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

Kaihong Wang[1], Donghyun Kim[2], Rogerio Feris[3], Margrit Betke[1]
[1]Boston University, [2]Korea University, [3]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.

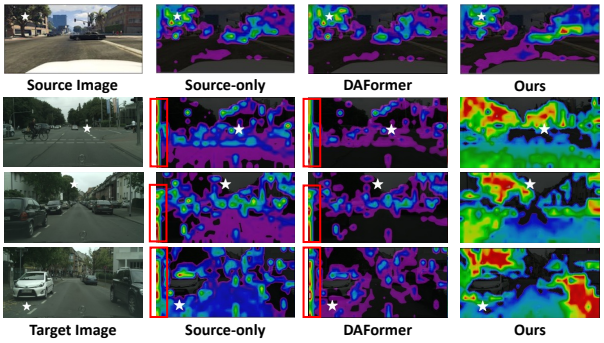

Figure 1: Attention map visualization for the query pixel ☆ in source and target domains from different methods. The attention maps on the source image tend to highlight regions sharing similar semantic classes. However, the attention maps on the target image from Source-only and prior work focus on spurious regions (*e.g.*, left red boxes), which can be caused by a domain gap. In comparison, CDAC encourages attention-level adaptation and learns from more diverse and informative signals from both domains.

We found that the attention maps from self-attention on the target images can focus on spurious regions and therefore fails 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 at 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 Segmenter [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 $D_s = \{(x_s^i, y_s^i)\}_{i=1}^{N^s}$ with corresponding labels and an unlabeled target domain dataset $D_t = \{x_t^i\}_{i=1}^{N^t}$, our task is to learn a student semantic segmentation model F with its self-ensemble teacher G and optimize the student's performance on the target domain. At step t, the teacher's weights ϕ_t are updated as the moving average of the student's weight θ_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 value vector V sharing the same dimensions of $N \times C$. Thus, the self-attention can be obtained by:

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_{head}}})V, \qquad (1)$$

where d_{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

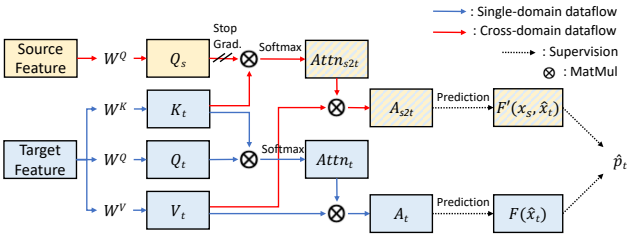

Figure 2: Illustration of our consistency learning with Cross-domain Attention module for UDA semantic segmentation on the target and source-to-target branches. We swap the query vectors so the cross-domain branches take features from both domains. The computation for the source and target-to-source branches is symmetrical, except that the supervision here is the pseudo-label $\hat{p}_t$ generated from the target image $\hat{x}_t$ while that for the source and target-to-source branches is from the source label y_s.

F on a source image x_s and its corresponding label y_s:

$$L_s^{(i)} = -\sum_{j=1}^{H\times W}\sum_{c=1}^{C} y_s^{(i,j,c)} \log F(x_s^i)^{(j,c)}, \qquad (2)$$

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:

$$L_t^{(i)} = -\sum_{j=1}^{H\times W}\sum_{c=1}^{C} q^i \hat{p}_t{}^{(i,j,c)} \log F(\hat{x}_t^i)^{(j,c)}, \qquad (3)$$

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 = Attention(Q_s, K_s, V_s) and A_t = 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. $F'(\hat{x}_t, x_s)$ extracts initial features from the source image x_s and produces output based on the target-to-source attention A_{t2s} = Attention(Q_t, K_s, V_s). Symmetrically, $F'(x_s, \hat{x}_t)$ extracts initial features from the target image $\hat{x}_t$ and computes the output based on the source-to-target attention A_{s2t} = 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:

$$
\begin{aligned}
L_{t2s}^{(i)} &= -\sum_{j=1}^{H\times W}\sum_{c=1}^{C} y_s^{(i,j,c)} \log F'(\hat{x}_t^i, x_s^i)^{(j,c)} \\
L_{s2t}^{(i)} &= -\sum_{j=1}^{H\times W}\sum_{c=1}^{C} q^i \hat{p}_t{}^{(i,j,c)} \log F'(x_s^i, \hat{x}_t^i)^{(j,c)}.
\end{aligned}
\tag{4}
$$

