Supplementary for CoPT: Unsupervised Domain Adaptive Segmentation using Domain-Agnostic Text Embeddings

Cristina Mata, Kanchana Ranasinghe, and Michael S. Ryoo

Stony Brook University, Stony Brook, NY, USA
{cfmata, kranasinghe, mryoo}@cs.stonybrook.edu

A Additional Related Works

A.1 Unsupervised Domain Adaptation for Semantic Segmentation

Self-supervised losses are used by the bulk of previous methods to learn stronger feature representations. In particular, contrastive losses promote learning appearance of objects in the target domain by connecting them to object representations learned in the source domain [3,14,17–19]. Recently, masked image modeling was shown to give large improvements in performance by Hoyer et al. [11]. To recover from errors on rare classes, some works learn priors of the source class distribution and use this to estimate the target class distribution [2, 9, 26, 31].

Domain adaptation methods encompass several data access settings to emulate different real-world applications. In some applications, source data can be proprietary, meaning clients only have access to a source-pretrained model and must adapt it to unlabeled target data. The source-free domain adaptation task was introduced to emulate this setting [12]. Kundu et al. [15] use data augmentation in the source-free domain adaptation setting. They separate the model training into a stage with access to only the source domain, where they perform data augmentation, then an adaptation stage with access only to unlabeled target data. Image translation, in which source pixels are directly operated on to transform to a target appearance, has been applied to source data and added to the training pipeline [1, 6]. In [4], data from an additional modality, depth, is used. Network architectures and training strategies have been optimized for the UDA segmentation task as well [10].

Adversarial learning is attractive because it directly modifies latent features, but can be difficult to optimize and does not currently lead to state of the art performance. Tsai et al. [27] use adversarial learning in the output/prediction space, modifying the latent features of the segmentation encoder directly to remove domain-specific information. Wang et al. [30] apply an adversarial loss between high and low confidence regions within pseudo-labeled target images. Adversarial learning has recently been revived by Chen et al. [2] who show that when used with self-training it leads transformer models to outperform CNNs on the task.

A.2 Domain Adaptation using Vision-Language Embeddings

A common theme in vision-language learning for domain adaptation is to combine text embeddings with image embeddings to imbue semantic details about the target domain into the image representation [7, 16, 28, 29]. Min et al. [20] train an LSTM-based text generator that outputs explanations for an image classifier's prediction, but this requires ground truth text descriptions of images. In tasks such as video-text retrieval where a joint vision-language embedding space must be learned, Hao et al. [8] perform domain adaptation by maximizing the similarity of target video-text embedding pairs, without the use of CLIP.

B Implementation Details

B.1 LLM Domain Template Attributes

We include the complete set of queries to ChatGPT [21] whose outputs are used to generate domain descriptors in Table 3. The format of the queries is shown in the top column as "What makes a <DOMAIN> image look <DOMAIN>?" where the <DOMAIN> can be filled in as synthetic, real, day-time, night-time, snow, fog, and rain depending on the style of the source and target domains. In the GTA→CS [5, 22] and Synthia→CS [23] benchmarks, the source domain is synthetic and the target domain is real. In the CS→DZ [24] benchmark, the source domain is day-time and the target domain is night-time. CS→ACDC [25] encompasses fog, rain, snow and night-time domains as target and day-time as source. To generate the domain-agnostic text embedding for a class, the class is formatted into the attributes following Equation 4. The attributes are fed to the frozen text encoder and averaged together.

B.2 Auxiliary Losses

In our main experiments we implement CoPT on top of MIC [11] because of its state of the art performance on UDA for semantic segmentation. During CoPT training, we also leave the auxiliary losses implemented by default in MIC: pixel-wise cross entropy loss on source, masked pseudo-label self-training using an EMA teacher, ImageNet feature distance regularization, and strongly augmented self-training. The pixel-wise cross entropy loss over source detail is explained in Equation 1. We strongly encourage the reader to refer to Hoyer et al. for details on the other auxiliary losses but give a brief summary of the self-training losses here.

