UNSUPERVISED DOMAIN ADAPTATION VIA PSEUDO-LABELS AND OBJECTNESS CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pseudo label self-training has emerged as a dominant approach to unsupervised domain adaptation (UDA) for semantic segmentation. Despite recent advances, this approach is susceptible to erroneous pseudo labels arising from confirmation bias that ultimately leads to sub-optimal segmentation. To mitigate the effect of noisy pseudo-labels, we propose regularising conventional self-training objectives with constraints that are derived from structure-preserving modalities, such as depth. Towards this end, we introduce a *contrastive image-level objectness constraint* that pulls the pixel representations of the same object *instance* closer while pushing those from different object *categories* apart. To identify pixels within an object, we subscribe to a notion of objectness derived from depth maps, that are robust to photometric variations, as well as superpixels, that are obtained via unsupervised clustering over the raw image space. Crucially, the objectness constraint is agnostic to the ground-truth semantic segmentation labels and, therefore, remains appropriate for unsupervised adaptation settings. In this paper, we show that our approach of leveraging multi-modal constraint improves top performing self-training methods in various UDA benchmarks for semantic segmentation. We make our code and data-splits available in the supplementary material.

1 INTRODUCTION

Semantic segmentation is a crucial and challenging task for such application as autonomous driving (Zhang et al., 2020; Araslanov et al., 2021; Vu et al., 2019; Zhang et al., 2019; Hoffman et al., 2018), where the goal is to label each pixel of a given image, according to a set of semantic classes. In the past few years, advances in deep learning models (Chen et al., 2018) have led to impressive performance on this task. However, an important limitation arises from the excessive cost and time taken to annotate images at the pixel level, in order to generate training data for these deep models. For instance, the time taken to densely annotate Cityscapes Cordts et al. (2016), a popular dataset with 5000 images, was reported to be 1.5 hours per image. Further, a small dataset like Cityscapes does not cover enough variations in outdoor scenes such as weather conditions and country-specific traffic layouts that can be crucial for, e.g.,large scale deployment of autonomous vehicles. Training models that cater to such scene variations could significantly add to the cost of annotation.

To address the annotation problem, synthetic datasets curated from 3D simulation environments like GTA (Richter et al., 2016) and SYNTHIA (Ros et al., 2016) have been proposed as cost-effective alternatives. While large amounts of data can be generated and automatically annotated using these engines, there is often a significant gap in the visual characteristics of generated and real images. Such a shift in the data domains can adversely impact the real-world performance (Tsai et al., 2018; Hoffman et al., 2018) of a model trained on the synthetic data. This issue of domain shift, in addition to problem of costly annotation, has motivated wide research in unsupervised domain adaptation for semantic segmentation. Among the various domain adaptation approaches Ganin et al. (2017); Tsai et al. (2018); Bousmalis et al. (2016a); Zou et al. (2018); Zhang et al. (2019; 2020); Araslanov et al. (2021) proposed in the recent years, self-training has emerged as a particularly promising direction. In the context of domain adaptation theory Ben-David et al., self-training provides a significant edge over prior line of works Ganin et al. (2017); Tsai et al. (2018); Bousmalis et al. (2016a); Zou et al. (2018); Hoffman et al. (2018); Shu et al. (2018) in that it can handle conditional distribution shifts in a simple and effective manner. The basic process behind self-training based unsupervised adaptation involves pseudo labelling (unlabelled) target data using a *seed* model solely trained on

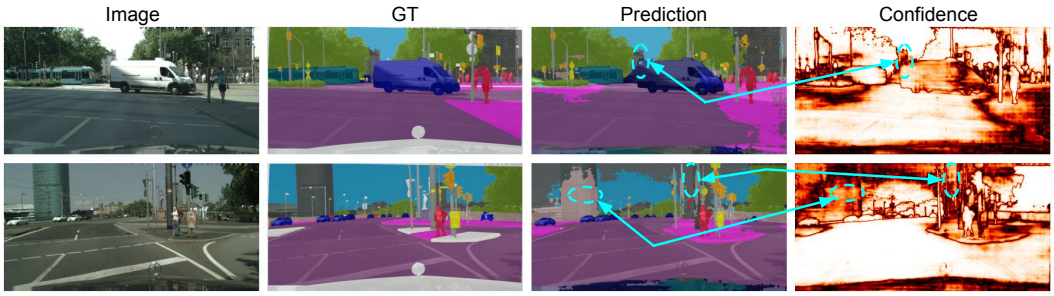

Figure 1: **Motivation for Objectness Constraints:** Shows two examples of an image, its ground-truth (GT) segmentation label, its prediction from a source-initialized model and the confidence of prediction (brighter regions are more confident). The blue dashed-circles highlight the high confidence regions that are mis-classified by the model. The goal of the our objectness constraints is to implicitly penalize such mis-classifiations

the source data. The target labels are assigned only if the model predictions exceed a predefined threshold. Training the model further with source data augmented with the pseudo-labelled target data ultimately improves target domain accuracy.

An important limitation of the self-training approach is its susceptibility to erroneous pseudo labels arising from confirmation bias IJCNN (2020) of the seed model. In the context of domain adaptation, the seed model is solely trained on the source domain. As a result, this model might learn to infer semantic label based on nuisance factors specific to the source domain images that may lead to highly confident but erroneous predictions on target domain images. For instance, if the source domain primarily consists of scenes captured in daylight (a nuisance factor), then bright spots on target objects might frequently be classified as the sky, irrespective of the actual semantic label (illustration in Figure 1). Continued training of the seed model on such spurious predictions can ultimately lead to suboptimal target performance. We argue that in current methods, confirmation bias primarily arises from heavy reliance on photometric cues for predicting the semantic label. However, important structural cues that can disambiguate spurious predictions are largely overlooked. Thus, in this work we seek to improve self-training performance by regularizing it with a constraint based on one such cue, i.e., objectness.

