

SUPPLEMENTARY MATERIAL

Anonymous authors

Paper under double-blind review

1 APPENDIX

1.1 EXPERIMENTAL SETTINGS

Datasets We conduct experiments on the two standard benchmark settings, namely, “GTAV Richter et al. (2016) to Cityscapes Cordts et al. (2016)” and “SYNTHIA Ros et al. (2016) to Cityscapes Cordts et al. (2016)”, where GTAV Richter et al. (2016) and SYNTHIA Ros et al. (2016) are adopted as labeled source domain, and Cityscapes Cordts et al. (2016) is taken as unlabeled target domain to evaluate the adaptation performance. GTAV Richter et al. (2016) is generated from a game environment and contains 24,966 synthetic images with resolution 1914×1052. Cityscapes-style annotation are adopted with 19 common classes. Synthia Ros et al. (2016) consists of 9,400 synthetic images with resolution 1280×760, labeled with Cityscapes-style annotation (16 common classes). Cityscapes Cordts et al. (2016) is a driving dataset for semantic segmentation containing 2975 training and 500 validation images with resolution 2048×1024. All three datasets were downloaded from the official website and used with the permission of the authors. There is no personally identifiable information or offensive content in three datasets.

Implementation Details We employ the Mix Transformer Xie et al. (2021) as the encoder network pre-trained on the Imagenet-1K dataset. The decoder network is borrowed from the DAFormer, since the context information of multi-scales features are considered. Adam optimizer with learning rate 6e-5 are used for training student model in propose STCT framework. The framework is trained with 40,000 iterations. We utilize the Cosine-Annealing learning rate scheduler with warm up to stabilize the training. Coordination weight $\alpha = 2e - 4$ is set as default. Images are resized to (1280, 720) and randomly cropped by (640, 640) as inputs. To solve the class imbalance, we introduce the rare-semantic-grouping and thing-class ImageNet feature distance proposed in DAFormer. Image resizing, horizontal flipping, color jitter, and gaussian blur is used as augmentation. We do not use DACS Tranheden et al. (2021) augmentations like DAFormer. Hoyer et al. (2022). All experiments are conducted on one NVIDIA Tesla V100. Codes will be released before Sep. 1st on <https://anonymous.4open.science/r/STCT-C0E0>

1.2 PROPORTION OF FOUR TYPES OF ATTENTION WEIGHTS

Besides, we also conduct experiments to investigate the effect of the proportion of features in the MHA module. The MHA module with semantic grouping strategy is formulated as follows:

$$Attn_{hybrid}(\tilde{Q}_i, \tilde{K}_s, \tilde{K}_t, \tilde{V}_s, \tilde{V}_t) = Softmax\left(\frac{\tilde{Q}_i \cdot [\tilde{K}_s; \tilde{K}_t]^\top}{\sqrt{d}}\right)[\tilde{V}_s; \tilde{V}_t], \quad \text{where } i \in \{s, t\}. \quad (1)$$

$$Attn_{hybrid}(\tilde{Q}_i^c, \tilde{K}_s^c, \tilde{K}_t^c, \tilde{V}_s^c, \tilde{V}_t^c) = Softmax\left(\frac{\tilde{Q}_i^c \cdot [\tilde{K}_s^c; \tilde{K}_t^c]^\top}{\sqrt{d}}\right)[\tilde{V}_s^c; \tilde{V}_t^c], \quad \text{where } i \in \{s, t\}. \quad (2)$$

We take the features \tilde{Q}_i , \tilde{K}_t , and \tilde{V}_i in semantic group (Eq. 1) as an example. Features \tilde{Q}_i^c , \tilde{K}_t^c , and \tilde{V}_i^c in complementary group (Eq. 2) follows the same way. There are four types of attention weights in the MHA module, *i.e.*, source domain attention weight ($\tilde{Q}_s \tilde{K}_s^\top$ and $\tilde{Q}_t \tilde{K}_s^\top$), target domain attention weight ($\tilde{Q}_t \tilde{K}_t^\top$ and $\tilde{Q}_s \tilde{K}_t^\top$), intra-domain attention weight ($\tilde{Q}_s \tilde{K}_s^\top$ and $\tilde{Q}_t \tilde{K}_t^\top$), inter-domain attention weight ($\tilde{Q}_s \tilde{K}_t^\top$ and $\tilde{Q}_t \tilde{K}_s^\top$). The source domain attention weights adopt the source feature to rebuild the source feature ($\tilde{Q}_s \tilde{K}_s^\top$) or target feature ($\tilde{Q}_t \tilde{K}_s^\top$). The target domain attention weights adopt the target feature to rebuild the target feature ($\tilde{Q}_t \tilde{K}_t^\top$) or source feature ($\tilde{Q}_s \tilde{K}_t^\top$). The inter-domain attention weights construct the interaction between different domains, while the intra-domain attention weights construct the interaction between the same domain.

