Cluster-to-adapt: Few Shot Domain Adaptation for Semantic Segmentation across Disjoint Labels

Tarun Kalluri   Manmohan Chandraker
University of California San Diego

Abstract

Domain adaptation for semantic segmentation across datasets consisting of the same categories has seen several recent successes. However, a more general scenario is when the source and target datasets correspond to non-overlapping label spaces. For example, categories in segmentation datasets change vastly depending on the type of environment or application, yet share many valuable semantic relations. Existing approaches based on feature alignment or discrepancy minimization do not take such category shift into account. In this work, we present Cluster-to-Adapt (C2A), a computationally efficient clustering-based approach for domain adaptation across segmentation datasets with completely different, but possibly related categories. We show that such a clustering objective enforced in a transformed feature space serves to automatically select categories across source and target domains that can be aligned for improving the target performance, while preventing negative transfer for unrelated categories. We demonstrate the effectiveness of our approach through experiments on the challenging problem of outdoor to indoor adaptation for semantic segmentation in few-shot as well as zero-shot settings, with consistent improvements in performance over existing approaches and baselines in all cases.

1. Introduction

In this work, we address the problem of knowledge transfer across domains with disjoint labels for semantic segmentation. In spite of massive strides in computer vision performance using deep learning [21], models trained on a large-scale labeled dataset are not guaranteed to generalize to data that lies outside the training distribution. This difficulty is amplified for applications like semantic segmentation, where collecting pixel level labeled data for all geographies, environments and weather conditions is restrictive, expensive or simply not feasible due to many practical and social implications [19].

Unsupervised domain adaptation emerged as a feasible alternative to transfer knowledge from a labeled source domain
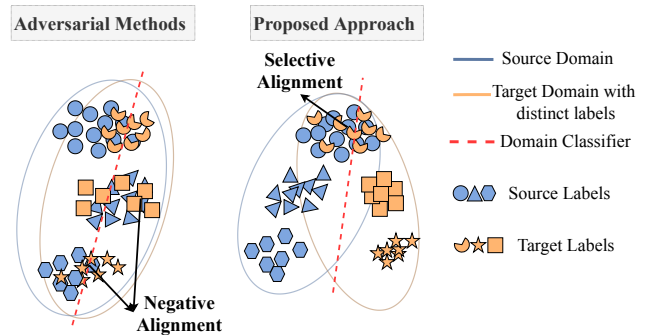

Figure 1. We illustrate the major challenge in our setting arising due to negative transfer. In our setting where source and target do not contain shared label space, traditional adversarial methods relying on global adaptation might lead to unrelated classes aligning with each other. In contrast, our method encourages only related classes to align with each other (selective alignment), while preventing negative transfer among unrelated classes.

to unlabeled target domains by minimizing some notion of divergence between the domains [4, 12, 22, 23, 42, 48]. Prior works in domain adaptation are based on a global distribution alignment objective, assuming that the source and target datasets share the same label space so that domain alignment would invariably result in learning transferable feature representations.

In many cases, the source and target labels might be completely distinct and share only high level geometric and semantic relationships. This makes it hard, yet necessary in few-shot settings, to perform useful knowledge transfer. In particular we show this in case of adaptation between outdoor datasets, where synthetic datasets are readily available, and indoor scenes, where we have few labeled data and it is considered difficult to render or maintain synthetic datasets. To address this challenging setting of outdoor to indoor adaptation, we propose a novel framework for adaptation across disjoint labels. For disjoint labels, we posit that a more suitable objective is to achieve *domain invariance* with respect to related categories and *domain equivariance* with respect to unrelated categories between source and target,
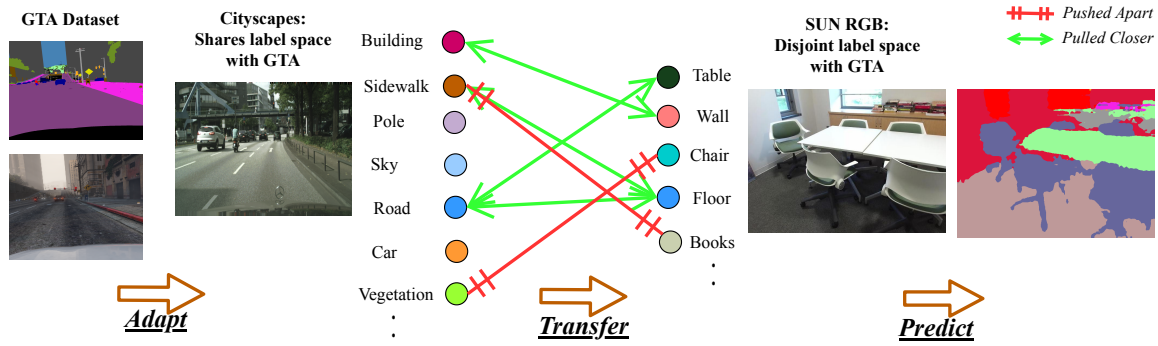
Figure 2. Overview of the proposed adaptation across disjoint labels using unlabeled bridge domains. Note that the source and target domains have completely disjoint label spaces here, which is a very realistic setting for real world transfer learning scenarios. We tackle this challenging setting in our problem using an unlabeled *bridge* domain. In this case, where synthetic driving data is acting as the labeled source domain with real indoor scenes as the target domain, real driving scenes can be used as the bridge as it shares properties with both of these, enabling few shot adaptation to the target domain.

thus avoiding negative transfer. For example, the categories that frequently occur in an indoor environment like wall, floor, ceiling and chair are completely distinct from any outdoor categories, yet we show how we can leverage useful discriminative information through implicit geometric and semantic correspondences.

