A Theory of Label Propagation for Subpopulation Shift

Tianle Cai[1, 2], Ruiqi Gao[1, 2], Jason D. Lee[1], Qi Lei[1]

[1]Princeton University
[2]Zhongguancun Haihua Institute for Frontier Information Technology

ICML 2021

Background

- In many machine learning tasks we encounter *distribution shifts* and often lots of data we have are *unlabeled.*

- *Unsupervised domain adaptation*: Source distribution S with labeled data (x, y), target distribution T with unlabeled data x.

Example Dataset: Office-31

(Saenko, et al., 2010)

Example Dataset: BREEDS

Source

Target

Entity30-Passerine

Source

Target

Entity30-Tableware

(Santurkar et al., 2021)

Classic Methods

- Traditional method: Reweight/resample based on the density ratio of S and T.
- Caveat: Only works when the support of S and T are equal.

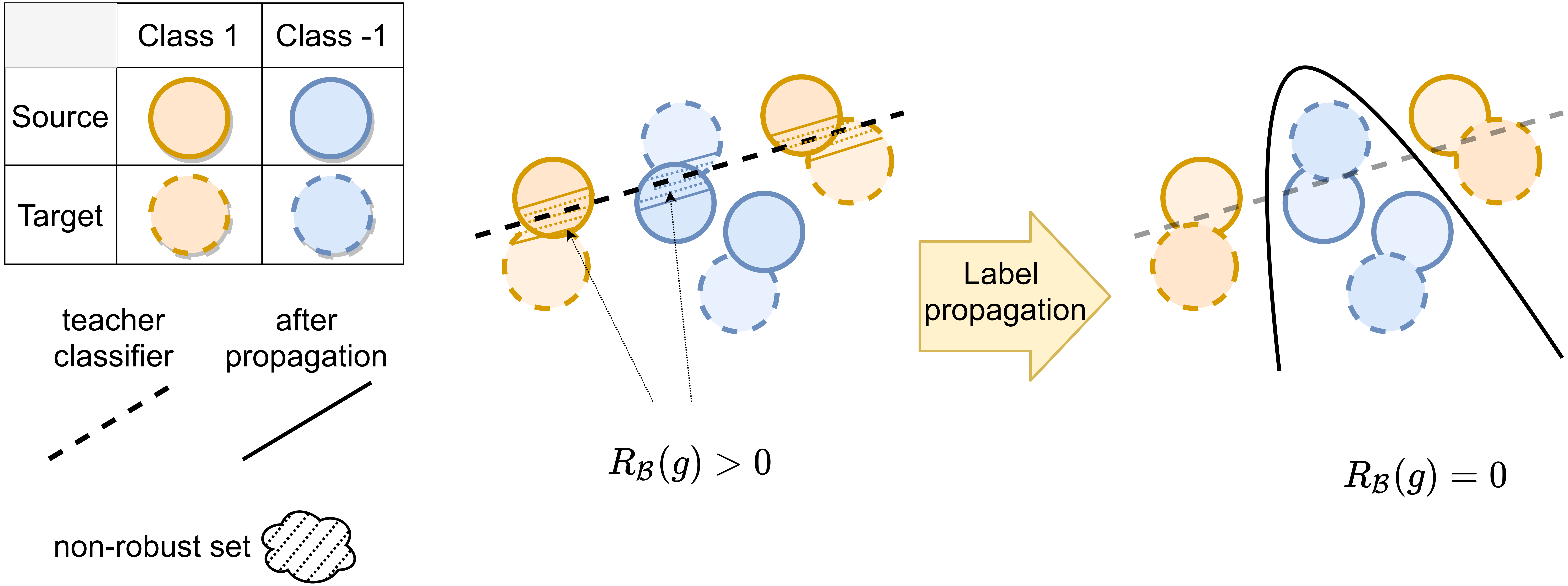
- Classic method for deep learning: Distributional matching[1], which learns a representation z on which the distribution of $z(x)$ for $x \sim S$ and $x \sim T$ are the same, while performing classification by $x \to z \to \hat{y}$.
- Caveat: Forcing representation to match may not preserve the right information for y.

([1]Ben-David et al, 2010)

