DADA: Depth-Aware Domain Adaptation in Semantic Segmentation (Extended Abstract)

Tuan-Hung Vu Himalaya Jain Maxime Bucher Patrick Perez
valeo.ai
Paris, France

Matthieu Cord
valeo.ai
Sorbonne University, Paris, France

Segmentation is essential for image analysis tasks and the semantic segmentation is the process of associating each pixel of an image with a set of labels like in our case, the labels are cars, bicycles, trees, roads, buildings, etc. Annotation to predict the semantic label of each pixel of the scene has always one of the popular areas of research for AI researchers. The first paper was published in 2015 for Semantic segmentation and was based on Adversarial Gradient Reversal [2]. It mainly meant for classification. This work gave an idea about using high-level training for domain adaptation. The mainstream of CNN or deep network started from the encoder to the deep representation which was followed by prediction or classification and the last part of the system was domain identification. The next paper was Adversarial feature alignment [3], this paper first gave the idea of semantic segmentation. For the source you have labels and for the target domain, you don't have labels, this paper uses domain adversarial training for semantic segmentation. After these several papers were published, like 2 networks which are trained to translate the data from one domain to another domain then converting source image to look like target domain(domain adaptation)CyCADA[3], some proposed label predictor for source domain and calculates the loss on comparing it to ground truth[1]. The novel method proposed in the paper[4] is the use of depth as privileged information along with an unsupervised domain adaptation method for semantic segmentation. Privileged information is the additional information used on source data at the time of training. It is similar to humans learning new things with the help of a teacher's comment or explanation. Depth, operating as an additional source domain supervision in their framework, thus a new depth-aware adversarial training protocol based on the fusion of network outputs has been introduced. The key objective is to categorize 'humans' and other objects. The main idea is to train a discriminator for predicting the domain of the data(source or target) while segmentation tries to fool it along with the supervised segmentation task on the source. The backbone CNN features are consecutively fed into three encoding convolutional layers, followed by an average pooling layer to output depth map predictions. On the residual path back to the main branch, the encoded features (before the depth pooling) are decoded by a convolutional layer and fused with the backbone features. For the feature-level fusion, they have adopted an element-wise product, indicated as "Feat Fusion" Fig.1. To produce segmentation predictions, they feed-forwarded the fused features through the remaining classification modules. The source domain supervised training model is trained with supervised segmentation and depth losses on the source domain [1][2][3][4]. To learn the model, the source and target is aligned so that the domain discriminator is unable to discriminate between source and target[5][6][7]. For this alignment they have produced soft segmentation map by the segmentation network on image 'x'. The depth space implicitly bridges the domain gaps of the shared lower-level CNN representation which had some improvements on the task performed on the target domain. The produced soft segmentation map P_x is further weighted. This weighted information is then fused with depth information Z_x is further fused to produce a depth aware map $\hat{I}_x$. This fused map they termed as DADA fusion which is further provided for adversarial adaptation of source domain over target domain. The losses associated with producing Soft-Segmentation map Fig.1. DADA Architecture and losses associated with adversarial learning . Obtained results in shown in Figure 2. SYNTHIA dataset used of training and testing results.

$$\pounds_{seg}(x_s, y_s) = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} y_s^{(h,w,c)} log P_{\infty_s}^{(h,w,c)} \tag{1}$$

$$\pounds_{dep}(x_s, z_s) = -\sum_{h=1}^{H}\sum_{w=1}^{W} berHu(Z_{x_s}^{(h,w)} - z_s^{(h,w)}) \tag{2}$$

$$berHu(e_z) = \begin{cases} |e_z|, & \text{if } |e_z| \le c, \\ \frac{e_z^2 + c^2}{2c} & \text{otherwise,} \end{cases} \tag{3}$$

