Supplementary Material of Penalizing Top Performers: Conservative Loss for Semantic Segmentation Adaptation

Xinge Zhu[1], Hui Zhou[2], Ceyuan Yang[1], Jianping Shi[2], Dahua Lin[1]

[1]CUHK-SenseTime Joint Lab, CUHK, Hong Kong S.A.R.
[2]SenseTime Research, Beijing, China
zhuxinge123@gmail.com

1 Network Blocks

In this section, we show the detailed architectures of different network blocks. As mentioned in the experimental setting, the backbone of the encoder is FCN8s-VGG16 [1]. For the generator and discriminator, we display the detailed architectures in Figure 1.

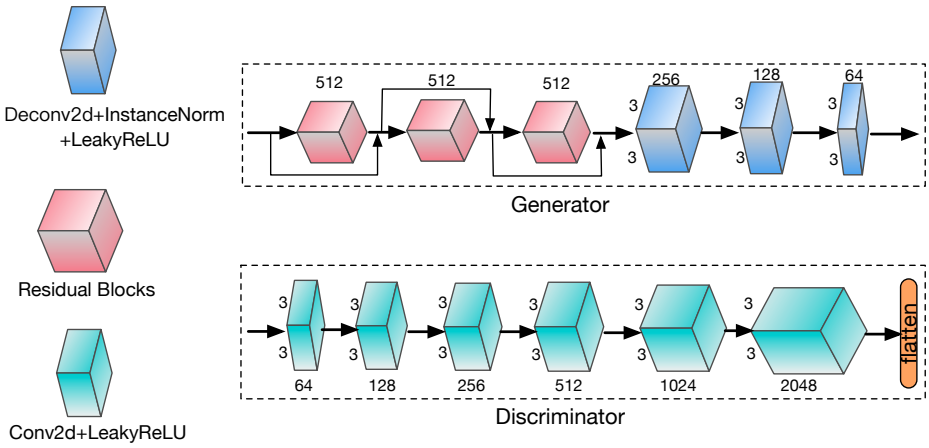

Fig. 1: Detailed architectures of network blocks in our proposed framework, including generator and discriminator. Basic components are shown in the left column. Note that the shape of each layer is labeled

2 More Training Details

Warm Start. We first use the cross-entropy loss to train the full model about 30k iterations, then use the relatively small learning rate (1e-6 $\sim$ 1e-7) to train the full model with conservative loss.

Data Augmentation. We utilize the common techniques, including crop, random scale, flip, color augmentation and random gaussian blur.

3 More Theoretical Analysis

At training time, the encoder aims to produce the feature embedding that maximize the loss of discriminator, which makes the distributions of two domains as similar as possible. At the same time, the Conservative Loss seeks a balance between the discriminativeness and domain-invariant. Formally, the overall objective function can be formulated as:

$$\mathbb{F}(\theta_e, \theta_g, \theta_d) = \mathcal{L}_d(D(G(E(\mathbf{x}; \theta_e); \theta_g); \theta_d)) + \mathcal{L}_s(S(E(\mathbf{x}; \theta_e))), \tag{1}$$

where $\theta_e, \theta_g, \theta_d$ denote the parameters of encoder, generator and discriminator. E, G, D, S represent the encoder, generator, discriminator and segmentation classifier. $\mathbf{x}$ is the input image. Note that for the segmentation part, $\mathbf{x}$ is the image from source domain, while for the GAN part $\mathbf{x}$ are source and target domain images. $\mathcal{L}_d$ and $\mathcal{L}_s$ are the loss for discriminator and loss for the segmentation classifier. Since the sign of $\mathcal{L}_s$ depends on p_t, where the high p_t produces a negative loss value and the low p_t makes a positive one, the overall objective can be restated as:

$$\mathbb{F}(\theta_e, \theta_g, \theta_d) = \begin{cases} \mathcal{L}_d(D(G(E(\mathbf{x}; \theta_e); \theta_g); \theta_d)) - \mathcal{L}_s(S(E(\mathbf{x}))), & p_t > \frac{1}{a}, \\ \mathcal{L}_d(D(G(E(\mathbf{x}; \theta_e); \theta_g); \theta_d)) + \mathcal{L}_s(S(E(\mathbf{x}))), & p_t \leq \frac{1}{a}, \end{cases} \tag{2}$$

In this setting, we are seeking the optimal parameters $\theta_e, \theta_g, \theta_d$ that deliver a balance of the function eq 2. When p_t is high, $\mathcal{L}_s$ penalizes the bias to source domain, which aims to improve the domain-invariant property via gradient ascend method. When p_t is low, $\mathcal{L}_s$ attempts to shape the discriminative features on source domain by gradient descend method. The zero point $\frac{1}{a}$ of the Conservative Loss controls the balance of the discriminativeness and domain-invariant.

