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

Kaihong Wang, Donghyun Kim, Rogerio Feris, Margrit Betke
Boston University, Korea University, MIT-IBM Watson AI Lab
{kaiwkh, betke}@bu.edu, d_kim@korea.ac.kr, rsferis@us.ibm.com

Abstract

While transformers have greatly boosted performance in semantic segmentation, domain adaptive transformers are not yet well explored. We identify that the domain gap can cause discrepancies in self-attention. Due to this gap, the transformer attends to spurious regions or pixels, which deteriorates accuracy on the target domain. We propose Cross-Domain Attention Consistency (CDAC), to perform adaptation on attention maps using cross-domain attention layers that share features between source and target domains. Specifically, we impose consistency between predictions from cross-domain attention and self-attention modules to encourage similar distributions across domains in both the attention and output of the model, i.e., attention-level and output-level alignment. We also enforce consistency in attention maps between different augmented views to further strengthen the attention-based alignment. Combining these two components, CDAC mitigates the discrepancy in attention maps across domains and further boosts the performance of the transformer under unsupervised domain adaptation settings. Our method is evaluated on various widely used benchmarks and outperforms the state-of-the-art baselines, including GTAV-to-Cityscapes by 1.3 and 1.5 percent point (pp) and Synthia-to-Cityscapes by 0.6 pp and 2.9 pp when combining with two competitive Transformer-based backbones, respectively. Our code will be publicly available at https://github.com/wangkaihong/CDAC.

1. Introduction

The Transformer model has shown remarkable performance on various computer vision tasks (e.g., [7, 2, 42, 44]) and often exhibits an outstanding prediction capacity compared to convolutional networks. Meanwhile, it is also relatively “data-hungry” and therefore expects a large amount of training data in order to achieve strong performance [29]. However, curating a large-scale annotated dataset could be a prohibitively expensive engineering task, especially for those problems that require pixel-level labeling, including semantic segmentation. Furthermore, deep models often generalize poorly to new domains such as different cities or weather in driving scenes. To overcome these issues, the use of unsupervised domain adaptation (UDA) has been proposed. UDA allows knowledge transfer from synthetic data (source domain), where pixel-level annotations are more cheaply available, to real-world data (unlabeled target domain). Recently, under the UDA setting for semantic segmentation, a Transformer-based method (DAFormer [13]) outperforms previous CNN-based UDA methods across diverse benchmark datasets.

The key component for the success in Transformer is self-attention, which learns to attend to certain regions of its input that could be informative to predicting the semantic class of a pixel. However, it is still not clear if the domain gap is completely resolved under the UDA setting.
We found that the attention maps from self-attention on the target images can focus on spurious regions and therefore fail to assist the prediction on the target domain as in Fig. 1, which suggests that the domain gap still remains. We aim to improve the robustness of self-attention by using cross-domain attention, which computes attention scores across different domain images. Previous work [43] explores cross-domain attention in image classification. However, this method requires a sophisticated search for finding positive cross-domain pairs, which may not be directly applicable to semantic segmentation.

To more effectively mitigate the domain discrepancy and improve the robustness of Transformers for semantic segmentation, we propose our cross-domain prediction consistency loss to encourage consistency in the prediction (i.e., output-level) and attention map (i.e., attention-level). To be more specific, we first compute predictions based on self-attention and cross-domain attention respectively. Then we supervise the predictions based on the self-attention module and the cross-domain attention module with the same label from either the source domain label or the target domain pseudo-label. The benefits of this design are twofold: firstly, the consistency allows attention-level alignment that helps pull the distributions of the self-attention and cross-domain attention closer as in Fig. 1. Secondly, the operation also acts as a perturbation and regularization on the attention maps that encourage output-level domain alignment, i.e., facilitates consistent predictions when the attention maps differ due to the different input query vectors. In the end, the model benefits from both the alignment at the attention level between the self-attention and cross-domain attention as well as the more robust predictions on the output level.

Inspired by consistency learning [28, 9], enforced on the output of models through different augmentations of an image, we propose our cross-domain attention consistency loss to learn attention-level consistency and induce a model to generate robust attention maps. The idea is rather straightforward: The attention maps from two different augmented views from the same image should always be consistent. With this objective, the model can learn to predict more consistent attention maps on both source and target images without supervision and further contribute to attention-level domain alignment.

Combining our cross-domain prediction consistency loss and cross-domain attention consistency loss, CDAC not only facilitates alignment on the output level but also brings the alignment to the attention level and improves the segmentation accuracy under UDA settings. To summarize, our key contributions in this work include:

1. We introduce a cross-domain prediction consistency loss that encourages robust attention as well as prediction across domains and thus helps with the attention-level and output-level domain alignment;

2. We propose to enforce attention-level consistency via our cross-domain attention consistency loss to further assist the alignment of attention-level discrepancy;

3. CDAC is verified in extensive experiments to be effective, reliable and generalizable under different scenarios and achieves state-of-the-art performance on several important UDA semantic segmentation benchmarks including GTAV-to-Cityscapes, Synthia-to-Cityscapes and Cityscapes-to-ACDC.

2. Related Work

Unsupervised Domain Adaptation (UDA) for Semantic Segmentation. UDA methods aim to leverage a labeled source domain dataset to learn on a target domain dataset without corresponding labels and reduce the domain shift. There are two major categories of existing methods in UDA semantic segmentation, the first one is to align domains into one distribution. Different approaches are proposed to achieve this goal including adversarial training [12, 11, 32, 35, 33, 35, 8, 21, 4, 39] on different levels including image, feature and output level, as well as non-adversarial approaches that such as style transfer [41, 16, 36], Fourier transformation [45]. Another category of methods focuses on the learning decision boundary on the target domain in an unsupervised manner, including self training [49, 50, 18, 45, 47, 31] that learns from pseudo-labels, entropy minimization [35] and consistency regularization [4, 36, 1]. Recently, with DAFormer [13], the Transformer model and adequate training strategies were introduced to the task of UDA semantic segmentation. DAFormer greatly improves the state-of-the-art performance over the existing convnet based methods, while HRDA [13] further improves performance under high-resolution scenarios. We take DAFormer as our backbone and focus on attention-level alignment via cross-domain attention within Transformer rather than proposing a new architecture or training strategy.

