C-SFDA: A Curriculum Learning Aided Self-Training Framework for Efficient Source Free Domain Adaptation

Nazmul Karim*, Niluthpol Chowdhury Mithun†, Abhinav Rajvanshi†, Han-pang Chiu†
Supun Samarasekera†, Nazanin Rahnavard*

*Department of Electrical Engineering, Orlando, FL, USA †SRI International, Princeton, NJ, USA
*nazmul.karim18@knights.ucf.edu †firstname.lastname@sri.com

Abstract

Unsupervised domain adaptation (UDA) approaches focus on adapting models trained on a labeled source domain to an unlabeled target domain. In contrast to UDA, source-free domain adaptation (SFDA) is a more practical setup as access to source data is no longer required during adaptation. Recent state-of-the-art (SOTA) methods on SFDA mostly focus on pseudo-label refinement based self-training which generally suffers from two issues: i) inevitable occurrence of noisy pseudo-labels that could lead to early training time memorization, ii) refinement process requires maintaining a memory bank which creates a significant burden in resource constraint scenarios. To address these concerns, we propose C-SFDA, a curriculum learning aided self-training framework for SFDA that adapts efficiently and reliably to changes across domains based on selective pseudo-labeling. Specifically, we employ a curriculum learning scheme to promote learning from a restricted amount of pseudo labels selected based on their reliabilities. This simple yet effective step successfully prevents label noise propagation during different stages of adaptation and eliminates the need for costly memory-bank based label refinement. Our extensive experimental evaluations on both image recognition and semantic segmentation tasks confirm the effectiveness of our method. C-SFDA is readily applicable to online test-time domain adaptation and also outperforms previous SOTA methods in this task.

1. Introduction

Deep neural network (DNN) models have achieved remarkable success in various visual recognition tasks [15, 20, 41, 43]. However, even very large DNN models often suffer significant performance degradation when there is a distribution or domain shift [54, 77] between training (source) and test (target) domains. To address the problem of domain shifts, various Unsupervised Domain Adaptation (UDA) [17, 29] algorithms have been developed over recent years. Most UDA techniques require access to labeled source domain data during adaptation, which limits their application in many real-world scenarios, *e.g*. source data is private, or adaptation in edge devices with limited computational capacity. In this regard, source-free domain adaptation setting has recently gained significant interest [33, 34, 84], which considers the availability of only source pre-trained model and unlabeled target domain data.

Recent state-of-the-art SFDA methods (e.g., SHOT [42], NRC [83], G-SFDA [85], AdaContrast [4]) mostly rely on the self-training mechanism that is guided by the source pre-trained model generated pseudo-labels (PLs). PL refinement using the knowledge of per-class cluster structure in feature space is recurrently used in these methods. At early stages of adaptation, the label information formulated based on cluster structure can be severely misleading or noisy; shown in Fig. 1. As the adaptation progresses, this label noise can negatively impact the subsequent cluster structure as the key to learning meaningful clusters hinges on the quality of pseudo-labels itself. Therefore, the inevitable presence of label noise at early training time is a critical issue in SFDA and requires proper attention. Furthermore, distributing cluster knowledge among neighbor samples requires a memory bank [4, 42] which creates a significant burden in resource-constraint scenarios. In addition, most memory bank dependent SFDA techniques are not suitable for online test-time domain adaptation [73,75]; an emerging area of UDA that has gained traction in recent times. *Designing a memory-bank-free SFDA approach that can guide the self-training with highly precise pseudo-labels is a very challenging task and a major focus of this work.*

In our work, we focus on increasing the reliability of generated pseudo-labels without using a memory-bank and clustering-based pseudo-label refinement. Our analysis shows that avoiding early training-time memoriza-

*Work done during Nazmul Karim's internship with SRI International.

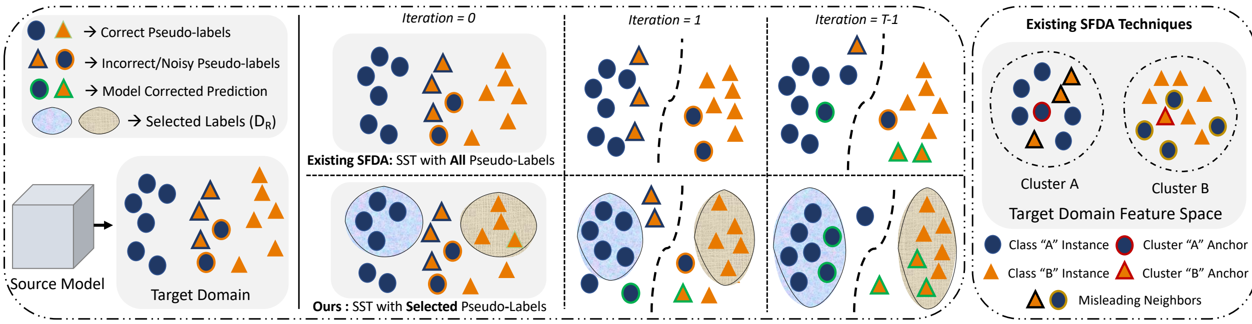


Figure 1. *Left:* In source-free domain adaptation, we only have a source model that needs to be adapted on the target data. Among the source-generated pseudo-labels, a large portion is noisy which is important to avoid during supervised self-training (SST) with regular cross-entropy loss. Instead of using all pseudo-labels, we choose the most reliable ones and effectively propagate high-quality label information to unreliable samples. As the training progresses, the proposed selection strategy tends to choose more samples for SST due to the improved average reliability of pseudo-labels. Such a restricted self-training strategy creates a model with better discrimination ability and eventually corrects the noisy predictions. Here, T is the total number of iterations. *Right:* While existing SFDA techniques leverages cluster structure knowledge in the feature space, there may exist many misleading neighbors–neighbors' pseudo labels that are different from the anchors' true label. Therefore, clustering-based label propagation inevitably suffers from label noise in subsequent training.

tion (ETM) of noisy PLs encourages noise-free learning in subsequent stages of adaptation. We further analyze that even with an expensive label refinement technique in place, learning equally from all labels eventually leads to label-noise memorization. Therefore, we employ a curriculum learning-aided self-training framework, C-SFDA, that prioritizes learning from easy-to-learn samples first and hard samples later on. We show that one can effectively identify the group of easy samples by utilizing the reliability of pseudo-labels, *i.e.* prediction confidence and uncertainty. We then follow a carefully designed curriculum learning pipeline to learn from highly reliable (easy) pseudo-labels first and gradually propagate more refined label information among less reliable (hard) samples later on. In addition to the label-guided self-training, we facilitate unsupervised contrastive representation learning that does not require any label information and helps us prevent the ETM phenomenon. Our main contributions can be summarized as follows:

- We introduce a novel SFDA technique that focuses on noise-free self-training exploiting the reliability of generated pseudo-labels. With the help of curriculum learning, we aim to prevent early training time memorization of noisy pseudo-labels and improve the quality of subsequent self-training as shown in Fig. 1.
- By prioritizing the learning from highly reliable pseudo-labels first, we aim to propagate *refined and accurate label information* among less reliable samples. Such a selective self-training strategy eliminates the requirement of a computationally costly and memory-bank dependent label refinement framework.
- C-SFDA achieves state-of-the-art performance on major benchmarks for image recognition and semantic segmentation. Being highly memory-efficient, the proposed method is readily applicable to online test-time adaptation settings and obtains SOTA performance.