4 More Experimental Results

Experimental Setting. During training, each mini-batch consists of two images, including one source image and one target image. During evaluation, the generator and discriminator are removed and the left part is a regular FCN8s-VGG16 [1]. It thus introduces no extra computational overhead.

Detailed Results. We present the detailed results of each category in Table 1 and Table 2. It can be observed that our method achieves the best results and obtains the largest performance gain, in which our model gets **8.1** points and **9.3** points gain, respectively.

Table 1: Results of domain adaptation from GTAV → Cityscapes. 19 common object categories are reported. The bold values denote the best scores in the column. NoAdapt denotes that the model is trained with source domain only

Methods	Base	road	sidewalk	building	wall	fence	pole	t-light	t-sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
NoAdapt [2]	dilated	31.9	18.9	47.7	7.4	3.1	16.0	10.4	1.0	76.5	13.0	58.9	36.0	1.0	67.1	9.5	3.7	0.0	0.0	0.0	21.1
FCNWild [2]	dilated	70.4	32.4	62.1	14.9	5.4	10.9	14.2	2.7	79.2	21.3	64.6	44.1	4.2	70.4	8.0	7.3	0.0	2.5	0.0	27.1
NoAdapt [3]	FCN8s	18.1	6.8	64.1	7.3	8.7	21.0	14.9	16.8	45.9	2.4	64.4	41.6	17.5	55.3	8.4	5.0	**6.9**	4.3	13.8	22.3
CDA [3]	FCN8s	74.9	22.0	71.7	6.0	11.9	8.4	16.3	11.1	75.7	13.3	66.5	38.0	9.3	55.2	18.8	18.9	0.0	16.8	**16.6**	28.9
Tsai *et al.* [4]	FCN8s	**87.3**	29.8	78.6	21.1	18.2	22.5	21.5	11.0	79.7	29.6	71.3	46.8	6.5	80.1	23.0	**26.9**	0.0	10.6	0.3	35.0
NoAdapt	FCN8s	78.7	15.2	73.5	15.3	12.9	18.7	22.9	**23.4**	72.5	15.0	**74.1**	38.4	4.6	58.7	10.2	10.1	0.0	12.7	14.3	30.0
Ours	FCN8s	85.6	**38.3**	**78.6**	**27.2**	**18.4**	**25.3**	**25.0**	17.1	**81.5**	**31.3**	70.6	**50.5**	**22.3**	**81.3**	**25.5**	21.0	0.1	**18.9**	4.3	**38.1**

Table 2: Results of domain adaptation from Synthia → Cityscapes. 16 common object categories are reported

Methods	Base	road	sidewalk	building	wall	fence	pole	t-light	t-sign	veg	sky	person	rider	car	bus	mbike	bike	mIoU
NoAdapt [2]	dilated	6.4	17.7	29.7	1.2	0.0	15.1	0.0	7.2	30.3	66.8	51.1	1.5	47.3	3.9	0.1	0.0	17.4
FCNWild [2]	dilated	11.5	19.6	30.8	4.4	0.0	20.3	0.1	11.7	42.3	68.7	**51.2**	3.8	**54.0**	3.2	0.2	0.6	20.2
NoAdapt [3]	FCN8s	5.6	11.2	59.6	**8.0**	0.5	21.5	8.0	5.3	72.4	**75.6**	35.1	9.0	23.6	4.5	0.5	18.0	22.0
CDA [3]	FCN8s	65.2	26.1	**74.9**	0.1	**0.5**	10.7	3.5	3.0	**76.1**	70.6	47.1	8.2	43.2	**20.7**	0.7	13.1	29.0
NoAdapt	FCN8s	53.0	22.8	53.6	2.5	0.0	10.4	0.1	1.4	71.7	73.5	39.4	1.4	50.8	12.5	0.0	4.7	24.9
Ours	FCN8s	**80.0**	**31.4**	72.9	0.4	0.0	**22.4**	**8.1**	**16.7**	74.8	72.2	50.9	**12.7**	53.9	15.6	**1.7**	**33.5**	**34.2**

5 Qualitative Segmentation Results

We display some qualitative examples in Figure 2 and 3. We show the results of FCN8s-NoAdapt, FCN8s+GAN and our full model. It can be observed that our full model performs much better than baselines.