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}:

$$
L_{pred} = \frac{1}{2}(L_s + L_{t2s}) + \frac{1}{2}(L_t + L_{s2t}). \tag{5}
$$

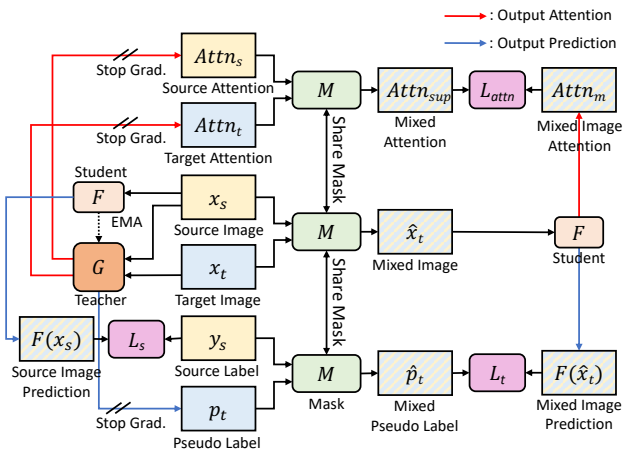

Figure 3: Illustration of our cross-domain attention consistency loss and its relation to L_s and L_t. With the mean-teacher framework that guides the training of the student model by the prediction of the teacher on an unlabeled target image through augmentations (DACS), we propose to enforce the consistency on the intermediate attention maps in a similar way.

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 $\{Attn_s^i\}_{i=1}^L$ and $\{Attn_t^i\}_{i=1}^L$ via L attention modules of the teacher model G from the source image x_s and the target image x_t, as well as $\{Attn_m^i\}_{i=1}^L$ via the student model F from the mixed image $\hat{x}_t$ (in Sec. 3.1). The source and the target attention maps $\{Attn_s^i\}_{i=1}^L$ and $\{Attn_t^i\}_{i=1}^L$ will be mixed by M that produces $\hat{x}_t$. Then we compute the mixed attention maps $\{Attn_{sup}^i\}_{i=1}^L$ to regularize the consistency on attention maps

$$
Attn_{sup}^i[:,:,j,:] = \begin{cases} Attn_s^i[:,:,j,:], & M'^j = 1, \\ Attn_t^i[:,:,j,:], & M'^j = 0, \end{cases} \tag{6}
$$

where M' is resized from M and j represents the spatial location. Finally, we minimize the Kullback–Leibler divergence to guide $Attn_m$ towards $Attn_{sup}$. Additionally, we apply a valid mask M_v to avoid misleading $Attn_m$ to attend

Table 1: Comparison with baselines on the GTAV-to-Cityscapes benchmark. The mIoU represents the average of individual mIoUs among all 19 categories between GTAV 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.

Method	Arch.	Road	Sidewalk	Building	Wall	Fence	Pole	TrLight	TrSign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motor.	Bicycle	mIoU
Source only	C	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
CBST [49]	C	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
DACS [31]	C	89.9	39.7	87.9	30.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1
CorDA [37]	C	94.7	63.1	87.6	30.7	40.6	40.2	47.8	51.6	87.6	47.0	89.7	66.7	35.9	90.2	48.9	57.5	0.0	39.8	56.0	56.6
ProDA [47]	C	87.8	56.0	79.7	46.3	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
Source only	T	71.5	18.0	84.2	34.4	30.9	33.4	44.3	23.5	87.4	41.3	86.6	64.0	22.5	88.3	44.5	39.1	2.3	35.2	31.6	46.5
DAFormer [13]	T	95.7	70.2	89.4	53.5	48.1	49.6	55.8	59.4	89.9	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.3
DAFormer + CDAC	T	96.5	73.9	89.5	56.8	48.9	50.7	55.8	63.3	89.9	49.1	91.2	72.2	45.4	92.7	78.3	82.9	67.5	55.2	63.4	69.6 (**+1.3**)
HDRA [14]	T	96.4	74.4	91.0	**61.6**	51.5	57.1	63.9	69.3	91.3	48.4	94.2	79.0	**52.9**	93.9	84.1	85.7	75.9	63.9	67.5	73.8
HRDA + CDAC	T	**97.1**	**78.7**	**91.8**	59.6	**57.1**	**59.1**	**66.1**	**72.2**	**91.8**	**53.1**	**94.5**	**79.4**	51.6	**94.6**	**84.9**	**87.8**	**78.7**	**64.9**	**67.6**	**75.3** (**+1.5**)
Oracle	T	98.0	84.2	92.6	59.4	59.7	61.9	66.7	76.6	92.5	66.4	94.9	79.6	60.7	94.6	84.0	88.6	81.2	63.2	75.0	77.9