Towards this end, we design a multimodal objectness constraint that is derived from depth information and image-based clustering. Most datasets in semantic segmentation are usually accompanied with depth maps that are aligned with the corresponding raw images. We use these maps to compute depth histograms Dinh et al. (2014) and combine it with SLIC Achanta et al. (2012) to define *superpixels*. These superpixels oversegment an image in a way that respects object boundaries based on depth as well as visual similarity and thus, form the basis of our notion of objectness. Finally, the constraint is formulated as a contrastive objective Sohn (2016); Chen et al. (2020); Khosla et al. (2020) that pulls together pixel-wise representations of the same object *instance* while pushes away those belonging to different object *categories*. Such a contrastive objective enforces overall object consistency that can resolve photometric ambiguities. To facilitate contrastive learning in the target domain, we extract complementary information from the superpixels as well as pseudo labels. As an important feature, our contrastive objective is agnostic to target domain labels that dovetails with the unsupervised domain adaptation setting. To summarise our contributions,

- We propose a novel objectness constraint that is derived using multi-modal information and regularises self-training based adaptation. Given its agnostic nature to ground truth categories, such a constraint *normalizes*, to some extent, statistical effects of different semantic classes based on their frequency in the source domain.
- We show that regularizing existing self-training approaches can lead to superior target domain performance on popular benchmarks such as `GTA` $\rightarrow$ `Cityscapes` and `SYNTHIA` $\rightarrow$ `Cityscapes`.
- Further, we empirically demonstrate that our proposed constraint is general enough to be plugged into any base self-training method and improve performance.

2 RELATED WORK

Unsupervised domain adaptation. Unsupervised domain adaptation (UDA) is of particular importance in complex structured-prediction problems, such as semantic segmentation in autonomous driving, where the domain gap between a source domain (e.g., an urban driving dataset) and target domain (real-world driving scenarios) can have devastating consequences on the efficacy of deployed models. Several approaches have been proposed for learning domain invariant representations, e.g., through adversarial feature alignment (Ganin et al., 2017; Bousmalis et al., 2016b; Tzeng et al., 2017), which addresses the domain gap by minimising a distance metric that characterises the divergence between the two domains (Pan et al., 2011; Long et al., 2015; 2017; Baktashmotlagh et al., 2013; Shalit et al., 2017; Courty et al., 2017; Si et al., 2010; Muandet et al., 2013). Problematically, such approaches address only shifts in the marginal distribution of the covariates or the labels and, therefore, prove insufficient for handling the more complex shifts in the conditionals (Johansson et al., 2019; Zhao et al., 2019; Wu et al., 2019). Self-training approaches have been proposed to induce category-awareness (Zhang et al., 2019) or cluster density-based assumptions (Shu et al., 2018), in order to anchor or regularise conditional shift adaptation, respectively. We build upon these works, in this paper, by jointly introducing category-awareness through the use of pseudo-labeling strategies and regularisation through the definition of contrastive depth-based objectness constraints.

Self-training with pseudo-labels. Applications of self-supervised learning (as in *self-training*) have become popular in the sphere of domain adaptation for semantic segmentation (Zou et al., 2018; Li et al., 2019; Zhang et al., 2019). Here, pseudo-labels are assigned to observations from the target domain, based on the semantic classes of high-confidence (e.g., the closest or least-contrastive) category centroids (Zhang et al., 2019; Xie et al., 2018), prototypes (Chen et al., 2019), cluster centers (Kang et al., 2019), or superpixel representations (Zhang et al., 2020) that are learned by a model trained on the source domain. Crucially, because these approaches make use the source model for generating pseudo-labels, they can fall prey to a *confirmation bias*, wherein highly-confident but false-positive class predictions are assigned to target observations, particularly for majority classes like 'sky' and 'road', thereby degrading the adaption performance. Moreover, despite auxiliary modalities (e.g., depth information) being largely available in autonomous driving applications, these approaches rely on transformations of the same modality that is being adapted, yielding vulnerability to the same sources of distribution shift and negative transfer (Tian et al., 2020; Tsai et al., 2020). Stekovic et al. (2020) propose the use of depth information to register and reconcile indoor room geometry constraints, as an anchoring basis for semantic segmentation. While we are inspired by this work, we observe that the learned geometric constraints apply best to other indoor environments that have similar geometrical regularity as those in which the model was originally trained; this conflicts with our focus on semantic segmentation for outdoor urban driving scenes.

Adaptation with multiple modalities. Learning and adaptation in multimodal contexts presents an opportunity for leveraging complementarity between different views of the input space, to improve model robustness and generalisability. In the context of unsupervised domain adaptation for semantic segmentation, additional modalities (e.g., depth information) can play a crucial role in distinguishing between two types of feature variation in the primary modality used for prediction (i.e., RGB camera images): photometric variation and structural variation. Vu et al. (2019) propose the use of depth as an auxiliary prediction objective, with the intention of learning structure-aware features, combining them with the primary semantic segmentation prediction branch, then performing adversarial adaptation on top of this fused representation. While this work shares similar underlying motivation for our use of auxiliary information, direct fusion of multimodal context in adaptation can yield noisy observations that are subject to shifts in the data distribution. In contrast to this method, we propose the use of contrastive depth-based objectness constraints, which remain robust to the photometric variation in the RGB signal, whilst alleviating concerns of distribution shift in depth-based feature extractors or multi-task prediction heads.