In practice, the distribution shift between such source and target domains arise from both low-level (lighting, contrast, object density etc.) and high-level (category, geometric orientation, pose etc.) variations [4]. To ease this extreme case of adaptation, we introduce an additional unlabeled auxiliary domain, which shares properties with both the source and target datasets and would act as a bridge to improve the adaptation. For instance, adaptation from synthetic outdoor to real indoor datasets can benefit from unlabeled images from real outdoor scenes, as explained in Figure 2.

To automatically discover related and unrelated categories across datasets, we propose a novel clustering based alignment approach called Cluster-to-adapt (C2A). C2A stems from the intuition that related categories from source and target should lie close to each other in the feature space for effective knowledge transfer. We realize this during training through a deep constrained clustering framework by posing the alignment as a clustering objective in a transformed feature space, which would force related categories to group close to each other while leaving room for unrelated categories to form independent clusters, as shown in Figure 1.

2. Related Work

Unsupervised Domain Adaptation (UDA)   UDA is used to transfer knowledge from a large labeled source domain to an unlabeled target domain. Large body of works that perform adaptation from labeled source to unlabeled target

rely on adversarial generative [3, 15, 31, 38, 58] or discriminative [12, 16, 47, 48] approaches to learn domain agnostic feature representations. A common assumption in most of these approaches is that the source and target label spaces completely overlap, so that a classifier learnt on the source domain can be directly applied on the target data. However, in most real world applications this assumption is invalid, and in most general case, the categories might be completely different. Very few works exist which address this more challenging setting. Previous works like open set adaptation [37], partial set adaptation [5, 6] and universal adaptation [19, 57] assume some degree of label overlap, [25] performs adaptation between distinct label spaces with few target labeled data using pairwise similarity constraints, while [40] addresses adaptation for verification tasks which is different from our focus on semantic segmentation. Similarly, more recent works for domain adaptation suited for semantic segmentation tasks [2, 26, 30, 32, 50, 55] achieve state of the art results for the case of completely overlapping label spaces in the source and target domains, and are not applicable in our setting of outdoor to indoor adaptation. In contrast to these existing works, we propose an efficient method to align only visually similar features across source and target domains which can have completely non-intersecting label spaces without re-annotation [20], while preventing potential negative transfer, specifically suited to cross domain semantic segmentation.

Deep Clustering   Although clustering algorithms like k-means [27] are extremely useful in automatically discovering structure from unlabeled data [1], they work directly on the high dimensional input space like images which is often ineffective for classification. Recent works propose jointly

learning a suitable feature representation of data along with clustering assignments. For example, [17] uses pairwise similarity based constraints, while [14] uses self-training objective on the cluster assignment scores to successfully perform unsupervised transfer across categories from the same domain. Other works make use of deep clustering to learn more discriminative clusters [52] useful for classification, or as suitable pretext tasks in self supervised learning [7, 8, 54]. While deep clustering based approaches have been previously applied in the case of unsupervised category discovery [14], we extend this idea to additionally account for the domain shift between the source and target datasets.

Also, note that many prior works that use clustering for adaptation consider the classical setting of completely matching source and target domains [44, 45, 51] or partial overlap in open world setting [13], and hence use clustering as a means to achieve one-to-one alignment between source and target. In contrast, we use clustering to selectively align source and target across completely disjoint label spaces.

3. Framework

We now explain our proposed approach, which addresses the most general case of knowledge transfer between domains with different, and non-overlapping label spaces. Denote using $\mathbb{D}_s$ the completely labeled source domain data with label space $\mathbb{Y}_s$, where $\mathbb{D}^s \sim p_s$(source distribution). The labeled target domain data is denoted by $\mathbb{D}_t$, with label space $\mathbb{Y}_t (\neq \mathbb{Y}_s)$, and $\mathbb{D}_t \sim p_t$(target distribution). We assume that a small subset $\mathbb{D}_t^l$ of the target data is labeled, for learning some task specific information like classifier boundaries, and the rest $\mathbb{D}_t^u$ as unlabeled, making our setting that of *few-shot adaptation across domains with disjoint labels*. We denote this small fraction of labeled samples by $\sigma = |\mathbb{D}_t^l|/(|\mathbb{D}_t^u| + |\mathbb{D}_t^l|)$. Following the nomenclature of [33], we henceforth call this as cross-task adaptation and the source and target as different tasks. In section 4, we show results varying σ from 0.01 to 0.3. In our case, the domain gap between the target data and source data comes from two factors, namely domain shift due to $p_s \neq p_t$ as well as label shift due to $\mathbb{Y}_s \neq \mathbb{Y}_t$. Furthermore, we do not assume any partial overlap between the label spaces unlike other partial or open set adaptation approaches which makes our setting more challenging. To ease the adaptation process across these widely different datasets, we introduce another completely unlabeled auxiliary domain $\mathbb{D}_a$, which serves as a useful bridge between the source and target datasets. For example, $\mathbb{D}_a$ could share *task/content* properties with $\mathbb{D}_s$ and *style* properties with $\mathbb{D}_t$. We show how to best exploit this completely unlabeled intermediate domain to achieve our primary goal of learning transferable features from source to target.