regions masked out in $\hat{x}_t$.

$$L_{attn} = \frac{1}{L}\sum_{i=1}^{L} \mathrm{KL}(Attn^i_{sup}, Attn^i_m) \odot M_v$$
$$M_v[j,:] = \begin{cases} M, & M'^j = 1, \\ 1 - M, & M'^j = 0. \end{cases} \quad (7)$$

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

$$L = L_{pred} + \lambda_{attn} L_{attn}. \quad (8)$$

In 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 from 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 6×10^{-5} for the encoder, 6×10^{-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. GTAV [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 GTAV 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.

Method	Arch.	Road	Sidewalk	Building	Wall	Fence	Pole	Tr.Light	Tr.Sign	Veget.	Sky	Person	Rider	Car	Bus	Motor.	Bicycle	mIoU
Source only	C	64.3	21.3	73.1	2.4	1.1	31.4	7.0	27.7	63.1	67.6	42.2	19.9	73.1	15.3	10.5	38.9	34.9
CBST [49]	C	68.0	29.9	76.3	10.8	1.4	33.9	22.8	29.5	77.6	78.3	60.6	28.3	81.6	23.5	18.8	39.8	42.6
DACS [31]	C	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	90.8	67.6	38.3	82.9	38.9	28.5	47.6	48.3
CorDA [37]	C	**93.3**	61.6	85.3	19.6	5.1	37.8	36.6	42.8	84.9	90.4	69.7	41.8	85.6	38.4	32.6	53.9	55.0
ProDA [47]	C	87.8	45.7	84.6	37.1	0.6	44.0	54.6	37.0	**88.1**	84.4	74.2	24.3	88.2	51.1	40.5	45.6	55.5
Source only	T	51.5	20.3	79.2	19.3	1.8	40.9	29.9	22.7	79.1	82.4	63.0	24.9	75.8	33.7	18.9	24.9	35.2
DAFormer [13]	T	84.5	40.7	88.4	41.5	6.5	50.0	55.0	54.6	86.0	89.8	73.2	48.2	87.2	53.2	53.9	61.7	60.9
DAFormer + CDAC	T	83.7	42.9	87.4	39.8	7.5	50.7	55.7	53.5	85.9	90.9	74.5	47.2	86.0	60.2	57.8	60.8	61.5 (**+0.6**)
HRDA [14]	T	85.2	47.7	88.8	49.5	4.8	57.2	**65.7**	60.9	85.3	92.9	79.4	52.8	89.0	**64.7**	63.9	64.9	65.8
HRDA + CDAC	T	93.1	**68.5**	**89.8**	**51.2**	**8.9**	**59.4**	65.5	**65.3**	84.7	**94.4**	**81.2**	**57.0**	**90.5**	56.9	**66.8**	**66.4**	**68.7** (**+2.9**)
Oracle	T	98.0	84.2	92.6	59.4	59.7	61.9	66.7	76.6	92.5	94.9	79.6	60.7	94.6	88.6	63.2	75.0	78.0

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.