3 PRELIMINARIES

We begin by introducing preliminary concepts on self-training based adaptation. These concepts serve as a bases for introducing multi-modal constraints in Section 4 that is then used to regularise the self-training methods. We refer to our complete framework as `PAC-UDA` that uses **P**seudo-labels **A**nd objectness **C**onstraints for self-training in **U**nsupervised **D**omain **A**daptation for semantic

segmentation. Although, we propose a specific form of self-training for formalising PAC-UDA, our regulariser can be treated as an independent module that can encompass other formulations as well (see experiments).

Unsupervised Domain Adaptation (UDA) for Semantic Segmentation: Consider a dataset $D^s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ of input-label pairs sampled from a source domain distribution, $P^s_{X\times Y}$. The input and label share the same spatial dimensions, $H \times W$, where each pixel of the label is assigned a class $c \in \{1, \ldots, C\}$ and represented via a C dimensional one-hot encoding. We also have access to a dataset $D^t = \{(x_i^t, \cdot)\}_{i=1}^{N_t}$ sampled from a target distribution, $P^t_{X\times Y}$ where the corresponding labels, $\{y_i^t\}$ are *unobserved*. Here, the target domain is separated from the source domain due to domain shift i.e., $P^s_{X\times Y} \neq P^t_{X\times Y}$. Under such a shift, the goal of unsupervised domain adaptation is to leverage D^s and D^t to learn a parametric model that performs well in the target domain. The model is a composition of an encoder, $E_\phi : X \to \mathcal{Z}$ and a classifier, $G_\psi : \mathcal{Z} \to \mathcal{Z}_P$ where, $\mathcal{Z} \in \mathbb{R}^{H\times W\times d}$ represents the space of d-dimensional spatial embeddings and $\mathcal{Z}_P \in [0,1]^{H\times W\times C}$ gives the distribution over the C classes at each spatial location, where, $\{\phi, \psi\}$ are the model parameters. In the recent years, a particularly effective way of learning the model parameters has been to jointly optimize a pixel-wise classification objective on the source domain,

$$L^s_{\text{cls}} = -\sum_{i=1}^{N_s}\sum_{l=1}^{H}\sum_{m=1}^{W}\sum_{c=1}^{C} y^s_{ilmc} \log G_\psi(z^s_{ilm})|_c, \;\; z^s_{ilm} = E_\phi(x^s_i)|_{l,m} \tag{1}$$

and a domain adaptation objective over the source and target domains as described next.

Pseudo-label self-training (PLST): Inspired by CAG-UDA Zhang et al. (2019), we propose a distance-based PLST method that leverages a model pretrained on labelled source data, D^s using (4) to *pseudo* label unlabelled target data, D^t via confidence thresholding. To obtain pseudo labels, we first compute class-specific prototypes, μ^s_c using the source embeddings, $z \in \mathcal{Z}$ associated with the pretrained encoder, E_{ϕ_0}

$$\mu^s_c = \frac{1}{|I^s_c|}\sum_{(i,l,m)\in I^s_c} z^s_{ilm}, \quad I^s_c = \{(i,l,m)|\; y^s_{ilmc} = 1\} \tag{2}$$

These prototypes are then used in conjunction with a pre-defined distance metric like cosine and a threshold, δ to assign pixel-wise pseudo labels in the unlabelled target dataset. More specifically, let the closest prototype to the i^{th} target embedding at spatial position (l, m) be μ_c such that $c = \underset{c'\in\{1,\ldots,C\}}{\arg\max}\; \cos(z^t_{ilm}, \mu_{c'})$, then the one-hot encoding for the corresponding pseudo label is defined as

$$\tilde{y}^t_{ilmc'} = \begin{cases} 1 & \text{if } c' = c \text{ and } \Delta^{'}(z^t_{ilm}) - \Delta^{''}(z^t_{ilm}) \geq \delta \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Here, $c' \in \{1, \ldots, C\}$ and the terms $\Delta^{'}(z^t_{ilm})$ and $\Delta^{''}(z^t_{ilm})$ denote the cosine similarity of z^t_{ilm} to its nearest and second-nearest prototype respectively. Such confidence thresholding ensure that only the most confident predictions contribute to successive self-training. These pseudo labels are then used to define the self-training loss

$$L^t_{\text{st}} = -\sum_{i=1}^{N_s}\sum_{l=1}^{H}\sum_{m=1}^{W}\sum_{c'=1}^{C} \tilde{y}^t_{ilmc'} \log G_\psi(z^t_{ilm})|_{c'}, \;\; z^t_{ilm} = E_\phi(x^t_i)|_{l,m} \tag{4}$$

Compared to the complex multi-objective loss used inZhang et al. (2019), our self-training loss is much simpler with just one probability based loss. Henceforth, we refer to our self-training method as Simplified-CAG.

4 Self-Training with Objectness Constraints

An important issue with the self-training scheme described in §3 is that it is usually prone to confirmation bias that leads to compounding errors in target model predictions due to training on spurious pseudo labels. Such pseudo labels are the confident but mis-classified predictions on target data made by the pretrained model. One of the reasons for erroneous pseudo labeling in semantic segmentation is due to the complete reliance on input pixel intensities. For example, a model pretrained