Transformers. Inspired by Transformer models used for NLP [34], Vision Transformers (e.g., ViT [7], DeiT [30], Swin [19], CSWin [6]) have been proposed that yield remarkable results for various computer vision tasks, including the task of semantic segmentation such as Segmener [27], SETR [48], and Segformer [42]. In multimodal tasks (especially in language-vision), cross-modal attention between image and text has been widely exploited for aligning image and text representations (e.g., [17, 3, 46]).

We aim to learn cross-attention across domains for knowledge transfer, where we do not have direct correspondences between source and target domains. The most relevant work is CDTrans [43], which shares features across
domains in cross-domain attention for image classification. However, they require two-way labeling to find corresponding positives (i.e., the same category) in cross-attention, which is not directly applicable to the task of semantic segmentation. A concurrent work BCAT [38] concatenates features from both self-attention and cross-domain attention and uses the feature to perform domain alignment based on MMD [10].

In contrast to image classification, CDAC is proposed for semantic segmentation where multiple categories exist in a single image and therefore unsupervised learning for this task is more challenging. Our method does not require positive label pairing across domains but uses arbitrary (pseudo) labels to guide the consistent (output-level) prediction from the self-attention module as well as the noisy cross-domain attention module to improve robustness to the domain gap. In addition, we also address the attention-level discrepancy more directly with our cross-domain attention consistency loss.

3. Method

In this section, we first give a brief overview of the UDA setting and self-attention mechanism in Sec. 3.1, and then introduce our cross-domain prediction consistency loss in Sec. 3.2 and the attention-consistency loss in Sec. 3.3.

3.1. Preliminaries

Formally, given a source domain dataset \( D_s = \{(x_i^s, y_i^s)\}_{i=1}^{N^s} \) with corresponding labels and an unlabeled target domain dataset \( D_t = \{(x_i^t)\}_{i=1}^{N^t} \), our task is to learn a student semantic segmentation model \( F \) with a self-ensemble teacher \( G \) and optimize the student’s performance on the target domain. At step \( t \), the teacher’s weights \( \phi_t \) are updated as the moving average of the student’s weight \( \theta_t \) as \( \phi_t = \alpha \phi_{t-1} + (1 - \alpha) \theta_t \).

In a multi-head self-attention module, an input feature map at the dimension of \( N \times d \), where \( N = H \times W \), will be passed through a query weight \( W^Q \), a key weight \( W^K \) and a value weight \( W^V \) with the shape of \( d \times C \) to produce a query vector \( Q \), a key vector \( K \), and a value vector \( V \) sharing the same dimensions of \( N \times C \). Thus, the self-attention can be obtained by:

\[
\text{Attention}(Q, K, V) = \text{Softmax} \left( \frac{QK^T}{\sqrt{d_{\text{head}}}} \right)V,
\]

where \( d_{\text{head}} \) is the number of dimensions of one head.

Typical UDA segmentation pipelines, including DAFormer [13], consist of a supervised branch for the source domain and an unsupervised branch for the target domain. The supervised learning branch computes the cross-entropy loss \( L_s \) between the prediction of the student

\[
F \text{ on a source image } x_s \text{ and its corresponding label } y_s:
\]

\[
L_s^{(i)} = - \sum_{j=1}^{H \times W} \sum_{c=1}^{C} y_{i,j,c} \log F(x_s^i)_{(j,c)},
\]

where \( H \) and \( W \) are the height and width of an input image, respectively, while \( C \) is the number of categories shared between the source and the target domain. We follow DACS [31] to generate a binary mixing mask \( M \): Given a source image with pixel-level labels, we randomly select half of the label categories appearing in the image and set the corresponding pixels in \( M \) to be 1 and the rest 0. Then, a target image \( x_t \) and its pseudo-label \( p_t \) predicted by \( G \) (the teacher) will be mixed with \( x_s \) and \( y_s \) through the mask \( M \) to obtain the augmented image \( \hat{x}_t \) and label \( \hat{p}_t \) for the target domain supervision. If a pixel at \( j \) from the mask \( M^j = 1 \), then \( \hat{x}_t^{i,j} = x_s^{i,j} \) and \( \hat{p}_t^{i,j} = y_s^{i,j} \), otherwise \( \hat{x}_t^{i,j} = x_t^{i,j} \) and \( \hat{p}_t^{i,j} = p_t^{i,j} \). This yields the loss for the unsupervised branch:

\[
L_t^{(i)} = - \sum_{j=1}^{H \times W} \sum_{c=1}^{C} q_i^{(i,j,c)} \log F(\hat{x}_t^i)_{(j,c)},
\]

where \( q_i \) is the weight mask for \( \hat{x}_t^i \) as in DAFormer. We take DAFormer as our backbone model.