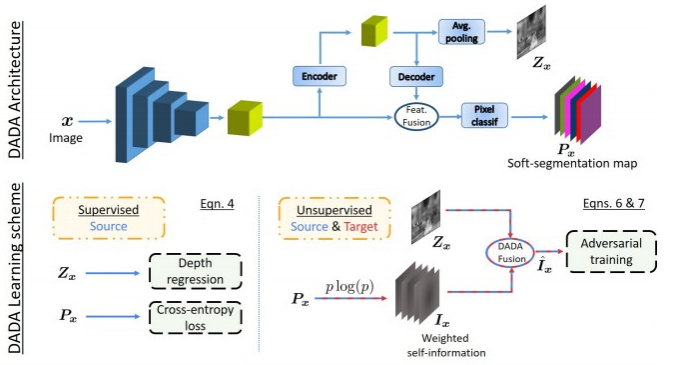

Figure 1: DADA architecture (top) and DADA learning scheme (bottom)

SYNTHIA → Cityscapes (16 classes)

Models	Depth	road	sidewalk	building	wall*	fence*	pole*	light	sign	veg	sky	person	rider	car	bus	mbike	bike	mIoU	D-Gain	mIoU*
SPIGAN-no-PI [18]		69.5	29.4	68.7	4.4	0.3	32.4	5.8	15.0	81.0	78.7	52.2	13.1	72.8	23.6	7.9	18.7	35.8	-	41.2
SPIGAN [18]	✓	71.1	29.8	71.4	3.7	0.3	**33.2**	6.4	**15.6**	81.2	78.9	52.7	13.1	75.9	25.5	10.0	20.5	36.8	1.0	42.4
AdaptSegnet [35]		79.2	37.2	78.8	-	-	-	9.9	10.5	78.2	80.5	53.5	19.6	67.0	29.5	21.6	31.3	-	-	45.9
AdaptPatch [36]		82.2	39.4	79.4	-	-	-	6.5	10.8	77.8	82.0	54.9	21.1	67.7	30.7	17.8	32.2	-	-	46.3
CLAN [23]		81.3	37.0	80.1	-	-	-	**16.1**	13.7	78.2	81.5	53.4	21.2	73.0	32.9	**22.6**	30.7	-	-	47.8
AdvEnt [39]		87.0	44.1	79.7	**9.6**	**0.6**	24.3	4.8	7.2	80.1	83.6	**56.4**	**23.7**	72.7	32.6	12.8	33.7	40.8	-	47.6
DADA	✓	**89.2**	**44.8**	**81.4**	6.8	0.3	26.2	8.6	11.1	**81.8**	**84.0**	54.7	19.3	**79.7**	**40.7**	14.0	**38.8**	**42.6**	**1.8**	**49.8**

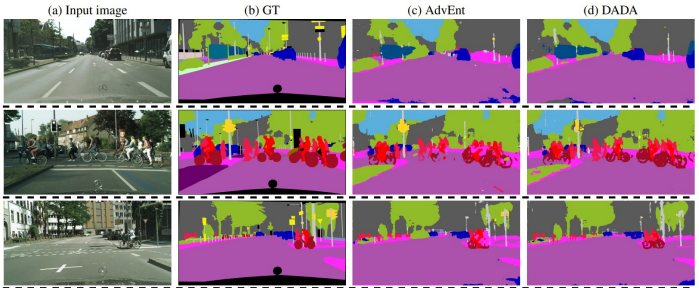

Figure 2: Results

$$\min_{\theta_{DADA}} \frac{1}{|\tau_s|}\sum_{\tau_s} \pounds_{seg}(x_s, y_s) + \lambda_{dep}\pounds_{dep}(x_s, y_s) \tag{4}$$

$$I_x^{(h,w,c)} = -P_x^{(h,w,c)}.logP_x^{(h,w,c)} \tag{5}$$

$$\min_{\theta_D} \frac{1}{|\tau_s|}\sum_{\tau_s} \pounds_D(\hat{I}_x, 1) + \frac{1}{|\chi_t|}\sum_{\chi_t} \pounds_D(\hat{I}_x, 0)] \tag{6}$$

$$\min_{\theta_{DADA}} \frac{1}{|\chi_t|}\sum_{\chi_t} \pounds_D(\hat{I}_x, 1) \tag{7}$$

References

[1] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes, 2017.

[2] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks, 2015.

[3] J. Hoffman, S. Gupta, and T. Darrell. Learning with side information through modality hallucination. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 826–834, 2016.

[4] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation, 2019.