Overview We present the overview of the proposed network architecture in Figure 3. We use a shared encoder

$\mathcal{E} : (H, W, 3) \rightarrow (H', W', f_d)$ across all the datasets which aggregates spatial features across multiple resolutions of the input image x, and outputs a downsampled encoder map $\mathcal{E}(x)$. f_d is the size of features in the encoder map. Since shallow level features are known to be more task agnostic and transferable [56], a shared encoder helps us to learn generic features useful across source and target datasets. Task specific decoders $\mathcal{G}_s$ and $\mathcal{G}_t$ then upsample the output of the encoder and compute class assignment probabilities for each pixel of the input image over the label space $\mathbb{Y}_s$ and $\mathbb{Y}_t$ respectively. Individual decoders for source and target helps us to make predictions over respective label spaces. The supervised loss computed using the labeled data from source and target datasets is given by

$$\mathcal{L}_{sup} = \mathcal{L}_{sup}(\mathbb{D}_s) + \mathcal{L}_{sup}(\mathbb{D}_t^l), \tag{1}$$

where

$$\mathcal{L}_{sup}(\mathbb{D}_s) = \frac{1}{N_s}$$
$$\sum_{(x,y)\in\{\mathbb{D}_s\}} \frac{-1}{HW} \sum_{h,w} \log(\mathcal{G}_s(\mathcal{E}(x))^y(h,w)) \tag{2}$$

and N_s is the number of labeled samples in the source dataset, and H, W are the height and width of the output feature map respectively. The target supervised loss $\mathcal{L}_{sup}(\mathbb{D}_t^l)$ is defined similarly.

Next, we decouple the source to target alignment into two different objectives. The first is a *within task* alignment between $\mathbb{D}_s$ and $\mathbb{D}_a$, and the second is the *cross task alignment* objective between $\mathbb{D}_a$ and $\mathbb{D}_t$, as explained next.

3.1. Within Task Domain Alignment

We introduce the within task alignment objective between the source and intermediate domains $\mathbb{D}_s$ and $\mathbb{D}_a$. We assume that the domains share the same label space and exhibit only low-level differences, and use an adversarial alignment strategy using a domain discriminator $\mathcal{D}$.

Following the idea presented in [46], we send the output probability maps $\mathcal{P}_s(x) = \mathcal{G}_s(\mathcal{E}(x))$ to the discriminator $\mathcal{D}$ as opposed to the encoder maps. This helps in better within-task alignment for pixel level prediction tasks and, as we found out, faster convergence during training. We train the discriminator $\mathcal{D} : (H, W, |\mathbb{Y}_s|) \rightarrow \{0, 1\}$ which takes as input the output map of the generator, to output the probability of the map coming from source data. The generator is then trained to produce outputs from $\mathbb{D}_a$ which are good enough to trick the discriminator into classifying them as coming from source. This alternative min-max optimization would then result in domain invariant output maps leading to successful feature alignment. The adversarial loss, using LS-GAN [28], is given by

$$\mathcal{L}_{adv} = \mathrm{E}_{x\sim\mathbb{D}_a}(\mathcal{D}(\mathcal{P}_s(x)))^2 \tag{3}$$
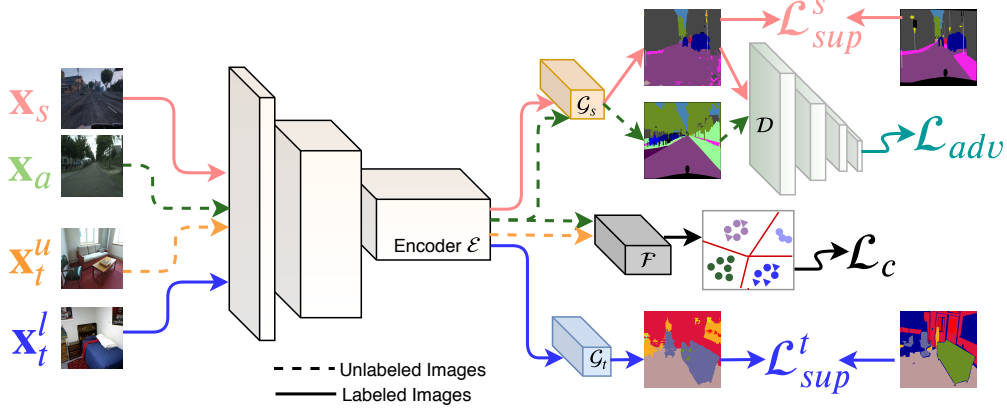
Figure 3. **C2A: Proposed Architecture.** The labeled data from the source ($\mathbb{D}_s$, in red) and target ($\mathbb{D}_t^l$, in blue) are used to train the task specific decoders using supervised losses $\mathcal{L}_{sup}^s$ and $\mathcal{L}_{sup}^t$. The unsupervised images from auxiliary domain ($\mathbb{D}_a$, in green) is used in adversarial adaptation from source to intermediate domain using loss $\mathcal{L}_{adv}$. The target unlabeled data ($\mathbb{D}_t^u$, in orange), along with $\mathbb{D}_a$, is used to compute the clustering loss $\mathcal{L}_c$ in a feature space transformed by $\mathcal{F}$. The encoder $\mathcal{E}$ is common for all the images.

and the discriminator objective $\mathcal{L}_D$ is given by

$$\mathcal{L}_{\mathcal{D}} = \mathrm{E}_{x \sim \mathbb{D}_s}(\mathcal{D}(\mathcal{P}_s(x)))^2 + \mathrm{E}_{x \sim \mathbb{D}_a}(\mathcal{D}(\mathcal{P}_s(x)) - 1)^2 \quad (4)$$

Although we use this adversarial adaptation strategy for within-task alignment, we note that our method can also be applied in combination with any other adaptation strategy based on generative modeling or distribution matching [31, 49, 55] for within task alignment.