3.2. Cross-Domain Prediction Consistency Loss

While existing Transformers rely on self-attention that attends regions in the same image that will be informative, we found that the attention often could be spurious and focus on non-informative regions. This becomes more serious when there are distributional shifts in test data as shown in Fig. 1. Some prior works [15, 40] were proposed to improve self-attention, but their methods do not work well in our adaptation studies.
Since the model mentioned in the previous section is mostly supervised with source labels in the initial training stages, it naturally overfits to the source domain. For the target domain from different distributions, self-attention can be noisy and attend to less informative regions. We compute the prediction based on the cross-domain attention and enforce its consistency to that of the self-attention. The enforcement of consistency not only helps bridging distributional shifts in attention across domains, but also leads a model to be robust even when the attention map is noisy. This can also be understood as regularization of the self-attention mechanism with our cross-domain attention layers in a similar spirit to other regularization methods such as Dropout [26] and Excitation Dropout [51] that randomly suppress important neuron activation, enabling a model to learn alternative paths and discover other important neurons, thereby preventing learning shortcuts or bias. Furthermore, instead of simply dropping tokens/neurons, CDAC encourages robust predictions against perturbed attention maps, as we will discuss in Sec. 4.4. This enhances the model’s ability to handle noisy attention from the domain gap (Fig. 1).

Specifically, through the same query weight $W^Q$, key weight $W^K$ and value weight $W^V$, we extract source domain vectors $Q_s$, $K_s$, and $V_s$ from source image $x_s$ and $Q_t$, $K_t$, and $V_t$ from target image $x_t$. In opposition to the typical self-attention module in a regular model $F$ that takes these vectors only from the same image as $A_s = \text{Attention}(Q_s, K_s, V_s)$ and $A_t = \text{Attention}(Q_t, K_t, V_t)$, we swap the query vectors to produce cross-domain attention. A cross-domain model $F'$ sharing the same architecture and weights with the student $F$ is applied to take a pair of images from different domains. $F'(\tilde{x}_t, x_s)$ extracts initial features from the source image $x_s$ and produces output based on the target-to-source attention $A_{t2s} = \text{Attention}(Q_t, K_s, V_s)$. Symmetrically, $F'(x_s, \tilde{x}_t)$ extracts initial features from the target image $\tilde{x}_t$ and computes the output based on the source-to-target attention $A_{s2t} = \text{Attention}(Q_s, K_t, V_t)$. This procedure is visualized in Fig. 2. Since we consider the attention from different query feature vectors in the other domain can be noisy, we use gradient stopping on the query features vectors from the other domain. The target-to-source prediction will be guided by the source label (Eq. 2) while the source-to-target prediction will be guided by the pseudo-label from $G$ (Eq. 3), yielding our proposed losses:

$$L_{t2s}^{(i)} = -\sum_{j=1}^{H_sW_s} \sum_{c=1}^{C} y_s^{(i,j,c)} \log F'(\tilde{x}_t, x_s)(j,c)$$

$$L_{s2t}^{(i)} = -\sum_{j=1}^{H_sW_s} \sum_{c=1}^{C} q^t \hat{p}_t^{(i,j,c)} \log F'(x_s, \tilde{x}_t)(j,c).$$

Finally, we take the average of the losses from the source and target supervision and construct our cross-domain prediction consistency loss into $L_{pred}$:

$$L_{pred} = \frac{1}{2}(L_s + L_{t2s}) + \frac{1}{2}(L_t + L_{s2t}).$$

### 3.3. Cross-Domain Attention Consistency Loss

We further regularize a model by consistency on attention maps. We enforce the model to produce consistent attention maps from different augmentations to strengthen the alignment on attention-level. Fig. 3 illustrate the overview of our cross-domain attention consistency loss. Firstly, we extract attention maps $\{\text{Attn}_{t}^{s}\}_{i=1}^{L}$ and $\{\text{Attn}_{s}^{t}\}_{i=1}^{L}$ via $L$ attention modules of the teacher model $G$ from the source image $x_s$ and the target image $x_t$, as well as $\{\text{Attn}_{m}^{t}\}_{i=1}^{L}$ via the student model $F$ from the mixed image $\tilde{x}_t$ (in Sec. 3.1). The source and the target attention maps $\{\text{Attn}_{s}^{t}\}_{i=1}^{L}$ and $\{\text{Attn}_{t}^{s}\}_{i=1}^{L}$ will be mixed by $M$ that produces $\tilde{x}_t$. Then we compute the mixed attention maps $\{\text{Attn}_{sup}^{t}\}_{i=1}^{L}$ to regularize the consistency on attention maps

$$\text{Attn}_{sup}^{t[i,:;j,:]} = \begin{cases} \text{Attn}_{s}^{t[i,:;j,:]}, & M' = 1, \\ \text{Attn}_{t}^{s[i,:;j,:]}, & M' = 0, \end{cases}$$

where $M'$ is resized from $M$ and $j$ represents the spatial location. Finally, we minimize the Kullback–Leibler divergence to guide $\text{Attn}_{m}^{t}$ towards $\text{Attn}_{sup}^{t}$. Additionally, we apply a valid mask $M$ to avoid misleading $\text{Attn}_{m}^{t}$ to attend..
Table 1: Comparison with baselines on the GTA V-to-Cityscapes benchmark. The mIoU represents the average of individual mIoUs among all 19 categories between GTA V and Cityscapes. Arch. represents the architecture of the model. C represents convnet-based model while T represents Transformer. The best results are highlighted in bold while the second best results are underlined.

