SENTRY: Selective Entropy Optimization via Committee Consistency for Unsupervised Domain Adaptation

Viraj Prabhu **Shivam Khare** **Deeksha Kartik** **Judy Hoffman**
Georgia Institute of Technology
`{virajp,skhare31,dkartik3,judy}@gatech.edu`

Abstract

Many existing approaches for unsupervised domain adaptation (UDA) focus on adapting under only data distribution shift and offer limited success under additional cross-domain label distribution shift. Recent work based on self-training using target pseudolabels has shown promise, but on challenging shifts pseudolabels may be highly unreliable and using them for self-training may lead to error accumulation and domain misalignment. We propose Selective Entropy Optimization via Committee Consistency (`SENTRY`), a UDA algorithm that judges the reliability of a target instance based on its predictive consistency under a committee of random image transformations. Our algorithm then selectively minimizes predictive entropy to increase confidence on highly consistent target instances, while maximizing predictive entropy to reduce confidence on highly inconsistent ones. In combination with pseudolabel-based approximate target class balancing, our approach leads to significant improvements over the state-of-the-art on 27/31 domain shifts from standard UDA benchmarks as well as benchmarks designed to stress-test adaptation under label distribution shift. Our code is available at https://github.com/virajprabhu/SENTRY.

1. Introduction

Unsupervised domain adaptation (UDA) learns to transfer a predictive model from a labeled source domain to an unlabeled target domain. The particular instantiation of learning under covariate shift has been extensively studied within the computer vision community [13, 18, 25, 34, 44, 45]. However, many modern UDA methods, such as distribution matching based techniques, implicitly assume that the task label distribution does not change across domains, i.e $P_{\mathcal{S}}(y) = P_{\mathcal{T}}(y)$. When such an assumption is violated, distribution matching is not expected to succeed [22, 49].

In many real-world adaptation scenarios, one may encounter data distribution (*i.e.* covariate) shift across domains together with label distribution shift (LDS). For instance, a source dataset can be curated to have a balanced label distribution while a naturally arising target dataset may follow a power law label distribution, as some categories naturally occur more often than others (*e.g.* DomainNet [31], LVIS [16], and MSCOCO [23]). In order to make domain adaptation broadly applicable, it is critical to develop UDA algorithms that can operate under joint data and label distribution shift.

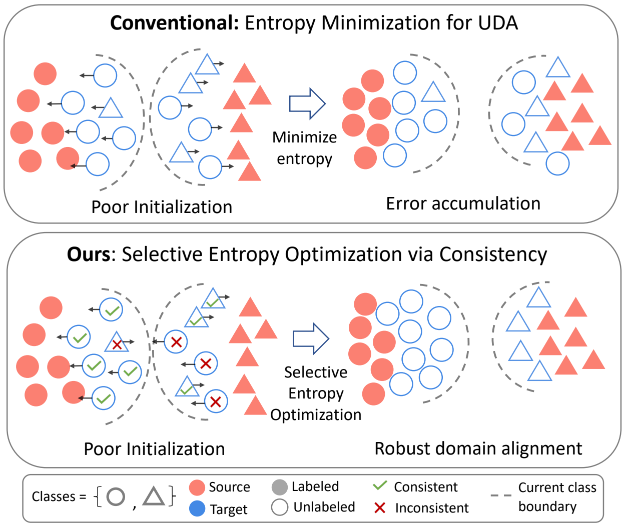

Figure 1: **Top**: Conventional entropy-minimization based approaches for unsupervised domain adaptation (UDA) operate by increasing model confidence on unlabeled target instances. Under strong distribution shifts, some instances may initially be misaligned and entropy minimization can lead to error accumulation. **Bottom**: We propose Selective Entropy Optimization via Committee Consistency (`SENTRY`), a UDA algorithm that i) identifies reliable target instances based on their predictive consistency under a set of random image transformations, and ii) selectively optimizes model entropy on these instances to induce domain alignment.

Recent works have attempted to address the problem of joint data and label distribution shift [22, 43], but these approaches can be unstable as they rely on self-training using often noisy pseudo-labels or conditional entropy minimization [22] over potentially miscalibrated predictions [15, 39]. Thus, when learning with unconstrained self-training, early mistakes can result in error accumulation [6] and significant domain misalignment (see Figure 1, top).

To address the problem of error accumulation arising from unconstrained self-training, we propose Selective Entropy Optimization via Committee Consistency (SENTRY), a novel *selective* self-training algorithm for UDA. First, rather than using model confidence which can be miscalibrated under a domain shift [39], we identify reliable target instances for self-training based on their *predictive consistency* under a committee of random, label-preserving image transformations. Such consistency checks have been found to be a reliable way to detect model errors [2]. Having identified reliable and unreliable target instances, we then perform selective entropy optimization: we consider a highly consistent target instance as likely correctly aligned, and increase model confidence by minimizing predictive entropy for such an instance. Similarly, we consider an instance with high predictive inconsistency over transformations as likely misaligned, and reduce model confidence by *maximizing* predictive entropy. See Figure 1 (bottom).

Contributions. We propose SENTRY, an algorithm for unsupervised adaptation under simultaneous data and label distribution shift. We make the following contributions:

1. A novel selection criterion that identifies reliable target instances for self-training based on predictive consistency over a committee of random, label-preserving image transformations.
2. A selective entropy optimization objective that minimizes predictive entropy (increasing confidence) on highly consistent target instances, and maximizes it (reducing confidence) on highly inconsistent ones.
3. We propose using class-balanced sampling on the source (using labels) and target (using pseudolabels), and find it to complement adaptation under LDS.
4. SENTRY sets a new state-of-the-art on 27/31 domain shifts belonging to both standard and LDS versions of several DA benchmarks for classification, including DomainNet [31], OfficeHome [46], and VisDA [32].

2. Related Work

Unsupervised Domain Adaptation (UDA). The task of transferring models from a labeled source to an unlabeled target domain has seen considerable progress [13, 18, 34, 45]. Many approaches align feature spaces via directly minimizing domain discrepancy statistics [20, 25, 45]. Recently, distribution matching (DM) via domain-adversarial learning has become a prominent UDA paradigm [13, 26, 36, 44, 53]. Such DM-based methods however achieve limited success in the presence of additional label distribution shift (LDS).

