Supplementary Materials for CDAC: Cross-domain Attention Consistency in Transformer for Domain Adaptive Semantic Segmentation

1. Combination with HDRA

The HRDA model trains high-resolution (x^d) and low-resolution (x^c) crops from source (x_s) and target (x_t) images simultaneously to capture both local and global features effectively. Similarly, our base model combined with DAFormer extracts source (x_s^d and x_s^c), target-to-source (x_{t2s}^d and x_{t2s}^c), target (x_t^d and x_t^c), and source-to-target (x_{s2t}^d and x_{s2t}^c) features from the high and low-resolution crops using an encoder. Then we apply supervised and unsupervised losses from HRDA to the four cross-domain features in parallel. However, given the high computation cost of HRDA, we optimize the process by randomly selecting between source features and target-to-source features, and between target features and source-to-target features. This results in only two branches of learning being conducted instead of four, and we extend the training iteration by two times for sufficient training. Finally, since we generate pseudo-labels in a sliding window manner, we do not apply our cross-domain attention consistency loss in this case.

2. Training Computation Complexity

During training, we added a source-to-target and a target-to-source branch of (pseudo-)supervision over the baselines (*e.g.*, DAFormer) (see Eq. 4) with a source and a target branch. Our complexity is $O(4LN^2)$ *vs*. DAFormer's $O(2LN^2)$, where L is the number of the attention modules and N the sequence length. Please note that the inference cost is the same.

3. Entropy of Attention Maps

In the main paper, we claim that our attention map can attend wider regions and learns from more diverse and informative signals. We measure the entropy of attention produced by different baselines and compare it with ours. Low entropy represents that the attention map is biased to small regions or fewer pixels, and high entropy represents that the attention map focuses on wider regions.

To be more specific, we calculate the average value of entropy of attention maps extracted from the first attention module of each stage on the entire training set of GTAV dataset as well as the validation set of the Cityscapes dataset. Since the length of the output sequence in MiT-B5 is always 512, the theoretical maximal entropy of an attention map is $-\sum \frac{1}{512} \log \frac{1}{512} \approx 6.238$. As shown in Fig. 1, our method generally produce attention maps with higher entropy in the first two stages, while our baselines' attention maps are comparably more sparse. Combining the visualization from Fig. 1 and Fig. 4 in the main paper, we can confirm that our method can effectively attend larger regions, which will be helpful to avoid spurious regions in attention maps and helps achieve attention-level domain adaptation. As for the attention maps from stage 3 and stage 4, we present their visualization in Fig. 2 on the same images used in Fig. 4 in the main paper. From the figure, we may find that compared to the results from the main paper obtained in stage 1, the attention module in stage 3 generally attends all pixels in the whole image, while that in stage 4 always attends the *upper-left corner* regardless of the input. Such observations are also supported by the results shown in Fig. 1 that the entropy values generally achieve the maximal entropy value at stage 3, and plummet to the lowest level at stage 4. We believe it is an important next step to further investigate this behavior in future work.

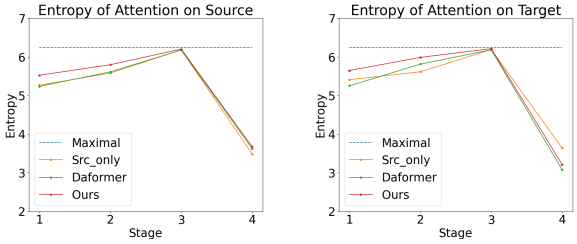

Figure 1: Entropy of attention maps at each stage from baselines and our model on the source domain (left) and the target domain (right). Higher entropy represents that the model attends wider regions. Our method can attend to wider regions compared to DAFormer. Maximal represents the theoretical maximal entropy value.

4. Attention Maps Visualization

We visualize the attention maps in models with the code adopted from the official release of TransPose [2]. Given a

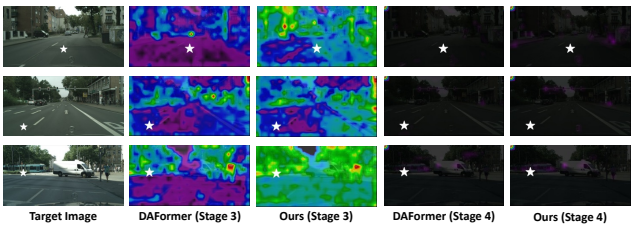

Figure 2: Attention maps from stage 3 and stage 4 from DAFormer and our model. Generally, attention maps from stage 3 in both models attend almost all pixels in the image, while that from stage 4 always attend only the upper-left corner.

series of attention maps $\{Attn^i\}_{i=1}^{L}$ extracted from an image x with a resolution of $H \times W$ via L attention modules, we select the attention module at the beginning of the first stage in MiT-B5 [1] to visualize the attention maps. Each attention map has the dimension of $C \times N \times N'$, where C is the number of features of an attention head, N is the number of the input sequence, and N' is the number of the output sequence. To visualize the attention maps for a query pixel from x, we at first locate its correspondence along the dimension of the input sequence with necessary resize and flatten operations. Then we take the average over the C heads and acquire the final attention score with a length of N', which is finally resized and upsampled to the original shape of the image x.