<table>
<thead>
<tr>
<th>Method</th>
<th>Arch.</th>
<th>Road</th>
<th>Building</th>
<th>Pedestrian</th>
<th>Traffic Light</th>
<th>Sign</th>
<th>Vegetation</th>
<th>Sky</th>
<th>Person</th>
<th>Rider</th>
<th>Car</th>
<th>Truck</th>
<th>Bus</th>
<th>Train</th>
<th>Moped</th>
<th>Bicycle</th>
<th>Synthia</th>
<th>Cityscapes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source only</td>
<td>T</td>
<td>71.5</td>
<td>18.0</td>
<td>84.2</td>
<td>34.4</td>
<td>30.9</td>
<td>44.3</td>
<td>23.5</td>
<td>87.4</td>
<td>41.3</td>
<td>86.6</td>
<td>64.0</td>
<td>22.5</td>
<td>88.3</td>
<td>39.1</td>
<td>2.3</td>
<td>35.2</td>
<td>31.6</td>
</tr>
<tr>
<td>CBST [69]</td>
<td>C</td>
<td>91.8</td>
<td>53.5</td>
<td>80.5</td>
<td>32.7</td>
<td>21.0</td>
<td>25.5</td>
<td>30.1</td>
<td>20.4</td>
<td>83.9</td>
<td>34.2</td>
<td>80.9</td>
<td>53.1</td>
<td>24.0</td>
<td>82.7</td>
<td>30.3</td>
<td>35.9</td>
<td>16.0</td>
</tr>
<tr>
<td>DACS [31]</td>
<td>C</td>
<td>89.9</td>
<td>39.7</td>
<td>87.9</td>
<td>30.7</td>
<td>39.5</td>
<td>38.5</td>
<td>46.4</td>
<td>52.8</td>
<td>88.0</td>
<td>44.0</td>
<td>88.8</td>
<td>67.2</td>
<td>35.8</td>
<td>84.5</td>
<td>45.7</td>
<td>50.2</td>
<td>0.0</td>
</tr>
<tr>
<td>CoDA [37]</td>
<td>C</td>
<td>94.7</td>
<td>63.1</td>
<td>87.6</td>
<td>30.7</td>
<td>40.6</td>
<td>40.2</td>
<td>47.8</td>
<td>51.6</td>
<td>87.6</td>
<td>47.0</td>
<td>89.7</td>
<td>66.7</td>
<td>35.9</td>
<td>90.2</td>
<td>48.9</td>
<td>57.5</td>
<td>0.0</td>
</tr>
<tr>
<td>ProDA [47]</td>
<td>C</td>
<td>87.8</td>
<td>56.0</td>
<td>79.7</td>
<td>46.3</td>
<td>44.8</td>
<td>45.6</td>
<td>53.5</td>
<td>53.5</td>
<td>88.6</td>
<td>45.2</td>
<td>82.1</td>
<td>70.7</td>
<td>39.2</td>
<td>88.8</td>
<td>45.5</td>
<td>59.4</td>
<td>1.0</td>
</tr>
<tr>
<td>DAFormer + CDAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source only</td>
<td>T</td>
<td>95.7</td>
<td>70.2</td>
<td>89.4</td>
<td>53.5</td>
<td>48.1</td>
<td>49.6</td>
<td>55.8</td>
<td>59.4</td>
<td>89.9</td>
<td>47.9</td>
<td>92.5</td>
<td>72.2</td>
<td>44.7</td>
<td>97.3</td>
<td>74.5</td>
<td>78.2</td>
<td>65.1</td>
</tr>
<tr>
<td>HRDA [14]</td>
<td>T</td>
<td>96.5</td>
<td>73.9</td>
<td>89.5</td>
<td>56.8</td>
<td>48.9</td>
<td>50.7</td>
<td>55.8</td>
<td>63.3</td>
<td>89.9</td>
<td>49.1</td>
<td>91.2</td>
<td>72.2</td>
<td>45.4</td>
<td>92.7</td>
<td>78.3</td>
<td>82.9</td>
<td>67.5</td>
</tr>
<tr>
<td>HRDA + CDAC</td>
<td>T</td>
<td>98.2</td>
<td>78.7</td>
<td>91.8</td>
<td>59.6</td>
<td>57.1</td>
<td>57.1</td>
<td>63.9</td>
<td>69.3</td>
<td>91.3</td>
<td>48.4</td>
<td>94.2</td>
<td>79.0</td>
<td>52.9</td>
<td>93.9</td>
<td>84.1</td>
<td>85.7</td>
<td>75.9</td>
</tr>
<tr>
<td>Oracle</td>
<td>T</td>
<td>98.0</td>
<td>84.2</td>
<td>92.6</td>
<td>59.4</td>
<td>59.7</td>
<td>61.9</td>
<td>66.7</td>
<td>76.6</td>
<td>92.5</td>
<td>66.4</td>
<td>94.9</td>
<td>79.6</td>
<td>60.7</td>
<td>94.6</td>
<td>84.0</td>
<td>88.6</td>
<td>81.2</td>
</tr>
</tbody>
</table>

$$L_{attn} = \frac{1}{L} \sum_{i=1}^{L} KL(Att^n_{sup}, Att_n^{m}) \odot M_v$$

$$M_v[j,:] = \begin{cases} M, & M_{ij} = 1, \\ 1 - M, & M_{ij} = 0. \end{cases}$$

Finally, our learning objective $L$ includes the combined cross-domain prediction consistency loss and the cross-domain attention consistency loss:

$$L = L_{pred} + \lambda_{attn} L_{attn}.$$ 

Inference, we only use self-attention modules for the target domain.

4. Experiments

In this chapter, we demonstrate experimental results under different settings in order to thoroughly verify the effectiveness and reliability of our proposed method. We introduce detailed experimental protocols in Sec. 4.1 and compare our method with other existing baselines on different benchmarks in Sec. 4.2. We also analyze the contribution of each of our vital model components in ablation studies in Sec. 4.3, and further discuss the effects regarding our cross-domain attention in Sec. 4.4 and the effects of the capacity of the encoder in Sec. 4.5. We present qualitative results and demonstrate the effectiveness of our attention-level alignment by showing the visualization of the attention maps and predictions obtained with our method and baseline in Sec. 4.6. Finally, we verify our model’s robustness when different hyper-parameter values are used, and its generalizability to different benchmarks in Sec. 4.7.