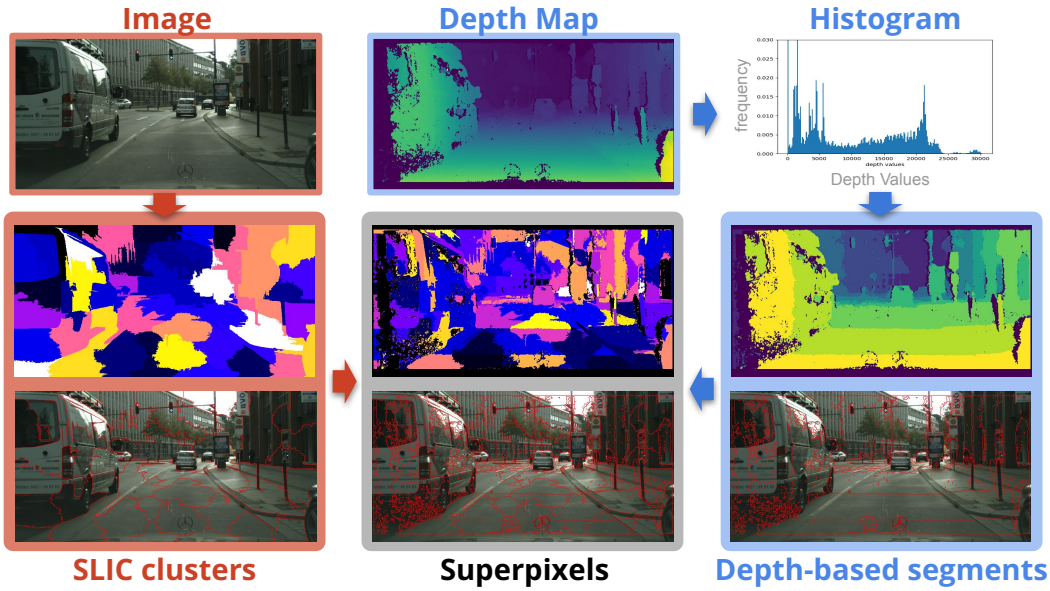

Figure 2: **Superpixel generation:** Overall pipeline for generating multi-modal superpixels (middle-column, bottom two rows) from image and depth. Each type of unsupervised segmentation map (middle-row) is accompanied with a corresponding boundary map (bottom-row) overlayed on the image for clear comparison. Multi-modal superpixels, in comparison with individual modality segments, yield segments that are most consistent with actual objects.

solely on RGB inputs may misclassify a "bright region" (high intensities in all three channels) of a car as the sky since sky is usually bright (see Figure 1). Using such a pseudo label for subsequent self-training can lead the model to wrongly correlate bright regions with sky irrespective of the ground truth class.

In order to mitigate such pitfalls, we propose to regularize self-training with an objectness constraint that enforces pixel embeddings belonging to the same object *instance* to lie close to each other while those belonging to different object *categories* to lie far away. The important question here is how to define an object region consistent with an actual object. In the next section, we describe how we extract the notion of objectness from RGB input and associated depth map that is then used to enforce objectness in embedding space via contrastive learning.

4.1 Supervision For Objectness Constraint

Depth: Segmentation datasets are often accompanied with depth maps registered with the RGB images. Visually, these depth maps reveal a great deal about object shapes and their area of occupancy in a 2D scene. To formalize the notion of objectness, we assign unique labels to each object instance such that every pixel within the object share the label. To compute this label, we first compute a histogram of depth values with predefined b number of bins. The range of depth of most objects in the types of scenes we explore is much smaller than the depth range of the entire scene. This property of objects translates into high density regions (or peaks) in the histogram. Among these peaks, the ones with prominence Llobera (2001) above a threshold, δ_{peak} are used as centers to cluster the histograms into discrete regions with unique labels. These labels are then assigned to every pixel whose depth values lie in the associated region. An example of the resulting depth-based segmentation is visualized in Figure 2.

RGB: An important form of self supervision for segmentation tasks is RGB-input based clustering. In this work, we adopt SLIC as a fast algorithm for partitioning images into multiple segments that respect image boundaries. The SLIC method applies k-means clustering in the pixel space to group together adjacent pixels that are visually similar. An important design decision for SLIC is the number of segments, k_s. A small k_s leads to large cluster sizes that may cause pixels from distinct objects to be grouped together, thus, violating the notion of objectness. On the other hand, a large

k_s will oversegment the scene, and hence the objects in the scene, resulting in a trivial objectness constraint. Moreover, we found that even with the large k_s, pixels from distinct objects can still be assigned to the same segment.

Thus, to formulate a non-trivial constraint with sufficiently small k_s that also respects object boundaries, we propose a two step multi-modal segmentation procedure that leverages both depth-based and RGB-based segmentation. First, we obtain k_s segments using SLIC over the RGB image. Then, we partition each segment according to the depth-based segmentation resulting in structurally consistent *superpixels*. The process, visualized in Figure 2 highlights the importance of our multimodal approach. While on one hand, segments derived from just depth maps fail to distinguish between objects that lie at similar depths, on the other, segments based solely on RGB-input can fail to distinguish objects even when they are at completely different depths. However, superpixels derived from a combination of both modalities can lead to a more consistent segmentation and a effective objectness constraint. In fact, we quantitatively verify the effectiveness in our experiments.

4.2 Imposing Objectness Constraint through Contrast

Our objectness constraint formulated using a contrastive objective that pulls pixel representations belonging to the same object *instance* close to each other while pushes those belonging to different object *categories* far away. We leverage both superpixels and pseudo labels to formalize the notion of object instances and object category. Formally, we assign each pixel at a spatial location (l, m) of an input image, x_i, a superpixel index, $\kappa_{lm} \in \{k_1, \ldots, k_s\}$ that is in turn assigned a superpixel label, l_k based on a majority voting over pixel-wise pseudo labels (§3) within the superpixel. In practice, noisy pseudo labels can lead to superpixel labeling that is inconsistent with true segmentation labels. To minimize such inconsistencies, we introduce a threshold, τ_p to allow the contribution of only those superpixel labels whose proportion of pixels with majority label as the pseudo label is greater than τ_p. This results in a potentially smaller set of valid superpixels, $\mathcal{K} = \{k_1, \ldots, k_{\tilde{s}}\}$ where $\tilde{s} \leq s$.