2. Related Work

Unsupervised Domain Adaptation: UDA for visual recognition tasks has been widely studied in the literature [9, 74]. Adversarial learning [21, 46, 70, 72], image-to-image translation [21,36,50], cross-domain divergence minimization [3, 39, 64, 67],and optimal transport [5, 11, 81, 81] are popular techniques across prior works on UDA. Self-training [48, 79, 86, 92] has recently been a dominant trend in UDA, which uses labeled source data and pseudo-labeled target data (typically generated using a teacher model) to iteratively train a student model. Most of the UDA approaches consider continued access to labeled source domain data during domain adaptation training, which leads to several practical concerns (e.g., data privacy). To cope with this, SFDA setting has drawn significant interest, which does not consider using source data during adaptation.

Source-Free UDA: In recent years, several approaches [4, 13, 33, 37, 42, 45, 56, 69, 78, 82, 85] addressing the source-free domain adaptation problem has been proposed. SHOT [42] utilizes a centroid-based label refinement technique that guides the self-training. G-SFDA [85] and NRC [83] follow a similar strategy with further measures for refining pseudo-labels by encouraging consistent predictions between local neighbor samples. In addition to label refinement, AdaContrast [4] leverages MoCo [19] like contrastive feature learning for SFDA by excluding the same class negative pairs detected by the pseudo-labels. However, pseudo-labels generated at early training stage can be noisy, a fact that has not been well-addressed in these works. Moreover, almost all of these label refinement strategies require having a large memory queue which is undesirable in many real-world scenarios. Other methods such as A^{2}Net [78] utilize a new target classifier and adversarial training to align source and target domains. SFDA-DE [13] tries to align 2 domains through source distribution estimation. These approaches also suffer from cluster structure-

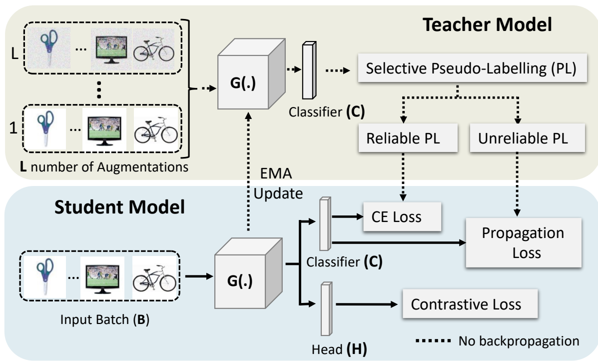


Figure 2. Overview of our proposed method where we use teacher predictions to train the student model. At early stage of training, we only consider reliable pseudo-labels for CE loss as well as contrastive loss for unsupervised feature representation learning. As the model gets more confident, we use label propagation loss to distribute high-quality label information, learned from reliable pseudo-labels, among unreliable samples.

induced label noise. In contrast, our proposed method reduces the risk of label noise memorization without any rigorous and expensive pseudo-label refinement technique.

3. Problem Setting

We start by defining the source domain dataset, $\mathcal{D}_{\mathsf{s}} = \{(x_{\mathsf{s}}^i, y_{\mathsf{s}}^i)\}_{i=1}^{N_{\mathsf{s}}}$ containing N_s labeled samples distributed over K number of classes. Here, $y \in \mathcal{Y}_{\mathsf{s}} \subseteq \mathbb{R}^K$ denotes the one-hot ground-truth label for sample x. Let $f_{\theta_{\mathsf{s}}} : x_s \rightarrow y_s$ be a source model trained on $\mathcal{D}_{\mathsf{s}}$. In general, model f consists of a feature extractor $\mathbf{G}$ and a fully connected (FC)-layer based classifier $\mathbf{C}$. We denote the target domain dataset, $\mathcal{D}_{\mathsf{t}} = \{(x_{\mathsf{t}}^i)\}_{i=1}^{N_{\mathsf{t}}}$ containing N_{t} unlabeled samples. We have access to the target labels $\{y_{\mathsf{t}}^i)\}_{i=1}^{N_{\mathsf{t}}}$ for evaluation only. $\mathcal{D}_{\mathsf{s}}$ and $\mathcal{D}_{\mathsf{t}}$ have same underlying label distribution with a common label set $\mathcal{C} = \{1, 2, \cdots, K\}$. In this paper, we focus on SFDA problem where access to $\mathcal{D}_{\mathsf{s}}$ is unavailable while adapting $f_{\theta_{\mathsf{s}}}$ on $\mathcal{D}_{\mathsf{t}}$.

4. Proposed Method

We employ a pseudo-labeling based self-training mechanism for SFDA. Consider a self-training framework shown in Fig. 2 where we use a teacher model ($f_{\hat{\theta}_{\mathsf{t}}}$) for generating the target pseudo-label for a student model ($f_{\theta_{\mathsf{t}}}$). At the start of training, both models share the same weights, *i.e.* $\theta_{\mathsf{t}} = \hat{\theta}_{\mathsf{t}} = \theta_{\mathsf{s}}$. For generating the target pseudo-label, we simply use the teacher model prediction, $\hat{y}_{\mathsf{t}}$. Considering a batch size of B, we use the batch of pseudo-labels $\mathcal{Y}_{\mathsf{t}} = \{\hat{y}_{\mathsf{t}}^{(i)}\}_{i=1}^{B}$ to update θ_{t} using a cross-entropy (CE) loss,

$$\mathcal{L}_{ce} = -\frac{1}{B}\sum_{i=1}^{B} \hat{y}_{\mathsf{tc}}^i \cdot \log f_{\theta_{\mathsf{t}}}(x_{\mathsf{t}}^i)), \quad (1)$$

where $\hat{y}_{\mathsf{tc}}^i$ is one hot encoding of $\hat{y}_{\mathsf{t}}^i$. Minimizing $\mathcal{L}_{ce}$ enforces consistency between student and teacher predictions.

In our self-training scenario, there is an inevitable presence of noisy labels in $\mathcal{Y}_{\mathsf{t}} \in \mathbb{R}^B$ which often leads to non-optimal model performance [1]. As an initial label-noise preventive measure, we intend to produce stable and high-quality pseudo-labels following these two simple steps: *i) augmentations averaged prediction, and ii) weight averaged teacher model*. We execute the first step by taking the augmentation-averaged teacher predictions,

$$\hat{y}_{\mathsf{t}} = \arg\max \frac{1}{L}\sum_{l=1}^{l=L} \hat{h}_t^l = \arg\max \frac{1}{L}\sum_{l=1}^{l=L} f_{\hat{\theta}_{\mathsf{t}}}(\hat{x}_{\mathsf{t}}^l). \quad (2)$$

Here, $\hat{x}_{\mathsf{t}}^l$ is the l^{th} augmented copy of the input x_{t} and L *(=12 in our case)* is the total number of augmentations. Data augmentation [10] has been widely used to improve model generalization performance to unseen data [32, 65]. In our case, however, we aim to generate consistent teacher model predictions over augmented input distributions. Although augmentations can be dataset-specific and manually designed [32, 42, 65], we use a general augmentation policy that is applicable to multiple datasets. For the second step, we simply take the exponential moving average (EMA) of student weights to update the teacher model during each iteration ($j \rightarrow j+1$),

$$\hat{\theta}_t^{j+1} = \gamma\hat{\theta}_t^j + (1-\gamma)\theta_t^{j+1}, \quad (3)$$

where γ is a smoothing factor that controls the degree of change we want during each iteration.