Some prior work has studied the problem of UDA under LDS, proposing class-weighted domain discrepancy measures [47, 51], generative approaches for pair-wise feature matching [42], or asymmetrically-relaxed distribution alignment [52]. Some prior work in UDA under LDS additionally assumes that the conditional input distribution does not change across domains i.e. $p_{\mathcal{S}}(y) \neq p_{\mathcal{T}}(y), p_{\mathcal{S}}(x|y) = p_{\mathcal{T}}(x|y)$ (referred to as “label shift” [1, 24, 41]). We tackle the problem of UDA under simultaneous covariate and label distribution shift, without making additional assumptions.

Self-training for UDA. Recently, training on model predictions or *self-training* has proved to be a promising approach for UDA under LDS [22, 43]. This typically involves supervised training on confidently predicted target pseudolabels [43], confidence regularization [54], or conditional entropy minimization [14] on target instances [22]. However, unconstrained self-training can lead to error accumulation. To address this, we propose a *selective* self-training strategy that first identifies reliable instances for self-training and selectively optimizes model entropy on those.

Predictive Consistency. Predictive consistency under augmentations has been found to be useful in several capacities – as a regularizer in supervised learning [10], self-supervised representation learning [7], semi-supervised learning [3, 38, 40, 50], and UDA [22]. Bahat *et al*. [2] find consistency under image transformations to be a reliable indicator of model errors. Unlike prior work which optimizes for invariance across augmentations, we use predictive consistency under a committee of random image transforms to *detect* reliable instances for alignment, and selectively optimize model entropy on such instances.

3. Approach

We address the problem of unsupervised domain adaptation (UDA) of a model trained on a labeled source domain to an unlabeled target domain. In addition to covariate shift across domains, we focus on the practical scenario of additional cross-domain label distribution shift (LDS), and present a selective self-training algorithm for UDA that leads to reliable domain alignment in such a setting.

3.1. Notation

Let $\mathcal{X}$ and $\mathcal{Y}$ denote input and ouput spaces, with the goal being to learn a CNN mapping $h : \mathcal{X} \rightarrow \mathcal{Y}$ parameterized by Θ. In unsupervised DA we are given access to labeled source instances $(\mathbf{x}_{\mathcal{S}}, y_{\mathcal{S}}) \sim \mathcal{P}_{\mathcal{S}}(\mathcal{X}, \mathcal{Y})$, and unlabeled target instances $\mathbf{x}_{\mathcal{T}} \sim \mathcal{P}_{\mathcal{T}}(\mathcal{X})$, where $\mathcal{S}$ and $\mathcal{T}$ correspond to source and target domains. We consider DA in the context of C-way image classification: the inputs $\mathbf{x}$ are images, and labels y are categorical variables $y \in \{1, 2, .., C\}$. For an instance $\mathbf{x}$, let $p_{\Theta}(y|\mathbf{x})$ denote the final probabilistic output from the model. For each target instance $\mathbf{x}_{\mathcal{T}} \sim \mathcal{P}_{\mathcal{T}}(\mathcal{X})$, we estimate a pseudolabel $\hat{y} = \text{argmax}\, p_{\Theta}(y|\mathbf{x}_{\mathcal{T}})$.

3.2. Preliminaries: UDA via entropy minimization

Unsupervised domain adaptation typically follows a two-stage training pipeline: source training, followed by target adaptation. In the first stage, a model is trained on the labeled source domain in a supervised fashion, minimizing a cross-

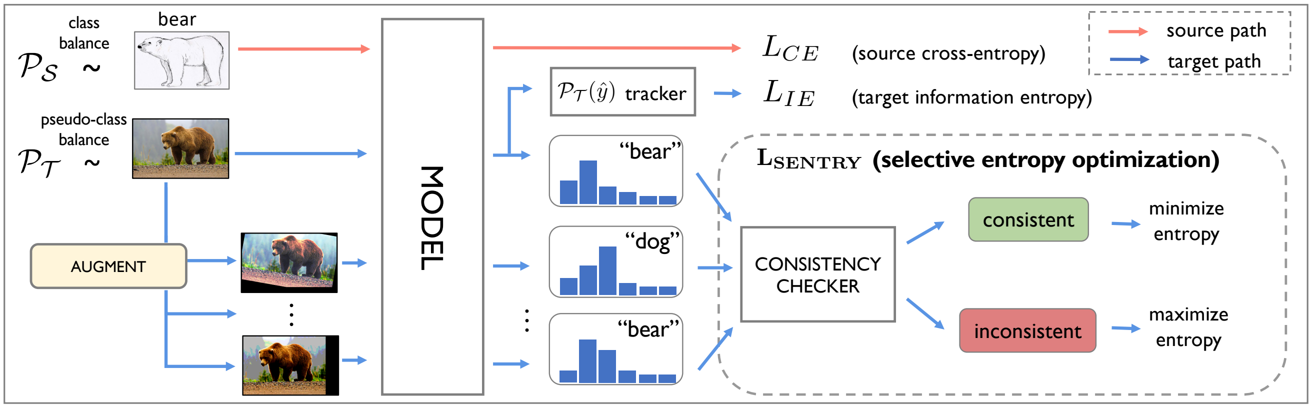

Figure 2: We propose Selective Entropy Optimization via Committee Consistency (SENTRY) for unsupervised DA. For each target instance, we generate a committee of random, label-preserving image transformations. A consistency checker then computes the consistency between model predictions for the original and augmented versions. The algorithm then minimizes predictive entropy (increasing model confidence) on highly consistent target instances, and maximizes predictive entropy (reducing model confidence) on highly inconsistent ones.

entropy loss with respect to ground truth labels.

$$\mathcal{L}_{CE} = \mathbb{E}_{(\mathbf{x}_\mathcal{S}, y_\mathcal{S})\sim\mathcal{P}_\mathcal{S}}[\mathcal{L}_{CE}(h(\mathbf{x}_\mathcal{S}), y_\mathcal{S})] \tag{1}$$