Further, we represent the k^{th} valid superpixel using the set of corresponding pixel-wise representation as $\nu_{ik} = \frac{1}{|U_k|} \sum_{(l,m)\in U_k} z_{ilm}$ where, $U_k = \{(l, m)|\kappa_{lm} = k\}$. Using the above notation we can define the pixel-wise objectness constraint as

$$L^t_{\text{obj}}(l, m) = -\log \left(\frac{\exp(z_{ilm} \cdot \nu_{i\kappa_{lm}})}{\sum\limits_{k\in\mathcal{K}\setminus\{\kappa_{lm}\}} \exp(z_{ilm} \cdot \nu_{ik})} \right) \tag{5}$$

The image-level objectness constraint is then simply $L^t_{\text{obj}} = \sum\limits_{k\in\mathcal{K}} \sum\limits_{(l,m)\in U_k} L^t_{\text{obj}}(l, m)$. Finally, we define the overall regularized self-training objective as $L_{\text{pac}} = L^s_{\text{cls}} + \alpha_{\text{st}} * L^t_{\text{st}} + \alpha_{\text{obj}} * L^t_{\text{obj}}$, where $\alpha_{\text{st}}, \alpha_{\text{obj}}$ are the relative weighting coefficients. Note that the objectness constraints are only computed for the target domain images since we are interested in improving target domain performance using self-training.

4.3 Learning and Optimization

To train the PAC-UDA model, we follow a stage-wise procedure similar to Zhang et al. (2019). The initialization stage involves as two-step warmup that trains a source segmentation model using only L^s_{cls} followed by adversarial adaptation Tsai et al. (2018) to the target domain. The self-training is then performed in an iterative fashion, wherein each stage pseudo labels are computed using the model from prior stage that are then used to finetune the full-model using PAC objective, L_{pac}.

5 Experiments

Datasets and Evaluation Metric: We evaluate the PAC-UDA framework in two common scenarios, `GTARichter et al. (2016)`→`CityScapesCordts et al. (2016)` and `SYNTHIARos et al. (2016)`→`CityScapesCordts et al. (2016)`. GTA5 is composed of $24,966$ synthetic images with resolution 1914×1052 and has annotations for 19 classes that are compatible with the categories in Cityscapes. Similarly, SYNTHIA consists of $9,400$ synthetic

images of urban scenes at resolution 1280×760 with annotations for only 16 common categories. Cityscapes has of $5,000$ real images and aligned depth maps of urban scenes at resolution 2048 × 1024 and is split into three sets of $2,975$ train, 500 validation and $1,525$ test images. Of the $2,975$, we use $2,475$ randomly selected images for self-training and remaining 500 images for validation. We report the final test performance of our method on the 500 images of the official validation split. The data-splits are consistent with prior works Araslanov et al. (2021); Zhang et al. (2020). The performance metrics used are per class Intersection over Union (IoU) and mean IoU (mIoU) over all the classes.

Base PLST methods and Architecture: The proposed PAC-UDA is a general regularization framework for self-training approaches and is agnostic to the base method. As a test of generality, we evaluate its performance on two base methods, namely, Simplified-CAG (see §3) and SAC Araslanov et al. (2021) that generate pseudo labels in very distinct manner. While Simplified-CAG uses a distance based thresholding to generate confident pseudo labels on the target domain, SAC relies on multi-scale fusion of model predictions. In the case of Simplified-CAG, we use the same variant of DeepLabv2 Chen et al. (2018) architecture Zhang et al. (2019) as the segmentation model and the same 5-layer CNN as the discriminator adversarial warmup. However, unlike the Euclidean metric used in Zhang et al. (2019), we adopt Cosine metric as it was found to generate more reliable pseudo-labels. For the SAC method, we use the official implementation with DeepLabv2 architecture and default configuration for all but one hyperparameter, `GROUP_SIZE` (we reduce the `GROUP_SIZE` from default value of 4 to 2 following GPU constraints). Finally, we use the output of ASPP module (classifier logits) in DeepLabv2 to compute the regularizer that is than weighted with α_{obj} and added to the SAC objective.

Implementation Details: When using Simplified-CAG as the base method, training images were randomly resized by 1 to 1.5 times and then randomly cropped to a resolution of 1280×640. Training batch size was set to 2 for both source and target domains and batch-normalization layer weights were frozen following GPU constraints. We initialize our PAC based self-training with the warmup model checkpoint provided by Zhang et al. (2019) and select the best model at each self-training stage according to our validation split of Cityscapes. We use SGD optimizer with initial learning rate of $2.5e-4$ that is decayed by the poly policy with power 0.9. The weight decay, momentum, α_{st} and α_{obj} were set to $1e-4$, 0.9, 1.0, and 1.0, respectively. Distance threshold, δ was set to 0.05. With SAC as the base method, we initialize our PAC based self-training with the warmed up model checkpoints released by Araslanov et al. (2021). The warmup here replaces adversarial training (Zhang et al. (2019)) with Adaptive Batch Normalization for initial adaptation. PAC training proceeds with batch size of 4 for both source and target domain images at 1024 × 512 image resolution. Augmented inputs are used for both segmentation self-supervised loss and PAC constraint that are assigned a weight of 5 and 1.0 respectively. An exhaustive list of hyperparameters is presented in the supplementary.