References

1. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015) 3431–3440 1, 2
2. 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) 3
3. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: ICCV. Volume 2. (2017) 6 3
4. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR. (2018) 3

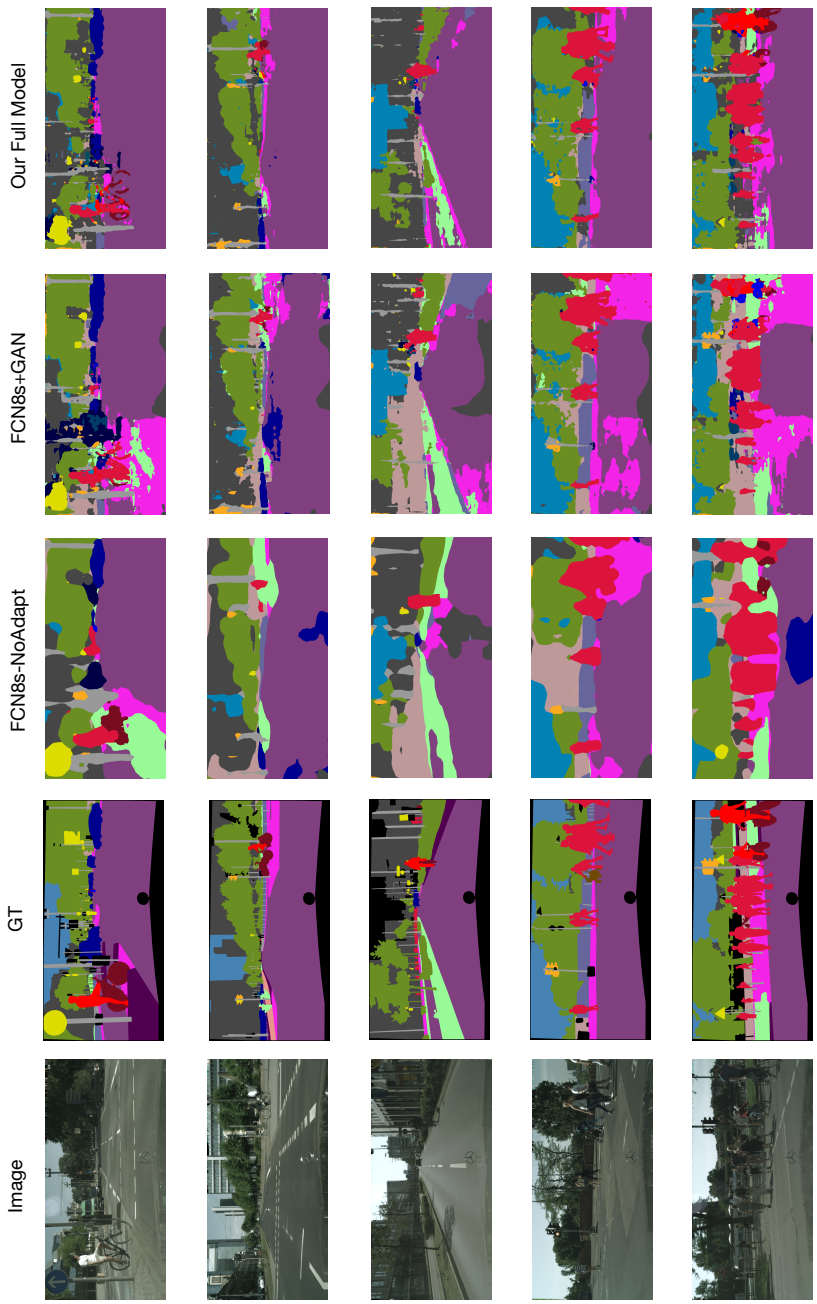

Fig. 2: We show the results of FCN8s-NoAdapt, FCN8s+GAN and our full model. It can be observed that our full model achieves the best results and the FCN8s+GAN performs better than the FCN8s trained on source domain only

