Active Domain Adaptation with Multi-level Contrastive Units for Semantic Segmentation

Hao Zhang[1,2⋆] and Ruimao Zhang[⋆⋆1[0000−0001−9511−7532]]

[1] Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong(Shenzhen), China
[2] University of Illinois Urbana-Champaign
haoz19@illinois.edu
ruimao.zhang@ieee.org

1 Technique Details about Adaptive Sampling

There exist several simple methods to control the annotation percentage to a specific level that we expect. Firstly, we can achieve it at the image-level by specifying that select the same percentage of pixels to label in each image so that the percentage of labeling on the entire dataset are controlled at the same ratio. While the disadvantage of this method is that some images may contain more outliers that need to be labeled. A straightforward way to solve such a problem is to use the memory to store all of the pixels, sorted by their uncertainty score. However, it's difficult to achieve in practice since it requires a huge memory cost, and it's hard to be progressive because of the time-consuming.

To tackle the above issue, in practice, we introduce a sample yet effective way to adjust π_{high}, which is shown in Algorithm 1. In the beginning, we initialize π_{high} with a large value. Then, in each training epoch, if the number of annotated pixels M_a is still lower than the expected quantity, we decrease π_{high} until M_a achieves the number of expected labeled high uncertainty M_e. For example, assume we set M_e equals 5% of total pixels in the target domain. Firstly we initialize π_{high} with a big number, and as a result, only around 1% of samples are selected as G_h. Then we reduce π_{high}, which will lead to the increment in the size of G_h [3]. When the M_a equals to $M_e/2$, which is about 2.5% in our assumption, we stop selecting samples from G_h. Then we directly select another $M_e/2$ samples uniformly from G_m for annotation. After that, we stop giving annotations and only apply predictions to those samples in G_l as their pseudo labels.

Effectiveness of MCUs. To investigate the effectiveness of MCUs, we further use the following two degradation models. (1) **AL(w MCU$_i$)** indicates the model that we employ active learning strategy with adaptive sampling to select

⋆ Research done when Hao Zhang was a RA at SRIBD and CUHK,SZ
⋆⋆ Corresponding Author
[3] Note that the size of G_h doesn't equal M_a, because we only sample half of the pixels in G_h uniformly to annotate, so the size of G_h equals $2M_a$

GTA5 to Cityscapes																				
	road	sidewalk	building	wall	fence	pole	light	sign	vege	terrace	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
AL(w AS)	96.4	75.7	86.8	40.3	42.0	47.4	46.1	65.4	87.9	44.0	84.3	68.6	44.9	91.5	66.7	72.6	53.9	41.9	64.7	64.3
AL(w p2p)	97.3	76.6	87.5	44.6	43.7	47.6	47.2	66.1	87.4	46.1	83.9	71.3	48.1	91.2	67.3	72.9	55.5	42.6	65.2	65.4
AL(w p2p+c2c)	97.1	77.4	87.8	42.1	43.9	48.1	47.4	65.3	87.4	55.1	82.9	72.1	49.1	91.2	70.4	73.1	55.3	45.7	66.3	66.1
Full Model	**97.2**	**78.3**	**88.4**	**46.0**	42.9	**48.5**	**48.6**	**66.5**	**89.2**	54.9	**89.3**	70.3	**49.7**	**92.1**	**70.9**	72.2	49.0	**46.4**	**67.0**	**66.7**

Table 1. Evaluation of different components of proposed method on GTA5-to-Cityscapes, with 20% labeled pixels.

annotated pixels and construct image-level contrastive units. (2) **AL(w MCU$_d$)** denotes the model that we construct both image-level and domain-level MCUs. At last, **Full Model** denotes our entire schemes, which further adds dynamic categories correlation matrix (DCCM) to the setting. For more ablation studies about MCUs, please refer to the Appendix.

As shown in Table. 1, firstly, by considering intra-image and cross-image relations in each domain, **AL(w MCU$_i$)** achieves a substantial performance gain, *i.e.* 1.2%, compared with **AL(w AS)**. Secondly, we investigate the effectiveness of domain-level contrast. **AL(w MCU$_d$)** in Table. 1 shows a further performance gain of 0.6%. Finally, we investigate the effectiveness of DCCM by comparing the results with and without DCCM. After we leverage DCCM to adjust the weight inside MCUs, we obtain a performance gain of 0.6% (*i.e.*, **Full Model**). We could find that the performance of `road` and `sidewalk`, which are always being misclassified, is improved.

As shown in Fig. 1, active learning based supervision in (b) can make the features of each category being separated, while adding MCUs in (c) can enforce features from the same category to be more compact and further separated from the features from other categories. Additionally, the features from some categories (*e.g.*, the wathet cluster at the top of Fig. 1 (b)) are already separated from the others. Thus we hope the model paying less attention to such categories, but paying more attention to those that are not separated well (*e.g.*, the red and the blue clusters). In practice, combining MCUs with DCCM could well address such an issue. Just as shown in Fig. 1 (c), the red cluster is well separated from the blue one. The above visualization results have further demonstrated the effectiveness of proposed MCUs.