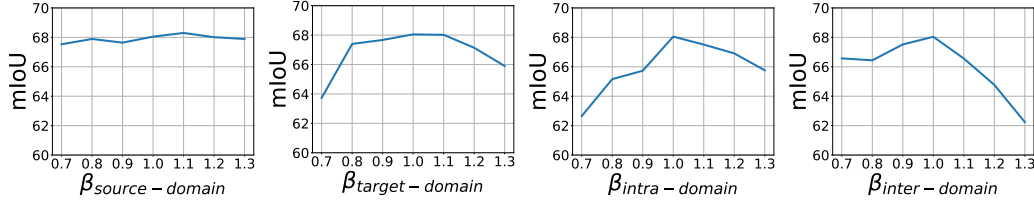


Figure 1: Effects of the proportions of four types of attention weights. From left to right, we evaluate the proportions of the source domain, target domain, intra-domain, and inter-domain attention weight, respectively. The $\beta = 1$ means no artificial proportion is imposed.

For the proportion of the inter-domain attention weights, we rescale the attention weights $\tilde{Q}_s \tilde{K}_t^\top$ and $\tilde{Q}_t \tilde{K}_s^\top$ based on a scalar β as follows:

$$Attn_{hybrid}(\tilde{Q}_s, \tilde{K}_s, \tilde{K}_t, \tilde{V}_s, \tilde{V}_t) = Softmax\left(\frac{\tilde{Q}_s \cdot [\tilde{K}_s; \beta \tilde{K}_t]^\top}{\sqrt{d}}\right)[\tilde{V}_s; \tilde{V}_t], \quad (3)$$

$$Attn_{hybrid}(\tilde{Q}_t, \tilde{K}_s, \tilde{K}_t, \tilde{V}_s, \tilde{V}_t) = Softmax\left(\frac{\tilde{Q}_t \cdot [\beta \tilde{K}_s; \tilde{K}_t]^\top}{\sqrt{d}}\right)[\tilde{V}_s; \tilde{V}_t]. \quad (4)$$

where β indicates the proportion of the inter-domain attention weight. The Eq. 1 is split to Eq. 3 of source domain and Eq. 4 of target domain. The complementary group features in Eq. 2 can be rescaled as the same manner.

For the proportion of the other three types of attention weights, the source domain attention weight is rescaled in $\tilde{Q}_s \tilde{K}_s^\top$ and $\tilde{Q}_t \tilde{K}_s^\top$:

$$Attn_{hybrid}(\tilde{Q}_s, \tilde{K}_s, \tilde{K}_t, \tilde{V}_s, \tilde{V}_t) = Softmax\left(\frac{\tilde{Q}_s \cdot [\beta \tilde{K}_s; \tilde{K}_t]^\top}{\sqrt{d}}\right)[\tilde{V}_s; \tilde{V}_t], \quad (5)$$

$$Attn_{hybrid}(\tilde{Q}_t, \tilde{K}_s, \tilde{K}_t, \tilde{V}_s, \tilde{V}_t) = Softmax\left(\frac{\tilde{Q}_t \cdot [\beta \tilde{K}_s; \tilde{K}_t]^\top}{\sqrt{d}}\right)[\tilde{V}_s; \tilde{V}_t]. \quad (6)$$

Similarly, the target domain attention weight is rescaled in $\tilde{Q}_s \tilde{K}_t^\top$ and $\tilde{Q}_t \tilde{K}_t^\top$:

$$Attn_{hybrid}(\tilde{Q}_s, \tilde{K}_s, \tilde{K}_t, \tilde{V}_s, \tilde{V}_t) = Softmax\left(\frac{\tilde{Q}_s \cdot [\tilde{K}_s; \beta \tilde{K}_t]^\top}{\sqrt{d}}\right)[\tilde{V}_s; \tilde{V}_t], \quad (7)$$

$$Attn_{hybrid}(\tilde{Q}_t, \tilde{K}_s, \tilde{K}_t, \tilde{V}_s, \tilde{V}_t) = Softmax\left(\frac{\tilde{Q}_t \cdot [\tilde{K}_s; \beta \tilde{K}_t]^\top}{\sqrt{d}}\right)[\tilde{V}_s; \tilde{V}_t]. \quad (8)$$