3.2. Cross Task Semantic Transfer

Training $\mathcal{E}$ and $\mathcal{G}$ with task specific supervised loss and adversarial alignment losses alone is insufficient to transfer useful semantic content to target dataset, since we do not explicitly transfer any semantic relations between the tasks. Naive adversarial training of yet another discriminator to distinguish outputs from two tasks would not work well, as we only want categories that share semantic cues to align with each other (selective alignment) as opposed to global alignment (Figure 1). Luo et. al. [25] propose using an entropy minimization objective after computing pairwise similarity of the features, but computing such pairwise similarity is computationally infeasible for pixel level prediction tasks. Towards this goal, we propose a novel deep clustering based approach, which lies at the core of our approach.

Constrained Clustering Objective Following the assumption that deep features form discriminative clusters in the feature space useful for classification tasks [9], we believe that better knowledge transfer would happen across tasks if the features of categories which share semantic information also form coherent clusters closer to each other. A major challenge with incorporating this constraint in deep neural

networks is the lack of information regarding the correspondence between categories of the datasets useful in preventing negative transfer effects. We use a clustering based objective to discover the similarity across categories, and enforce the clustering constraint by performing k-means clustering of the feature vectors. This encourages the features corresponding to similar categories across tasks to form a single "meta-cluster", while leaving room for unrelated categories to form independent clusters.

We first pass the outputs of the shared encoder $\mathcal{E}$ through a feature transfer module $\mathcal{F} : (H', W', f_d) \rightarrow (H', W', f_e)$, where f_e is the feature dimension in the transformed space. $\mathcal{F}$ is necessary because the features learnt specific to a task might not be suitable for cross-task semantic transfer directly in the feature space. A learnable transformation function would, instead, find the best subspace amenable for alignment. Also, since $f_e \ll f_d$, the feature transformation would result in efficient computation of centers and similarity metrics for k-means. We formulate our constrained clustering objective using the cross-entropy loss, given by

$$\mathcal{L}_c = \sum_{x \in \{\mathbb{D}_a, \mathbb{D}_t^u\}} \sum_{v_j \in \mathcal{F}(\mathcal{E}(x))} -\log(\max_k p(\mu_k | v_j)) \quad (5)$$

where $p(\mu_k | v_j)$ is the probability score that a feature vector v_j belongs to a cluster with center μ_k, and

$$p(\mu_k | v_j) \propto \exp\left(\frac{v_j \cdot \mu_k}{||v_j||_2 ||\mu_k||_2}\right) \quad (6)$$

Avoiding Trivial Solution Direct optimization of Eq. (5) would quickly lead to a trivial solution where all the vectors are mapped to a single cluster. We found that initializing the cluster centers using features computed from pretrained

network on labeled target data alone ($\mathbb{D}_t^l$, trained offline) would reduce this problem to a large extent. Additionally, we follow the idea proposed in [52], and add a self-training constraint which encourages uniformity among the clusters and the cluster assignment probabilities are forced to be equal to an auxiliary target distribution. Specifically, we would like to have the target distribution $q(\mu_k|v_j)$ to hold the property that

$$q(\mu_k|v_j) \propto p(\mu_k|v_j) \cdot p(v_j|\mu_k)$$

The first term on the RHS would improve the association of correct points to clusters, while the second term would discourage very large clusters. Applying bayes rule would give us the form of the target distribution as

$$q(\mu_k|v_j) = \frac{p(\mu_k|v_j)^2/\sum_j p(\mu_k|v_j)}{\sum_{k'} p(\mu_{k'}|v_j)^2/\sum_j p(\mu_{k'}|v_j)}. \quad (7)$$

The constraint is now enforced in the form of a KL-Loss between the source distribution and the auxiliary target distribution.

$$\begin{aligned} \mathcal{L}_{kl} &= KL(p||q) \\ &= \sum_j \sum_k q(\mu_k|v_j) \log\left(\frac{q(\mu_k|v_j)}{p(\mu_k|v_j)}\right) \end{aligned} \quad (8)$$

The final training objective for the model can be summarised as follows,

$$\underset{\mathcal{E},\mathcal{G}_s,\mathcal{G}_t,\mathcal{F}}{\arg\min} \quad \mathcal{L}_{sup} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_c(\mathcal{L}_c + \mathcal{L}_{kl})$$

$$\underset{\mathcal{D}}{\arg\min} \quad \mathcal{L}_{\mathcal{D}} \quad (9)$$

where λ_{adv} and λ_c are the coefficients which control the relative importance of the adversarial loss and clustering loss respectively. The optimization is done by alternating between the two objectives within every iteration. A crucial factor in our method is the initialization of cluster centers before training, and we discuss our strategy followed for it next.

