^{I_{t\rightarrow s}}}\mathbf{1}_{pq}^{I_t I_{t\rightarrow s}} log(\frac{e_{pq}^{I_t I_{t\rightarrow s}}}{\sum_{k=1}^{N^{I_{t\rightarrow s}}} e_{pk}^{I_t I_{t\rightarrow s}}}) \quad (4)$$

Since we do not have the labels for target domain images, we use pseudo labels obtained from the first model to define positive and negative pairs.

After summing this loss with cross entropy on translated source data and entropy on target data, the overall loss for training of the second model becomes

Method	*road*	*sidewalk*	*building*	*wall*	*fence*	*pole*	*light*	*sign*	*vegetation*	*terrain*	*sky*	*person*	*rider*	*car*	*truck*	*bus*	*train*	*motocycle*	*bicycle*	mIoU
VGG16 backbone																				
CBST [23]	66.7	26.8	73.7	14.8	9.5	28.3	25.9	10.1	75.5	15.7	51.6	47.2	6.2	71.9	3.7	2.2	5.4	18.9	32.4	30.9
SIBAN [10]	83.4	13.0	77.8	20.4	17.5	24.6	22.8	9.6	81.3	29.6	77.3	42.7	10.9	76.0	22.8	17.9	5.7	14.2	2.0	34.2
Cycada [6]	85.2	37.2	76.5	21.8	15.0	23.8	22.9	21.5	80.5	31.3	60.7	50.5	9.0	76.9	17.1	28.2	4.5	9.8	0	35.4
AdvEnt [19]	86.9	28.7	78.7	28.5	25.2	17.1	20.3	10.9	80.0	26.4	70.2	47.1	8.4	81.5	26.0	17.2	18.9	11.7	1.6	36.1
DCAN [20]	82.3	26.7	77.4	23.7	20.5	20.4	30.3	15.9	80.9	25.4	69.5	52.6	11.1	79.6	24.9	21.2	1.30	17.0	6.70	36.2
CLAN [11]	88.0	30.6	79.2	23.4	20.5	26.1	23.0	14.8	81.6	34.5	72.0	45.8	7.9	80.5	26.6	29.9	0.0	10.7	0.0	36.6
LSD [15]	88.0	30.5	78.6	25.2	23.5	16.7	23.5	11.6	78.7	27.2	71.9	51.3	19.5	80.4	19.8	18.3	0.9	20.8	18.4	37.1
BDL [8]	89.2	40.9	81.2	29.1	19.2	14.2	29.0	19.6	83.7	35.9	80.7	54.7	23.3	82.7	25.8	28.0	2.3	25.7	19.9	41.3
FDA-MBT [21]	86.1	35.1	80.6	30.8	20.4	27.5	30.0	26.0	82.1	30.3	73.6	52.5	21.7	81.7	24.0	30.5	29.9	14.6	24.0	42.2
Ours	92.79	55.48	74.99	7.17	21.35	11.39	25.65	39.98	80.51	26.35	84.28	38.34	32.72	70.8	18.95	25.74	28.76	37.91	37.03	**42.64**
ResNet101 backbone																				
AdaStruct [17]	86.5	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5	41.4
DCAN [20]	85.0	30.8	81.3	25.8	21.2	22.2	25.4	26.6	83.4	36.7	76.2	58.9	24.9	80.7	29.5	42.9	2.5	26.9	11.6	41.7
DLOW [5]	87.1	33.5	80.5	24.5	13.2	29.8	29.5	26.6	82.6	26.7	81.8	55.9	25.3	78.0	33.5	38.7	0.0	22.9	34.5	42.3
Cycada [6]	86.7	35.6	80.1	19.8	17.5	38.0	39.9	41.5	82.7	27.9	73.6	64.9	19	65.0	12.0	28.6	4.5	31.1	42.0	42.7
CLAN [11]	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
ABStruct [2]	91.5	47.5	82.5	31.3	25.6	33.0	33.7	25.8	82.7	28.8	82.7	62.4	30.8	85.2	27.7	34.5	6.4	25.2	24.4	45.4
AdvEnt [19]	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
BDL [8]	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
FDA	90.0	40.5	79.4	25.3	26.7	30.6	31.9	29.3	79.4	28.8	76.5	56.4	27.5	81.7	27.7	45.1	17.0	23.8	29.6	44.6
FDA-MBT [21]	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	**50.45**
Ours	92.68	53.3	78.22	23.44	17.81	25.29	15.67	36.52	79.85	34.42	82.8	45.88	31.82	83.67	47.95	42.76	30.61	17.59	41.66	46.42

Table 1. Quantative comparison on GTA-5→Cityscapes with VGG16 and ResNet 101 backbone where FDA-BMT refers to the entire method proposed in [21] and FDA refers to single setting training.

$$\mathcal{L}_{total} = \mathcal{L}_{CE}^{s \to t} + \mathcal{L}_{t,t \to s} + \lambda_{ent} \mathcal{L}_{ent} \tag{5}$$

The overall idea of our method illustrated in Fig. 2. For ENet backbone, comparison of our method with FDA is given in the second part of Table 2. Using ground truth labels in contrastive learning provides us an upper bound and pseudo labels gives a similar performance. Since ENet is a relatively small model and trained from scratch, minority classes such as lights, have 0 mIoU score in this table. Results for VGG16 and DeepLab are given in Table 1.