The intra-domain attention weight is rescaled as follows:

$$Attn_{hybrid}(\tilde{Q}_s, \tilde{K}_s, \tilde{K}_t, \tilde{V}_s, \tilde{V}_t) = Softmax\left(\frac{\tilde{Q}_s \cdot [\beta \tilde{K}_s; \tilde{K}_t]^\top}{\sqrt{d}}\right)[\tilde{V}_s; \tilde{V}_t], \quad (9)$$

$$Attn_{hybrid}(\tilde{Q}_t, \tilde{K}_s, \tilde{K}_t, \tilde{V}_s, \tilde{V}_t) = Softmax\left(\frac{\tilde{Q}_t \cdot [\tilde{K}_s; \beta \tilde{K}_t]^\top}{\sqrt{d}}\right)[\tilde{V}_s; \tilde{V}_t]. \quad (10)$$

We scale the proportion of each weight from 0.7 to 1.3, and the experimental results of four attention weights are reported in Fig. 1. We find the model is robust to the proportion of the source and target domain attention weights, as illustrated in the first and second images of Fig. 1. Besides, we identify that a balanced proportion between the inter-domain and intra-domain attention weights, *i.e.*, $\beta = 1$ in the third and fourth images of Fig. 1, is vital for effective feature alignment. This phenomenon verifies that the intra-domain self-attention and inter-domain cross-attention mechanism are indispensable and complementary in the proposed MHA module.

1.3 COMPARISON OF FEATURE GROUPING STRATEGIES

In the main paper, we introduce six feature grouping strategies: "non-grouping" strategy, "random-grouping" strategy, "HVG-grouping" strategy, "HVR-grouping" strategy, "cutout-grouping",

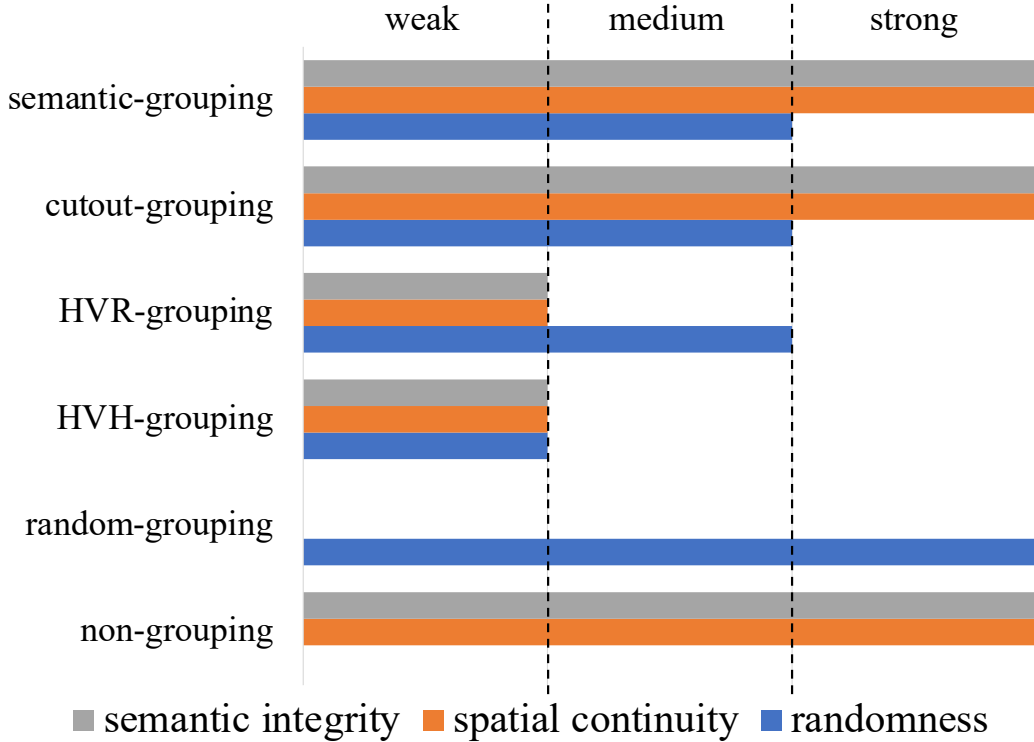


Figure 2: Proportion of three factors in feature grouping strategies. The longer the bar length, the greater the proportion of the corresponding factor in the strategy. We use weak, medium, and strong to represent the proportion of the three factors in the strategy. Since all three factors are integrated with the maximum degree, "semantic-grouping" strategy is adopted in our method.

"semantic-grouping" strategy. Through designing five grouping strategies from "random-grouping" to "semantic-grouping", we progressively introduce three factors (randomness, spatial continuity, and semantic integrity) into the feature grouping strategies to measure the role of three factors in feature alignment. It is worth noting that the six grouping strategies do not coexist in our method. Only the semantic-grouping strategy is adopted in our method, since all three factors are considered with the maximum degree. We illustrate the proportion of three factors in six strategies in Fig. 2.