However, regardless of these measures, adopting an entire batch of pseudo-labels for self-training eventually lead to the memorization of noisy labels. To alleviate this, we propose a curriculum learning aided self-training strategy that encourages learning from high reliable pseudo-labels.

4.1. Selective Pseudo-Labelling

We describe the process of reliable label selection first as it is a vital part of the proposed curriculum learning strategy. To measure the label reliability, we intend to exploit two widely-used [14, 23, 57, 76] statistics: *i) prediction confidence and ii) average entropy or uncertainty*. In general, if a model is well-calibrated, the accuracy of that model can be strongly related to prediction confidence [18]. Therefore, prediction confidence can be a reliable measure of pseudo-label accuracy in SFDA settings. In addition to the confidence score, [18] shows that difference in entropy can be a reliable estimate for different types of domain shift. While [18] assumes the availability of both source and target domains, we are restricted to target domain data only. Therefore, we use a carefully designed augmentation policy to create a virtual distribution shift among target domain data. Prediction variance or uncertainty over augmented distributions should give us a close estimate of actual domain shift. To this end, we assign a binary reliability score

(r^i) to each target sample based on their prediction confidence and prediction uncertainty (g_u^i),

$$r^i = \begin{cases} 1, & \text{if } conf(\hat{h}_{\mathrm{t}}^i) \geq \tau_c \text{ and } g_u^i \leq \tau_u \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

We calculate g_u^i by taking the standard deviation (*std.*) over augmentation-based predictions, $g_u^i = std\{conf(\hat{h}_{\mathrm{t}}^l)\}_{l=1}^{l=L}$. We particularly consider aleatoric uncertainty [26] here since it better addresses the concern of domain shift. The pair of selection thresholds τ_c and τ_u can be estimated as

$$\tau_c = \frac{1}{B}\sum_{i=1}^{i=B} conf(\hat{h}_{\mathrm{t}}^i); \quad \tau_u = \frac{1}{B}\sum_{i=1}^{i=B} g_u^i. \tag{5}$$

Taking the prediction average as a threshold eliminates the requirement of per-dataset hyper-parameter tuning and makes our selection process highly adaptive. Note that, the proposed selection strategy is also applicable to fully test-time adaptation [4, 73] without any modification to the proposed method.

4.2. Loss Functions

After getting the reliability score for each sample, we separate the input batch $\mathbb{D}$ into *more reliable (R)* and *less reliable (U)* groups, $\mathbb{D}_R = \{(x_{\mathrm{t}}^i, \hat{y}_{\mathrm{t}}^i) : r_i = 1\}_{i=1}^B$ and $\mathbb{D}_U = \{(x_{\mathrm{t}}^i, \hat{y}_{\mathrm{t}}^i) : r_i = 0\}_{i=1}^B$. While this gives us a good estimate of reliable samples, $\mathbb{D}_R$ may lack diverse samples (sometimes, missing some categories completely). As a potential remedy to this, we choose a few samples from $\mathbb{D}_U$ based on another metric: *Top-2 confidence score difference (DoC)* and consider them as reliable. Finally, we employ class-balanced cross-entropy loss for $\mathbb{D}_R$ ($\mathcal{L}_{ce}^R$) with an inverse frequency loss-weighting factor (λ_k) that accounts for the label imbalance in $\mathbb{D}_R$. Details of DoC and λ_k are in supplementary. For $\mathbb{D}_U$, we employ label propagation loss [90] as follows,

$$\mathcal{L}_{\mathcal{P}} = \frac{1}{2|\mathbb{D}_U|}\sum_{i=1}^{|\mathbb{D}_U|} \|f_{\theta_{\mathrm{t}}}(x_{\mathrm{t}}^i) - \hat{y}_{\mathrm{tc}}^i\|^2. \tag{6}$$

Due to the transductive property of $\mathcal{L}_{\mathcal{P}}$, it propagates label information from $\mathbb{D}_R$ to $\mathbb{D}_U$.

Note that both $\mathcal{L}_{CE}^R$ and $\mathcal{L}_P$ require pseudo-label which may lead to memorization; depending on the success in selection stage. In addition to supervised self-training, learning useful representations of images in an unsupervised manner may reduce the risk of memorization. One such approach is *fully unsupervised contrastive learning (CL)* where meaningful representation learning becomes possible by enforcing similarity between two augmented copies of each sample x_t, ${x_t}^{aug,1}$ and ${x_t}^{aug,2}$. To this end, we employ a projection head $\mathbf{H}$ to obtain feature projections $q_i = \mathbf{H}(\mathbf{G}({x_t}^{aug,1}))$, and $q_j = \mathbf{H}(\mathbf{G}({x_t}^{aug,2}))$ that gives us the contrastive criterion [7, 30] as

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(q_i, q_j)/\kappa)}{\sum_{b=1}^{2B} 1_{b\neq i} \exp(\mathrm{sim}(q_i, q_b)/\kappa)}\ , \tag{7}$$

$$\mathcal{L}_{\mathcal{C}} = \frac{1}{2B}\sum_{b=1}^{2B}[\ell_{2b-1,2b} + \ell_{2b,2b-1}], \tag{8}$$

where $1_{b\neq i}$ is an indicator function that gives a 1 if $b \neq i$, κ is a temperature constant and $\mathrm{sim}(q_i, q_j)$ is the cosine similarity between q_i and q_j. Even though label-dependent contrastive learning has been employed for SFDA [4], we focus on label-independent CL to minimize the effect of label noise; especially at an early stage of training. Finally, the total loss can be expressed as

$$\mathcal{L}_{tot} = \mu_r \mathcal{L}_{ce}^R + (1-\mu_r)\mathcal{L}_{\mathcal{P}} + \mu_c \mathcal{L}_{\mathcal{C}}, \tag{9}$$

where μ_r and μ_c are loss coefficients that dictate the pace of curriculum learning we propose next.

4.3. Curriculum Learning

Curriculum Learning [2, 91] promotes the strategy of learning from easier samples first and harder samples later. Our selection strategy in Section 4.1 provides us with an estimation of easy and hard groups. Since pseudo-labels in $\mathbb{D}_R$ are most likely to be correct, DNN finds it easier to learn from them. On the other hand, learning from $\mathbb{D}_U$ should be more restricted due to the presence of a higher noise level. Therefore, we set an update equation for μ_r as

$$\mu_r^j = \mu_r^{j-1}(1 - \alpha e^{-\frac{1}{d^j}}), \tag{10}$$

where $d^j = \frac{\tau_u}{\tau_c}$ is difficulty score of current batch of samples and μ_r^{j-1} is the labeled loss coefficient at previous iteration. We set α and μ_r^0 to 0.01 and 1, respectively, to restrict the learning from $\mathbb{D}_U$ since pseudo-labels are mostly noisy during the early stages of training. As the training progresses and overall reliability improves, we start learning from $\mathbb{D}_U$ by gradually decreasing μ_r. In addition, the change in μ_r is directly controlled by the difficulty in learning the current batch of samples. If the batch of samples at iteration j is hard-to-learn, (*i.e.* d^j is high), we keep the change in μ_r to minimal. Similarly, we exponentially decrease the contrastive loss coefficient μ_c as

$$\mu_c^j = \mu_c^{j-1} e^{-\beta}. \tag{11}$$

We set initial μ_c^0 to 0.5 as unsupervised feature learning helps more at early stages of training. β is set to be 1e-4.