Our New Framework: Subpopulation Shift

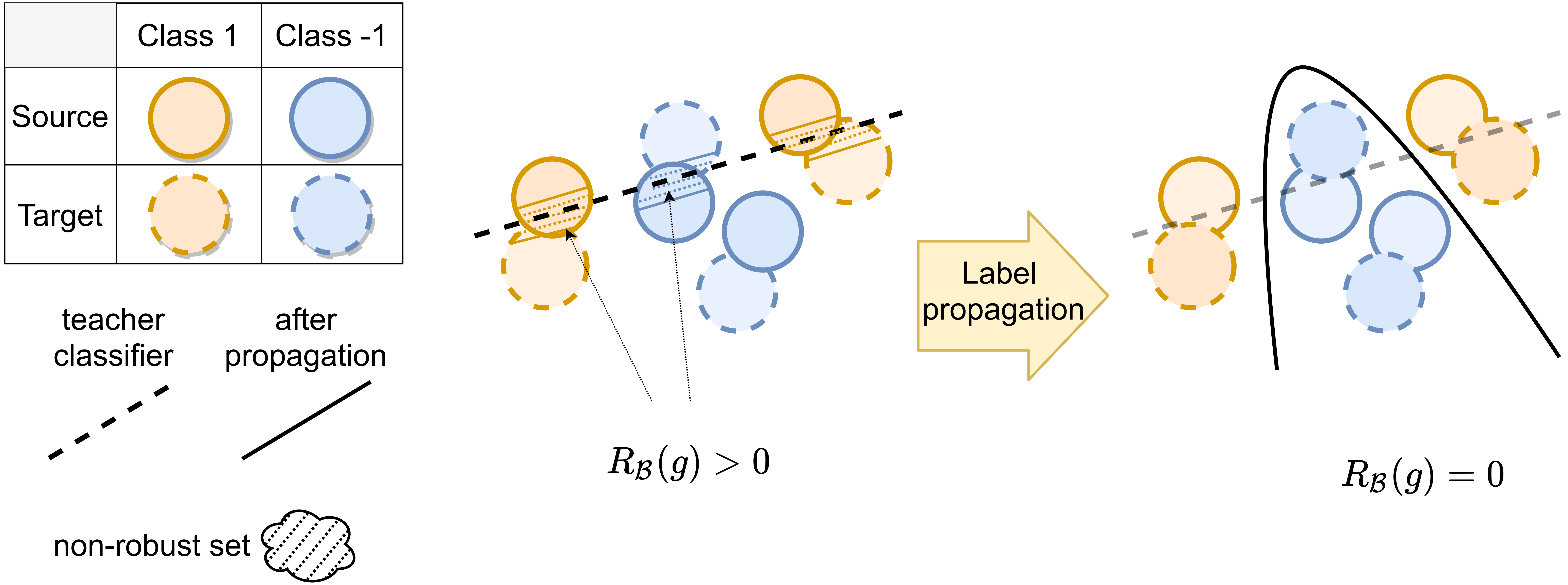
- A new model and framework for distribution shift.
- Characterize source and target by $S = S_1 \cup \cdots \cup S_m$ and $T = T_1 \cup \cdots \cup T_m$, where each S_i and T_i are correspondent in a certain sense.
- Subpopulation shift is ubiquitous in practical tasks, e.g. “Poodles eating dog food” in the source and “Labradors eating meat” in the target. The previous BREEDS is the classic dataset for subpopulation shifts of this form.
- Subpopulation shift can also be more implicit and hard to describe by words.

Algorithmic Framework



- Suppose there is a (possibly noisy) teacher classifier g_{tc} on S. Goal: Propagate the label information from S to T based on unlabeled data.
- In this toy illustration, each $S_i \cup T_i$ forms a regular connected component.

Algorithmic Framework



- Consistency regularizer $R_B(g)$ measures the amount of non-robust set of g, i.e. points whose predictions by g is inconsistent in a small neighborhood.

- g_{tc} + A proper consistency regularization = Label propagation!

Algorithm

- We expect the predictions to be stable under a suitable set of input transformations $B(x) \subset X$, and use the following consistency regularization:

$$R_B(g) := P_{x \sim \frac{1}{2}(S+T)}[\,\exists x' \in B(x), \text{s.t. } g(x) \neq g(x')].$$

- B can be a distance-based neighborhood set or some data augmentations A, and can take the general form $B(x) = \{x' : \exists A \text{ such that } d(x', A(x)) \leq r\}$.

- Define $L_{01}^S(g, g_{tc}) := P_{x \sim S}[g(x) \neq g_{tc}(x)]$, our algorithm is

$$g = \operatorname{argmin}_{g:X \rightarrow Y, g \in G} L_{01}^S(g, g_{tc}) \text{ s.t. } R_B(g) \leq \mu,$$

where μ is a constant satisfying $R_B(g^*) < \mu$, which is expected to be small.

Technical Assumption: Expansion

- The expansion property proposed in [1], some geometric regularity on $S_i \cup T_i$ w.r.t. B, is needed for local consistency regularization to propagate globally.

- Define the neighborhood set $N(x) := \{x' \,|\, B(x) \cap B(x') \neq \emptyset\}$ (informal), and for a set $A \subset X$ define $N(A) := \cup_{x \in A} N(x)$.

- Definition of (a, c)-multiplicative expansion: For $a \in (0,1)$, $c > 1$, any i, any $A \in S_i \cup T_i$ with $P_{\frac{1}{2}(S+T)}[A] \leq a$, we have $P_{\frac{1}{2}(S_i+T_i)}[N(A)] \geq \min(cP_{\frac{1}{2}(S_i+T_i)}[A], 1)$.

[1] Wei, C., Shen, K., Chen, Y., and Ma, T. (2021). Theoretical analysis of self-training with deep networks on unlabeled data.