1.4 QUALITATIVE COMPARISON

To make a fair comparison, we select the first image from the three cities ("frankfurt_000000_000294_leftImg8bit.png", "lindau_000000_000019_leftImg8bit.png", and "munster_000000_000019_leftImg8bit.png") in the Cityscapes Cordts et al. (2016) validation set to illustrate the prediction. We list the qualitative results of the ablation studies in Fig. 3. For example, compared to the visualization of the ablation studies, our method can give a more complete prediction on "sidewalk" classes.

The qualitative results of our method and three state-of-the-art (SOTA) methods, *i.e.*, CorDA Wang et al. (2021), ProDA Zhang et al. (2021), DAFormer Hoyer et al. (2022), are shown in Fig. 4. Compared to the three SOTA methods, our method achieves finer predictions on the "rider" class.

1.5 LIMITATIONS

Domain adaptive semantic segmentation aims to transfer knowledge from the source domain to the target domain. Our approach achieves superior performance on datasets where both the source and target domains are urban street scenes. However, the performance of our method may degrade if the scenes in the source and target domains are significantly different. One limitation of this work is that it cannot be employed for the privacy-preserving source-free DA setting Fleuret et al. (2021);

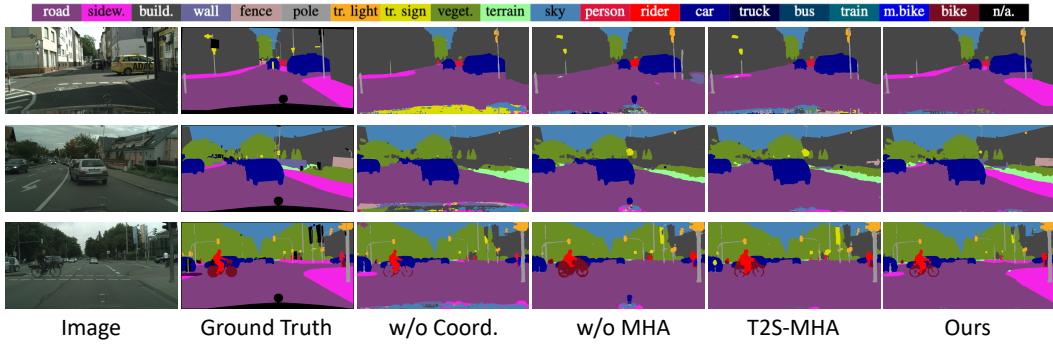


Figure 3: Qualitative results of semantic segmentation on the Cityscapes dataset. From left to right: image, ground truth, our method without coordination weight, our method without the MHA module, our method with unidirectional (target-to-source) cross-attention mechanism, our method.

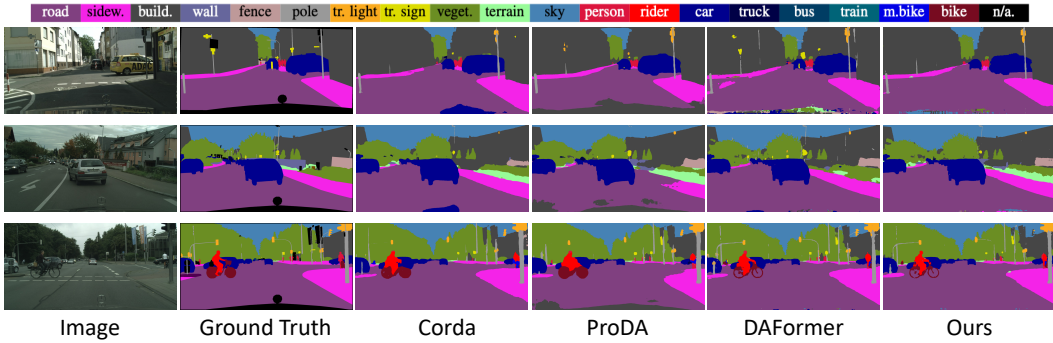


Figure 4: Qualitative comparison between state-of-the-art methods on the Cityscapes dataset.

Kundu et al. (2021) since cross-domain attention requires simultaneous access to both source and target data. Another limitation is that our method is currently not applicable to multi-source and multi-target domain adaptive settings, which is left as future work.

1.6 POTENTIAL NEGATIVE SOCIETAL IMPACT

Due to the lack of supervisory signals in the target domain, objects might not be segmented in a way that we are used to or problematic biases in the data might become apparent. Lack of monitoring, could have potential negative impacts in areas such as autonomous driving and virtual reality.

REFERENCES

- 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.
- Francois Fleuret et al. Uncertainty reduction for model adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9613–9623, 2021.
- Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7046–7056, 2021.

- Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pp. 102–118. Springer, 2016.
- 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3234–3243, 2016.
- Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1379–1389, 2021.
- Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8515–8525, 2021.
- Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021.
- Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12414–12424, 2021.