In the second stage, the trained source model is adapted to the target with the use of unlabeled target and labeled source data. Recently, self-training via conditional entropy minimization (CEM) [14] has been shown to lead to strong performance for domain adaptation [35]. This approach optimizes model parameters to minimize conditional entropy on unlabeled target data $\mathcal{H}_\Theta(y|\mathbf{x})$. The entropy minimization objective $\mathcal{L}_{ENT}$ is given by:

$$\begin{aligned}\mathcal{L}_{ENT} &= \mathbb{E}_{\mathbf{x}_\mathcal{T}\sim\mathcal{P}_\mathcal{T}}\left[\mathcal{H}_\Theta(y|\mathbf{x}_\mathcal{T})\right] \\ &= \mathbb{E}_{\mathbf{x}_\mathcal{T}\sim\mathcal{P}_\mathcal{T}}\left[\sum_{c=1}^{C} -p_\Theta(y=c|\mathbf{x}_\mathcal{T})\log p_\Theta(y=c|\mathbf{x}_\mathcal{T})\right]\end{aligned} \tag{2}$$

However, in many real-world scenarios, in addition to covariate shift, *label distributions* across domains might also shift. Further, there might also be significant label imbalance within the target domain. In such cases, naive CEM has been found to potentially encourage trivial solutions of only predicting the majority class [22]. Li *et al.* [22] regularize CEM with an "information-entropy" objective $\mathcal{L}_{IE}$ that encourages the model to make diverse predictions over unlabeled target instances. This is achieved by computing a distribution over classes predicted by the model for the last-Q instances, denoted by $q(\hat{y})$, and updating parameters to maximize entropy over these predictions. This method is shown to help with domain alignment in the presence of label-distribution shift [22] [1]. $\mathcal{L}_{IE}$ is defined as:

$$\mathcal{L}_{IE} = \mathbb{E}_{\mathbf{x}_\mathcal{T}\sim\mathcal{P}_\mathcal{T}}\left[\sum_{c=1}^{C} p_\Theta(y=c|\mathbf{x}_\mathcal{T})\log q(\hat{y}=c)\right] \tag{3}$$

CEM and error accumulation. While conditional entropy minimization has been a part of many successful approaches for semi-supervised learning [3, 14], few-shot learning [12], and more recently, UDA [22, 35], it suffers from a key challenge in the case of domain adaptation. Intuitively, conditional entropy minimization encourages the model to make confident predictions on unlabeled target data. This makes its success highly dependent on its initialization. Under a good initialization, categories may be reasonably aligned across source and target domains after source training, and such self-training works well. However, under strong domain shifts, several categories may initially be misaligned across domains, often systematically so, and entropy minimization will only lead to *reinforcing* such errors.

3.3. SENTRY: Selective Entropy Optimization via Committee Consistency

Predictive consistency-based selection. To address the problem of error accumulation under CEM, we propose *selective* optimization on well-aligned instances. The question then becomes: how can we identify reliable instances? One possibility is to use top-1 softmax confidence (or alternatively, predictive entropy), and only self-train on highly confident instances, as done in prior work [43]. However, under a distribution shift, such confidence measures tend to be miscalibrated and are often unreliable [39]. Instead, we propose using *predictive consistency* under a committee of label-preserving image transformations as a more robust measure for instance selection.

For a target instance $\mathbf{x}_\mathcal{T} \sim \mathcal{P}_\mathcal{T}$, we generate a committee of k transformed versions $\{a_1(\mathbf{x}_\mathcal{T}), a_2(\mathbf{x}_\mathcal{T}), ..., a_k(\mathbf{x}_\mathcal{T})\}$. We make predictions for each of these k transformed versions, and measure *consistency* between the model's prediction for the original image and for each of its k augmented versions. In practice, we use a simple majority voting scheme: if the model's prediction for a majority of

[1] The objective is referred to as "mutual information maximization" in Li *et al.* [22]

augmented versions matches its prediction on the original image, we consider the instance as "consistent". Similarly, if the prediction for a majority of augmented versions does not match the original prediction, we mark it as "inconsistent".

Selective Entropy Optimization. Having identified consistent and inconsistent instances, we perform Selective Entropy Optimization (SENTRY). First, for an instance marked as consistent, we *increase model confidence* by minimizing predictive entropy [14] with respect to one of its consistent augmented versions.

As described previously, some target instances may be misaligned under a domain shift. Entropy minimization on such instances would increase model confidence, *reinforcing such errors*. Instead, having identified such an instance via predictive inconsistency, we *reduce model confidence* by *maximizing* predictive entropy [33] with respect to one of its inconsistent augmented versions. While the former encourages confident predictions on highly consistent instances, the latter reduces model confidence on highly inconsistent and likely misaligned instances. In Sec. 4.6, we provide further intuition into the behavior of entropy maximization by illustrating its similarity to a binary cross-entropy loss with respect to the ground truth label for an incorrectly classified example in the binary classification case.

Without loss of generality, we minimize/maximize entropy with respect to the last consistent/inconsistent transformed version in our experiments. Our selective entropy optimization objective $\mathcal{L}_{\texttt{SENTRY}}$ is given by:

$$\mathcal{L}_{\texttt{SENTRY}}(\mathbf{x}_{\mathcal{T}}) = \begin{cases} +\mathcal{H}_{\Theta}(y|a_i(\mathbf{x}_{\mathcal{T}})) & \text{if consistent} \\ -\mathcal{H}_{\Theta}(y|a_j(\mathbf{x}_{\mathcal{T}})) & \text{if inconsistent} \end{cases} \quad (4)$$

Here i and j denote the index of the last consistent and inconsistent transformed version, respectively.

Such an approach may raise two concerns: First, that entropy minimization only on consistent instances might lead to the exclusion of a large percentage of target instances. Second, that indefinite entropy maximization on inconsistent target instances might prove detrimental to learning. Both of these concerns are addressed via the augmentation invariance regularizer built into our objective, which leads to an adaptive selection strategy that we now discuss.

Adaptive selection via augmentation invariance regularization. For instances marked as consistent, our approach minimizes entropy with respect to its last consistent *augmented* version rather than with respect to the original image itself. This yields two benefits: First, this builds data augmentation into the entropy minimization objective, which helps reduce overfitting. Second, it encourages invariance to the same set of augmentations that is used for selecting instances. We find that this makes our selection strategy *adaptive*, wherein an increasing percentage of target instances are selected for entropy minimization over the course of training, and consequently a decreasing percentage of target instances are selected for entropy maximization.