For both methods, we use SGD optimizer with initial learning rate of $2.5e-4$ that is then decayed by poly policy with power 0.9 for Simplified-CAG. Weight decay and momentum is set to $5e-4$ and 0.9 respectively. The inference follows the usual procedure of a single forward pass through the segmentation network at the original image resolution without any post-processing. Experiments were conducted on 4×11GB RTX 2080 Ti GPUs with PyTorch implementation. Code will be made publicly available.

5.1 Main Results

In this section, we compare the the effect of regularization on the two self-training approaches, Simplified-CAG and SAC* and also compare the results to prior state of the art methods. For Simplified-CAG, we report results after single stage of self-training as they were already close to the best performance achieved with multiple stages. Here, SAC* denotes a lightweight configuration of the original SAC that fits within our GPU constraints.

`GTA`→`CityScapes` (Table 1): In the case of Simplified-CAG, our PAC regularization improves the mIoU of the base method by a significant margin of 1.4% and outperforms it in 16 out of 19 class-wise IoUs. In the case, of SAC*, our regularization improves the mIoU of the base method by 0.6% and outperforms it in 13 of the 19, sometimes with significant margin (upto 12.2%). Compared to recent works, our SAC*+PAC method outperforms top-performing approaches like IAST Mei et al.

	road	sidewalk	building	wall	fence	pole	light	sign	vege.	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
AdvEnt Vu et al. (2018)	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
DISE Chang et al. (2019)	91.5	47.5	82.5	31.3	25.6	33.0	33.7	25.8	82.7	28.8	82.7	62.4	30.8	85.2	27.7	34.5	6.4	25.2	24.4	45.4
Cycada Hoffman et al. (2018)	86.7	35.6	80.1	19.8	17.5	38.0	39.9	41.5	82.7	27.9	73.6	64.9	19.0	65.0	12.0	28.6	4.5	31.1	42.0	42.7
BLF Li et al. (2019)	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
CAG-UDA Zhang et al. (2019)	90.4	51.6	83.8	34.2	27.8	38.4	25.3	48.4	85.4	38.2	78.1	58.6	34.6	84.7	21.9	42.7	**41.1**	29.3	37.2	50.2
PyCDA† Lian et al.	90.5	36.3	84.4	32.4	28.7	34.6	36.4	31.5	86.8	37.9	78.5	62.3	21.5	85.6	27.9	34.8	18.0	22.9	49.3	47.4
CD-AM Yang et al. (2021)	91.3	46.0	84.5	34.4	29.7	32.6	35.8	36.4	84.5	43.2	83.0	60.0	32.2	83.2	35.0	46.7	0.0	33.7	42.2	49.2
FADA Wang et al.	92.5	47.5	85.1	**37.6**	32.8	33.4	33.8	18.4	85.3	37.7	83.5	63.2	**39.7**	87.5	32.9	47.8	1.6	34.9	39.5	49.2
FDA Yang & Soatto (2020)	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
SA-I2I Musto & Zinelli (2020)	91.2	43.3	85.2	38.6	25.9	34.7	41.3	41.0	85.5	**46.0**	86.5	61.7	33.8	85.5	34.4	48.7	0.0	36.1	37.8	50.4
PIT Lv et al. (2020)	87.5	43.4	78.8	31.2	**30.2**	36.3	39.9	42.0	79.2	37.1	79.3	65.4	37.5	83.2	46.0	45.6	25.7	23.5	49.9	50.6
IAST Mei et al. (2020)	**93.8**	57.8	85.1	39.5	26.7	26.2	43.1	34.7	84.9	32.9	**88.0**	62.6	29.0	87.3	39.2	49.6	23.2	34.7	39.6	51.5
RPT† Zhang et al. (2020)	89.2	43.3	86.1	39.5	29.9	40.2	**49.6**	33.1	87.4	38.5	86.0	64.4	25.1	**88.5**	36.6	45.8	23.9	**36.5**	**56.8**	52.6
SACAraslanov et al. (2021)	90.4	53.9	86.6	**42.4**	27.3	45.1	48.5	42.7	87.4	40.1	86.1	**67.5**	29.7	**88.5**	49.1	**54.6**	9.8	26.6	45.3	**53.8**
Simplified-CAG††	87.0	44.6	82.9	32.1	35.7	40.6	38.9	45.5	82.6	23.5	78.7	64.0	27.2	84.4	17.5	34.8	35.8	26.7	32.8	48.2
Simplified-CAG†† + PAC (ours)	86.3	45.7	84.5	30.5	35.5	38.9	40.3	49.9	86.0	33.5	81.1	64.1	25.5	84.5	21.3	32.9	36.3	26.7	40.0	49.6
SAC*	89.9	54.0	86.2	37.8	28.9	**45.9**	46.9	47.7	88.0	44.8	85.5	66.4	30.3	88.6	**50.5**	**54.5**	1.5	17.0	39.3	52.8
SAC* + PAC (ours)	93.3	**63.6**	**87.2**	42.0	25.4	44.9	49.0	**50.6**	**88.1**	45.2	87.6	64.0	28.1	83.6	37.5	43.9	13.7	20.1	46.2	53.4

Table 1: **GTA → Cityscapes results:** IoU (per-class and mean) comparison of our PAC-UDA with prior works. †† denotes the use of DeepLabV2 variant introduced by Zhang et al. (2019), † denotes the use of PSPNet Zhao et al. (2017), * denotes a different configuration (batch size = 4) than the original SAC method (batch size = 8). All other methods use DeepLabV2Chen et al. (2018) architecture.

