Differential Treatment for Stuff and Things: A Simple Unsupervised Domain Adaptation Method for Semantic Segmentation

Zhonghao Wang[1], Mo Yu[2], Yunchao Wei[3], Rogerio Feris[2], Jinjun Xiong[2], Wen-mei Hwu[1], Thomas S. Huang[1], Honghui Shi[4,1]

[1]C3SR, UIUC, [2]IBM Research, [3]ReLER, UTS, [4]University of Oregon

Abstract

We consider the problem of unsupervised domain adaptation for semantic segmentation by easing the domain shift between the source domain (synthetic data) and the target domain (real data) in this work. Based on the observation that stuff categories usually share similar appearances across images of different domains while things (i.e. object instances) have much larger differences, we propose to improve the semantic-level alignment with different strategies for stuff regions and for things: 1) for the ***stuff*** *categories, we generate feature representation for each class and conduct the alignment operation from the target domain to the source domain; 2) for the* ***thing*** *categories, we generate feature representation for each individual instance and encourage the instance in the target domain to align with the most similar one in the source domain. In addition to our proposed method, we further reveal the reason why the current adversarial loss is often unstable in minimizing the distribution discrepancy and show that our method can help ease this issue by minimizing the most similar stuff and instance features between the source and the target domains. We conduct extensive experiments in two unsupervised domain adaptation tasks,* i.e. *GTA5 → Cityscapes and SYNTHIA → Cityscapes, and achieve the new state-of-the-art segmentation accuracy.*

1. Introduction

Semantic segmentation [11] enables image scene understanding at the pixel level, which is crucial to many real-world applications such as autonomous driving. However, collecting data with pixel-level annotation is costly in terms of both time and money. To address the problem of high-cost annotation, unsupervised domain adaptation methods are proposed for semantic segmentation [13, 14]. To address the domain shift problem, existing methods use the GAN [6] architectures to minimize the distribution discrepancy of the features extracted by a feature extractor [16, 7] between the source domain and the target domain

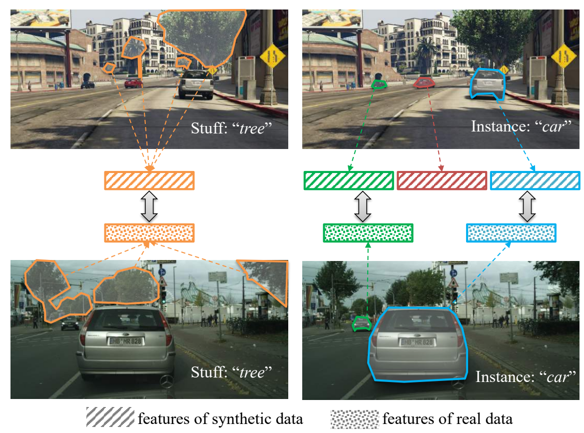

Figure 1. Illustration of the proposed Stuff Instance Matching (SIM) structure. By matching the most similar stuff regions and things (i.e., instances) with differential treatment, we can adapt the features more accurately from the source domain to the target domain.

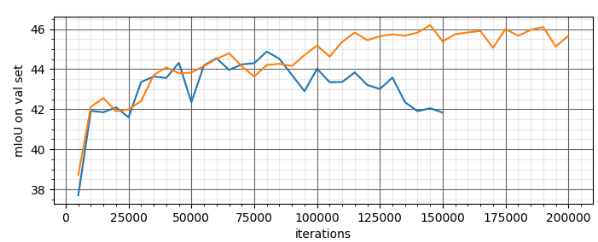

Figure 2. mIoU comparison on the validation set of Cityscapes by adapting from GTA5 dataset to Cityscapes dataset. The blue line corresponds to the output space adversarial adaptation strategy [17]. The orange line corresponds to the output space adversarial adaptation combined with our proposed SIM structure. The model performance is tested every 5000 iterations.

In the previous GAN-style approaches, the adversarial loss is essentially a binary cross-entropy about whether the generated feature is from the source domain. We observe that such a global training signal is usually weak for the segmentation task. First, the alignments between stuff regions