For masked pseudo-label self-training, an unlabeled target sample $\mathbf{x}^t \in T$ is passed to an EMA-updated teacher model $\mathcal{T}_\alpha$ with the same architecture as $\mathcal{E}_\psi$ to get pseudo label $\mathbf{y}^t = \mathcal{T}_\alpha(\mathbf{x}^t)$. Then a patch in $\mathbf{x}^t$ is uniformly sampled and masked to get $\mathbf{x}^t_m$. The student model's prediction on the masked image is $\hat{\mathbf{y}}^t_m = \mathcal{E}_\psi(\mathbf{x}^t_m)$. The final masked loss is

$$\mathcal{L}_m = q^T \mathcal{L}_{ce}(\hat{\mathbf{y}}^t_m, \mathbf{y}^t) \tag{1}$$

Table 1: Ablation experiment results for memory bank decay are reported as % mIOU on GTA→CS over 19 classes.

MemBank Decay	mIOU
0.01	70.52
0.1	75.05
0.5	**76.08**

Table 2: Ablation experiment results for CoPT's training scheme are reported as % mIOU on GTA→CS over 19 classes.

Method	mIOU
Finetune	37.34
Joint Training	**76.08**

where q^t is a segmentation quality estimate. In Hoyer et al. q^T is calculated as the ratio of pixels exceeding a threshold of the maximum softmax probability.

In strongly augmented self-training, an unlabeled target sample $\mathbf{x}^t$ is strongly augmented to get $\tilde{\mathbf{x}}^t$. The original sample is fed to the EMA teacher to get $\hat{\mathbf{y}}^t = \mathcal{T}_\alpha(\mathbf{x}^t)$. Then a cross entropy loss is applied to the student's output on the strongly augmented image,

$$\mathcal{L}_{st} = \mathcal{L}_{ce}(\mathcal{E}_\psi(\tilde{\mathbf{x}}^t), \hat{\mathbf{y}}^t) \tag{2}$$

C Training Details

We re-iterate that in our main experiments, CoPT is implemented on top of the MIC [11] code without changes to the default training hyperparameters. This means that the largest image resolution used for training is (1024, 1024), training is done for 40,000 iterations with batch size 2, the image encoder is MiT-B5 [32] and decoder is a DAFormer [9] head. We chose to build on MIC since it is the state of the art for UDA for segmentation and has public code that is simple to run.

D Ablation and Analysis Details

D.1 Memory Bank Decay

We use a grid search over three values shown in Table 1 to choose the decay parameter λ in Equation 9 in our pixel feature memory bank. The results indicate that a high decay parameter, where information from previous training iterations is kept for longer, is advantageous to CoPT. Owing to the low batch size, it takes many iterations for the model to see enough samples to give a class enough pixels to draw diverse, class-representative features from. CoPT benefits from diverse pixel features because its goal is to re-orient the average feature of the class in pixel latent space.

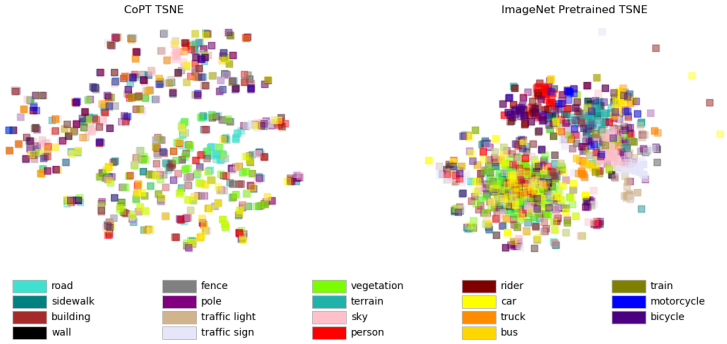

Fig. 1: T-SNE visualization on Cityscapes val set of CoPT (left) and ImageNet-pretrained class pixel features (right). Best viewed in color

D.2 Joint Training vs. Finetuning

We evaluate CoPT's optimal training scheme in Table 2 on the GTA→CS benchmark. In the joint training row we add CoPT with weight 1 to the other losses during MIC training, and in the finetune row we take a fully trained MIC model and finetune it with just CoPT and source domain cross entropy loss. Our results show that joint training far outperforms finetuning. This implies that learning domain-agnostic features in pixel space using CoPT risks losing discriminative information learned through the self-supervised losses.