4.4. Semantic Segmentation

Up to now, we have only considered the classification task where each input sample is associated with a single label. However, semantic segmentation is a multi-label classification task where we assign a label to each pixel. Consider a target domain image $x \in \mathbb{R}^{H\times W}$ where x_{ij} indicates the pixel of i^{th} row and j^{th} column. The task at hand is to assign one of K semantic labels, $y_{ij} \in 1, 2, \ldots K$ to each x_{ij}. For x, we use a model (f) to produce a probabilistic output prediction $p \in \mathbb{R}^{H\times W\times K}$ over K classes. The map of pseudo-labels ($\hat{y} \in \mathbb{R}^{H\times W}$)can be estimated as $\hat{y} = \arg\max_k p; k \in 1, 2, \ldots K$. As some predictions are more reliable than others, using similar selection criteria (as image classification) to separate pixels makes sense. However, instead of using one single threshold for all pixels, we instead choose per-category thresholds. To this end, we estimate a pair of thresholds for each category k. Given a batch, we accumulate all confidence scores and select the per-category confidence threshold, τ_c^k, as the P-th percentile confidence score. Similarly, we select P-th percentile uncertainty score for the uncertainty threshold, τ_c^k. In our work, we set the value of P to 55. After choosing the thresholds, we follow eq. 4 to assign a per-pixel reliability score, r_{ij}. As for loss functions, we consider cross-entropy loss ($\mathcal{L}_{ce}^R$) for the reliable labels $\hat{y}^R$, and to promote diverse predictions, we minimize the prediction entropy loss,

$$\mathcal{L}_E = -\frac{1}{HW}\sum_{i=1,j=1}^{H,W} p_{ij}\cdot\log(p_{ij}). \tag{12}$$

Finally, we update our model by minimizing the total loss,

$$\mathcal{L}_{tot} = \mathcal{L}_{ce}^R + \mu_e\mathcal{L}_E, \tag{13}$$

Where μ_e is the entropy loss coefficient. We follow a similar update equation as 11 for μ_E with an initial value of, μ_e^0 = 1e-3. Note that, we only update BN layers and freeze other parameters. For uncertainty measures, we use ColorJitter and Gaussian noise as augmentation transformations. More details are in the supplementary.

5. Experiments

5.1. Datasets

Image Classification Datasets: *Office-31* [60] is a small-scale benchmark with images from 31 categories across 3 domains, **A**mazon (2,817), **D**SLR (498) and **W**ebcam (795). *Office-Home* [71] has a total of 15.5K images from 65 classes collected from 4 different image domains: **Ar**tistic, **Cl**ipart, **Pr**oduct, and **R**eal-world. We consider 12 transfer tasks for this dataset. *VisDA* [55] contains 2 different domains, synthetic and real, with 12 classes in both domains. The synthetic or source domain contains 150K rendered 3D images with different poses. The corresponding real or target domain contains about 55K real-world images. For evaluation, we consider per-class accuracy and the average (Avg.) over them. *DomainNet* [54] is another large-scale dataset with 6 domains containing over 500K images from 126 classes. We consider 4 domains (**R**eal, **S**ketch, **C**lipart, **P**ainting), as [61] identify severe noisy labels in the dataset. We evaluate the methods on 7 transfer tasks between 4 domains and report top-1 accuracy.

Table 1. Classification accuracy (%) under UDA and SFDA settings on **Office-31** dataset (ResNet50 backbone). We report Top-1 accuracy on 6 domain shifts (→) and take the average (Avg.) over them. The best results under the SFDA setting are shown in bold font.

Method	source -free	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
MCC [28]	×	95.6	95.4	72.6	98.6	73.9	100	89.4
GSDA [22]	×	94.8	95.7	73.5	99.1	74.9	100	89.7
CAN [29]	×	95.0	94.5	78.0	99.1	77.0	99.8	90.6
SRDC [68]	×	95.8	95.7	76.7	99.2	77.1	100	90.8
SFDA [31]	✓	92.2	91.1	71.0	98.2	71.2	99.5	87.2
SHOT [42]	✓	94.0	90.1	74.7	98.4	74.3	99.9	88.6
3C-GAN [37]	✓	92.7	93.7	75.3	98.5	77.8	99.8	89.6
A^{2}Net [78]	✓	94.5	94.0	76.7	**99.2**	76.1	**100**	90.1
SFDA-DE [13]	✓	96.0	**94.2**	76.6	98.5	75.5	99.8	90.1
C-SFDA (Ours)	✓	**96.2**	93.9	**77.3**	98.8	**77.5**	99.7	**90.5**

Semantic Segmentation Datasets: For segmentation, we consider GTA5→Cityscapes, SYNTHIA→Cityscapes & CityScapes→Dark-Zurich adaptations tasks. GTA5 [58] contains ∼25k synthetic images, with a resolution of 1914× 1052, generated from GTA5 video frames. Cityscapes [8] provides 3,975 daytime street scenes, with a resolution of 2048 × 1024, from 50 different cities. Following prior work [21, 72, 93], we consider splitting Cityscapes images into train-val splits and report 19-way classification performance over the validation split. SYNTHIA [59] is another synthetic dataset with 9400 scenes of size 1280x760. As SYNTHIA and Cityscapes have overlaps only for 16 categories, we report 16-way and 13-way performances for SYNTHIA→Cityscapes. Dark Zurich [62] is a large dataset with 2,416 nighttime unlabeled images of 1080p resolution.

5.2. Implementation Details

We use ResNet50 [20] backbone for Office-31, Office-Home, DomainNet and ResNet-101 [20] for VisDA. Following SHOT [42], we replace the fully connected (FC) layer with a 256-dimensional bottleneck layer and task-specific FC classification layer. We use batch normalization [27] after bottleneck and apply WeightNorm [63] on the classifier. For source training, we initialize the models with ImageNet-1K [12] pre-trained weights. Following [42], we split the source dataset into the train (90%) and validation (10%) sets. Compared to the backbone, we employ a 10 times higher learning rate for bottleneck and classifier. For target domain adaptation, we use similar training settings for all datasets. For Office datasets [60, 71], we use SGD optimizer with a learning rate of 5e-3 and a momentum of 0.9 with a weight decay of 1e-4. We use a batch size of 256. For VisDA and DomainNet, we use a learning rate of 5e-4 with cosine annealing [4]. For VisDA, we train for 20 epochs with a batch size of 128. We consider a larger batch size of 512 for DomainNet to prevent severe class imbalance in $\mathbb{D}_R$ and train for 25 epochs. For the EMA update, we set η to 0.98 for all datasets.