4.1. Experiment Protocols

We implemented using the MMSegmentation library. We take DAFormer as our backbone model and followed the same configurations such as learning rate, etc. We adopted MiT-B5 [42] pretrained on ImageNet as our encoder, and decoder from DAFormer. We also followed the optimization policies and chose AdamW [20] as the optimizer with the learning rate of $6 \times 10^{-5}$ for the encoder, $6 \times 10^{-4}$ for the decoder with weight decay of 0.01 and batch size of 2. DACS [31], Learning rate warmup, rare class sampling (RCS), thing-class ImageNet feature distance (FD) and other tricks applied in DAFormer were also kept with the original configurations including hyper-parameters. We also note that our cross-domain consistency learning can be easily combined with any Transformer-based model, e.g., HRDA [14] that learns detailed patches from images with higher resolution. More information regarding the combination with HRDA is available in our supplementary material.

Dataset. GTA V [22] is a synthetic street scene dataset rendered by a game engine with pixel-level annotations. It includes 24,966 images with resolution 1914 × 1052. Synthia [23] is another synthetic street scene dataset containing 9,400 images with the corresponding annotation with a resolution 1280 × 760. Cityscapes [5] is a real-world street scene dataset that includes 2,975 training images and 500 held-out images for evaluation with resolution of 2048 × 1024. ACDC [24] is also a real-world street scene dataset sharing 19 common categories with Cityscapes under adverse conditions including foggy, nighttime, rainy, and snowy scenarios. There are 1,600 training images, 406 validation images, and, more importantly, 2,000 held-out testing images with resolution 1920 × 1080. We follow our baselines to resize GTA images to resolution 1280 × 720, Synthia images to resolution 1280 × 760, Cityscapes images to resolution 1024 × 512, and ACDC images to 960 × 540 and crop images to resolution 512 × 512 during the training.

Baselines and Metric. We choose representative and state-of-the-art UDA CNN-based baselines including CBST [49], ADVENT [35], DACS [31], CorDA [37], MGCDA [25],
Table 2: Comparison with baselines on the Synthia-to-Cityscapes benchmark. The mIoU represents the average of individual mIoUs among all 16 categories between Synthia and Cityscapes. Arch. represents the architecture of the model. C represents convnet-based model while T represents Transformer. The best results are highlighted in bold while the second best results are shown by underline.

<table>
<thead>
<tr>
<th>Method</th>
<th>Arch.</th>
<th>Road</th>
<th>Sidewalk</th>
<th>Building</th>
<th>Wall</th>
<th>Fences</th>
<th>Pole</th>
<th>Tr.Lights</th>
<th>Vege</th>
<th>Person</th>
<th>Rider</th>
<th>Car</th>
<th>Bus</th>
<th>Motor</th>
<th>Bicycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source only</td>
<td>C</td>
<td>64.3</td>
<td>21.3</td>
<td>73.1</td>
<td>2.4</td>
<td>1.1</td>
<td>31.4</td>
<td>7.0</td>
<td>27.7</td>
<td>63.1</td>
<td>67.6</td>
<td>42.2</td>
<td>19.9</td>
<td>73.1</td>
<td>51.3</td>
</tr>
<tr>
<td>CBST [40]</td>
<td>C</td>
<td>68.0</td>
<td>29.9</td>
<td>76.3</td>
<td>10.8</td>
<td>1.4</td>
<td>33.9</td>
<td>22.8</td>
<td>29.5</td>
<td>77.6</td>
<td>78.3</td>
<td>60.6</td>
<td>28.3</td>
<td>81.6</td>
<td>23.5</td>
</tr>
<tr>
<td>DACS [31]</td>
<td>C</td>
<td>80.6</td>
<td>25.1</td>
<td>81.9</td>
<td>21.5</td>
<td>2.9</td>
<td>37.2</td>
<td>22.7</td>
<td>24.0</td>
<td>83.7</td>
<td>90.8</td>
<td>67.6</td>
<td>38.3</td>
<td>82.9</td>
<td>28.5</td>
</tr>
<tr>
<td>CorDA [37]</td>
<td>C</td>
<td>93.3</td>
<td>61.6</td>
<td>85.3</td>
<td>19.6</td>
<td>5.1</td>
<td>37.8</td>
<td>36.6</td>
<td>42.8</td>
<td>84.9</td>
<td>90.4</td>
<td>69.7</td>
<td>41.8</td>
<td>85.6</td>
<td>38.4</td>
</tr>
<tr>
<td>ProDA [47]</td>
<td>C</td>
<td>87.8</td>
<td>45.7</td>
<td>84.6</td>
<td>37.1</td>
<td>0.6</td>
<td>44.0</td>
<td>54.6</td>
<td>37.0</td>
<td>88.1</td>
<td>84.4</td>
<td>74.2</td>
<td>24.3</td>
<td>88.2</td>
<td>51.1</td>
</tr>
<tr>
<td>Source only</td>
<td>T</td>
<td>51.3</td>
<td>20.3</td>
<td>79.2</td>
<td>19.3</td>
<td>1.8</td>
<td>40.9</td>
<td>29.9</td>
<td>22.1</td>
<td>79.1</td>
<td>82.4</td>
<td>63.0</td>
<td>24.9</td>
<td>75.8</td>
<td>33.7</td>
</tr>
<tr>
<td>DAFormer [13]</td>
<td>T</td>
<td>84.5</td>
<td>40.7</td>
<td>88.4</td>
<td>41.5</td>
<td>6.5</td>
<td>50.0</td>
<td>55.0</td>
<td>54.6</td>
<td>86.0</td>
<td>89.8</td>
<td>73.2</td>
<td>48.2</td>
<td>87.2</td>
<td>53.2</td>
</tr>
<tr>
<td>DAFormer + CDAC</td>
<td>T</td>
<td>83.7</td>
<td>42.9</td>
<td>87.4</td>
<td>39.8</td>
<td>7.5</td>
<td>50.7</td>
<td>55.7</td>
<td>53.5</td>
<td>85.9</td>
<td>90.9</td>
<td>74.5</td>
<td>47.2</td>
<td>86.0</td>
<td>60.2</td>
</tr>
<tr>
<td>HRDA [14]</td>
<td>T</td>
<td>85.2</td>
<td>47.7</td>
<td>88.8</td>
<td>49.5</td>
<td>4.8</td>
<td>57.2</td>
<td>65.7</td>
<td>60.9</td>
<td>85.3</td>
<td>92.9</td>
<td>79.4</td>
<td>52.8</td>
<td>89.0</td>
<td>64.7</td>
</tr>
<tr>
<td>HRDA + CDAC</td>
<td>T</td>
<td>93.1</td>
<td>68.5</td>
<td>89.8</td>
<td>51.2</td>
<td>8.9</td>
<td>59.4</td>
<td>65.3</td>
<td>65.3</td>
<td>84.7</td>
<td>94.4</td>
<td>81.2</td>
<td>70.0</td>
<td>90.5</td>
<td>56.9</td>
</tr>
<tr>
<td>Oracle</td>
<td>T</td>
<td>98.0</td>
<td>84.2</td>
<td>92.6</td>
<td>39.4</td>
<td>59.1</td>
<td>61.9</td>
<td>66.7</td>
<td>76.6</td>
<td>92.5</td>
<td>94.9</td>
<td>79.6</td>
<td>60.7</td>
<td>88.6</td>
<td>63.2</td>
</tr>
</tbody>
</table>