D.3 T-SNE Visualization

Figure 1 visualizes t-SNE plots of class pixel features extracted from an MiT-B5 [32] encoder loaded with ImageNet-pretrained weights at the start of training (right) and at the end of training using CoPT (left) implemented on top of MIC [11]. Images from Cityscapes val set are fed to the models. CoPT encourages feature embeddings to be more distinct from each other as can be seen by the greater separation between data points. There are fewer outliers in CoPT versus ImageNet e.g. the traffic sign, building and wall classes showing the regularization of the latent space.

D.4 Mistral 7B Sentence Embedding

When using Mistral 7B [13] as CoPT's text embedding method in Section 5.2, we generate a single text embedding for each prompt by averaging together Mistral's embedding for each word in the prompt. Mistral 7B outputs a single embedding for each word because it was trained for next token prediction and not designed to generate sentence embeddings.

Table 3: The complete list of ChatGPT [21] output attributes that are used to generate the domain descriptions as part of the LLM Domain Template process. The domain in the left column is formatted into the ChatGPT Query template. Table continued on next page.

ChatGPT Query: "What makes a <DOMAIN> image look <DOMAIN>"?	
DOMAIN	Attribute Outputs
synthetic	1. lack of realism 2. unusual colors and lighting 3. perfect symmetry 4. repetitive elements 5. lack of organic variation 6. tiling and tiling artifacts 7. overly sharp or soft focus 8. inconsistent shadows and reflections 9. distinctive noise patterns 10. regular patterns and grids 11. lack of depth and perspective 12. looks like an inorganic object 13. an artistic style or stylization 14. exaggerated features
real	1. natural colors and lighting 2. high resolution 3. depth and perspective 4. organic variation 5. complex textures 6. authentic shadows and reflections 7. blurred bokeh background 8. lens flare and glare 9. natural poses and expressions 10. environmental integration 11. realistic props and objects 12. accurate reflections in water and glass 13. natural compression artifacts 14. weather and atmospheric effects 15. unscripted candid moments 16. a tangible sense of scale
day-time	1. an abundance of natural light 2. bright and vibrant colors 3. soft shadows 4. warmth in illumination 5. dynamic range 6. natural sky colors 7. distinctive sun position 8. clear visibility 9. minimal artificial lighting 10. lively and active atmosphere 11. natural textures and patterns 12. shimmering water bodies 13. pleasant weather conditions 14. outdoor shadows 15. minimal noise and grain
night-time	1. low light conditions 2. dark shadows 3. diminished color saturation 4. warm artificial lighting 5. with contrast between light and dark 6. point light sources 7. visible light trails 8. with glowing skylines 9. with silhouettes and outlines 10. with noise and grain 11. astrophotography elements 12. long shadows 13. reflective surfaces 14. distant atmospheric haze 15. a sense of mystery and atmosphere

ChatGPT Query: "What makes a <DOMAIN> image look <DOMAIN>"?	
DOMAIN	Attribute Outputs
foggy	1. soft, diffused light 2. hazy atmosphere 3. low contrast 4. diminished sharpness and clarity 5. gradient of opacity 6. desaturation of colors 7. loss of detail in distance 8. diffused light sources 9. visible water droplets 10. muffled soundscape 11. ethereal and dreamlike atmosphere 12. silhouettes and silhouetted forms 13. moisture on surfaces 14. elongated shadows
rainy	1. raindrops on surfaces 2. wet and reflective surfaces 3. muted colors 4. glossy textures 5. diminished contrast 6. blurred backgrounds 7. dynamic water movement 8. puddles and reflections 9. umbrellas and rain gear 10. dramatic sky 11. wet vegetation 12. water ripples and disturbances 13. smeared or distorted lights 14. misty atmosphere
snowy	1. white blanket of snow 2. soft, diffused light 3. cool color palette 4. snowflakes in motion 5. crystalline texture 6. cold, wintry atmosphere 7. footprints and tracks 8. snow-covered trees and branches 9. icy surfaces and frozen water 10. winter clothing and gear 11. frost and ice crystals 12. blurred backgrounds and depth of field 13. cold breath and condensation 14. winter sports and activities