Algorithm 1 SENTRY Optimization

```
Input: $\mathcal{X}_{\mathcal{S}}$, $\mathcal{Y}_{\mathcal{S}}$, $\mathcal{X}_{\mathcal{T}}$, $Q$, $\Theta$
for all $x_T^{(i)} \in \mathcal{X}_{\mathcal{T}}$ do                          ▷ Init target pseudo-labels
    $\hat{\mathcal{Y}}_T^{(i)} \leftarrow \operatorname{argmax} p_{\Theta}(y|x_T^{(i)})$
SrcLoader ← ClassBalancedSampler($\mathcal{X}_{\mathcal{S}}, \mathcal{Y}_{\mathcal{S}}$)
TgtLoader ← ClassBalancedSampler($\mathcal{X}_{\mathcal{T}}, \hat{\mathcal{Y}}_{\mathcal{T}}$)
$q \leftarrow$ Queue(size=Q)
for epoch ← 1 to MAX_EPOCH do
    for $x_{\mathcal{S}}, y_{\mathcal{S}}$ in SrcLoader and $x_{\mathcal{T}}$ in TgtLoader do
        $\hat{y}_{\mathcal{T}} \leftarrow \operatorname{argmax} p_{\Theta}(y|x_{\mathcal{T}})$          ▷ Clean prediction
        $\{a_1(x_{\mathcal{T}}), \ldots, a_k(x_{\mathcal{T}})\} \leftarrow$ RandAugment($x_{\mathcal{T}}$)
        C ← $\{a_i(x_{\mathcal{T}}) | \hat{y}_{\mathcal{T}} = \operatorname{argmax} p_{\Theta}(y|a_i(x_{\mathcal{T}}))\}_{i=1}^{k}$
        IC ← $\{a_i(x_{\mathcal{T}}) | \hat{y}_{\mathcal{T}} \neq \operatorname{argmax} p_{\Theta}(y|a_i(x_{\mathcal{T}}))\}_{i=1}^{k}$
        if len(C) > len(IC) then                           ▷ Consistent
            $\mathcal{L}_{\texttt{SENTRY}} = \mathcal{H}_{\Theta}(y|\text{C.last()})$
        else                                               ▷ Inconsistent
            $\mathcal{L}_{\texttt{SENTRY}} = -\mathcal{H}_{\Theta}(y|\text{IC.last()})$
        Update($\hat{Y}_{\mathcal{T}}, \hat{y}_{\mathcal{T}}$)
        $q$.enqueue($\hat{y}_{\mathcal{T}}$)                        ▷ Update pseudo-label queue
    Minimize $\mathcal{L}_{\texttt{SENTRY}} + \mathcal{L}_{IE}(q) + \mathcal{L}_{CE}(x_{\mathcal{S}}, y_{\mathcal{S}})$
    TgtLoader ← ClassBalancedSampler($\mathcal{X}_{\mathcal{T}}, \hat{\mathcal{Y}}_{\mathcal{T}}$)
```

3.4. Overcoming LDS via pseudo class balancing

Under LDS, methods often have to adapt in the presence of severe label imbalance. While label imbalance on the source domain often leads to poor performance on tail classes [11, 48], adapting to an imbalanced target often results in poor performance on head classes [22, 49]. To overcome this, we employ a simple class-balanced sampling strategy. On the source domain, we perform class-balanced sampling using ground truth labels. On the target domain, we approximate the label distribution via *pseudolabels*, and perform approximate class-balanced sampling [55].

Such balancing also complements the target information-entropy loss L_{IE} [22] (Eq. 3). To recap, L_{IE} encourages a uniform distribution over predictions. Under severe label imbalance, it is possible to sample highly label-imbalanced batches (with most instances belonging to head classes) and so encouraging a uniform distribution over predictions can adversely affect learning. However, our class-balanced sampling strategy reduces the probability of such a scenario, and we find that it consistently improves performance.

Algorithm 1 details our full approach. The complete objective we optimize is given by:

$$\begin{aligned} \underset{\Theta}{\operatorname{argmin}} \quad & \mathop{\mathbb{E}}_{(\mathbf{x}_{\mathcal{S}}, y_{\mathcal{S}}) \overset{\text{bal}}{\sim} \mathcal{P}_{\mathcal{S}}} \mathcal{L}_{CE} \quad + \\ & \mathop{\mathbb{E}}_{\mathbf{x}_{\mathcal{T}} \overset{\text{pbal}}{\sim} \mathcal{P}_{\mathcal{T}}} \lambda_{IE} \mathcal{L}_{IE} + \lambda_{\texttt{SENTRY}} \mathcal{L}_{\texttt{SENTRY}} \end{aligned} \quad (5)$$

where the λ's denote loss weights, and $\overset{\text{bal}}{\sim}$ and $\overset{\text{pbal}}{\sim}$ denote balanced and pseudo class-balanced sampling.

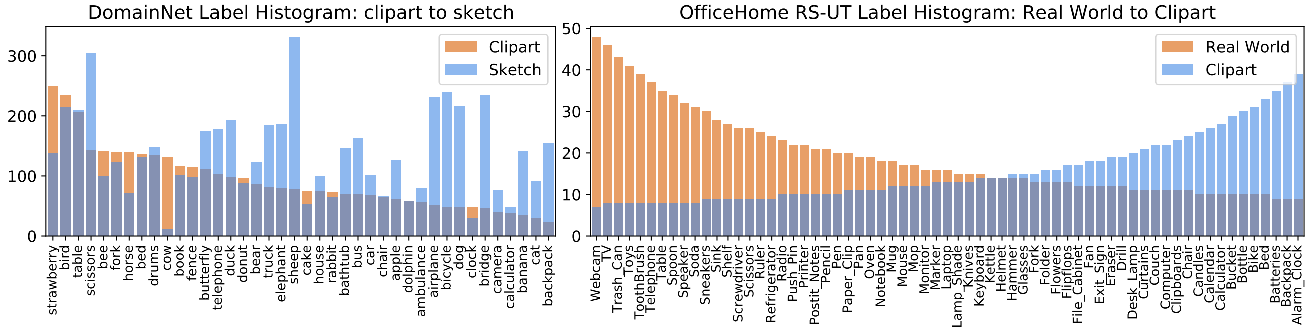

Figure 3: **Left**: Natural label distribution shift (LDS) on the Clipart→Sketch shift from DomainNet. **Right:** Manually generated LDS on the Real World→Clipart shift from OfficeHome RS-UT [43].