	road	sidewalk	building	wall	fence	pole	light	sign	vegetable	sky	person	rider	car	bus	motor	bike	mIoU
DCAN Wu et al. (2018)	82.8	36.4	75.7	5.1	0.1	25.8	8.0	18.7	74.7	76.9	51.1	15.9	77.7	24.8	4.1	37.3	38.4
DISE Chang et al. (2019)	**91.7**	**53.5**	77.1	2.5	0.2	27.1	6.2	7.6	78.4	81.2	55.8	19.2	82.3	30.3	17.1	34.3	41.5
AdvEnt Vu et al. (2018)	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	41.2
CAG-UDA Zhang et al. (2019)	84.7	40.8	81.7	7.8	0.0	35.1	13.3	22.7	84.5	77.6	64.2	27.8	80.9	19.7	22.7	48.3	44.5
PIT Lv et al. (2020)	83.1	27.6	81.5	8.9	0.3	21.8	26.4	**33.8**	76.4	78.8	64.2	27.6	79.6	31.2	31.0	31.3	44.0
PyCDA† Lian et al.	75.5	30.9	83.3	20.8	0.7	32.7	27.3	33.5	84.7	85.0	64.1	25.4	85.0	45.2	21.2	32.0	46.7
FADA Wang et al.	84.5	40.1	83.1	4.8	0.0	34.3	20.1	27.2	84.8	84.0	53.5	22.6	85.4	43.7	26.8	27.8	45.2
IAST Mei et al. (2020)	81.9	41.5	83.3	17.7	**4.6**	32.3	30.9	28.8	83.4	85.0	**65.5**	30.8	86.5	38.2	**33.1**	52.7	49.8
RPT† Zhang et al. (2020)	88.9	46.5	84.5	15.1	0.5	38.5	39.5	30.1	85.9	85.8	59.8	26.1	**88.1**	**46.8**	27.7	**56.1**	51.2
SAC Araslanov et al. (2021)	89.3	47.2	**85.5**	26.5	1.3	**43.0**	**45.5**	32.0	**87.1**	89.3	63.6	25.4	86.9	35.6	30.4	53.0	**52.6**
Simplified-CAG††	87.0	41.0	79.0	9.0	1.0	34.0	15.0	11.0	81.0	81.0	55.0	16.0	77.0	17.0	2.0	47.0	40.8
Simplified-CAG†† + PAC (ours)	87.0	42.0	80.0	12.0	3.0	30.0	17.0	17.0	80.0	88.0	57.0	5.0	75.0	20.0	1.0	52.0	41.7
SAC*	**91.7**	52.7	85.1	22.6	1.5	42.2	44.1	30.9	82.5	73.8	63.0	20.9	84.9	29.5	26.9	52.2	50.3
SAC* + PAC (ours)	83.2	40.5	**85.4**	**30.0**	2.0	**43.0**	42.2	**33.8**	86.3	**89.8**	65.3	**33.5**	85.1	35.2	29.9	55.3	**52.5**

Table 2: **SYNTHIA → Cityscapes results:** IoU (per-class and mean) comparison of our PAC-UDA with prior works. †† denotes the use of DeepLabV2 variant introduced by Zhang et al. (2019), † denotes the use of PSPNet Zhao et al. (2017). All other methods use DeepLabV2 Chen et al. (2018) architecture.

(2020) and RPT Zhang et al. (2020) by large margins (0.8%–1.9%). Additionally, compared to a more GPU-intensive configuration of SAC Araslanov et al. (2021) with batch size 8, our limited configuration of batch size 4 yields comparable results.

An interesting observation is that compared to SAC, our SAC*+PAC model improves classwise IoUs for both dominant categories like road and sidewalk (by upto 9.4%) as well as less frequent categories like traffic-sign and terrain. We hypothesise that the use of complementary modalities like Depth in our PAC framework normalises the effect of image statistic based on class frequency. Additionally, for classes like sidewalk, we suspect that structural constraints based on PAC reduces contextual bias Shetty et al. (2019), responsible for coarse boundaries.

SYNTHIA→CityScapes (Table 2): In this adaptation setting, we observe similar trends where PAC regularisation improves the both base methods by significant amount (upto 2.2%). More importantly, our SAC*+PAC outperforms most prior approaches Zhang et al. (2020); Mei et al. (2020); Wang et al. with competitive margin and performs comparable to the state-of-the-art Araslanov et al. (2021).

5.2 Ablations and Qualitative results

In this section, we provide an important ablation in Table 3 that deconstructs our multimodal objectness constraint and determine the effect of individual modalities on final performance. To conduct these ablations, we chose our Simplified-CAG + PAC method and GTA→CityScapes settings. Comparing the two modalities, we observe that the model regularized with only Depth based objectness constraints seem to be more effective than the one with only Image-based clustering constraints. This is intuitive and evident from Figure 2 where Depth-based segments reveal a lot of object information in the scene and hence, is more useful for computing PAC constraints. While both individual

Configuration	road	sidewalk	building	wall	fence	pole	light	sign	vege.	terrace	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
Only-Image	87.7	47.1	82.5	34.0	35.7	40.6	40.2	45.8	82.9	23.2	76.3	64.2	25.4	85.7	23.7	41.8	24.9	29.0	34.8	48.7
Only-Depth	85.5	43.3	83.6	34.8	35.0	41.2	38.1	48.8	85.4	28.3	80.1	65.0	25.5	85.6	25.9	38.6	33.5	24.9	35.1	49.4
Both	86.3	45.7	84.5	30.5	35.5	38.9	40.3	49.9	86.0	33.5	81.1	64.1	25.5	84.5	21.3	32.9	36.3	26.7	40.0	49.6

Table 3: **Ablation:** Comparing the effect of generating superpixels via individual versus combined modalities. Clearly, multi-modal generation (last row) yield the best performance in terms of mIoU. Model: Simplified-CAG + PAC, Datastet settings: `GTA` $\rightarrow$ `Cityscapes`.