For all semantic segmentation tasks, we use DeepLabV2 [6] with a ResNet101 [20] backbone and

Table 2. Classification performance (%) under UDA and SFDA settings on **Office-Home** dataset (ResNet50 backbone). We report Top-1 accuracy on 12 domain shifts (→) and take the average (Avg.) over them. Our method achieves SOTA performance on 8 of these shifts.

Method	SF	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
RSDA [17]	×	53.2	77.7	81.3	66.4	74.0	76.5	67.9	53.0	82.0	75.8	57.8	85.4	70.9
TSA [40]	×	57.6	75.8	80.7	64.3	76.3	75.1	66.7	55.7	81.2	75.7	61.9	83.8	71.2
SRDC [68]	×	52.3	76.3	81.0	69.5	76.2	78.0	68.7	53.8	81.7	76.3	57.1	85.0	71.3
FixBi [51]	×	58.1	77.3	80.4	67.7	79.5	78.1	65.8	57.9	81.7	76.4	62.9	86.7	72.7
SFDA [31]	✓	48.4	73.4	76.9	64.3	69.8	71.7	62.7	45.3	76.6	69.8	50.5	79.0	65.7
G-SFDA [85]	✓	57.9	78.6	81.0	66.7	77.2	77.2	65.6	56.0	82.2	72.0	57.8	83.4	71.3
SHOT [42]	✓	57.1	78.1	81.5	68.0	78.2	78.1	67.4	54.9	82.2	73.3	58.8	84.3	71.8
A^{2}Net [78]	✓	58.4	79.0	82.4	67.5	79.3	78.9	**68.0**	56.2	82.9	**74.1**	60.5	85.0	72.8
SFDA-DE [13]	✓	59.7	79.5	82.4	**69.7**	78.6	**79.2**	66.1	57.2	82.6	73.9	60.8	85.5	72.9
C-SFDA (Ours)	✓	**60.3**	**80.2**	**82.9**	69.3	**80.1**	78.8	67.3	**58.1**	**83.4**	73.6	**61.3**	**86.3**	**73.5**

Table 3. Source-free (SF) domain adaptation performance on **VisDA** dataset (ResNet-101 backbone) shown by per-class accuracy (%) and their average (Avg.). Our method improves the average accuracy by 1% compared to the previous SOTA, Adacon [4]. C-SFDA also achieves a significant performance gain (3.5% in Avg.) for online test-time domain adaptation settings.

Method	SF	plane	bike	bus	car	horse	knife	mcycle	person	plant	sktbrd	train	truck	Avg.
SFAN [80]	×	93.6	61.3	84.1	70.6	94.1	79.0	91.8	79.6	89.9	55.6	89.0	24.4	76.1
SWD [35]	×	90.8	82.5	81.7	70.5	91.7	69.5	86.3	77.5	87.4	63.6	85.6	29.2	76.4
MCC [28]	×	88.7	80.3	80.5	71.5	90.1	93.2	85.0	71.6	89.4	73.8	85.0	36.9	78.8
STAR [47]	×	95.0	84.0	84.6	73.0	91.6	91.8	85.9	78.4	94.4	84.7	87.0	42.2	82.7
RWOT [81]	×	95.1	80.3	83.7	90.0	92.4	68.0	92.5	82.2	87.9	78.4	90.4	68.2	84.0
SE [16]	×	95.9	87.4	85.2	58.6	96.2	95.7	90.6	80.0	94.8	90.8	88.4	47.9	84.3
Source only	-	57.2	11.1	42.4	66.9	55.0	4.4	81.1	27.3	57.9	29.4	86.7	5.8	43.8
3C-GAN [37]	✓	94.8	73.4	68.8	74.8	93.1	95.4	88.6	84.7	89.1	84.7	83.5	48.1	81.6
SHOT [42]	✓	94.3	88.5	80.1	57.3	93.1	94.9	80.7	80.3	91.5	89.1	86.3	58.2	82.9
A^{2}Net [78]	✓	94.0	87.8	85.6	66.8	93.7	95.1	85.8	81.2	91.6	88.2	86.5	56.0	84.3
G-SFDA [85]	✓	96.1	88.3	85.5	**74.1**	97.1	95.4	89.5	79.4	95.4	92.9	89.1	42.6	85.4
SFDA-DE [13]	✓	95.3	**91.2**	77.5	72.1	95.7	**97.8**	85.5	**86.1**	**95.5**	**93.0**	86.3	61.6	86.5
AdaCon [4]	✓	97.0	84.7	84.0	77.3	96.7	93.8	91.9	84.8	94.3	93.1	**94.1**	49.7	86.8
C-SFDA (Ours)	✓	**97.6**	88.8	**86.1**	72.2	**97.2**	94.4	**92.1**	84.7	93.0	90.7	93.1	**63.5**	**87.8**
AdaCon [4] (Online)	✓	95.0	68.0	82.7	**69.6**	94.3	80.8	90.3	79.6	**90.6**	69.7	**87.6**	36.0	78.7
C-SFDA (Online)	✓	**95.9**	**75.6**	**88.4**	68.1	**95.4**	**86.1**	**94.5**	**82.0**	89.2	**81.4**	87.3	**43.8**	**82.2**

Table 4. Classification accuracy (%) on **DomainNet** for source-free domain adaptation (ResNet-50 backbone). Considering 7 domain shifts, the proposed method achieves the best results on 5 of them. The performance for online test-time adaptation is also superior to SOTA AdaCon [4].

Method	SF	R→C	R→P	P→C	C→S	S→P	R→S	P→R	Avg.
MCC [28]	×	44.8	65.7	41.9	34.9	47.3	35.3	72.4	48.9
Source only	-	55.5	62.7	53.0	46.9	50.1	46.3	75.0	55.6
TENT [73]	✓	58.5	65.7	57.9	48.5	52.4	54.0	67.0	57.7
SHOT [42]	✓	67.7	68.4	66.9	60.1	66.1	59.9	**80.8**	67.1
AdaCon [4]	✓	70.2	69.8	**68.6**	58.0	65.9	61.5	80.5	67.8
Ours	✓	**70.8**	**71.1**	68.5	**62.1**	**67.4**	**62.7**	80.4	**69.0**
AdaCon [4] (online)	✓	61.1	66.9	60.8	53.4	62.7	54.5	**78.9**	62.6
Ours (online)	✓	**61.6**	**67.4**	**61.3**	**55.1**	**63.2**	**54.8**	78.5	**63.1**

initialize models with ImageNet-1K [12] pre-trained weights. For source GTA5 and SYNTHIA source models training, we use SGD optimizer with a 1e-4 learning rate and a momentum of 0.9 with a weight decay of 5e-4. We train the model for 20 epochs with a batch size of 8 and apply different weather augmentations [49] during training. For Cityscapes, we follow the settings in [62] and use a learning rate of 2.5e-4. For target domain training, we use a learning rate of 1e-4 to tune only the batch normalization (BN) parameters. With a batch size of 8, we train the model for 50K steps. Note that we only consider online adaptation for Cityscapes→Dark-Zurich and train the model for 1 epoch. Similar to Image classification, we also consider EMA update for segmentation and set η to be 0.995. Please see supplementary for more details.