Method	Arch.	Road	Sidewalk	Building	Wall	Fence	Pole	Tr.Light	Tr.Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motor.	Bicycle	mIoU
ADVENT [35]	C	72.9	14.3	40.5	16.6	21.2	9.3	17.4	21.3	63.8	23.8	18.3	32.6	19.5	69.5	36.2	34.5	46.2	26.9	36.1	32.7
BDL [18]	C	56.0	32.5	68.1	20.1	17.4	15.8	30.2	28.7	59.9	25.3	37.7	28.7	25.5	70.2	39.6	40.5	52.7	29.2	38.4	37.7
CLAN [21]	C	79.1	29.5	45.9	18.1	21.3	22.1	35.3	40.7	67.4	29.4	32.8	42.7	18.5	73.6	42.0	31.6	55.7	25.4	30.7	39.0
FDA [45]	C	74.6	**73.2**	70.1	**63.3**	**59.0**	54.7	52.3	47.0	44.9	44.8	43.3	39.5	34.7	29.5	28.6	28.5	28.3	28.2	24.8	45.7
MGCDA [25]	C	73.4	28.7	69.9	19.3	26.3	36.8	**53.0**	53.3	**75.4**	32.0	84.6	51.0	26.1	77.6	43.2	45.9	53.9	32.7	41.5	48.7
DAFormer [13]	T	58.4	51.3	84.0	42.7	35.1	50.7	30.0	57.0	74.8	52.8	51.3	58.3	32.6	82.7	58.3	54.9	82.4	44.1	50.7	55.4
DAFormer + CDAC	T	57.6	43.7	85.1	43.5	33.9	50.1	42.9	53.9	72.8	52.9	52.2	59.4	34.7	83.6	60.4	68.7	84.3	41.4	53.0	56.5 (**+1.1**)
HRDA* [14]	T	67.1	64.8	**86.7**	52.9	39.7	**56.5**	41.1	**62.5**	74.0	**58.8**	61.7	**69.8**	42.4	87.6	**71.7**	**82.0**	88.7	46.8	**61.8**	64.0
HRDA* + CDAC	T	**87.0**	56.7	84.5	53.5	34.3	54.6	43.6	51.4	71.7	58.6	**85.4**	68.7	**45.7**	**89.0**	70.9	81.5	**90.1**	**47.6**	59.0	**64.9**(**+0.9**)

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, GTAV-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 λ_{attn} 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 Citsycapes-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

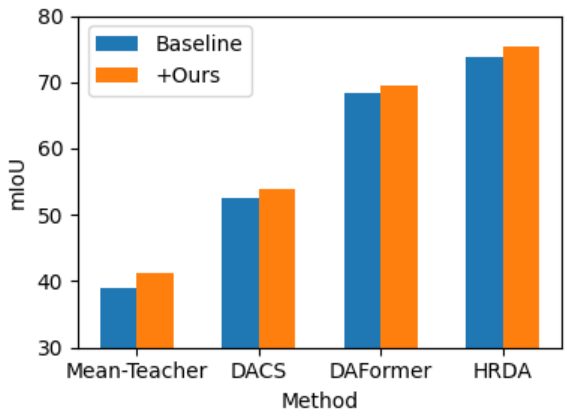

Figure 4: Performance of CDAC combined with baselines with different levels of complexity.

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.

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).

	S&T	T2S	S2T	Attn	mIoU
1	✓				68.3 ± 0.7
2	✓	✓			68.1 ± 0.8
3	✓		✓		68.5 ± 0.2
4	✓	✓	✓		69.1 ± 0.6
5	✓			✓	68.6 ± 0.2
6	✓	No GS	No GS	✓	69.1 ± 0.8
7	✓	✓	✓	✓	69.6 ± 0.5