3.3. Cluster Initialization

The cluster centers are initialized using networks pre-trained on the limited labeled data. Specifically, we use the same architecture as described in the paper to train a model on the labeled source data $\mathbb{D}_s$ as well as sparsely labeled target data $\mathbb{D}_t^l$ using pixel level cross entropy loss. We then pass the unlabeled images from the target $\mathbb{D}_t^u$ and collect all the encoder maps corresponding to all the images. Each encoder map is of size $(H/8, W/8, 2048)$ for our ResNet-101 backbone. To match the dimension of the FTN output, which is 128 in our case, we apply PCA over these feature vectors to reduce their dimension. Then, a clustering is performed using the classical k-means objective with K cluster centers, and the resulting centers are used to initialize $\mu_k's$ in the downstream adaptation approach.

Efficient Computation of Centers In traditional k-means, the centers μ_k are calculated using an iterative algorithm consisting of cluster assignment and centroid computation repeated until convergence. We mention a couple of issues persistent with this approach. Firstly, for dense prediction tasks like semantic segmentation, the encoder map consists multiple feature vectors which correspond to different patches of the input image. For example, an encoder map of size (H', W') has $H'W'$ vectors of size f_e. Performing k-means over these vectors collected over all images over all the tasks would demand huge storage and computation requirements. Secondly, switching between gradient based training of network parameters and iterative computation of cluster centers after every few iterations would lead to an inefficient procedure that is not end-to-end trainable. To counter these limitations, we follow the idea proposed in [14] and include μ_k as trainable parameters in the network, and update them after each iteration based on the gradients received from $\mathcal{L}_C$.

4. Experiments

4.1. Datasets

For the source dataset $\mathbb{D}_s$, we use synthetic images from the driving dataset GTA [35]. GTA consists of 24966 images synthetically generated from a video game consisting of outdoor scenes with rich variety of variations in lighting and traffic scenes. We also show results using the SYNTHIA-RAND-CITYSCPAES split from Synthia [36] dataset, which consists of 9600 synthetic images with labels compatible with Cityscapes. For the target dataset $\mathbb{D}_t$, we use real images from SUN-RGBD [41] consisting of images from indoor scenes. SUN-RGBD consists of 5285 training images and 5050 validation images containing pixel level labels of objects which frequently occur in an indoor setting like chair, table, floor, windows etc. We use the 13 class version from [29]. The background class is ignored during training and evaluation. Additionally, we use the 2975 training images from Cityscapes [11] dataset, which consists of outdoor traffic scenes captured from various cities in Europe, as the unlabeled auxiliary domain $\mathbb{D}_a$. Cityscapes shares its semantic categories with GTA, so that the variation between $\mathbb{D}_s$ and $\mathbb{D}_a$ is only due to synthetic and real appearance, while $\mathbb{D}_s$ and $\mathbb{D}_t$ have many low-level as well as high-level differences.

4.2. Training Details

We use the DeepLab [10] architecture with a resnet-101 backbone for the encoder framework $\mathcal{E}$. For the task-specific decoder $\mathcal{G}$, we use an ASPP convolution layer followed by an upsampling layer. The architecture of discriminator $\mathcal{D}$ is similar to DC-GAN [34] with four 4×4 convolution layers, each with stride 2 followed by a leaky ReLU non-linearity.

| | Few-shot settings (our goal) | | | |
Method	$\sigma = 0.01$ $N = 50$	$\sigma = 0.04$ $N = 200$	$\sigma = 0.1$ $N = 500$	$\sigma = 0.3$ $N = 1500$
Target Labeled Only	22.62	30.43	36.62	43.17
Fine Tune	21.44	29.46	34.84	**44.10**
AdaptSegNet* [46]	25.20	32.51	36.9	43.83
LET* [25]	25.19	32.44	35.87	42.96
UnivSeg [19]	22.21	31.32	36.08	42.10
AdvSemiSeg [18]	24.72	33.22	38.46	45.10
C2A (Ours, $\lambda_c = 0$)	24.10	32.22	35.89	43.08
C2A (Ours, full, K=10)	**25.98**±0.03	**33.37**±0.07	**37.41**±0.04	43.16±0.03

Table 1. **Few Shot Adaptation:** mIoU values on SUN-RGB validation set for the proposed few shot segmentation approach. σ is the fraction of labeled examples from the total number of images from SUNRGB dataset. Note that our method particularly shows improvement in cases when the amount of labeled examples is very low. K is the number of clusters. (* denotes our extension of the existing works to suit our task.)

The feature transformation module $\mathcal{F}$ is a 1×1 convolution layer with output channels equal to the embedding dimension, which is fixed as $f_e = 128$ for all the experiments. We use a default value for $\lambda_{adv} = 0.001$. Following [53], to suppress the noisy alignment during the initial iterations, we set $\lambda_c = \frac{2}{1+e^{-10*\delta}} - 1$, where δ changes from 0 to 1 over the course of training. The backbone architecture is trained using SGD objective, with an initial learning rate of 2.5×10^{-4}. For training the cluster centers, we follow a similar learning rate decay schedule, but start with a smaller learning rate of 2.5×10^{-5}. This is because the cluster centers are already initialized using networks trained on the labeled data, and we would ideally like the centers to not drift too far away from their initial values.