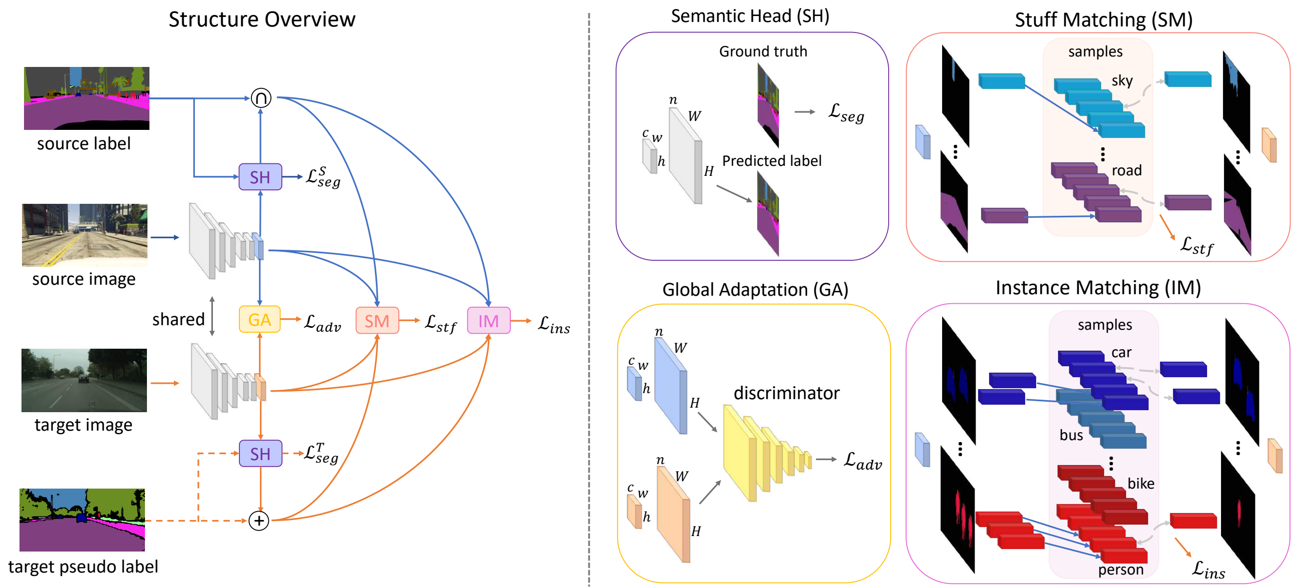

Figure 3. Framework. 1) The overall structure is shown on the left. The sub-module structures are shown on the right. Please refer to [19] for more details.

and between things require different treatments but the adversarial loss lacks such structural information. Second, the global GAN structure only adapts the feature distribution between two domains and does not necessarily adapt the target domain features towards the most likely space of source domain features. Therefore, as the semantic head gathers the features from the source domain with more training iterations, it becomes harder for the feature generator to adapt the target domain features exactly toward the source domain features. This leads to a performance drop on the target domain images as shown in figure 2.

This paper proposes a stuff and instance matching (SIM) framework to address the aforementioned difficulties. First, we treat the alignments between stuff regions and between instances of things with different guidance. The key idea is shown in figure 1. Second, we deal with the instability with the GAN training framework, we apply a L1 loss to explicitly minimize the distance between the target domain stuff and thing features with the most similar source domain counterparts. Finally, we propose to improve the SIM framework with a self-supervised learning strategy. We evaluate the proposed approach on two unsupervised domain adaptation tasks, the adaptation from GTA5 to Cityscapes and from SYNTHIA to Cityscapes, and achieve a new state-of-the-art performance on both tasks.

2. Method

Our training process is composed of two steps. The first step applies our SIM structure to the adaptation framework of [17] whose data flow is shown by the solid lines in Figure 3. The second step adds the pseudo labels obtained from the first step to supervise the semantic segmentation task whose data flow is shown by the solid and dash lines in Figure 3.

The backbone structure is shared for the source and the target domain to extract the image feature maps. Based on the extracted feature maps, we implement four sub modules shown on the right of Figure 3. The Semantic head supervises the model to generate correct semantic labels. It follows the operation in Deeplab V2 [2] to convolute and upsample the feature maps to the size of the ground truth label maps, and calculate the segmentation loss and back propagate it to the backbone structure. The generated label prediction maps are passed to the the stuff matching module and the instance matching module. The global adaptation module adapts the feature maps globally from the source domain to the target one. This module adopts a discriminator structure to make the features from the target domain more alike the ones from the source domain in an adversarial training strategy as [17].

We propose the Stuff and Instance Matching (SIM) module to better supervise the adaptation process. The stuff matching module inputs the feature maps from the source and the target domains, the ground truth label maps of the source domain, and the predicted label maps (the first training step) or the pseudo label maps (the second training step). By resizing the label maps to the size of the feature maps, we can overlap the them with the feature maps. Thus, the channel-wise feature vectors at each location of the feature maps can be classified to corresponding classes indicated by the label maps. By averaging the feature vectors belonging to the same class, we can get the representation for a specific class from this image. We store these feature vectors

Table 1. Comparison to the state-of-the-art results of adapting GTA5 to Cityscapes.