5.3. Baseline Methods

We consider a number of baselines that work with or without source data. SFAN [80], STAR [47] RWOT [81], SE [16] are among the source-dependent UDA techniques. For source-free settings, we consider SFDA [31], 3C-GAN [37], SHOT [42], A^{2}Net [78], G-SFDA [85], SFDA-DE [13], AdaCon [4]. For Segmentation, we consider SOTA SFDA techniques such as UR [66], SFDA [45], HCL [24]. For online semantic segmentation benchmarks, we consider Test Time BN [52], TENT [73], AUGCO [56].

5.4. Experimental Results

Evaluation on Image Classification Task: We compare the proposed method on Image Classification benchmarks in Table 1-4. We report the Top-1 accuracy for each domain shift and take their average. For Office-31 dataset, we achieve an average 0.4% accuracy improvement over the previous SOTA. We also achieve a similar improvement (0.6%) for the Office-Home dataset. We believe, avoiding the early training time label noise propagation, helps our method significantly to perform well. In VisDA, C-SFDA outperforms SOTA AdaCon [4] by 1% and obtains significant performance improvement for the rare classes such as "truck". Table 4 shows that the proposed method sees similar accuracy improvement (1.2%) over the previ-

Table 5. Performance evaluation on **GTA5→Cityscapes** (DeepLabV2 with ResNet101) where we report mean IoU (mIoU) over 19 categories on Cityscapes validations set. Our method achieves the best mIoU in SFDA and online test-time adaptation.

Method	SF	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg.	Terrain	Sky	PR	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
IDA [53]	×	90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	85.7	40.5	79.7	58.7	31.1	86.3	31.5	48.3	0.0	30.2	35.8	46.3
CrCDA [25]	×	92.4	55.3	82.3	31.2	29.1	32.5	33.2	35.6	83.5	34.8	84.2	58.9	32.2	84.7	40.6	46.1	2.1	31.1	32.7	48.6
ProDA [87]	×	91.5	52.4	82.9	42.0	35.7	40.0	44.4	43.3	87.0	43.8	79.5	66.5	31.4	86.7	41.1	52.5	0.0	45.4	53.8	53.7
CPSL [38]	×	91.7	52.9	83.6	43.0	32.3	43.7	51.3	42.8	85.4	37.6	81.1	69.5	30.0	88.1	44.1	59.9	24.9	47.2	48.4	55.7
Source Only	-	69.7	20.5	73.3	22.1	12.3	23.5	31.8	17.9	78.7	18.7	68.2	53.9	26.5	70.6	32.2	4.5	8.1	26.8	31.5	36.4
UR [66]	✓	92.3	55.2	81.6	30.8	18.8	37.1	17.7	12.1	84.2	35.9	83.8	57.7	24.1	81.7	27.5	44.3	6.9	24.1	40.4	45.1
SFDA [45]	✓	91.7	52.7	82.2	28.7	20.3	36.5	30.6	23.6	81.7	35.6	84.8	59.5	22.6	83.4	29.6	32.4	11.8	23.8	39.6	45.8
HCL [24]	✓	92.0	55.0	80.4	33.5	24.6	37.1	35.1	28.8	83.0	37.6	82.3	59.4	27.6	83.6	32.3	36.6	14.1	28.7	43.0	48.1
C-SFDA (ours)	✓	90.4	42.2	83.2	34.0	29.3	34.5	36.1	38.4	84.0	43.0	75.6	60.2	28.4	85.2	33.1	46.4	3.5	28.2	44.8	**48.3**
TENT [73] (Online)	✓	87.3	39.0	79.8	24.3	19.6	21.2	25.1	16.6	83.8	34.7	77.7	57.9	17.8	85.0	24.9	20.8	2.0	16.6	4.5	38.9
AUGCO [56] (Online)	✓	90.3	41.2	81.8	26.5	21.4	34.5	404.	33.3	83.6	34.6	79.7	61.4	19.3	84.7	30.3	39.5	7.3	27.6	34.6	45.9
C-SFDA (Online)	✓	84.7	37.8	82.4	29.7	28.0	31.8	34.8	29.3	83.7	43.8	76.9	58.8	28.4	84.9	33.5	44.1	0.5	24.5	39.1	**46.3**

Table 6. Performance evaluation on **SYNTHIA→Cityscapes**. We report mean IoU (mIoU) over 16 common categories between SYNTHIA and Cityscapes. mIoU* are calculated over 13 categories. Our method achieves SOTA performance in both mIoU and mIoU*.

Method	SF	Road	SW	Build	Wall*	Fence*	Pole*	TL	TS	Veg.	Sky	PR	Rider	Car	Bus	Motor	Bike	mIoU	mIoU*
IDA [53]	×	84.3	37.7	79.5	5.3	0.4	24.9	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	41.7	48.9
CrCDA [25]	×	86.2	44.9	79.5	8.3	0.7	27.8	9.4	11.8	78.6	86.5	57.2	26.1	76.8	39.9	21.5	32.1	42.9	50.0
ProDA [87]	×	87.1	44.0	83.2	26.9	0.7	42.0	45.8	34.2	86.7	81.3	68.4	22.1	87.7	50.0	31.4	38.6	51.9	58.5
CPSL [38]	×	87.3	44.4	83.8	25.0	0.4	42.9	47.5	32.4	86.5	83.3	69.6	29.1	89.4	52.1	42.6	54.1	54.4	61.7
Source Only	-	45.2	19.6	72.0	6.7	0.1	24.3	5.5	7.8	74.4	81.9	57.3	17.3	39.0	19.5	7.0	6.2	31.3	36.2
UR [66]	✓	59.3	24.6	77.0	14.0	1.8	31.5	18.3	32.0	83.1	80.4	46.3	17.8	76.7	17.0	18.5	34.6	39.6	45.0
SFDA [45]	✓	67.8	31.9	77.1	8.3	1.1	35.9	21.2	26.7	79.8	79.4	58.8	27.3	80.4	25.3	19.5	37.4	42.4	48.7
HCL [24]	✓	80.9	34.9	76.7	6.6	0.2	36.1	20.1	28.2	79.1	83.1	55.6	25.6	78.8	32.7	24.1	32.7	43.5	50.2
C-SFDA (Ours)	✓	87.0	39.0	79.5	12.2	1.8	32.2	20.4	24.3	79.5	82.2	51.5	24.5	78.7	31.5	21.3	47.9	**44.6**	**51.3**
TENT [73] (Online)	✓	88.1	44.9	74.4	4.3	0.1	21.8	2.0	7.8	77.3	82.8	52.9	9.7	77.6	7.5	0.2	15.8	35.5	41.6
AUGCO [56] (Online)	✓	74.8	32.1	79.2	5.0	0.1	29.4	3.0	11.1	78.7	83.1	57.5	26.4	74.3	20.5	12.1	39.3	39.2	45.5
C-SFDA (Online)	✓	85.9	38.1	79.2	11.9	1.1	32.0	17.1	22.9	79.7	89.4	46.6	22.0	78.4	29.6	17.4	46.0	**43.0**	**49.5**

Table 7. Evaluation on **Cityscapes→Dark-Zurich**. We report mean IoU (mIoU) over 19 common categories between theses datasets.