5. Datasets and Training

We try to extensively study and evaluate our method on a challenging synthetic-to-real world UDA task benchmark GTA5 [13] to Cityscapes [3]. Where GTA5 is synthetic (source) dataset with abundant semantic segmentation labels and Cityscapes is a real world semantic segmentation (target) dataset whose labels are not used during adaptation. GTA5 has 25k synthetic images taken from a video game, with the resolution of 1914×1052. we resize the images labels to 1280×720, also randomly cropping them to 1024×512. There are 33 classes in GTA5 of which 19 common classes to cityscapes are used for DA. CityScapes has 2,975 images from the training set as the target, resized to 1024×512, with no random cropping. We test on the 500 validation images with dense manual annotations.

Training: Our implementation is based in Pytorch Framework [12]. For all our experiments we use a GTX1080 Ti GPU where we set our batch size to 1 for all source and target dataloaders due to memory limitations. We train ENet backbone experiments for 30k iterations with learning rate of 2.5e-4, and adjusted according to the 'poly' learning rate scheduler with a power of 0.9, and weight decay 0.0005. To train DeepLabV2 with ResNet101 using SGD, we use same settings as ENet backbone. For FCN-8s with VGG16, we use ADAM with the initial learning rate 1e-5, which is decreased by the factor of 0.1 every 50000 steps until 150000 steps. The momentum for Adam is 0.9 and 0.99.

6. Results

GTA→Cityscapes: We benchmark our method on GTA5→Cityscapes using two different backbones VGG and Resnet101. The results for which are reported in the table 1. We observed that our complete method ourperforms FDA-MBT baseline on VGG backbone and outperforms single model FDA on Resnet101 backbone by significant margin. The results for Resnet101 backbone are poorer in comparison to FDA-MBT possibly because the final layer in Resnet101 is 65×129 hence we resize our pseudo labels to 65×129 for employing contrastive loss which could result into loss of finer details.

Pseudo labels: We try to evaluate and benckmark our training of the first model on the Cityscpaes validation set for GTA5→Cityscapes setting on ResNet101 backbone in table 2. We observe that evaluation of unseen target images in source style significantly outperforms the evaluation of target images in their original style. Roughy by

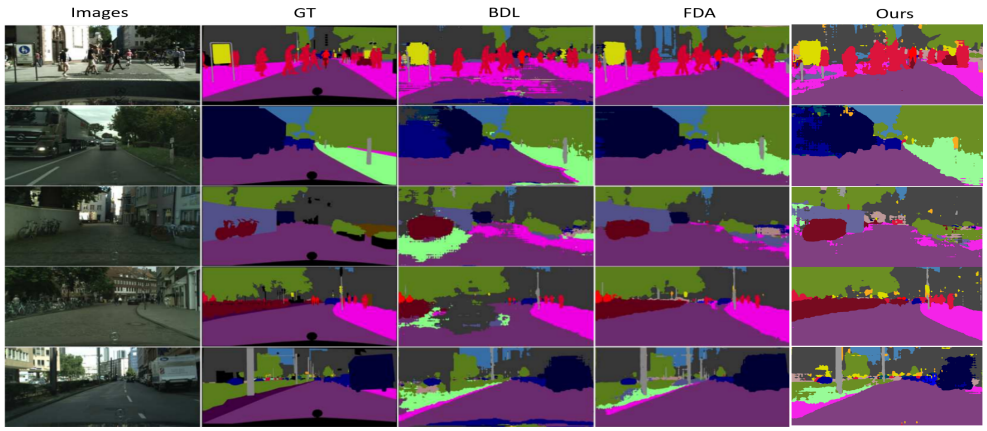

Figure 3. Visual Comparison. Left to right: Input image from CityScapes, ground-truth semantic segmentation, BDL [8], FDA-MBT [21] and Ours. Note that the predictions from Our model and are generally smoother than BDL.