GTA5 → Cityscapes																				
Method	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorbike	bike	mIoU
Wu et al.[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
Tsai et al.[17]	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
Saleh et al.[15]	79.8	29.3	77.8	24.2	21.6	6.9	23.5	44.2	80.5	38.0	76.2	52.7	22.2	83.0	32.3	41.3	**27.0**	19.3	27.7	42.5
Luo et al. [12]	88.5	35.4	79.5	26.3	24.3	28.5	32.5	18.3	81.2	40.0	76.5	58.1	25.8	82.6	30.3	34.4	3.4	21.6	21.5	42.6
Hong et al.[8]	89.2	**49.0**	70.7	13.5	10.9	38.5	29.4	33.7	77.9	37.6	65.8	**75.1**	32.4	77.8	39.2	45.2	0.0	25.5	35.4	44.5
Chang et al. [1]	**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
Du et al. [5]	90.3	38.9	81.7	24.8	22.9	30.5	37.0	21.2	84.8	38.8	76.9	58.8	30.7	85.7	30.6	38.1	5.9	28.3	36.9	45.4
Vu et al. [18]	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
Chen et al. [3]	89.4	43.0	82.1	30.5	21.3	30.3	34.7	24.0	85.3	39.4	78.2	63.0	22.9	84.6	36.4	43.0	5.5	**34.7**	33.5	46.4
Zou et al. [21]	89.6	58.9	78.5	33.0	22.3	**41.4**	**48.2**	**39.2**	83.6	24.3	65.4	49.3	20.2	83.3	39.0	48.6	12.5	20.3	35.3	47.0
Lian et al. [10]	90.5	36.3	84.4	32.4	**28.7**	34.6	36.4	31.5	**86.8**	37.9	78.5	62.3	21.5	**85.6**	27.9	34.8	18.0	22.9	**49.3**	47.4
Li et al. [9]	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
ours (ResNet101)	90.6	44.7	**84.8**	34.3	**28.7**	31.6	35.0	37.6	84.7	43.3	**85.3**	57.0	31.5	83.8	**42.6**	48.5	1.9	30.4	39.0	**49.2**
Du et al. [5]	88.7	32.1	79.5	29.9	22.0	23.8	21.7	10.7	80.8	29.8	72.5	49.5	16.1	82.1	23.2	18.1	3.5	24.4	8.1	37.7
Li et al. [9]	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
ours (VGG16)	88.1	35.8	83.1	25.8	23.9	29.2	28.8	28.6	83.0	36.7	82.3	53.7	22.8	82.3	26.4	38.6	0.0	19.6	17.1	42.4

Table 2. Comparison to the state-of-the-art results of adapting SYNTHIA to Cityscapes.

SYNTHIA → Cityscapes													
Method	road	sidewalk	building	light	sign	vegetation	sky	person	rider	car	bus	motorbike	bike	mIoU
Luo et al. [12]	82.5	24.0	79.4	16.5	12.7	79.2	82.8	58.3	18.0	**79.3**	25.3	17.6	25.9	46.3
Tsai et al.[17]	84.3	42.7	77.5	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	46.7
Du et al. [5]	84.6	41.7	**80.8**	11.5	14.7	**80.8**	**85.3**	57.5	21.6	82.0	36.0	19.3	34.5	50.0
Li et al. [9]	**86.0**	**46.7**	80.3	14.1	11.6	79.2	81.3	54.1	27.9	73.7	**42.2**	25.7	45.3	51.4
ours (ResNet101)	83.0	44.0	80.3	**17.1**	**15.8**	80.5	81.8	**59.9**	**33.1**	70.2	37.3	**28.5**	**45.8**	**52.1**

in a memory bank if they are from the source domain. We match the averaged feature vectors from the target domain to the most similar intra-class ones from the memory bank. By calculating their L1 distance, we can supervise the backbone to generate features closer to the ones from the source domain. The instance matching module adopts the same strategy for the matching procedure. However, when averaging instance feature vectors, we find the disconnected regions belonging to the same class and treate these regions as instances.

For more detailed method description and model description, please refer to [19].

3. Experiments

We compare the results of our model adapting from GTA5 [13] to Cityscapes [4] and from SYNTHIA [14] to Cityscapes [4] with those of other methods as shown in Table 1 and Table 2 respectively. We compare with the models using the same backbone structures as ours, either Resnet101 [7] or VGG16 [16]. For more results of ablation studies and visualization, please refer to [19].

4. Conclusions

We propose a stuff and instance matching (SIM) module for the unsupervised domain adaptation of semantic segmentation from a synthetic dataset to a real-image dataset. We (1) consider the difference of appearance variance between the stuff regions and the instances of things, and thus treat them differently in the adaptation process; (2) explicitly minimize the distance of the closest stuff and instance features between the source domain and the target domain, which enables the adaptation in a more accurate direction and stabilize the GAN training process at longer iterations. By combining our SIM module with self-training, our model achieves a new state-of-the-art on this task.

Acknowledgments This work is in part supported by IBM-Illinois Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM AI Horizons Network, and ARC DECRA DE190101315.

References

[1] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[2] 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 TPAMI*, 40(4):834–848, 2017.

[3] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[4] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[5] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *IEEE CVPR*, 2018.

[9] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[12] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] 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, *Computer Vision – ECCV 2016*, pages 102–118, Cham, 2016. Springer International Publishing.

[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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[15] Fatemeh Saleh, Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, and Jose M. Alvarez. Effective use of synthetic data for urban scene semantic segmentation. In *ECCV*, 2018.

[16] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv e-prints*, page arXiv:1409.1556, Sep 2014.

[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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[18] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[19] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerior Feris, Jinjun Xiong, Wen mei Hwu, Thomas S. Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. *arXiv preprint arXiv:2003.08040*, 2020.

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

[21] Yang Zou, Zhiding Yu, B.V.K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *The European Conference on Computer Vision (ECCV)*, September 2018.