2 Effectiveness of `c2c` / `p2p` Contrast

As mentioned in the main article, the **AL(w AS)** in Table **??** is about the experiment that we implement an active learning selection strategy with proposed adaptive sampling. **AL(w `p2p`)** denotes the method that only calculates pixel-to-pixel contrastive units in three levels based on the above experiment setting. **AL(w `p2p+c2c`)** indicates the method that we apply both `p2p` and `c2c` to construct the contrastive units. Note that **AL(w `p2p+c2c`)** has the same setting as

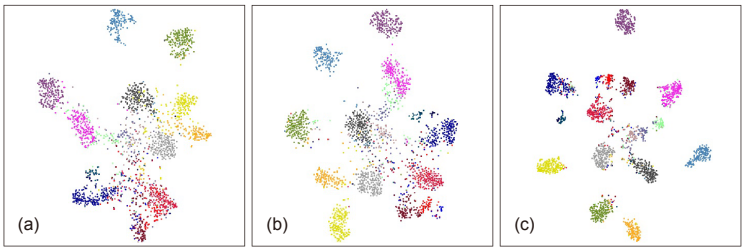

Fig. 1. Visualization of features learned (a) by UDA (AdvEnt), (b) using AL selection policy with adaptive sampling and (c) adding MCUs. 19 classes are reported on Cityscapes validation set.

AL(w $\mathbf{MCU_d}$) that is mentioned in the main article, and the DCCM is not adopted in such a case.

According to Table. **??**, we can see that using pixel-to-pixel contrast, *i.e.*, **AL(w `p2p`)**, can improve the performance by 1.1% compared with **AL(w AS)**. And further adding center-to-center contrast, *i.e.* **AL(w `p2p+c2c`)**, can further improve performance by 0.7%. Note that the method only uses the center-to-center contrast but without considering the pixel-to-pixel one could not achieve desired performance. According to the [38] and [39], a large set of negative samples is critical for contrastive representation learning. As the number of center representations in both source and target domain is limited, thus there are not enough negative samples to be used to calculate the contrastive losses, leading to poor accuracy. Thus we have not listed the corresponding results in Table **??**.

Additionally, we also investigate the effectiveness of pixel-to-center contrastive loss. However, the result of applying both `p2p` contrast and `p2c` contrast shows limited improvement compared with only adding `p2p` contrast. After careful analysis, we think it's because of the special form of InfoNCE Loss [29] defined as follows,

$$\mathcal{L}_I^{NCE} = -\log \frac{\exp(v \cdot v^+/\lambda)}{\exp(v \cdot v^+/\lambda) + \sum_{v^- \in N} \exp(v \cdot v^-/\lambda)}, \tag{1}$$

where v, v^-, and v^+ denote the anchor, negative sample, and positive sample. The operation $\cdot$ denote the vector dot product. N denotes the set of negative samples. And $\lambda > 0$ is a temperature hyper-parameter. When we calculate `p2p` contrastive loss using pixels from all three groups (*i.e.*, G_l, G_m and G_h), every anchor has negative/positive samples not only from G_h, G_m, but also from G_l. *The partial loss from `p2p`, which is calculated by a specific anchor and its corresponding samples from G_l, seem to have a similar effort to the `p2c` contrastive loss.* This is because the center representation of each category is aggregated from the pixels from G_l. As mentioned in the main article, we use those pixels from G_l (*i.e.* the low uncertainty group) with high predictive confidence to generate the category center. Intuitively, high confident samples always lie in the center of category clusters, *leading to a high density*. Thus the generated category centers would have very similar representations to the corresponding pixels in G_l that

is collected from various images. It contributes limited to further improve the performance.

Algorithm 1 The training pipeline of proposed ADA-MCU

Require: Labeled source dataset $D_s = \{(x_s^n, y_s^n)\}_{n=1}^{Ns}$, unlabeled target dataset $D_t = \{x_t^n\}_{n=1}^{Nt}$, segmentation network $\mathcal{F}$.
Parameter: Parameters of network: θ, number of training epochs: T, budget of expected annotated pixels in target domain M_e, number of already labeled pixels: M_a, dynamic correlation category matrix W.
Procedure:

Use CycleGAN to translate images in D_s to having a similar appearance as target images following [45].
Use D_s to train the segmentation network $\mathcal{F}$.
Set epoch variable $i = 0$.
while $i <= N_e$ **do**
  Randomly load two images from D_s, and two images from D_t,
  if $M_a < M_e/2$ **then**
    Update $\pi_{high} = \pi_{high} \times 0.9$ and G_h. Select and annotate pixels in G_h.
    Update M_a as the number of selected pixels.
  else if $M_a == M_e/2$ **then**
    Stop selecting pixels from G_h but select and annotate $M_e/2$ of pixels in G_m.
  else
    Sample $M_e/2$ pixels from G_l and using their predictions from $\mathcal{F}$ as the pseudo labels.
  end if
  Update W according to the Eqn. (8) in the main article.
  Calculate segmentation loss and multi-level contrastive losses.
  Update the network parameters θ and $i = i + 1$.
end while
return θ