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

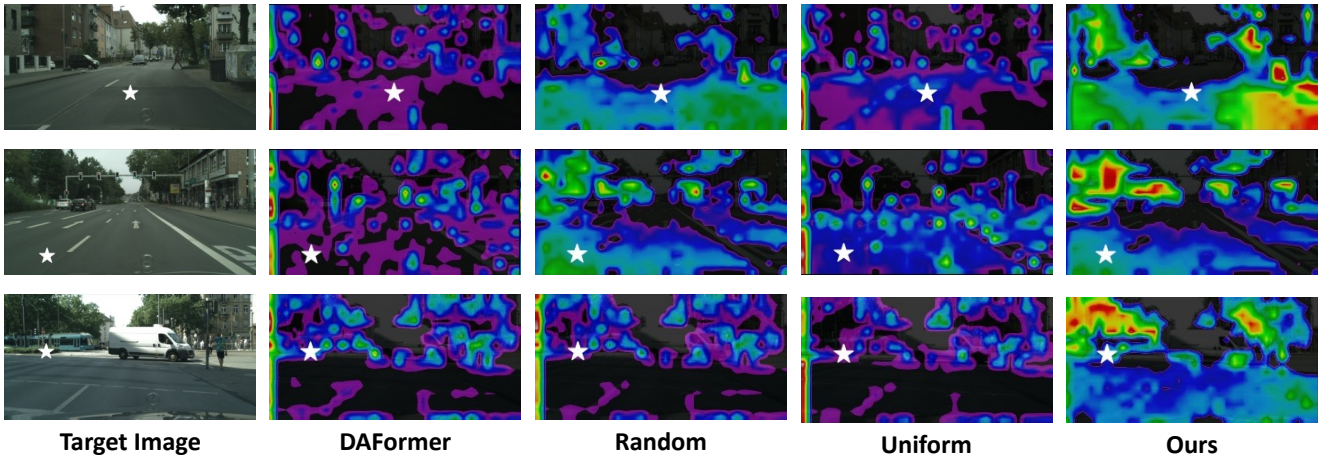

Figure 5: Attention map visualization for the query pixel ☆ from different models trained with a different types of regularization methods, including adding random noises or uniformity to the attention maps. Compared with our baseline DAFormer, we observe that the evidence of attention-level discrepancy, *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.

Benchmark	DAFormer	+ Random	+ Uniform	+ CDAC
GTAV-to-Cityscapes	68.3	67.3	66.9	69.6
Synthia-to-Cityscapes	60.9	60.4	59.1	61.5

of DAFormer with original configurations. The results for both the random and uniform attention versions are reported in Tab. 5, and the visualization of their attention maps are compared in Fig. 5.

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.

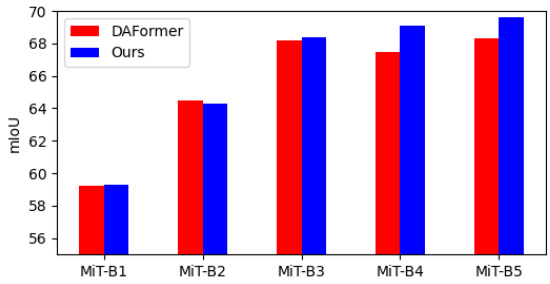

Figure 6: Comparison of the overall performance of models with different encoder sizes.

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, ap-

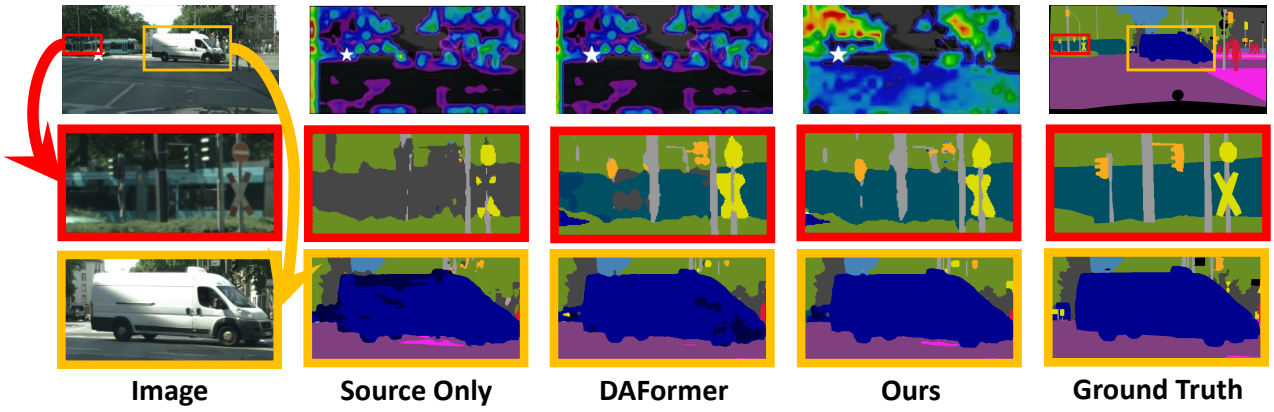

Figure 7: Illustration of the attention maps and predictions obtained from source-only, DAFormer, and our model (DAFormer+CDAC). With better attention-level alignment, our model effectively addresses the spurious artifact in the attention map (first row) that misleads the predictions and therefore makes more accurate segmentation predictions, *e.g.*, within the traffic light region (red box, second row) and the vehicle (orange box, third row)

pear in the attention maps from the baselines (first row), the predictions near the region (second row) and far from the region (third row) are also affected. In contrast, CDAC effectively addresses the attention-level discrepancy and improves the prediction at the same time. This intuitively reveals the correlation between attention-level discrepancies with the quality of predictions and showcases how CDAC overcomes the issue.