4. Experiments

We first describe our experimental setup: datasets and metrics (Sec. 4.1), implementation details (Sec. 4.2), and baselines (Sec. 4.3). We then present our results (Sec. 4.4), ablation studies (Sec. 4.5), and analyze our approach (Sec. 4.6).

4.1. Datasets and Metrics

We report results on a mix of standard UDA benchmarks and specialized benchmarks designed to stress-test UDA methods under label distribution shift.

DomainNet. DomainNet [31] is a large UDA benchmark for image classification, containing 0.6 million images belonging to 6 domains spanning 345 categories. Due to labeling noise prevalent in its full version, we instead use the subset proposed in Tan *et al*. [43], which uses 40-commonly seen classes from 4 domains: Real (**R**), Clipart (**C**), Painting (**P**), and Sketch (**S**). As seen in Fig. 3 (left), there exists a natural label distribution shift across domains, which makes it suitable for testing our method without manual subsampling.

OfficeHome. OfficeHome [46] is an image classification-based benchmark containing 65 categories of objects found in office and home environments, spanning 4 domains: Real-world (**Rw**), Clipart (**Cl**), Product (**Pr**), and Art (**Ar**). We report performance on two versions: i) standard: the original dataset proposed in Venkateswara *et al*. [46], and ii) RS-UT: The Reverse-unbalanced Source (RS) and Unbalanced-Target (UT) version from Tan *et al*. [43], wherein source and target label distributions are manually long-tailed to be reversed versions of one another (see Fig. 3 (right)).

VisDA. VisDA2017 [32] is a large dataset for synthetic→real adaptation with 12 classes and >200k images.

DIGITS. We use the SVHN [29]→MNIST [21] shift for 10-way digit recognition.

Metric. On LDS DA benchmarks (DomainNet and OfficeHome RS-UT), consistent with prior work in UDA under LDS [19, 43], we compute a mean of per-class accuracy on the target test split as our metric, that weights performance on all classes equally. On standard DA benchmarks (OfficeHome and VisDA2017) we report standard accuracy.

4.2. Implementation details

We use PyTorch [30] for all experiments. On DomainNet, OfficeHome, and VisDA2017, we modify the standard ResNet50 [17] CNN architecture to a few-shot variant used in recent DA work [8, 35, 43]: we replace the last linear layer with a $C-$ way (for C classes) fully-connected layer with Xavier-initialized weights and no bias. We then L_2-normalize activations flowing into this layer and feed its output to a softmax layer with a temperature $T = 0.05$. We match optimization details to Tan *et al*. [43]. On DIGITS, we make similar modifications to the LeNet architecture and use $T = 0.01$ [18]. For augmenting images for consistency checking, we use RandAugment [10], which sequentially applies N label-preserving image transformations randomly sampled from a set of 14 transforms. We set $N = 3$, use transformation severity $M = 2.0$, and use a committee of $k = 3$ transforms. We use class-balanced sampling on the source domain and pseudo class-balanced sampling on the target. We set λ_{IE} and λ_{SENTRY} to 0.1 and 1.0, and match InstaPBM to set Q=256 for the information entropy loss.

4.3. Baselines

As our primary baselines we use four state-of-the art UDA methods from prior work specifically designed for DA under LDS: i) **COAL** [43]: Co-aligns feature and label distributions, using prototype-based conditional alignment via MME [35], and self-training on confidently-predicted pseudo-labels. ii) **MDD + Implicit Alignment (I.A)** [19]: Uses target pseudolabels to construct $N-$way (# classes per-batch) $K-$shot (# examples per class) dataloaders that are "aligned" (i.e. sample the same set of classes within a batch for both source and target), in conjunction with Margin Disparity Discrepancy [53], a strong UDA method, iii) **InstaPBM [22]**: Proposes "predictive-behavior" matching, which entails matching properties of $p_\Theta(y|\mathbf{x})$ between source and target. This is achieved via optimizing a combination of mutual information maximization, contrastive, and mixup losses, and iv) **F-DANN** [49]: Proposes an asymmetrically-relaxed distribution matching-based version of DANN [13] to deal with LDS. COAL, InstaPBM, and

Method	R → C	R → P	R → S	C → R	C → P	C → S	P → R	P → C	P → S	S → R	S → C	S → P	AVG
source	65.75	68.84	59.15	77.71	60.60	57.87	84.45	62.35	65.07	77.10	63.00	59.72	66.80
BBSE [24]	55.38	63.62	47.44	64.58	42.18	42.36	81.55	49.04	54.10	68.54	48.19	46.07	55.25
PADA [4]	65.91	67.13	58.43	74.69	53.09	52.86	79.84	59.33	57.87	76.52	66.97	61.08	64.48
MCD [37]	61.97	69.33	56.26	79.78	56.61	53.66	83.38	58.31	60.98	81.74	56.27	66.78	65.42
DAN [25]	64.36	70.65	58.44	79.44	56.78	60.05	84.56	61.62	62.21	79.69	65.01	62.04	67.07
F-DANN [49]	66.15	71.80	61.53	81.85	60.06	61.22	84.46	66.81	62.84	81.38	69.62	66.50	69.52
UAN [52]	71.10	68.90	67.10	83.15	63.30	64.66	83.95	65.35	67.06	82.22	70.64	68.09	72.05
JAN [28]	65.57	73.58	67.61	85.02	64.96	67.17	87.06	67.92	66.10	84.54	72.77	67.51	72.48
ETN [5]	69.22	72.14	63.63	86.54	65.33	63.34	85.04	65.69	68.78	84.93	72.17	68.99	73.99
BSP [9]	67.29	73.47	69.31	86.50	67.52	70.90	86.83	70.33	68.75	84.34	72.40	71.47	74.09
DANN [13]	63.37	73.56	72.63	86.47	65.73	70.58	86.94	73.19	70.15	85.73	75.16	70.04	74.46
COAL [43]	73.85	75.37	70.50	89.63	69.98	71.29	89.81	68.01	70.49	87.97	73.21	70.53	75.89
InstaPBM [22]	80.10	75.87	70.84	89.67	70.21	72.76	89.60	74.41	72.19	87.00	79.66	71.75	77.84
SENTRY (Ours)	**83.89**	**76.72**	**74.43**	**90.61**	**76.02**	**79.47**	**90.27**	**82.91**	**75.60**	**90.41**	**82.40**	**73.98**	**81.39**

Table 1: Per-class average accuracies on DomainNet. Bold and underscore denote the best and second-best performing methods respectively.