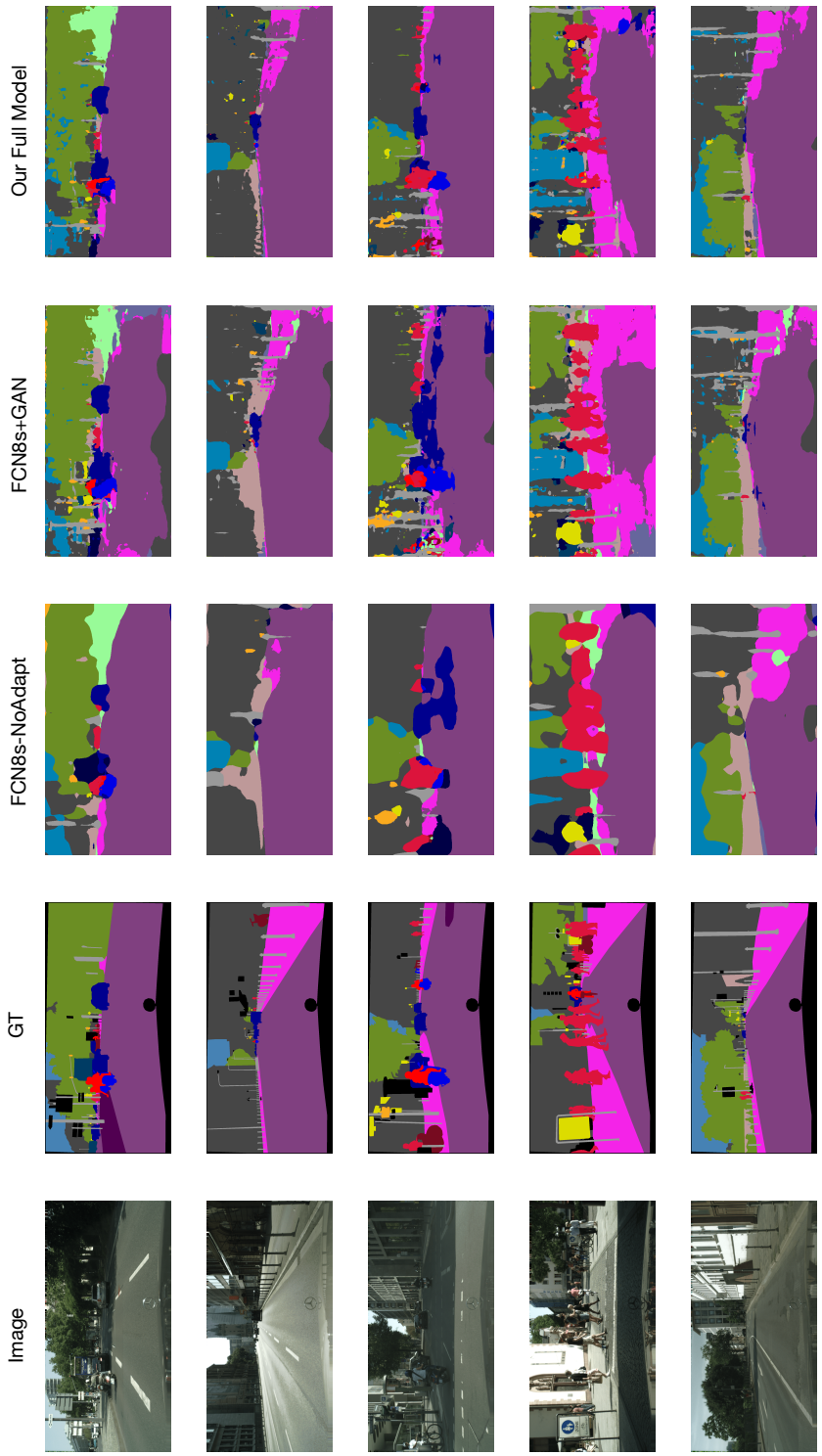

Fig. 3: We show the results of FCN8s-NoAdapt, FCN8s+GAN and our full model. It can be observed that our full model achieves the best results and the FCN8s+GAN performs better than the FCN8s trained on source domain only