Unsupervised Multi-source Domain Adaptation Without Access to Source Data (Supplementary Material)

Sk Miraj Ahmed[1,*], Dripta S. Raychaudhuri[1,*], Sujoy Paul[2,*,†], Samet Oymak[1], Amit K. Roy-Chowdhury[1]
[1] University of California, Riverside, [2] Google Research
{sahme047@, drayc001@, spaul003@, oymak@ece., amitrc@ece.}ucr.edu

[*]Equal Contribution
[†]Work done while SP was a PhD student at UC Riverside.

1. Proof of Lemma 1

Lemma 1. *Assume that the loss $L(\theta(x), y)$ is convex in its first argument and that there exists a $\lambda \in \mathbb{R}^n$ where $\lambda \geq 0$ and $\lambda^\top \mathbb{1} = 1$, such that the target distribution is exactly equal to the mixture of source distributions, i.e $Q_T = \sum_{i=1}^n \lambda_i Q_S^i$. Set the target predictor as the following convex combination of the optimal source predictors*

$$\theta_T(x) = \sum_{k=1}^{n} \frac{\lambda_k Q_S^k(x)}{\sum_{j=1}^n \lambda_j Q_S^j(x))} \theta_S^k(x).$$

Recall the pseudo-labeling loss (10). Then, for this target predictor, over the target distribution, the unsupervised loss induced by the pseudo-labels and the supervised loss are both less than or equal to the loss induced by the best source predictor. In particular,

$$\mathcal{L}(Q_T, \theta_T) \leq \min_{1 \leq j \leq n} \mathcal{L}(Q_T, \theta_S^j).$$

Proof. We can see that the left hand-side of the inequality can be upper-bounded by some loss as follows,

$$\begin{aligned}
\mathcal{L}(Q_T, \theta_T) = \int_x Q_T(x) L(\theta_T(x), y) &= \int_x Q_T(x) L\left(\sum_{i=1}^{n} \frac{\lambda_i Q_S^i(x)}{\sum_{j=1}^n \lambda_j Q_S^j(x))} \theta_S^i(x), y \right) dx \\
&\leq \int_x Q_T(x) \sum_{i=1}^{n} \frac{\lambda_i Q_S^i(x)}{\sum_{j=1}^n \lambda_j Q_S^j(x))} L(\theta_S^i(x), y) dx \quad \text{(from Jensen's inequality)} \\
&= \int_x Q_T(x) \sum_{i=1}^{n} \frac{\lambda_i Q_S^i(x)}{Q_T(x)} L(\theta_S^i(x), y) dx \quad \text{(from distribution assumption)} \\
&= \sum_{i=1}^{n} \lambda_i \int_x Q_S^i(x) L(\theta_S^i(x), y) dx \quad \text{(changing the order of summation)} \\
&= \sum_i \lambda_i \mathcal{L}(Q_S^i(x), \theta_S^i)
\end{aligned} \tag{1}$$

Now for the R.H.S. we can write this loss as follows,

$$\begin{aligned}
\mathcal{L}(Q_T, \theta_S^j) &= \int_x Q_T(x) L(\theta_S^j(x), y) dx \\
&= \int_x \sum_{i=1}^{n} \lambda_i Q_S^i(x) L(\theta_S^j(x), y) dx \\
&= \sum_{i=1}^{n} \lambda_i \int_x Q_S^i L(\theta_S^j(x), y) dx \\
&= \sum_{i=1}^{n} \lambda_i \mathcal{L}(Q_S^i(x), \theta_S^j)
\end{aligned} \tag{2}$$

Now recall from main paper that,

$$\theta_S^k = \arg\min_\theta \mathcal{L}(Q_S^k, \theta) \quad \text{for} \quad 1 \leq k \leq n.$$

.

This means θ_S^i is the best predictor for the source i, which has distribution Q_S^i. Thus we find that $\mathcal{L}(Q_S^i, \theta_S^i) \leq \mathcal{L}(Q_S^i, \theta_S^j)\ \forall j$, which implies $\sum_i \lambda_i \mathcal{L}(Q_S^i, \theta_S^i) \leq \sum_i \lambda_i \mathcal{L}(Q_S^i, \theta_S^j)$. This further implies that $\mathcal{L}(Q_T, \theta_T) \leq \mathcal{L}(Q_T, \theta_S^j)\ \forall j$, which in turn concludes the proof $\mathcal{L}(Q_T, \theta_T) \leq \min_{1 \leq j \leq n} \mathcal{L}(Q_T, \theta_S^j)$. Finally, suppose the entries of λ are strictly positive and let $\beta = \arg\min_j \mathcal{L}(Q_T, \theta_S^j)$. Observe that, if there is a source i such that the strict inequality $\mathcal{L}(Q_S^i, \theta_S^i) < \mathcal{L}(Q_S^i, \theta_S^\beta)$ holds, then the main claim of the lemma also becomes strict as we find

$$\mathcal{L}(Q_T, \theta_T) \leq \sum_i \lambda_i \mathcal{L}(Q_S^i, \theta_S^i) < \sum_i \lambda_i \mathcal{L}(Q_S^i, \theta_S^\beta) \leq \min_j \mathcal{L}(Q_T, \theta_S^j).$$