Method	Source	TTBN [52]	TENT [73]	AUGCO [56]	C-SFDA (Ours)
mIoU	28.8	28.0	26.6	32.4	**33.2**

ous SOTA for DomainNet. Although AdaCon [4] uses a large memory queue to refine the pseudo-labels, it still suffers from early training time memorization. Whereas utilizing a *label-selection technique* for curriculum training, C-SFDA eliminates the requirement of a *label-refinement technique* and still outperforms AdaCon [4]. We also consider several general UDA techniques considering continued source data access. We encouragingly find that the proposed C-SFDA performs better than most of these methods across datasets, even without source data access.

Evaluation on Semantic Segmentation Task: Table 5 shows the performance on GTA5→Cityscapes. For this adaptation, we resize the target scenes to 1024×512 and use DeepLabV2 for training. We choose this common architecture to be consistent with other recent works. The proposed method outperforms the state-of-the-art SFDA method HCL [24] with 19-way averaged mIoU of 48.3%. Note that, some classes in Cityscapes have very low initial pixel-level accuracy, *e.g*. *Train* category, and it is challenging to obtain satisfactory performance even with selective pseudo-labeling. It requires mentioning that HCL [24] employs historical contrastive loss enforcing additional memory overhead; an undesirable property in most adaptation scenarios. On the other hand, our method utilizes a simple pixel-level prediction reliability measure which is highly computationally efficient and leads to the best mIoU. As regular UDA techniques have the advantage of source data access and most employ highly sophisticated techniques specific to semantic segmentation, they usually perform better than SFDA techniques. However, we find C-SFDA performs comparably to several UDA techniques, *e.g*. CrCDA [25]. We also evaluate SYNTHIA→Cityscapes (Table 6) benchmark, where we use the same DeepLabV2 architecture and adaptation strategy. Compared to the baselines, C-SFDA performs significantly better, with a mIoU improvement of 1.1% over the previous SOTA.

5.4.1 Compatibility to Online Adaptation

As C-SFDA employs batch-wise selection instead of using the whole dataset, it is readily applicable to online fully test-time domain adaptation [56, 73]. In contrast to regular SFDA experiments, we only train the model for 1 epoch following prior works [56, 73], without any change to our training settings. In both image classification and semantic segmentation experiments, C-SFDA performs better than previous state-of-the-art methods (Table 3 -7). For instance, we achieve a 3.5% accuracy gain in VisDA image classification (Table 3). In segmentation, we improve the mIoU by 0.8% in Cityscapes→Dark Zurich (Table 7) and 4% in SYNTHIA→Cityscapes (Table 6). These gains can be attributed to the adoption of high-quality pseudo-labels right from the beginning of the training. Learning in this manner

Table 8. Ablation study with different components of our proposed method. Contrastive learning along with uncertainty plays a vital role in achieving SOTA average accuracy (%) in image classification benchmarks.

Selection Strategy			**Label Bal.**	**Loss**			**Accuracy (%)**			
Conf.	Unc.	DoC	λ_k	$\mathcal{L}_{ce}^R$	$\mathcal{L}_\mathcal{P}$	$\mathcal{L}_\mathcal{C}$	Office-31	Office-Home	VisDA	DomainNet
Self-training ($\mathcal{L}_{ce}$) with all pseudo-labels							81.1	62.3	57.2	52.6
✓			✓	✓	✓	✓	87.6	69.2	85.2	65.5
✓	✓		✓	✓	✓	✓	89.9	71.8	87.4	68.7
✓	✓	✓	×	✓	✓	✓	90.1	73.3	86.5	68.3
✓	✓	✓	✓	✓	×	×	88.7	71.6	85.9	67.3
✓	✓	✓	✓	✓	✓	×	88.9	72.3	86.4	67.9
✓	✓	✓	✓	✓	✓	✓	**90.5**	**73.5**	**87.8**	**69.0**

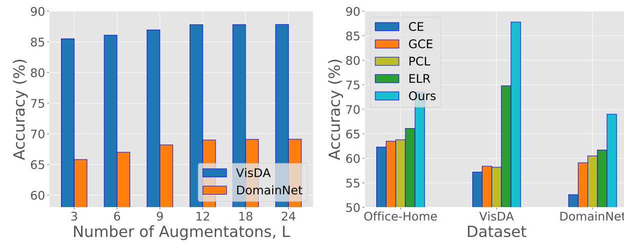
gives us a head start in producing reliable pseudo-labels for subsequent self-training.

5.4.2 Ablation Studies

Does Traditional Label Noise Learning Help? Since our method deals with noisy labels, we explore the literature on label noise learning (LNL) and apply them in the SFDA setting. We consider 3 widely used techniques: GCE [89], PCL [88], ELR [44] along with regular cross-entropy loss with all pseudo labels and compare their performance on 3 datasets. Fig. 3b shows that traditional LNL techniques may not be suitable for SFDA as they severely underperform compared to C-SFDA. One possible reason could be that SFDA contains unbounded label noise due to an unknown domain shift. In the case of *unbounded label noise*, noise rates are unknown and can be very high; which is in contrast to the general belief of *bounded label noise* where noise rate and type are known priors. In such a scenario, traditional LNL methods struggle to curate label noise. Our method can convincingly perform under this scenario without any prior knowledge of noise type, rate, *etc*.

Effect of Different Selection Criteria: Table 8 shows the ablation study with different elements of our proposed method. We first show the performance without curriculum learning where we use all pseudo-labels for fully supervised training with CE loss. Since implementing the curriculum learning requires pseudo-label selection, we analyze the impact of confidence, uncertainty, and DoC here first. It can be observed that each of these metrics can have a significant impact on the overall performance, especially *prediction uncertainty*. As the choice of augmentations plays a vital role in measuring uncertainty, we conduct a detailed study on different augmentation policies (details are in supplementary). In Fig. 3a, we also analyze the impact of augmentations number L on overall classification performance.

Effect of Different Loss Functions: We also analyze the impact of different loss functions in Table 8. It can be seen that using only CE loss produces quite satisfactory performance. This indicates learning only from reliable samples is good enough for SFDA. Interestingly, applying propagation loss without contrastive loss may experience performance degradation as we are still using label information. This



(a) Effect of Augmentation Number (b) LNL Techniques in SFDA

Figure 3. **(a)** Ablation with different L shows that we need to consider a sufficient number of augmentations for measuring prediction uncertainty as it plays a crucial role in obtaining SOTA average accuracy. **(b)** Performance of SOTA noisy label learning methods in SFDA. Due to the presence of unbounded label noise (*i.e.* high noise rates and unknown noise types), traditional LNL struggles to perform well in SFDA settings.

underlines the importance of CL in preventing noise memorization. We also show the impact of label balancing here. Note that, label balancing is only being considered for CE loss. We also study the effect of fixed and adaptive curriculum learning (ours) in supplementary.

6. Limitations

We proposed to utilize the reliability of generated labels in selective pseudo-labeling. Depending on the domain shift, the initial reliability often varies and can be severely misleading if the domain shift is too large. Such scenarios may require additional measures such as label noise robust self-supervised learning or strongly augmented source domain training. However, due to the hyper-parameter independence of our selection strategy, the proposed method should be over-restrictive in selecting labels whenever such an extreme situation appears. Furthermore, part of the reason we conduct extensive evaluations of our method is to show its applicability in a wide variety of domain shifts.