References

1. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.
2. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
3. J. Hoffman, D. Wang, F. Yu, and T. Darrell, "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv preprint arXiv:1612.02649*, 2016.
4. T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.
5. J. Huang, D. Guan, S. Lu, and A. Xiao, "Mlan: Multi-level adversarial network for domain adaptive semantic segmentation," *arXiv preprint arXiv:2103.12991*, 2021.
6. J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. PMLR, 2018, pp. 1989–1998.
7. J. Huang, D. Guan, A. Xiao, and S. Lu, "Fsdr: Frequency space domain randomization for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6891–6902.
8. Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 289–305.
9. J. Huang, D. Guan, A. Xiao, and S. Lu, "Cross-view regularization for domain adaptive panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 133–10 144.
10. Z. Wang, Y. Wei, R. Feris, J. Xiong, W.-M. Hwu, T. S. Huang, and H. Shi, "Alleviating semantic-level shift: A semi-supervised domain adaptation method for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 936–937.
11. S. Chen, X. Jia, J. He, Y. Shi, and J. Liu, "Semi-supervised domain adaptation based on dual-level domain mixing for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 018–11 027.
12. J. Huang, D. Guan, A. Xiao, and S. Lu, "Semi-supervised domain adaptation via adaptive and progressive feature alignment," *arXiv preprint arXiv:2106.02845*, 2021.
13. Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.
14. M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *arXiv preprint arXiv:2006.09882*, 2020.
15. P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *arXiv preprint arXiv:2004.11362*, 2020.
16. J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," *arXiv preprint arXiv:2010.04592*, 2020.

17. K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *arXiv preprint arXiv:2006.10511*, 2020.
18. W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," *arXiv preprint arXiv:2101.11939*, 2021.
19. D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Machine learning proceedings 1994*. Elsevier, 1994, pp. 148–156.
20. T. Scheffer, C. Decomain, and S. Wrobel, in *International Symposium on Intelligent Data Analysis*. Springer, 2001, pp. 309–318.
21. S. D. Jain and K. Grauman, "Active image segmentation propagation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2864–2873.
22. S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Semisupervised svm batch mode active learning with applications to image retrieval," *ACM Transactions on Information Systems (TOIS)*, vol. 27, no. 3, pp. 1–29, 2009.
23. A. Vezhnevets, J. M. Buhmann, and V. Ferrari, "Active learning for semantic segmentation with expected change," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3162–3169.
24. Y. Siddiqui, J. Valentin, and M. Nießner, "Viewal: Active learning with viewpoint entropy for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9433–9443.
25. W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," *arXiv preprint arXiv:2101.11939*, 2021.
26. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
27. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.
28. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
29. A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
30. G. French, T. Aila, S. Laine, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, high-dimensional perturbations," 2019.
31. Z. Feng, Q. Zhou, G. Cheng, X. Tan, J. Shi, and L. Ma, "Semi-supervised semantic segmentation via dynamic self-training and classbalanced curriculum," *arXiv preprint arXiv:2004.08514*, vol. 1, no. 2, p. 5, 2020.
32. K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8050–8058.
33. Z. Wang, Y. Wei, R. Feris, J. Xiong, W.-M. Hwu, T. S. Huang, and H. Shi, "Alleviating semantic-level shift: A semi-supervised domain adaptation method for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 936–937.

34. W. Liu, D. Ferstl, S. Schulter, L. Zebedin, P. Fua, and C. Leistner, "Domain adaptation for semantic segmentation via patch-wise contrastive learning," *arXiv preprint arXiv:2104.11056*, 2021.
35. Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4085–4095.
36. O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *International workshop on artificial intelligence and statistics*. PMLR, 2005, pp. 57–64.
37. T. Kasarla, G. Nagendar, G. M. Hegde, V. Balasubramanian, and C. Jawahar, "Region-based active learning for efficient labeling in semantic segmentation," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1109–1117.
38. Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.
39. K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.
40. O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *International workshop on artificial intelligence and statistics*. PMLR, 2005, pp. 57–64.
41. M. Ning, D. Lu, D. Wei, C. Bian, C. Yuan, S. Yu, K. Ma, and Y. Zheng, "Multi-anchor active domain adaptation for semantic segmentation," *arXiv preprint arXiv:2108.08012*, 2021.
42. L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.
43. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.
44. Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7472–7481.
45. T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.
46. G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3234–3243.
47. A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
48. R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018.
49. X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

50. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
51. L. Qi, J. Kuen, Y. Wang, J. Gu, H. Zhao, Z. Lin, P. Torr, and J. Jia, "Open-world entity segmentation," *arXiv preprint arXiv:2107.14228*, 2021.
52. L. Qi, J. Kuen, Z. Lin, J. Gu, F. Rao, D. Li, W. Guo, Z. Wen, and J. Jia, "Casp: Class-agnostic semi-supervised pretraining for detection and segmentation," *arXiv preprint arXiv:2112.04966*, 2021.