5. Cross-Domain Attention Mixing

To help understand our cross-domain attention consistency learning, we illustrate the pipeline with exemplar images and intermediate features from our model in Fig. 3. For each of the input sequences to the attention module, *i.e.*, each pixel from the input image, we decide whether we take the corresponding attention maps from the source or the target domain based on the mixing mask at that position. For example, in Fig. 3, the mixing mask M selects pixels from the target image at point A and the source image at point B. Correspondingly, the mixed attention maps $Attn_{sup}$ take attention maps from the target image at point A, *i.e.*, $Attn_{sup}[:,:,A,:] = Attn_t[:,:,A,:]$ and from the source image at point B, *i.e.*, $Attn_{sup}[:,:,B,:] = Attn_s[:,:,B,:]$ for supervision purposes. Moreover, to avoid forcing the model to attend invalid regions that are occluded in the mixed image, we create a valid mask M_v that filters out regions that are from the other domain. For instance, since point A is selected to take the pixel from the target domain, the valid mask M_v will mask out the supervision from $Attn_{sup}[:,:,A,:]$ on $Attn_m[:,:,A,:]$ to the regions representing the source domain in the mixed image $\hat{x}_t$. On the contrary, the valid mask M_v for point B will mask out the supervision from $Attn_{sup}[:,:,B,:]$ to $Attn_m[:,:,B,:]$ on the regions that take pixels from the target image.

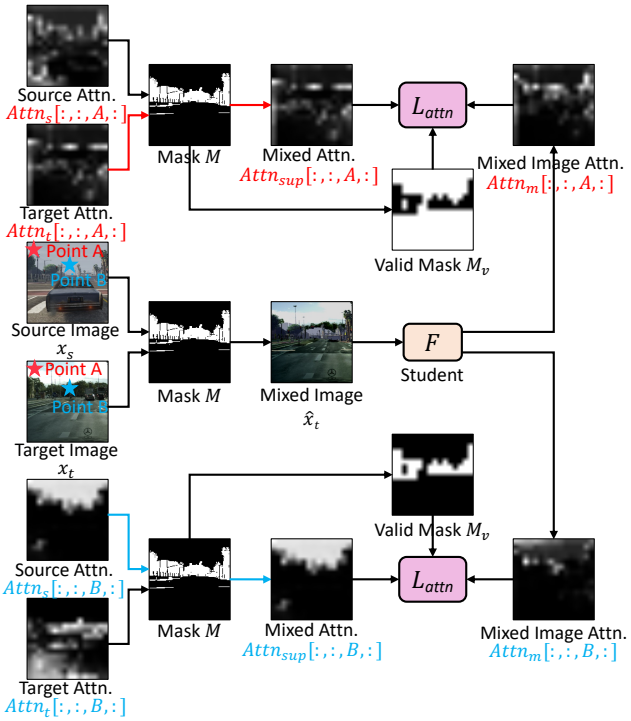

Figure 3: Illustration of the pipeline for our cross-domain attention consistency loss. If a point is selected by the mixing mask M from the source domain, *e.g.* point B, then the mixed attention map for supervision at that point should also be selected from the source attention map $Attn_{sup}[:,:,B,:] = Attn_s[:,:,B,:]$. On the other hand, for a point selected from the target domain, *e.g.* point A, its corresponding mixed attention map for supervision should be selected from the target attention map $Attn_{sup}[:,:,A,:] = Attn_t[:,:,A,:]$.

5.1. Qualitative Results

We present qualitative results in Fig. 4. In these samples, we observe that DAFormer and ours obtain higher quality predictions over the source-only model. At the same time, we also observe that our method generally obtains finer and better predictions than that of DAFormer (*e.g.*, the fence in the first row, the traffic lights in the second row, and the road in the third row highlighted by red boxes).

References

[1] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 2

[2] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *ICCV*, 2021. 1

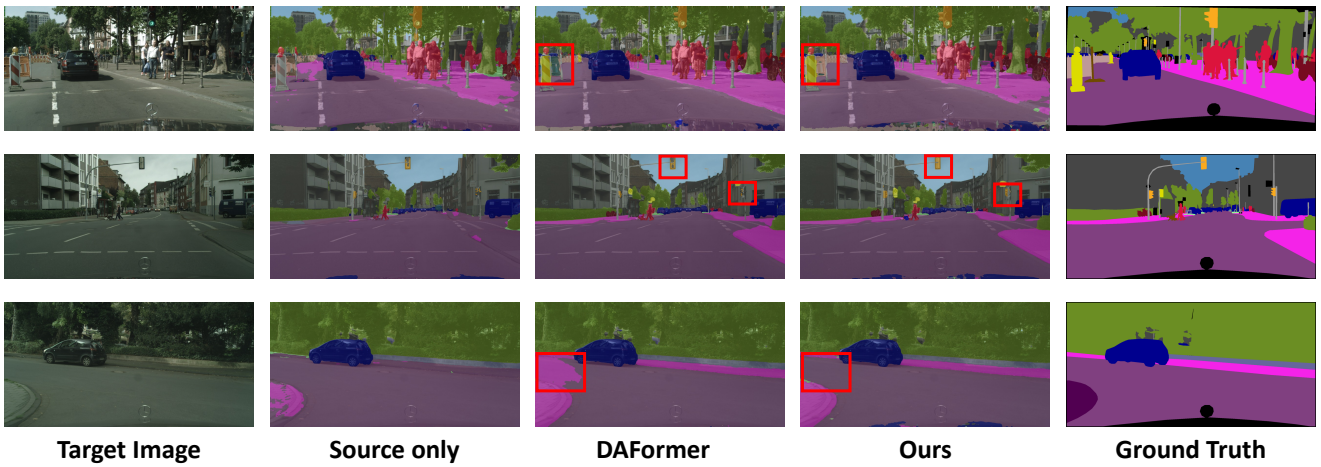

Figure 4: Qualitative results of our method versus the source-only baseline and the DAFormer.