Baselines and Ablations We provide ablation studies of the clustering module proposed in our approach and compare with the existing baselines. Specifically, we provide comparisons against the following. (i) **Target Labeled Only**: We train the segmentation encoder and decoder using only the limited labeled data from the target domain, $\mathbb{D}_t^l$, (ii) **Fine-tune**: We use a model trained on source dataset $\mathbb{D}_s$ till convergence, and finetune it on the labeled target data, (iii) **Ours (C2A),** $\lambda_c = 0$: Our cluster to adapt approach, without the clustering objective, and (iv) **Ours (C2A)**: our proposed approach with all the losses included.

Comparison with prior works We reiterate the paucity of existing works which tackle the same setting as ours, making direct comparison hard. Many traditional adaptation methods prevalent in literature for segmentation [16, 24, 43, 49] are not directly applicable in cases with disparate source and target label sets. Therefore, we compare against two competitive approaches that perform domain adaptation by

extending them as follows. (i) **AdaptSegNet*** [46]: We choose [46] as the backbone pixel level adaptation method for global adaptation across source and target datasets as it achieves high performance with a simple method. Since [46] is not directly applicable to our case due to different labels spaces, we extend their method to perform feature space adaptation. (ii) **LET*** [25]: We compare against adaptation proposed in [25] using entropy minimization criterion. We extend it to suit the segmentation task by using class prototypes in the feature space instead of pairwise enumeration which keeps the computation feasible.

GTA to SunRGB We show in Table 1 our results by varying the amount of supervision by choosing $|\mathbb{D}_t^l| = \{50, 200, 500, 1500\}$ images which corresponds to $\sigma = \{1\%, 4\%, 10\%, 30\%\}$ respectively. Our method based on a novel clustering objective consistently outperforms other approaches by considerable margins, more so in cases when there is extreme scarcity of labeled data. We see upto 15% and 10% relative increase in mIoU for $\sigma = 1\%$ and $\sigma = 4\%$ respectively compared to training only on the labeled target dataset. It is also evident that the clustering loss $\mathcal{L}_c$ is important for the objective to successfully carry selective alignments from source and intermediate domains to the target domain, as seen from improvements in our results compared to prior works like [46] and [25]. We also observed that 30% is already sufficient data for supervised fine-tuning to do well without any adaptation. In this work, our goal is focused on boosting the adaptation performance when enough labeled examples are not present in the target domain ($\sigma \ll 1$).

Role of intermediate bridge domain The intuition behind using an intermediate domain is to ease the adaptation pro-

Method	bed	books	ceil.	chair	floor	furn.	objs.	paint.	sofa	table	tv	wall	win.	pAcc.	mIoU
GTA to SunRGB															
Target Labeled	36.47	9.25	27.15	45.0	71.13	25.21	6.76	22.88	25.86	36.13	0.0	58.55	31.16	65.02	30.43
Fine tune	29.46	8.09	33.71	44.33	70.89	23.75	9.06	**25.4**	24.3	35.76	0.0	61.05	25.5	64.60	29.46
C2A [Ours]	**40.75**	**12.8**	**37.77**	**46.73**	**75.7**	**25.56**	**9.52**	23.3	**30.33**	**36.5**	0.0	**61.51**	**33.42**	**66.71**	**33.37**
Synthia to SUNRGB															
Target Labeled	36.47	9.25	27.15	45.0	71.13	25.21	6.76	**22.88**	25.86	36.13	0.0	58.55	31.16	65.02	30.43
Fine tune	31.87	9.76	34.98	43.89	71.44	25.67	7.76	15.74	24.44	36.12	0.0	61.30	**32.98**	65.08	30.46
C2A [Ours]	**40.84**	**14.32**	**36.39**	**47.76**	**73.33**	**25.95**	**12.11**	19.26	**31.16**	**39.98**	0.0	**63.92**	32.77	**67.37**	**33.68**

Table 2. Classwise IoU values for the 13 classes in SUN RGB validation set compared against baselines for $\sigma = 4\%(N = 200)$. Classes like floor, wall and ceiling which share rich geometric and semantic properties with categories in outdoor GTA benefit the most, thus validating our approach of selective knowledge transfer across indoor and outdoor datasets.

cess between the synthetic source domain data and real target domain data, which differ in both the appearance and the label spaces (categories). The use of an unlabeled domain bridge leads to no degradation with respect to a direct GTA to SUN adaptation for N = 200 (33.5% without and 33.4% with), while leading to a noticeable benefit for N = 50 (25.0% without and 26.0% with bridge domain).

Comparison with semi-supervised learning methods We also compare our work with existing semi-supervised segmentation algorithms in literature, namely AdvSemiSeg [18] and universal semi-supervised segmentation [19] and include results in Table 1. For [18], we use the unlabeled and labeled image sets from SUN-RGB to run the experiments. For [19], we follow the setting of their paper and use N=(50,200,500,1500) labeled examples from source and target, and use rest of images without annotations. From table Table 1, we note that our method delivers much better performance compared to semi-supervised learning methods for lower amounts of target supervision, as the latter do not leverage rich supervision available from a source domain.

Category-wise Performance We show the classwise mIoU results of our method in Table 2 for both cases of using GTA and Synthia as the source dataset. The proposed C2A approach outperforms the baselines, that do not make use of the alignment strategy, on most of the classes. The gains are especially significant on classes like floor, wall and ceiling, which share many geometric as well as semantic properties with categories in Cityscapes and GTA (Figure 4). For example, the patches corresponding to road in GTA dataset can help to successfully identify the parts of indoor images that correspond to floor since both occur mostly in the lower parts of images and share many other appearance and geometric relations.