Method	*road*	*sidewalk*	*building*	*wall*	*fence*	*pole*	*light*	*sign*	*vegetation*	*terrain*	*sky*	*person*	*rider*	*car*	*truck*	*bus*	*train*	*motocycle*	*bicycle*	mIoU
Trg in Trg	78.98	41.19	65.04	13.87	16.73	22.16	22.32	34.33	68.18	25.79	58.69	53.01	30.18	77.76	12.88	31.3	4.7	12.34	45.46	37.63
Trg in Src	89.42	40.66	76.9	13.96	22.18	24.03	21.51	39.23	80.78	24.08	81.87	48.06	33.16	80.88	34.43	42.23	26.33	15.59	45.99	44.28
Pseudo Labels	89.12	44.14	78.71	13.21	27.79	24.34	25.42	40.08	79.84	27.79	85.43	52.06	40.39	81.18	28.33	37.59	29.78	25.99	48.09	46.28
Results of the first model with DeepLab on Cityscapes																				
FDA ($\beta = 0.01, \lambda_{ent} = 0$)	29.45	14.82	40.5	0.17	0.0	0.11	0.0	0.0	16.47	0.48	65.48	0.0	0.0	12.97	0.28	0.0	0.0	0.0	0.0	9.51
FDA ($\beta = 0.01$)	82.42	2.96	59.78	0.04	0.0	0.02	0.0	0.0	64.96	18.46	61.65	0.0	0.0	48.11	0.87	0.0	0.0	0.0	0.0	17.86
FDA ($\beta = 0.05$)	82.41	13.73	56.1	0.01	0.0	0.77	0.0	0.0	65.06	5.01	62.1	0.0	0.0	33.86	2.25	0.0	0.0	0.0	0.0	16.91
FDA ($\beta = 0.09$)	72.07	22.07	56.32	0.66	0.0	2.17	0.0	0.0	63.37	12.83	57.97	0.0	0.0	30.11	1.46	0.0	0.0	0.0	0.0	16.79
CL using PL (Ours)	80.11	27.72	72.31	1.61	0.0	0.0	0.0	0.0	71.43	0.38	77.47	0.0	0.0	61.33	0.0	0.0	0.0	0.0	0.0	20.65
CL using GT	90.53	48.83	62.93	0.01	0.0	0.0	0.0	0.0	69.3	0.0	77.53	0.0	0.0	65.06	0.0	0.0	0.0	0.0	0.0	21.8
Ablation on ENet backbone																				

Table 2. Ablation of the first and the second model for GTA-5→Cityscapes

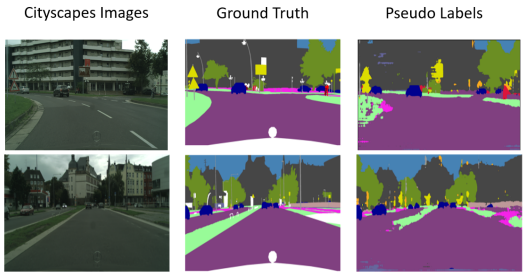

Figure 4. Qualitative comparison of pseudo labels with ground truth on Cityscapes dataset obtained from the first model.

6mIoU.

Ablation: In second part of Table 2 we analyse the effect of hyper-parameter β and training without entropy minimization $\lambda_{ent} = 0$ using ENet as a backbone. It can be observed that larger β improves the performance and for ENet and entropy minimization significantly improves the performance. Our model performs comparatively well as compared to contrastive loss with target ground truth data, which is upper-bound for our method. The models perform consistently poorly on minority classes due to no re-weighting being applied. Each component of our model adds to the performance of our model as demonstrated in

Components			mIoU
CE	CL	$\lambda_{ent} = 0$	
	✓		46.42
✓			45.42
			45.01
		✓	44.64

Table 3. Ablation on Deeplab backbone

Table 3. Where we see effect of contrastive loss and entropy minimization on overall performance. The contrastive loss uses pseudo labels (mIoU 44.28) from the first model and has an improvement of 1.5 mIoU over baseline.

Qualitative visualization: We perform qualitative analysis of our model on deeplab backbone for direct comparison with two recent methods BDL [8] and FDA-MBT [21] which is referred to as FDA in the Figure 3 for brevity. We observe that our model output is more consistent than BDL and qualitatively comparable to FDA in all settings. We also analyse pseudo labels generated by our model wrt. ground truth on cityscapes training data in Table 2.

7. Conclusion

We have propose a simple method for domain alignment using contrastive learning, and can be easily integrated into a learning system that transforms unsupervised do-

main adaptation into semi-supervised learning. Our method does not involve learning any mapping like Cycada [6] and is comparable to other style transfer methods BDL [8] & FDA [21].

References

[1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. In *Machine learning, 79(1-2)*, page 151–175, 2010. 1

[2] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *CVPR*, 2019. 3

[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. 3

[4] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 1

[5] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Domain flow for adaptation and generalization. In *CVPR*, 2019. 3

[6] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 3, 5

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[8] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, 2019. 3, 4, 5

[9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[10] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *ICCV*, 2019. 3

[11] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, 2019. 3

[12] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 3

[13] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 3

[14] 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 *CVPR*, 2016. 1

[15] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *CVPR*, 2018. 3

[16] Tasfia Shermin, Guojun Lu, Shyh Wei Teng, Manzur Murshed, and Ferdous Sohel. Adversarial network with multiple classifiers for open set domain adaptation. *IEEE Transactions on Multimedia*, 2020. 1

[17] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 3

[18] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 1

[19] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. Advent: Adversarial entropy min´imization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 3

[20] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *ECCV*, 2018. 3

[21] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 1, 3, 4, 5

[22] Xiangyun Zhao, Raviteja Vemulapalli, Philip Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label-efficient semantic segmentation. *arXiv preprint arXiv:2012.06985*, 2020. 2

[23] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 3