DADA: Depth-Aware Domain Adaptation in Semantic Segmentation (Critical Review)

Tuan-Hung Vu , Himalaya Jain[1]
valeo.ai
Maxime Bucher , Patrick Perez[1]
valeo.ai
Matthieu Cord[2]
valeo.ai

[1] Paris, France
[2] Sorbonne University, Paris, France

Abstract

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 paper has presented comparatively better solution from the previous solutions for semantic segmentation. They have used unsupervised domain adaptation for the target set and have used depth as privileged information which is fused with standard CNN to obtain some more meaningful information.

1 Introduction

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. With the help of such labeling, one can infer the classes contained in the dataset effectively if such labeling are done properly.

2 Review

Previous methods achieved semantic segmentation by several methods like unsupervised domain adaptation for semantic labelling, converting training data to target like data with the help of domain adaptation[2], adversarial training in feature space[1], adversarial training for class-level alignment on grid-wise soft pseudo-labels, generative networks to turn source domain samples into target-like images etc. The paper has discussed the semantic segmentation method along with depth as the privileged information which caused effect semantic labeling. This labeling information is then utilised to adapt the target so that classification can be done further. The aim was to classify objects like humans, cycle, cars, buildings, trees etc. The updates in this domain has achieved satisfactory results for classification of objects.

The most important issues associated with such tests are collection of the data to train the model and effective annotation on target domain. Depth Estimation has improved the semantic labeling by fusing the depth information to the standard CNN architecture. Adversarial training is used to minimise the domain gap between the source and target so that, the

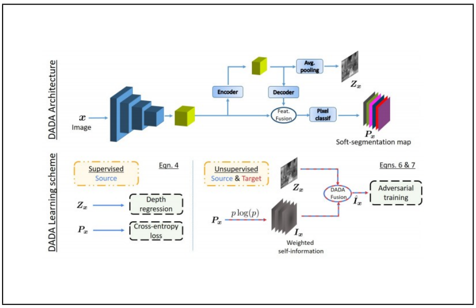
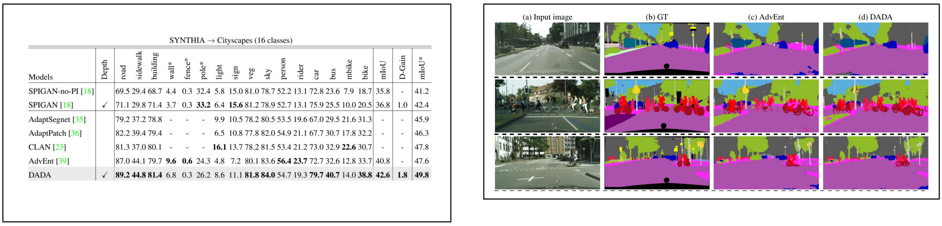


Figure 1: DADA architecture (top) and DADA learning scheme (bottom).



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**

(a) (b)

Figure 2: Left: Semantic segmentation performance mIoU (%) on Cityscapes validation set of different models trained on SYNTHIA, Right: Qualitative results in the SYNTHIACityscapes (16 classes) set-up

model can not differentiate between the source and the target. This was done by modifying the train data to look like the target data. The results obtained has impressively surpassed previous work and has classified around 80% of classes effective. Refer Figure 2.

2.1 Logic / Mathematics

For Soft Segmentation and fusion of depth information along with the standard CNN network for classification is give nas follows:

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

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

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

x_s refers to the sample S from input X, y_s refers to the Soft-Segmentation output Y associated with input S. h,w refers to the size of the image and c refers to the class from set C. The equation for segmentation loss and depth loss are given in equation 1 and 2 and the soft-segmentation cost function is given in equation 4. For learning the parameters and performing adversarial training, following equations have been used.

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

Learning Scheme associated equation are 5, 6 and 7.I_x refers to weighted self-information, Z_x depth prediction. P_x is the soft segmentation map.

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

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

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

References

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

[2] 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.