L_c ***and*** L_{KL} For the ablation into clustering losses, we found that removing KL Loss (using only L_c) drops per-formance to 24.24%, removing clustering loss (using only L_{KL}) drops to 23.32%, while using both these losses gives 25.98% for the case of N = 50 in Table 1. We can conclude that both the clustering loss as well as the KL divergence loss are necessary as they offer complementary benefits (discussed in Sec 3.3).

Number of clusters An important aspect of our formulation is the choice of number of clusters K for the clustering approach in Eq. (5). From Figure 3, K=10 clusters works well for the case of $N = 50$. This concurs well with our intuition that a small value of K, like 5, would adversely affect the discriminative performance of the original task while a large value of K, like 20, would not encourage the semantic transfer we are after, as even related categories might form distinct clusters with no overlap

Unsupervised Zero Shot Adaptation With a slight modification, our approach also works well for the problem of zero shot adaptation, when no (labeled, or unlabeled) examples are available from the target domain. For learning some task specific information, we assume availability of synthetically generated samples from the target domain. In our case, we render artificial indoor scenes from SceneNet [29] dataset and use it in conjunction with our approach. We use this synthetically generated data from SceneNet instead of SUN-RGB in the formulation, in Eq. (1) and Eq. (5) instead of $\mathbb{D}_t$.

We compare our approach against the baselines where we use a classifier trained on SceneNet directly on SUNRGB. We report the results in Table 4, and our method which jointly optimizes a clustering objective along with an adversarial objective performs much better than the baselines that use only synthetic images from SceneNet. We believe that is due to the complementary knowledge that the network is able to infer through our approach. Our method even improves upon plain joint training on labeled GTA and SceneNet datasets

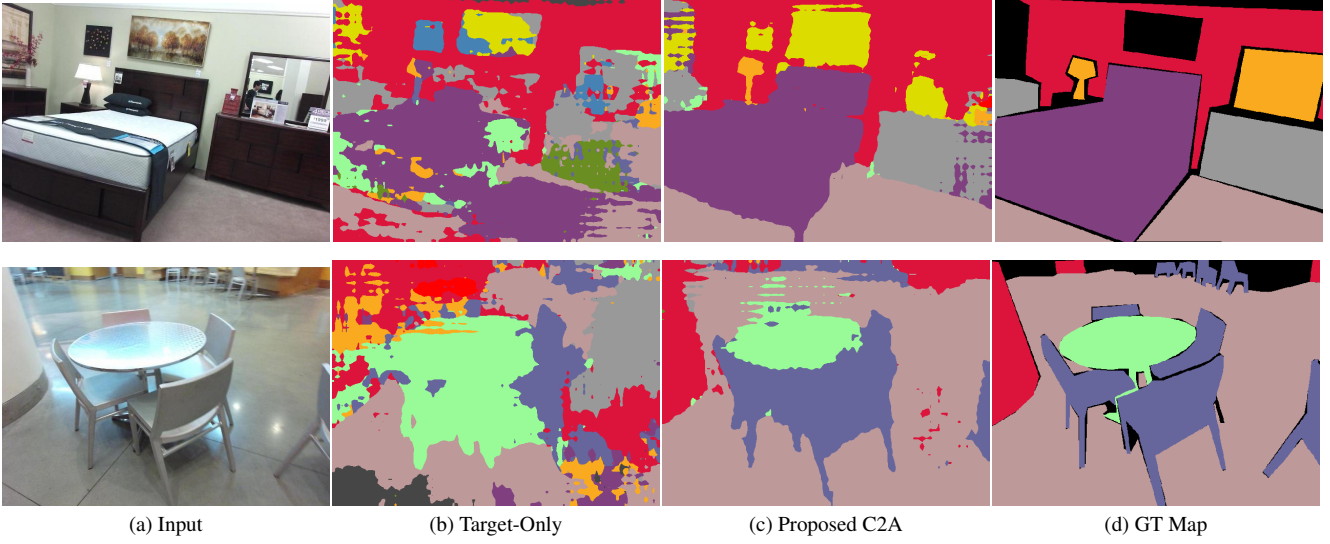| (a) Input | (b) Target-Only | (c) Proposed C2A | (d) GT Map |

Figure 4. Qualitative segmentation outputs for examples from the SUNRGB validation set. Compared to a baseline model that is only trained on the few-shot target domain data, the proposed model (C2A) consistently produces better segmentation maps compared to the baselines in all cases.
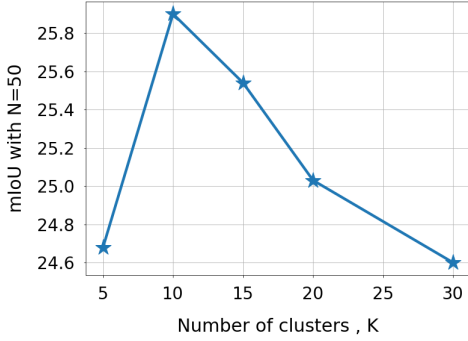

Table 3. **Effect of K:** Influence of the number of clusters K on the performance of segmentation in the case of N=50.