Method	Rw → Pr	Rw → Cl	Pr → Rw	Pr → Cl	Cl → Rw	Cl → Pr	AVG
source	70.74	44.24	67.33	38.68	53.51	51.85	54.39
BSP [9]	72.80	23.82	66.19	20.05	32.59	30.36	40.97
PADA [4]	60.77	32.28	57.09	26.76	40.71	38.34	42.66
BBSE [24]	61.10	33.27	62.66	31.15	39.70	38.08	44.33
MCD [37]	66.03	33.17	62.95	29.99	44.47	39.01	45.94
DAN [25]	69.35	40.84	66.93	34.66	53.55	52.09	52.90
F-DANN [49]	68.56	40.57	67.32	37.33	55.84	53.67	53.88
JAN [28]	67.20	43.60	68.87	39.21	57.98	48.57	54.24
DANN [13]	71.62	46.51	68.40	38.07	58.83	58.05	56.91
MDD [53]	71.21	44.78	69.31	42.56	52.10	52.70	55.44
COAL [43]	73.65	42.58	73.26	40.61	59.22	57.33	58.40
InstaPBM [22]	75.56	42.93	70.30	39.32	61.87	63.40	58.90
MDD+I.A [19]	76.08	50.04	**74.21**	45.38	61.15	63.15	61.67
SENTRY (Ours)	**76.12**	**56.80**	73.60	**54.75**	**65.94**	**64.29**	**65.25**

Table 2: Per-class average accuracies on OfficeHome RS→UT (right) benchmarks. Bold and underscore denote the best and second-best performing methods respectively.

MDD+I.A. all make use of target pseudolabels, and COAL and InstaPBM are self-training based approaches.

For completeness, we also include results for additional baselines from Tan *et al*. [43]: i) Conventional feature alignment-based UDA methods: DAN [25], JAN [28], DANN [13], MCD [35], and MDD [53], ii) BBSE [22] which only aligns label distributions, iii) Methods that assume non-overlapping labeling spaces: PADA [4], ETN [5], and UAN [52]. We also report results for FixMatch [40], a state-of-the-art self-training method for semi-supervised learning, on two benchmarks.

4.4. Results

Results on label-shifted DA benchmarks. We present results on 12 shifts from DomainNet (Table 1) and 6 shifts from OfficeHome RS→UT (Table 2). On DomainNet, SENTRY outperforms the next best performing method InstaPBM [22] on every shift, and by 3.55% mean accuracy averaged across shifts. On OfficeHome RS-UT, SENTRY outperforms the next best performing method MDD+I.A [19] on 5 out of 6 shifts, and on average by 3.58% mean accuracy. Our method also significantly outperforms F-DANN [49] (by 11.87% and 11.37%) and COAL [43] (by 5.50% and 6.85%), which are both UDA strategies designed for adaptation under LDS.

Method	acc (%)
Source	46.1
DAN [25]	56.3
DANN [13]	57.6
JAN [28]	58.3
CDAN [26]	65.8
BSP [9]	66.3
MDD [53]	68.1
FixMatch [40]	59.0
InstaPBM [22]	69.2
MDD+I.A [19]	69.5
SENTRY (Ours)	**72.2**

(a) OfficeHome (12 shift avg)

Method	acc (%)
Source	41.0
JAN [25]	61.6
MCD [37]	69.8
CDAN [26]	70.0
FixMatch [40]	64.9
MDD [53]	74.6
MDD+I.A [19]	75.8
InstaPBM [22]	76.3
SENTRY (Ours)	**76.7**

(b) VisDA2017

Table 3: Accuracies on standard DA benchmarks.

Results on standard DA benchmarks. Table 3 presents results on 2 standard DA benchmarks: OfficeHome and VisDA 2017. As seen, SENTRY improves mean accuracy over the next best method by 2.7% averaged over 12 shifts (full table in supp.) on OfficeHome, and by 0.4% on VisDA.

Varying degree of label imbalance. To perform a controlled study of adapting to targets with varying degrees of label imbalance, we use the SVHN→MNIST shift. Since MNIST is class-balanced, we manually long-tail its train-

Method	SVHN→MNIST-LT IF=1	IF=20	IF=50	IF=100	AVG
source	68.1	68.1	68.1	68.1	68.1
MMD [27]	53.4±0.9	56.7±1.2	56.2±1.4	55.1±0.7	55.4±1.1
DANN [13]	68.0±0.9	71.5±1.0	66.9±0.5	60.6±2.2	66.8±1.5
COAL [43]	78.8±1.0	67.1±1.4	70.2±1.5	70.0±1.8	71.5±1.4
InstaPBM [22]	90.7±0.2	77.9±3.5	68.9±1.3	65.9±2.2	75.9±1.8
SENTRY (Ours)	**92.9**±0.3	**93.9**±2.2	**85.6**±4.5	**85.6**±1.1	**89.5**±2.0

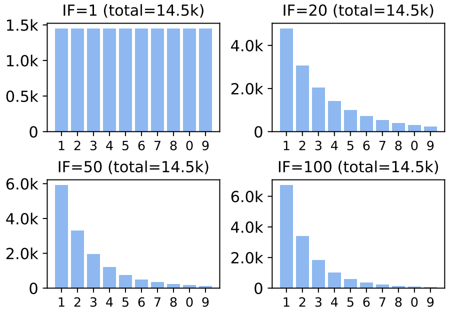

Table 4: **Left**: Per-class average accuracy after adapting from SVHN to manually long-tailed (-LT) training sets of MNIST (test set is unchanged). The degree of label imbalance is measured by the imbalance factor (IF). All long-tailed versions use an *identical* amount of data. For each IF, we construct 3 long-tailed versions and report mean and 1 standard deviation. **Right**: Label distribution for each IF.

select for entmin	select for entmax	DomainNet C→S	OH (RS-UT) Rw→Cl
all	none	68.8	44.9
consistent	none	77.7	55.3
consistent	inconsistent	79.5	56.8
correct	none	84.3	77.7
correct	incorrect	86.3	80.1

Table 5: Ablations of our selection strategy on DomainNet C→S and OfficeHome RS-UT Rw→Cl. Gray row is our method. Last two rows are oracle approaches that use target labels.