Table 3: Comparison with baselines on the Cityscapes-to-ACDC benchmark. The results are obtained on the held-out test set of ACDC whose annotation is not accessible. * represents results from our re-implementation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Arch.</th>
<th>Road</th>
<th>Sidewalk</th>
<th>Building</th>
<th>Wall</th>
<th>Fences</th>
<th>Pole</th>
<th>Tr.Lights</th>
<th>Vege</th>
<th>Person</th>
<th>Rider</th>
<th>Car</th>
<th>Bus</th>
<th>Motor</th>
<th>Bicycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADVENT [35]</td>
<td>C</td>
<td>72.9</td>
<td>14.3</td>
<td>40.5</td>
<td>10.6</td>
<td>21.2</td>
<td>9.3</td>
<td>17.4</td>
<td>7.4</td>
<td>18.3</td>
<td>32.6</td>
<td>19.5</td>
<td>69.5</td>
<td>56.2</td>
<td>34.5</td>
</tr>
<tr>
<td>BDL [18]</td>
<td>C</td>
<td>56.0</td>
<td>32.5</td>
<td>68.1</td>
<td>20.1</td>
<td>17.4</td>
<td>14.8</td>
<td>30.2</td>
<td>28.7</td>
<td>25.3</td>
<td>37.7</td>
<td>28.7</td>
<td>25.5</td>
<td>40.5</td>
<td>52.7</td>
</tr>
<tr>
<td>CLAN [21]</td>
<td>C</td>
<td>79.1</td>
<td>29.5</td>
<td>45.9</td>
<td>18.1</td>
<td>31.3</td>
<td>22.1</td>
<td>35.3</td>
<td>40.7</td>
<td>29.4</td>
<td>42.7</td>
<td>18.5</td>
<td>73.6</td>
<td>40.2</td>
<td>31.6</td>
</tr>
<tr>
<td>FDA [45]</td>
<td>C</td>
<td>74.6</td>
<td>73.2</td>
<td>70.1</td>
<td>63.3</td>
<td>50.9</td>
<td>54.7</td>
<td>52.3</td>
<td>47.0</td>
<td>44.9</td>
<td>48.3</td>
<td>39.5</td>
<td>34.7</td>
<td>29.5</td>
<td>28.6</td>
</tr>
<tr>
<td>MGCDMA [25]</td>
<td>C</td>
<td>73.4</td>
<td>28.7</td>
<td>69.9</td>
<td>19.3</td>
<td>26.3</td>
<td>36.8</td>
<td>53.0</td>
<td>53.3</td>
<td>75.4</td>
<td>32.0</td>
<td>84.6</td>
<td>51.0</td>
<td>26.1</td>
<td></td>
</tr>
<tr>
<td>ADVENT [35]</td>
<td>C</td>
<td>72.9</td>
<td>14.3</td>
<td>40.5</td>
<td>10.6</td>
<td>21.2</td>
<td>9.3</td>
<td>17.4</td>
<td>7.4</td>
<td>18.3</td>
<td>32.6</td>
<td>19.5</td>
<td>69.5</td>
<td>56.2</td>
<td>34.5</td>
</tr>
<tr>
<td>Oracle</td>
<td>C</td>
<td>98.0</td>
<td>84.2</td>
<td>92.6</td>
<td>39.4</td>
<td>59.1</td>
<td>61.9</td>
<td>66.7</td>
<td>76.6</td>
<td>92.5</td>
<td>94.9</td>
<td>79.6</td>
<td>60.7</td>
<td>88.6</td>
<td>63.2</td>
</tr>
</tbody>
</table>

ProDA [47], and transformer-based baselines DAFormer and HRDA [14]. Following [13], we report the mean Intersection over Union (mIoU) as our evaluation metrics averaged over three trials with different random seeds.

4.2. Comparison with Existing Methods

Comparative Evaluation. We compare our method, CDAC, combined with DAFormer and with HRDA, respectively, to other existing state-of-the-art methods on three benchmarks, GTAIV-to-Cityscapes (Tab. 1), Synthia-to-Cityscapes (Tab. 2), and Cityscapes-to-ACDC (Tab. 3).