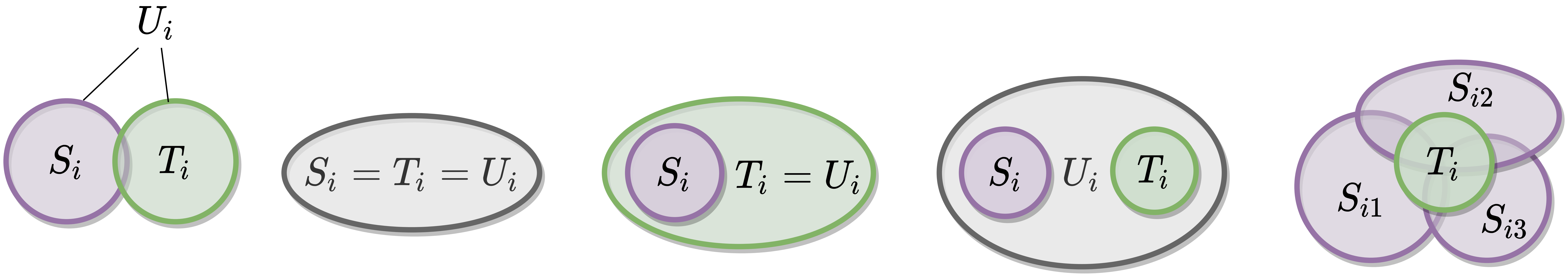
Main Theorem

- Based on $(1/2, c)$-multiplicative expansion, the target error of the classifier g returned by the algorithm $\epsilon_T(g) = P_{x \sim T}[g(x) \neq g^*(x)]$ is bounded by

$$\epsilon_T(g) \leq O\left(\frac{\mu}{c-1}\right).$$

- Remark: The accuracy of g can actually improve upon the accuracy on g_{tc}, supposing μ is small.

Generalized Subpopulation Shift



(a) Unsupervised domain adaptation

(b) Semi-supervised learning or self-supervised denoising

(c) Domain expansion

(d) Domain extrapolation

(e) Multi-source domain adaptation or domain generalization

- The previous result holds on the general setting where there is a general unlabeled dataset U that "covers" S and T, on which we perform label propagation.

Experiments: Subpopulation Shift Dataset

- ENTITY-30 task from BREEDS tasks.
- We use FixMatch, an existing consistency regularization method. We also leverage SwAV, an existing unsupervised representation learned from ImageNet, where there can be a better structure of subpopulation shift. We compare with popular distribution matching methods like DANN and MDD.

Method	Source Acc	Target Acc
Train on Source	91.91±0.23	56.73±0.32
DANN (Ganin et al., 2016)	92.81±0.50	61.03±4.63
MDD (Zhang et al., 2019)	92.67±0.54	63.95±0.28
FixMatch (Sohn et al., 2020)	90.87±0.15	72.60±0.51

Experiments: Classic Domain Adaptation Dataset

- Office-31 and Office-home.
- We add consistency regularization (FixMatch) to MDD, and observed improvement to the distribution matching method.

Method	A $\to$ W	D $\to$ W	W $\to$ D	A $\to$ D	D $\to$ A	W $\to$ A	Average
MDD	94.97$\pm$0.70	98.78$\pm$0.07	100$\pm$0	92.77$\pm$0.72	75.64$\pm$1.53	72.82$\pm$0.52	89.16
MDD+FixMatch	95.47$\pm$0.95	98.32$\pm$0.19	100$\pm$0	93.71$\pm$0.23	76.64$\pm$1.91	74.93$\pm$1.15	**89.84**

Table 2: Performance of MDD and MDD+FixMatch on Office-31 dataset.

Method	Ar $\to$ Cl	Ar $\to$ Pr	Ar $\to$ Rw	Cl $\to$ Ar	Cl $\to$ Pr	Cl $\to$ Rw	Pr $\to$ Ar	Pr $\to$ Cl	Pr $\to$ Rw	Rw $\to$ Ar	Rw $\to$ Cl	Rw $\to$ Pr	Average
MDD	54.9$\pm$0.7	74.0$\pm$0.3	77.7$\pm$0.3	60.6$\pm$0.4	70.9$\pm$0.7	72.1$\pm$0.6	60.7$\pm$0.8	53.0$\pm$1.0	78.0$\pm$0.2	71.8$\pm$0.4	59.6$\pm$0.4	82.9$\pm$0.3	68.0
MDD+FixMatch	55.1$\pm$0.9	74.7$\pm$0.8	78.7$\pm$0.5	63.2$\pm$1.3	74.1$\pm$1.8	75.3$\pm$0.1	63.0$\pm$0.6	53.0$\pm$0.6	80.8$\pm$0.4	73.4$\pm$0.1	59.4$\pm$0.7	84.0$\pm$0.5	**69.6**

Table 3: Performance of MDD and MDD+FixMatch on Office-Home dataset.

Takeaway Message

Consistency-based methods (e.g. semi-supervised learning methods like FixMatch) can help domain adaptation, especially in the presence of subpopulation shift!