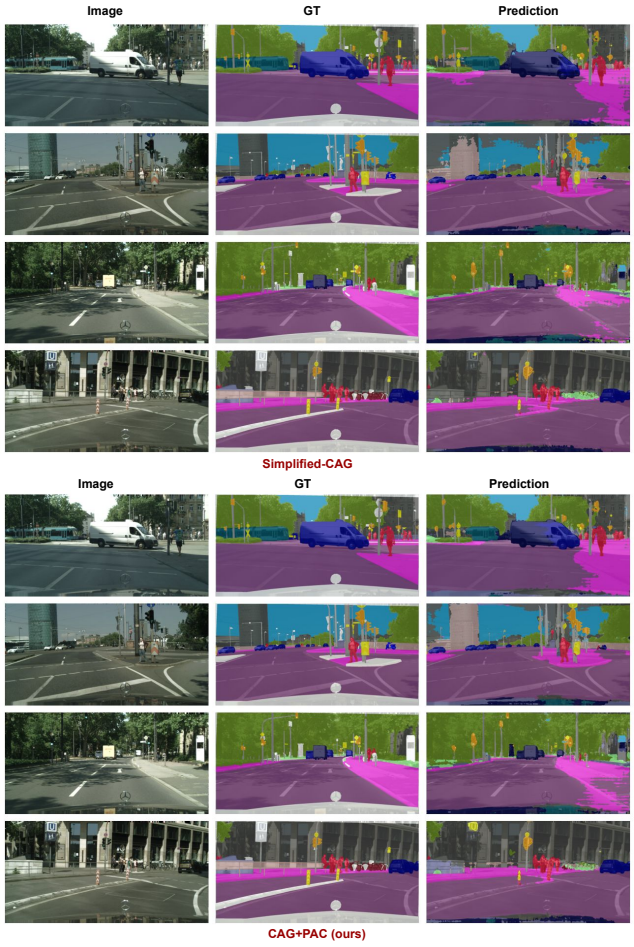

Figure 3: **Qualitative evaluations:** We observe that our regularised model (CAG+PAC) makes fewer mistakes. For example, in the second row of each method, the mis-classification of sky pixels as building is much benign in our method

modalities demonstrate reasonable performance, neither of them is sufficient. Combining the two, however, leads to the best results. This highlights the importance of multi-modal constraints. In Figure 3, we visualise and compare the predictions of Simplified-CAG and our CAG+PAC models on the `GTA`$\rightarrow$`CityScapes` settings.

6 Conclusion

In this work, we show that regularising self-training improves performance of the base method. We introduce multi-modal superpixels that present complementary information about segmentation in the form of objectness. These superpixels are then use to formulate the regularization in a way that is agnostic to photometric variations.

REFERENCES

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

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

M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann. Unsupervised domain adaptation by domain invariant projection. In *ICCV*, 2013.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NIPS*.

Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. *CoRR*, 2016a.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *NeurIPS*, 2016b.

Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. *CoRR*, 2019.

Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 627–636, 2019.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

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.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *NeurIPS*, 2017.

Tran Hiep Dinh, Minh Trien Pham, Manh Duong Phung, Due Manh Nguyen, Van Manh Hoang, and Quang Vinh Tran. Image segmentation based on histogram of depth and an application in driver distraction detection. In *ICARCV*, 2014.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Advances in Computer Vision and Pattern Recognition*, 2017.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018.

IJCNN. Pseudo-labeling and confirmation bias in deep semi-supervised learning. 2020.

Fredrik D. Johansson, David A Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *AISTATS*, 2019.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4893–4902, 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020.

Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6936–6945, 2019.

Qing Lian, Lixin Duan, Fengmao Lv, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *ICCV*.

Marcos Llobera. Building past landscape perception with gis: Understanding topographic prominence. *Journal of Archaeological Science - J ARCHAEOL SCI*, 2001.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. 2015.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, 2017.

Fengmao Lv, Tao Liang, Xiang Chen, and Guosheng Lin. Cross-domain semantic segmentation via domain-invariant interactive relation transfer. In *CVPR*, 2020.

Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*, 2020.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. *ICML*, 2013.

Luigi Musto and Andrea Zinelli. Semantically adaptive image-to-image translation for domain adaptation of semantic segmentation. In *BMVC*, 2020.

S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 2011.

Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *ICML*, 2017.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. Not using the car to see the sidewalk – quantifying and controlling the effects of context in classification and segmentation. In *CVPR*, June 2019.

Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. *arXiv*, 2018.

S. Si, D. Tao, and B. Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*, 2016.

Sinisa Stekovic, Friedrich Fraundorfer, and Vincent Lepetit. Casting geometric constraints in semantic segmentation as semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1854–1863, 2020.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *arXiv preprint arXiv:2005.10243*, 2020.

Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*, 2020.

E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. *arXiv*, 2018.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7364–7373, 2019.

Haoran Wang, Tong Shen, Wei Zhang, Lingyu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *ECCV*.

Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Chase Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. *ICML*, 2019.

Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *ECCV*, 2018.

Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International conference on machine learning*, pp. 5423–5432. PMLR, 2018.

Jinyu Yang, Weizhi An, Chaochao Yan, Peilin Zhao, and Junzhou Huang. Context-aware domain adaptation in semantic segmentation. In *WACV*, 2021.

Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020.

Qiming Zhang, Jing Zhang, Wenyu Liu, and D. Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. In *NeurIPS*, 2019.

Yiheng Zhang, Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Dong Liu, and Tao Mei. Transferring and regularizing prediction for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9621–9630, 2020.

Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon. On learning invariant representation for domain adaptation. *ICML*, 2019.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 289–305, 2018.