References

1. Chen, L., Wei, Z., Jin, X., Chen, H., Zheng, M., Chen, K., Jin, Y.: Deliberated domain bridging for domain adaptive semantic segmentation. In: NeurIPS (2022)
2. Chen, M., Zheng, Z., Yang, Y.: Transferring to real-world layouts: A depth-aware framework for scene adaptation (2023)
3. Chen, M., Zheng, Z., Yang, Y., Chua, T.S.: Pipa: Pixel- and patch-wise self-supervised learning for domain adaptative semantic segmentation. In: ACM MM (2023)
4. Chen, R., Rong, Y., Guo, S., Han, J., Sun, F., Xu, T., Huang, W.: Smoothing matters: Momentum transformer for domain adaptive semantic segmentation (2022)
5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
6. Ettedgui, S., Abu-Hussein, S., Giryes, R.: Procst: Boosting semantic segmentation using progressive cyclic style-transfer (2022)
7. Fahes, M., Vu, T.H., Bursuc, A., Pérez, P., de Charette, R.: PØda: Prompt-driven zero-shot domain adaptation. In: ICCV (2023)
8. Hao, X., Zhang, W., Wu, D., Zhu, F., Li, B.: Dual alignment unsupervised domain adaptation for video-text retrieval. In: CVPR (2023)
9. Hoyer, L., Dai, D., Gool, L.V.: Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: CVPR (2022)
10. Hoyer, L., Dai, D., Gool, L.V.: Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In: ECCV (2022)
11. Hoyer, L., Dai, D., Wang, H., Gool, L.V.: Mic: Masked image consistency for context-enhanced domain adaptation. In: CVPR (2023)
12. Hu, X., Wang, K., Zhang, K., Xia, L., Chen, A., Luo, J., Qiao, N., Zeng, X., Sun, M., Kuo, C.H., Sun, Y., Nevatia1, R.: Reclip: Refine contrastive language image pre-training with source free domain adaptation. In: WACV (2024)
13. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023)
14. Jiang, Z., Li, Y., Yang, C., Gao, P., Wang, Y., Tai, Y., Wang, C.: Prototypical contrast adaptation for domain adaptive semantic segmentation. In: ECCV (2022)
15. Kundu, J.N., Kulkarni, A., Singh, A., Jampani, V., Babu, R.V.: Generalize then adapt: Source-free domain adaptive semantic segmentation. In: ICCV (2021)
16. Lee, S., Park, H., Kim, D.U., Kim, J., Boboev, M., Baek, S.: Image-free domain generalization via clip for 3d hand pose estimation. In: WACV (2023)
17. Li, G., Kang, G., Liu, W., Wei, Y., Yang, Y.: Content-consistent matching for domain adaptive semantic segmentation. In: ECCV (2020)
18. Li, J., Wang, Z., Gao, Y., Hu, X.: Exploring high-quality target domain information for unsupervised domain adaptive semantic segmentation. In: ACM MM (2022)
19. Lu, Y., Luo, Y., Zhang, L., Li, Z., Yang, Y., Xiao, J.: Bidirectional self-training with multiple anisotropic prototypes for domain adaptive semantic segmentation. In: ACM MM (2022)
20. Min, S., Park, N., Kim, S., Park, S., Kim, J.: Grounding visual representations with texts for domain generalization. In: ECCV (2022)
21. OpenAI, :, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I.,

Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N.S., Khan, T., Kilpatrick, L., Kim, J.W., Kim, C., Kim, Y., Kirchner, J.H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C.M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H.P., Michael, Pokorny, Pokrass, M., Pong, V.H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M.B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., Zoph, B.: Gpt-4 technical report (2024)

22. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: ECCV (2016)
23. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR (2016)
24. Sakaridis, C., Dai, D., Gool, L.V.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: ICCV (2019)
25. Sakaridis, C., Dai, D., Gool, L.V.: Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: ICCV (2021)

26. Truong, T.D., Le, N., Raj, B., Cothren, J., Luu, K.: Fredom: Fairness domain adaptation approach to semantic scene understanding. In: CVPR (2023)
27. 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)
28. Vidit, V., Engilberge, M., Salzmann, M.: Clip the gap: A single domain generalization approach for object detection. In: CVPR (2023)
29. Wang, Z., Zhang, L., Wang, L., Zhu, M.: Landa: Language-guided multi-source domain adaptation (2024)
30. Wang, Z., Liu, X., Suganuma, M., Okatani, T.: Cross-region domain adaptation for class-level alignment (2022)
31. Xie, B., Li, S., Li, M., Liu, C.H., Huang, G., Wang, G.: Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. In: IEEE TPAMI (2023)
32. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: NeurIPS (2021)