We stress that it is important to evaluate and report model performance on a held-out test set. Most existing works in the UDA semantic segmentation literature focus only on the performance on the validation set of Cityscapes. Without testing on the held-out test set, however, we cannot ensure that model hyper-parameters are generalizable to other adaptation benchmarks without overfitting. During the experiments, we set $\lambda_{atts}$ to 1, which is the only hyper-parameter of CDAC, according to the performance on the validation set on ACDC, and applied the same value in the experiments with the other benchmarks. Further details about the hyper-parameter selection will be discussed in Sec. 4.7.

Evaluation Results. Tables 1–3 show that the introduction of DAFormer results in improvements over the CNN-based models when only source domain data are available. Notably, CDAC further improves the accuracy of the state-of-the-art method on all three benchmarks. On average, CDAC provides a 1.3 and 1.5 percent point (pp) improvement over the respective DAFormer and HRDA performance on GTAV-to-Cityscapes, 0.6 and 2.9 pp on Synthia-to-Cityscapes. This suggests that CDAC can flexibly combine with different Transformer-based models and universally mitigate the domain gap with attention-level and output-level consistency learning by effectively aligning domain gaps due to different imaging scenarios found in these datasets. Additionally, CDAC also outperforms the baseline on the Cityscapes-to-ACDC benchmark by 1.1 and 0.9 pp, indicating the improved generalization ability under different scenarios.

CDAC is modular and compatible with different existing DA baselines. For example, besides combining with the more advanced methods DAFormer and HRDA, CDAC also works on simpler baselines, like DACS, a critical component of these advanced methods, and even just Mean-Teacher, the basic model of all three previously mentioned methods. As illustrated in Fig 4, combined with Mean-Teacher, CDAC improves the accuracy from 39.0% to 41.18%. Similarly, combined with DACS, CDAC achieves
an accuracy improvement from 52.54% to 53.79%. These results show that our contribution is orthogonal to the existing methods and can therefore consistently benefit baselines with different levels of complexity.

4.3. Ablation Study

We compared the effectiveness of our final model based on DAFormer with its ablated versions that study the respective contributions of our cross-domain prediction consistency loss and cross-domain attention consistency loss (Tab. 4).

We report that adding the proposed combination of target-to-source (T2S) or source-to-target (S2T) losses (Eq. 4) yields a model mIoU performance that surpasses the baseline by 0.8 pp (row 1 vs. 4), while the use of the cross-domain attention consistency loss (Eq. 7) provides a 0.3 pp (row 1 vs. 5) improvement. Comparing the improvements brought by the cross-domain prediction consistency loss (row 1 vs. 4 and row 5 vs. 7) and cross-domain attention consistency loss (row 1 vs. 5 and row 4 vs. 7), we observe that the cross-domain prediction consistency loss yields relatively more improvements to the mean value of mIoU (0.8 pp from row 1 to 4 and 1.0 pp from row 5 to 7), while the cross-domain attention consistency loss contributes mainly on decreasing the standard deviation (-0.5 pp from row 1 to 5 and -0.1 pp from row 4 to 7). This could imply the different effects of our attention and output level alignment: the former helps better in narrowing down the random factors and strengthening the robustness of the training while the latter plays a more important role in directly boosting the accuracy. When we combined the cross-domain prediction consistency loss and the cross-domain attention consistency loss, our full model achieved the best performance.

We compared how our final model performs with and without query gradient stopping (row 6 vs. 7). We see that by stopping the gradient flowing back through the query in the cross-domain branches, we can improve the performance by 0.5 pp with even lower computation complexity. We conjecture that the improvement comes from the less aggressive alignment in the cross-domain branches, as the cross-domain attention map itself can be very noisy and possibly provide higher gradient signals to the wrong regions or pixels due to the domain discrepancy.

4.4. A Study of Perturbations on Attention Maps

We explained in earlier chapters that the effectiveness of our attention level cross-domain consistency learning could partially come from its perturbation to the self-attention in Transformer and therefore encourage more robust learning. We additionally tried several perturbations in the self-attention in DAFormer. We first tried to simply perturb the attention maps in both the source and target branches of the DAFormer by adding synthetic Gaussian noise (denoted by + Random). To better simulate attention maps, we generated noise from the normal distribution in a smaller resolution, upsampled it via bilinear interpolation to the actual size of an attention map to smoothen it, and then calculated the attention score through a softmax layer. Finally we took the average of the original attention maps with the synthetic noise to simulate the perturbation on original self-attention. Meanwhile we also followed another recent work [15] that advocates introducing uniform attention maps during the training phase (denoted by + Uniform). Distracted by the denser uniform attention, the model is forced to exploit more informative and crucial signals from the self-attention, and the model can therefore benefit from the concentrated attention to achieve better performance with barely extra computational cost. Technically, it requires only a context broadcasting (CB) module after the self-attention module to manually inject uniformity into self-attention. We also added the CB module on both the source and target branches.

Table 4: Ablation study on the GTAV-to-Cityscapes benchmark, reporting average and standard deviation of mIoU over three runs. S&T represents the performance of DAFormer, T2S and S2T represent the use of $L_{t2s}$ and $L_{s2t}$ in Eq. 4, and Attn represents use of $L_{Attn}$ in Eq. 7. No GS does not use gradient stopping on the query vectors (by default, we use gradient stopping).