	N=1	N=3	N=5
k=1	78.2	78.6	78.9
k=3	76.8	79.5	77.8
k=5	77.5	78.4	77.7

(a) C→S

	N=1	N=3	N=5
k=1	53.8	57.5	55.6
k=3	55.3	56.8	56.2
k=5	54.7	58.4	54.5

(b) Rw→Cl

voting	C→S	Rw→Cl
maj.	79.5	56.8
unan.	77.8	52.2

(c) Vary voting

Table 6: Ablating the consistency checker on C→S and Rw→Cl: **a-b)** Varying committee size (k) and num. consecutive transforms in RandAugment (N). **c)** Varying voting strategy: maj. and unan. denote majority and unanimous. Gray is our method.

ing split, and use it as our unlabeled target train set (test set is unchanged). The long-tailing is performed by sampling from a Pareto distribution and subsampling, with class cardinality following the same sorted order as the source label distribution for simplicity. To systematically vary the degree of imbalance, we modulate the parameters of the Pareto distribution so as to generate a desired Imbalance Factor (IF) [11], computed as the ratio of the cardinality of the largest and smallest classes. Larger IF's represent a higher degree of imbalance. We thus create 3 splits with IF $\in \{20, 50, 100\}$, corresponding to varying label imbalance but with an identical amount of data (=14.5k instances). Further, we create a *control* version that also has 14.5k instances but possesses a balanced label distribution. Table 4 (right) shows the resulting label distributions.

We report per-class average accuracies in Table 4 (left). As baselines, we include a domain discrepancy based method (MMD [27]), an adversarial DA method (DANN [13]), as well as COAL [43] and InstaPBM [22]. Across methods, performance at higher imbalance factors is worse, illustrating the difficulty of adapting under severe label imbalance. However, SENTRY significantly outperforms baselines, even at higher imbalance factors, achieving 13.6% higher mean accuracy than the next competing method.

4.5. Ablations

We now present ablations of SENTRY, our proposed approach, on the Clipart→Sketch from DomainNet and the Real World→Clipart shift from OfficeHome RS-UT.

Selective optimization helps significantly (Tab. 5). We first measure the effect of performing entropy minimization on *all samples*, as done in prior work. We find this to perform *significantly* worse (by 10.7%, 11.9%) than our method! Clearly, consistency-based selective optimization is crucial.

Entropy maximization helps consistently (Tab. 5). Next, we opt to only minimize entropy on consistent target instances, but *do not perform entropy maximization*. We find this to underperform against our min-max optimization (by 1.8%, 1.5%). Further, as an oracle approach, we use ground truth target labels to determine whether an instance is correctly or incorrectly classified, and perform two experiments: entropy minimization on correct instances (and no maximization), and min-max entropy optimization on correct and incorrect instances. Selective min-max optimization again outperforms just minimization by 2% and 2.4%, showing that reducing confidence on misaligned instances helps.

Ablating consistency checker. In Tables 6a, 6b, we vary the committee size k and number of RandAugment transforms N used by our consistency checker. We do not find our method to be very sensitive to either hyperparameter. In Table 6c, we vary the voting strategy used to judge committee consistency and inconsistency. We experiment with majority voting (atleast $\frac{k}{2} + 1$ votes needed) and unanimous voting (k votes needed), and find the former to generalize better.

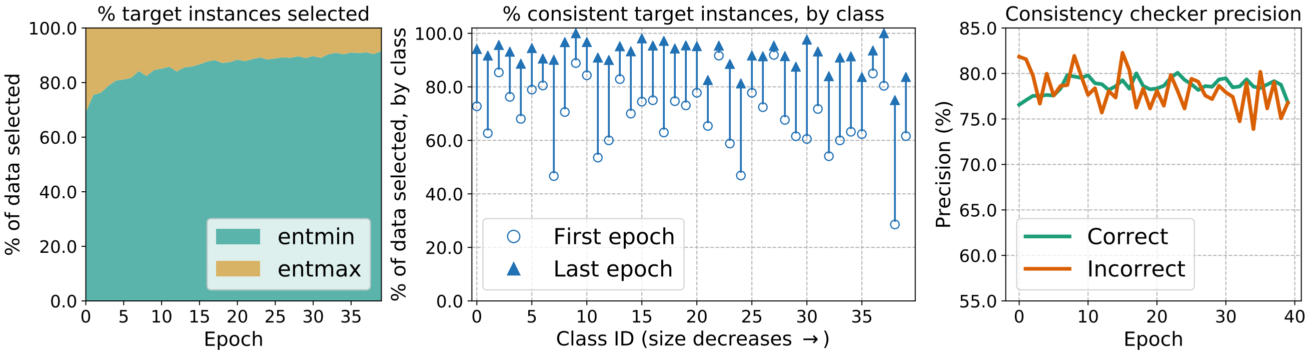

Figure 4: Analysis of SENTRY on Clipart→Sketch. **Left:** % of seen target instances selected for entropy minimization and maximization over epochs. **Middle:** % of seen target data chosen for entropy minimization at the end of first and last epochs of adaptation, broken down by class. **Right:** Ground truth precision of SENTRY's committee consistency strategy at identifying correct and incorrect instances over epochs.

Gains are not simply due to stronger augmentation. To verify this, we continue using RandAugment (with N=1) for consistency checking, but backpropagate on the original (rather than augmented) target instances, effectively removing data augmentation entirely. On C→S and Rw→Cl, this achieves 73.1% and 52.6%, which is still 0.3% and 2.6% better than the next best baseline on each shift, despite not using any data augmentation for optimization at all.