7. Conclusion

In this work, we introduced a novel source-free domain adaptation technique exploiting the phenomenon of early memorization of noisy pseudo-labels. Due to the inevitable presence of this phenomenon, we employ a curriculum learning-aided selective self-training strategy that prioritizes learning from highly reliable pseudo-labels and propagating label information to less reliable ones. This leads us to a hyper-parameter independent label selection technique that replaces the need for a label refinement technique. In addition, we utilize contrastive loss-based representation learning that helps generate consistent feature representation and better guides the overall adaptation. Due to the memory-efficient property of our method, C-SFDA can easily be adopted for online test-time domain adaptation scenarios. Extensive evaluations show that our method achieves superior performance on a wide range of image classification and semantic segmentation benchmarks.

References

[1] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 3

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 4

[3] Chao Chen, Zhihang Fu, Zhihong Chen, Sheng Jin, Zhaowei Cheng, Xinyu Jin, and Xian-Sheng Hua. Homm: Higher-order moment matching for unsupervised domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3422–3429, 2020. 2

[4] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022. 1, 2, 4, 5, 6, 7

[5] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553. PMLR, 2020. 2

[6] 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

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 4

[8] 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

[9] Gabriela Csurka. A comprehensive survey on domain adaptation for visual applications. *Domain adaptation in computer vision applications*, pages 1–35, 2017. 2

[10] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019. 3

[11] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2018. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5, 6

[13] Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7212–7222, 2022. 2, 5, 6

[14] Jiahua Dong, Zhen Fang, Anjin Liu, Gan Sun, and Tongliang Liu. Confident anchor-induced multi-source free domain adaptation. *Advances in Neural Information Processing Systems*, 34:2848–2860, 2021. 3

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[16] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, number 6, 2018. 6

[17] Xiang Gu, Jian Sun, and Zongben Xu. Spherical space domain adaptation with robust pseudo-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 6

[18] Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1134–1144, 2021. 3

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 5

[21] 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*, pages 1989–1998. Pmlr, 2018. 2, 5

[22] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Unsupervised domain adaptation with hierarchical gradient synchronization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4043–4052, 2020. 5

[23] Ran He Hu and Jiashi Feng. A balanced and uncertainty-aware approach for partial domain adaptation. 3

[24] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34:3635–3649, 2021. 6, 7

[25] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In *European conference on computer vision*, pages 705–722. Springer, 2020. 7

[26] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021. 4

[27] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 5

[28] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *European Conference on Computer Vision*, pages 464–480. Springer, 2020. 5, 6

[29] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019. 1, 5

[30] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 4

[31] Youngeun Kim, Donghyeon Cho, Kyeongtak Han, Priyadarshini Panda, and Sungeun Hong. Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence*, 2021. 5, 6

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 3

[33] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4544–4553, 2020. 1, 2

[34] Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, R Venkatesh Babu, et al. Towards inheritable models for open-set domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12376–12385, 2020. 1

[35] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10285–10295, 2019. 6

[36] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 2

[37] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2020. 2, 5, 6

[38] Ruihuang Li, Shuai Li, Chenhang He, Yabin Zhang, Xu Jia, and Lei Zhang. Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11593–11603, 2022. 7

[39] Shuang Li, Chi Liu, Qiuxia Lin, Binhui Xie, Zhengming Ding, Gao Huang, and Jian Tang. Domain conditioned adaptation network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11386–11393, 2020. 2

[40] Shuang Li, Mixue Xie, Kaixiong Gong, Chi Harold Liu, Yulin Wang, and Wei Li. Transferable semantic augmentation for domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11516–11525, 2021. 6

[41] Justin Liang, Namdar Homayounfar, Wei-Chiu Ma, Yuwen Xiong, Rui Hu, and Raquel Urtasun. Polytransform: Deep polygon transformer for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9131–9140, 2020. 1

[42] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020. 1, 2, 3, 5, 6

[43] Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, and Carlos Fernandez-Granda. Adaptive early-learning correction for segmentation from noisy annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2606–2616, 2022. 1

[44] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020. 8

[45] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021. 2, 6, 7

[46] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018. 2

[47] Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9111–9120, 2020. 6

[48] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *European conference on computer vision*, pages 415–430. Springer, 2020. 2

[49] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 6

[50] 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

[51] Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. Fixbi: Bridging domain spaces for unsupervised

domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1094–1103, June 2021. 6

[52] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020. 6, 7

[53] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020. 7

[54] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 1, 5

[55] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 5

[56] Viraj Uday Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Augmentation consistency-guided self-training for source-free domain adaptive semantic segmentation. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*. 2, 6, 7

[57] Fengchun Qiao and Xi Peng. Uncertainty-guided model generalization to unseen domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6790–6800, 2021. 3

[58] 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

[59] 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

[60] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 5

[61] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019. 5

[62] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7374–7383, 2019. 5, 6

[63] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016. 5

[64] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[65] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 3

[66] Prabhu Teja Sivaprasad and Francois Fleuret. Uncertainty reduction for model adaptation in semantic segmentation. In *2021 Ieee/Cvf Conference On Computer Vision And Pattern Recognition, Cvpr 2021*, number CONF, pages 9608–9618. IEEE, 2021. 6, 7

[67] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 2

[68] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8725–8735, 2020. 5, 6

[69] Jiayi Tian, Jing Zhang, Wen Li, and Dong Xu. Vdm-da: Virtual domain modeling for source data-free domain adaptation. *arXiv preprint arXiv:2103.14357*, 2021. 2

[70] 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. 2

[71] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 5

[72] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 2, 5

[73] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 1, 4, 6, 7

[74] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 2

[75] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 1

[76] Yuxi Wang, Junran Peng, and ZhaoXiang Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9092–9101, 2021. 3

[77] Haifeng Xia and Zhengming Ding. Structure preserving generative cross-domain learning. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4364–4373, 2020. 1

[78] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9010–9019, 2021. 2, 5, 6

[79] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 2

[80] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1426–1435, 2019. 6

[81] Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, and Jindong Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4394–4403, 2020. 2, 6

[82] Yixing Xu, Kai Han, Chang Xu, Yehui Tang, Chunjing Xu, and Yunhe Wang. Learning frequency domain approximation for binary neural networks. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[83] Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 34:29393–29405, 2021. 1, 2

[84] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Unsupervised domain adaptation without source data by casting a bait. *arXiv preprint arXiv:2010.12427*, 1(2):5, 2020. 1

[85] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8978–8987, 2021. 1, 2, 6

[86] Fei Yu, Mo Zhang, Hexin Dong, Sheng Hu, Bin Dong, and Li Zhang. Dast: Unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10754–10762, 2021. 2

[87] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021. 7

[88] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. *arXiv preprint arXiv:2103.07756*, 2021. 8

[89] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018. 8

[90] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16, 2003. 4

[91] Tianyi Zhou and Jeff Bilmes. Minimax curriculum learning: Machine teaching with desirable difficulties and scheduled diversity. In *International Conference on Learning Representations*, 2018. 4

[92] 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)*, pages 289–305, 2018. 2

[93] 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. 5