<table>
<thead>
<tr>
<th>Method</th>
<th>S&amp;T</th>
<th>T2S</th>
<th>S2T</th>
<th>Attn</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean-Teacher</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>68.3 ± 0.7</td>
</tr>
<tr>
<td>DACS</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>68.1 ± 0.8</td>
</tr>
<tr>
<td>DAFormer</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>68.5 ± 0.2</td>
</tr>
<tr>
<td>HRDA</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>69.1 ± 0.6</td>
</tr>
<tr>
<td>GS</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>68.6 ± 0.2</td>
</tr>
<tr>
<td>No GS</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>69.1 ± 0.8</td>
</tr>
<tr>
<td>+ GS</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>69.6 ± 0.5</td>
</tr>
</tbody>
</table>
Figure 5: Attention map visualization for the query pixel ⭐ from different models trained with a different types of regularization methods, including adding random noises or uniformity to the attention maps. Compared with our baseline DAFormer, we observe that the evidence of attention-level discrepancy, i.e., the spurious artifacts on the left margin, still exists when uniformity is added to the attention maps, while the random noises indeed partially eliminate the artifacts. In comparison, CDAC can more effectively address the issue.

Table 5: Comparison of mIoU with DAFormer backbone when attention maps are perturbed.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>DAFormer</th>
<th>+ Random</th>
<th>+ Uniform</th>
<th>+ CDAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTAV-to-Cityscapes</td>
<td>68.3</td>
<td>67.3</td>
<td>66.9</td>
<td>69.6</td>
</tr>
<tr>
<td>Synthia-to-Cityscapes</td>
<td>60.9</td>
<td>60.4</td>
<td>59.1</td>
<td>61.5</td>
</tr>
</tbody>
</table>

We can observe that both aforementioned perturbation methods fail to provide improvement compared to the performance of DAFormer on GTAV-to-Cityscapes (68.3%) and Synthia-to-Cityscapes (60.9%). From Fig. 5, the random perturbation alleviates spurious attention in some cases (first and second rows). However, uniform attention does not improve attention maps. The models trained with these synthetic perturbations still suffer from the attention-level domain gaps similar to DAFormer. The quantitative results in Tab. 5 also show that the use of random or uniform attention even hurts the performance of DAFormer. We believe that the reason is because those perturbations do not help attention-level alignment, and it is not enough to improve the performance under UDA settings by naively perturbing the self-attention without addressing the domain gap issues. In comparison, CDAC shares query vectors across domains in the attention module and thus manages to satisfy both adaptation and perturbation. It therefore better addresses the domain discrepancy on different levels, as shown qualitatively in Fig. 5 and quantitatively in Sec. 4.2.

Figure 6: Comparison of the overall performance of models with different encoder sizes.

4.5. Effects of Encoder Architecture

We tested DAFormer and our model with different encoder architectures, i.e. from MiT-B1 to MiT-B5, and studied the effects of their size on the overall performance. The results are presented in Fig. 6, which shows that the accuracy of both models significantly increases with respect to the increase of the capacity of the encoder. Nevertheless, as the encoder grows deeper, DAFormer’s accuracy exhibits a tendency to saturate. In contrast, our approach consistently delivers significant performance enhancements, revealing the effectiveness of CDAC.

4.6. Qualitative Results

We present qualitative results in Fig. 7. We observe that when the attention-level domain discrepancies, i.e. the spurious regions on the left margin of the attention maps, a-
Figure 7: Illustration of the attention maps and predictions obtained from source-only, DAFormer, and our model (DAFormer+CDAC). With better attention-level alignment, our model effectively addresses the spurious artifact in the attention map (first row) that misleads the predictions and therefore makes more accurate segmentation predictions, e.g., within the traffic light region (red box, second row) and the vehicle (orange box, third row).

4.7. Sensitivity Analysis

As we introduced in Sec. 4.2, the verification of the generalizability of UDA segmentation methods is regularly ignored in many existing works, and we cannot exclude the possibility that their methods are overtuned on a few specific adaptation tasks or target datasets. To this end, we tested the sensitivity of CDAC to its hyper-parameter, i.e., $\lambda_{\text{attn}}$. Specifically, we ran experiments with attention weight $\lambda_{\text{attn}} = 0.1$, 1, and 10 on GTAV-to-Cityscapes, Synthia-to-Cityscapes, and Cityscapes-to-ACDC. We report validation mIoU on Cityscapes and ACDC in Fig. 8.

Based on the results, we can see that our current setting of $\lambda_{\text{attn}} = 1$ stably achieves the best performance on the three datasets. When it is set lower, it does not provide strong enough help to attention-level alignment, which affects the overall performance. On the other hand, the accuracy plummets when the attention weight is set too high, as the consistency learning starts suppressing other learning objectives. More importantly, we observe a similar tendency of performance with respect to $\lambda_{\text{attn}}$ among all three benchmarks, which indicates that CDAC is generalizable, rather than dependent on a specific group of hyperparameters or benchmarks.

5. Conclusion

Despite the significant improvement by the recent introduction of Transformers to UDA semantic segmentation, these still struggle with noticeable attention-level domain discrepancy. In this work, we leverage consistency between the predictions from self-attention and cross-domain attention and address both attention-level and output-level domain shifts. Further assisted by the proposed cross-domain attention consistency loss, CDAC universally facilitates attention-level alignment and encourages Transformers to be more robust to features from different levels on both domains and therefore improves the performance on the target domain. Extensive experiments as well as ablation studies verify the effectiveness and reliability of CDAC, and the generalizability under diverse domains.

6. Acknowledgements

This work has been partially supported by ONR MURI grant N00014-19-1-2571 associated with AUSMURIB000001 and Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University).
References


[40] Haiyan Wu, Yuting Gao, Yinqi Zhang, Shaohui Lin, Yuan Xie, Xing Sun, and Ke Li. Self-supervised models are good teaching assistants for vision transformers. In ICML, 2022. 3

[41] 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. 2


[47] 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 CVPR, 2021. 2, 5, 6