Method	SUN	NYUv2
Train on SceneNet [29]	14.09	15.05
Joint train on SN+GTA	16.99	18.39
Ours (C2A); K=5	21.64	22.06
Ours (C2A); K=10	21.89	20.27
Ours (C2A); K=20	**22.79**	**23.08**

Table 4. **Zero Shot Unsupervised Adaptation:** Our approach significantly outperforms all the baselines, even in the extreme case of having 0 real target images during training.

from $\sim 17\%$ to $\sim 22.8\%$, indicating that the benefit is not only due to increase in labeled data, but due to alignment as well. Additionally, a network which learns without any real data from target domain should ideally generalize well to any similar datasets. Indeed, we observe from Table 4 that the improvement on performance is not restricted to SUNRGB alone, but also observed on NYUv2 [39] validation set across all the baselines.

5. Conclusion

We introduce C2A, a clustering based approach called C2A to study the most general, yet largely understudied setting of adaptation between domains with non-overlapping label spaces for feature alignment across source and target datasets with disjoint labels. C2A encourages positive alignment of visually similar feature representations while preventing negative transfer. We experimentally verify the effectiveness of our approach on the task of outdoor to indoor adaptation for semantic segmentation and demonstrate significant improvements over existing approaches and prevalent baselines in both fewshot and zeroshot adaptation settings.

References

[1] Charu C Aggarwal. *Data classification: algorithms and applications*. CRC press, 2014. 2

[2] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 15384–15394, 2021. 2

[3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017. 2

[4] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in neural information processing systems*, pages 343–351, 2016. 1, 2

[5] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, 2018. 2

[6] Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2985–2994, 2019. 2

[7] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 3

[8] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2959–2968, 2019. 3

[9] Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In *AISTATS*, volume 2005, pages 57–64. Citeseer, 2005. 4

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 5

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette,

Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 1, 2

[13] Rui Gong, Yuhua Chen, Danda Pani Paudel, Yawei Li, Ajad Chhatkuli, Wen Li, Dengxin Dai, and Luc Van Gool. Cluster, split, fuse, and update: Meta-learning for open compound domain adaptive semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8344–8354. Computer Vision Foundation / IEEE, 2021. 3

[14] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8401–8409, 2019. 3, 5

[15] 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. *arXiv preprint arXiv:1711.03213*, 2017. 2

[16] 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, 6

[17] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. *arXiv preprint arXiv:1711.10125*, 2017. 3

[18] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 65. BMVA Press, 2018. 6, 7

[19] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and CV Jawahar. Universal semi-supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5259–5270, 2019. 1, 2, 6, 7

[20] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. Mseg: a composite dataset for multi-domain semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2879–2888, 2020. 2

[21] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[22] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015. 1

[23] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2208–2217. JMLR. org, 2017. 1

[24] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019. 6

[25] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li F Fei-Fei. Label efficient learning of transferable representations acrosss domains and tasks. In *Advances in Neural Information Processing Systems*, pages 165–177, 2017. 2, 4, 6

[26] Fengmao Lv, Tao Liang, Xiang Chen, and Guosheng Lin. Cross-domain semantic segmentation via domain-invariant interactive relation transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4334–4343, 2020. 2

[27] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297. Oakland, CA, USA, 1967. 2

[28] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017. 3

[29] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2678–2687, 2017. 5, 7, 8

[30] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. *arXiv preprint arXiv:2008.12197*, 2020. 2

[31] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018. 2, 4

[32] Luigi Musto and Andrea Zinelli. Semantically adaptive image-to-image translation for domain adaptation of semantic segmentation. *arXiv preprint arXiv:2009.01166*, 2020. 2

[33] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 3

[34] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 5

[35] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 5

[36] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 5

[37] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–168, 2018. 2

[38] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018. 2

[39] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 8

[40] Kihyuk Sohn, Wenling Shang, Xiang Yu, and Manmohan Chandraker. Unsupervised domain adaptation for distance metric learning. *ICLR*, 2018. 2

[41] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 5

[42] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 1

[43] Ruoqi Sun, Xinge Zhu, Chongruo Wu, Chen Huang, Jianping Shi, and Lizhuang Ma. Not all areas are equal: Transfer learning for semantic segmentation via hierarchical region selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4360–4369, 2019. 6

[44] Hui Tang, Xiatian Zhu, Ke Chen, Kui Jia, and C. L. Philip Chen. Towards uncovering the intrinsic data structures for unsupervised domain adaptation using structurally regularized deep clustering. *CoRR*, abs/2012.04280, 2020. 3

[45] Marco Toldo, Umberto Michieli, and Pietro Zanuttigh. Unsupervised domain adaptation in semantic segmentation via orthogonal and clustered embeddings. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 1357–1367. IEEE, 2021. 3

[46] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 3, 6

[47] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015. 2

[48] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017. 1, 2

[49] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 4, 6

[50] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *European Conference on Computer Vision*, pages 642–659. Springer, 2020. 2

[51] 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 Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, volume 11209 of *Lecture Notes in Computer Science*, pages 535–552. Springer, 2018. 3

[52] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487, 2016. 3, 5

[53] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 5423–5432, 2018. 6

[54] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. *arXiv preprint arXiv:1912.03330*, 2019. 3

[55] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2, 4

[56] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014. 3

[57] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2720–2729, 2019. 2

[58] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019. 2