Pseudo class-balanced sampling helps. We find that class-balanced sampling using pseudolabels on the target improves per-class average accuracy over random sampling by 0.91% and 0.52% on C→S and Rw→Cl. In the absence of the target information entropy regularizer L_{IE}, this performance gap grows to 2.9% and 3.7%. As explained in Sec 3.4, both objectives contribute towards overcoming LDS in similar ways, and we find here that using both together works best.

4.6. Analysis

% of target instances selected over time. Fig. 4 (left) shows that the % of seen target instances selected for entropy minimization steadily increases over time, while that selected for entropy maximization decreases. This adaptive nature is a result of the augmentation invariance regularization built into our method (Sec. 3). In Fig. 4 (middle), we measure the % of target instances selected for entropy minimization, *per-class*, at the end of the first and last epoch of adaptation. Despite no explicit class-conditioning, we find that this measure increases for all classes.

Precision of consistency checker. Fig 4 (right) shows the precision of our consistency and inconsistency-based selection strategies at identifying instances that are actually correct and incorrect. As seen, committee-based consistency and inconsistency are both 75-80% precise at identifying correct and incorrect instances respectively.

Per-class accuracy change. In the supplementary, we report the per-class accuracy after adaptation using our method, and contrast it against InstaPBM [22]. On the C→S shift, SENTRY outperforms InstaPBM across 37/40 categories.

Computational efficiency. Compared to standard entropy minimization, SENTRY requires k (for committee size k) forward passes per iteration (to determine consistency), but no additional backward passes. SENTRY thus does not add a sizeable computational overhead over prior work.

Correctness of entropy maximization. For simplicity, consider 2-way classification. For output score p and true class y, entropy maximization loss $L_{\text{EM}} = p \log p + (1-p)\log(1-p)$, and binary cross-entropy (BCE) loss $L_{BCE} = -[y\log(p) + (1-y)\log(1-p)]$. Without loss of generality, assume an incorrect prediction with $y = 0$ and $0.5 \leq p < 1$. In Fig. 5 we show that in this case, gradients ($\nabla_p L$) for entropy maximization and BCE (wrt true class) are strongly correlated. Thus, after identifying misaligned target instances based on predictive inconsistency, entropy maximization has a similar effect as supervised training with respective to the true class.

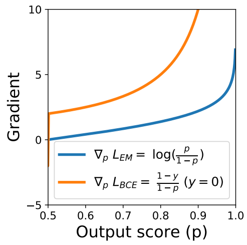

Figure 5: $\nabla_p L$ v/s p

5. Conclusion

We propose SENTRY, an algorithm for unsupervised domain adaptation (UDA) under simultaneous data and label distribution shift. Unlike prior work that suffers from error accumulation arising from unconstrained self-training, SENTRY first judges the reliability of a target instance based on its predictive consistency under a committee of random image transforms, and then selectively minimizes entropy (increasing confidence) on consistent instances, while maximizing entropy (reducing confidence) on inconsistent ones. We show that SENTRY significantly improves upon the state-of-the-art across 27/31 shifts from several UDA benchmarks.

Acknowledgements. This work was supported in part by funding from the DARPA LwLL project.

References

[1] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2018. 2

[2] Yuval Bahat, Michal Irani, and Gregory Shakhnarovich. Natural and adversarial error detection using invariance to image transformations. *arXiv preprint arXiv:1902.00236*, 2019. 2

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059, 2019. 2, 3

[4] 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. 6

[5] 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. 6

[6] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 627–636, 2019. 1

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2

[8] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2018. 5

[9] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1081–1090, 2019. 6

[10] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 2, 5

[11] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019. 4, 7

[12] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *International Conference on Learning Representations*, 2019. 3

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015. 1, 2, 5, 6, 7

[14] Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. In *CAP*, pages 281–296, 2005. 2, 3, 4

[15] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330, 2017. 1

[16] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 1

[17] 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. 5

[18] 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, 2018. 1, 2, 5

[19] Xiang Jiang, Qicheng Lao, Stan Matwin, and Mohammad Havaei. Implicit class-conditioned domain alignment for unsupervised domain adaptation. In *International Conference on Machine Learning*, 2020. 5, 6

[20] 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. 2

[21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[22] Bo Li, Yezhen Wang, Tong Che, Shanghang Zhang, Sicheng Zhao, Pengfei Xu, Wei Zhou, Yoshua Bengio, and Kurt Keutzer. Rethinking distributional matching based domain adaptation. *arXiv preprint arXiv:2006.13352*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[24] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pages 3122–3130, 2018. 2, 6

[25] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 1, 2, 6

[26] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018. 2, 6

[27] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 2200–2207, 2013. 7

[28] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017. 6

[29] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco,

Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 5

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019. 5

[31] 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 International Conference on Computer Vision*, pages 1406–1415, 2019. 1, 2, 5

[32] 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. 2, 5

[33] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *ICLR*, 2017. 4

[34] 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. 1, 2

[35] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8050–8058, 2019. 3, 5, 6

[36] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. In *International Conference on Learning Representations*, 2018. 2

[37] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 6

[38] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *arXiv preprint arXiv:1606.04586*, 2016. 2

[39] Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D Sculley, Joshua Dillon, Jie Ren, and Zachary Nado. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13969–13980, 2019. 1, 2, 3

[40] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 2, 6

[41] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[42] Ryuhei Takahashi, Atsushi Hashimoto, Motoharu Sonogashira, and Masaaki Iiyama. Partially-shared variational auto-encoders for unsupervised domain adaptation with target shift. In *The European Conference on Computer Vision (ECCV)*, 2020. 2

[43] Shuhan Tan, Xingchao Peng, and Kate Saenko. Class-imbalanced domain adaptation: An empirical odyssey. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2020. 1, 2, 3, 5, 6, 7

[44] 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

[45] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 1, 2

[46] 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. 2, 5

[47] Jindong Wang, Yiqiang Chen, Shuji Hao, Wenjie Feng, and Zhiqi Shen. Balanced distribution adaptation for transfer learning. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1129–1134. IEEE, 2017. 2

[48] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017. 4

[49] Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In *International Conference on Machine Learning*, pages 6872–6881, 2019. 1, 4, 5, 6

[50] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2020. 2

[51] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2272–2281, 2017. 2

[52] 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, 6

[53] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413, 2019. 2, 5, 6

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

[55] Yang Zou, Zhiding Yu, BVK Vijaya 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. 4