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.*, λ_{attn}. Specifically, we ran experiments with attention weight λ_{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 λ_{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 λ_{attn} among all three benchmarks, which indicates that CDAC is generalizable, rather than dependent on a specific group of hyperparameters or benchmarks.

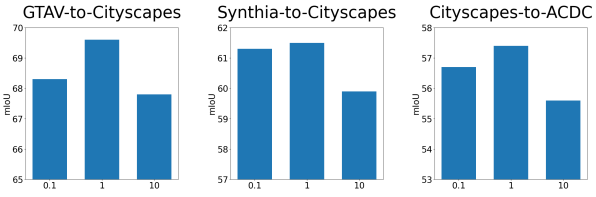

Figure 8: Sensitivity analysis on λ_{attn}. CDAC shows stable performance over hyper-parameters on each setting.

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 domians.

6. Acknowledgements

This work has been partially supported by ONR MURI grant N00014-19-1-2571 associated with AUS-MURIB000001 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

[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *CVPR*, 2021. 2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 2

[4] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *ICCV*, 2019. 2

[5] 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 *CVPR*, 2016. 5

[6] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *CVPR*, 2022. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 2

[8] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. SSF-DAN: Separated semantic feature based domain adaptation network for semantic segmentation. In *ICCV*, 2019. 2

[9] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *ICLR*, 2018. 2

[10] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *NeurIPS*, 2006. 3

[11] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 2

[12] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 2

[13] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022. 1, 2, 3, 5, 6

[14] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *ECCV*, 2022. 5, 6

[15] Nam Hyeon-Woo, Kim Yu-Ji, Byeongho Heo, Dongyoon Han, Seong Joon Oh, and Tae-Hyun Oh. Scratching visual transformer's back with uniform attention. *arXiv preprint arXiv:2205.01917*, 2022. 3, 7

[16] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *CVPR*, 2020. 2

[17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 2021. 2

[18] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, 2019. 2, 6

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[21] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, 2019. 2, 6

[22] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 5

[23] Germán Ros, Laura Sellart, Joanna Materzynska, David Vázquez, and Antonio M. López. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 5

[24] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, 2021. 5

[25] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *TPAMI*, 2022. 5, 6

[26] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 2014. 4

[27] Robin Strudel, Ricardo Garcia Pinel, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 2

[28] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *ICLR*, 2017. 2

[29] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 1

[30] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 2

[31] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. DACS: Domain adaptation via cross-domain mixed sampling. In *WACV*, 2021. 2, 3, 5, 6

[32] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 2

[33] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *ICCV*, 2019. 2

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[35] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 2, 5, 6

[36] Kaihong Wang, Chenhongyi Yang, and Margrit Betke. Consistency regularization with high-dimensional non-adversarial source-guided perturbation for unsupervised domain adaptation in segmentation. In *AAAI*, 2021. 2

[37] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *ICCV*, 2021. 5, 6

[38] Xiyu Wang, Pengxin Guo, and Yu Zhang. Domain adaptation via bidirectional cross-attention transformer. *arXiv preprint arXiv:2201.05887*, 2022. 3

[39] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogério Feris, Jinjun Xiong, Wen-Mei Hwu, Thomas S. Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *CVPR*, 2020. 2

[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

[42] 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. 1, 2, 5

[43] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. In *ICLR*, 2022. 2

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

[45] Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020. 2, 6

[46] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 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

[48] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 2

[49] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 2, 5, 6

[50] Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *ICCV*, 2019. 2

[51] Andrea Zunino, Sarah Adel Bargal, Pietro Morerio, Jianming Zhang, Stan Sclaroff, and Vittorio Murino. Excitation dropout: Encouraging plasticity in deep neural networks. *IJCV*, 2021. 4