Verbally, this strict inequality has a natural meaning that the model j is strictly worse than model i for the source data i. □

2. Detailed steps of combination rule under source distribution uniformity assumption

See the discussion after **Lemma 1** in the main paper for reference.

$$\begin{aligned}\theta_T(x) &= \sum_{k=1}^{n} \frac{\lambda_k Q_S^k(x)}{\sum_{j=1}^{n} \lambda_j Q_S^j(x)} \theta_S^k(x) \\ &= \sum_{k=1}^{n} \frac{\lambda_k c_k \mathcal{U}(x)}{\sum_{j=1}^{n} \lambda_j c_j \mathcal{U}(x)} \theta_S^k(x) \\ &= \sum_{k=1}^{n} \frac{\lambda_k c_k}{\sum_{j=1}^{n} \lambda_j c_j} \theta_S^k(x)\end{aligned} \tag{3}$$

3. Additional Experiments

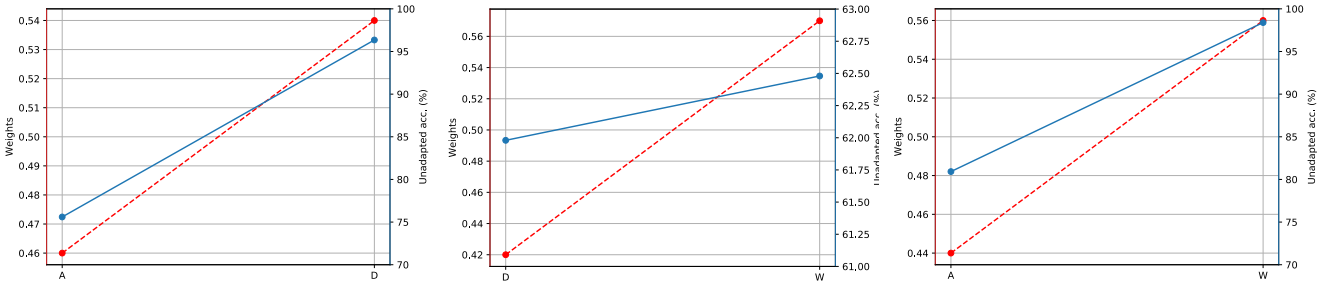

Figure 1: **Weights as model selection proxy.** The weights learnt by our framework on Office-31 correlates positively with the unadapted source model performance. (Left axis corresponds to the red plot and right to the blue plot, best viewed in color.)

From Figure 1, we can clearly see that for the model which gives higher accuracy for the unadapted scenario, it is automatically given higher weightage by our algorithm. As a result, we can easily infer about the quality of the source domain, in relation to the target, from the weights learnt by our framework.

Effect of weight on pseudo-labeling. We investigate the effect of the weight λ on $\mathcal{L}_{\text{pl}}$. We perform experiments on the Office dataset by varying the value of λ and plot the results in Figure 2. As shown in the plot, the proposed method performs best at $\lambda = 0.3$

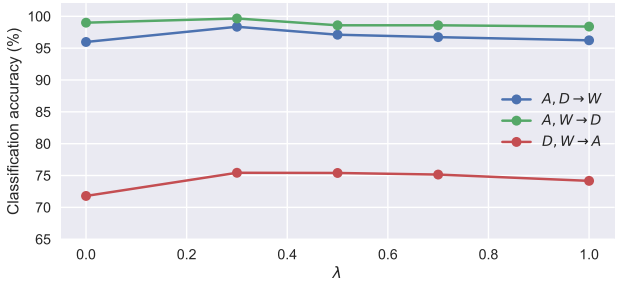

Figure 2: **Effect of λ.** The variations in classification as the weight on $\mathcal{L}_{\text{pl}}$ is varied. (Best viewed in color)

Effect of outlier source models. Our method is clearly robust to outlier source models. In Table 2 of the main paper, when *MNIST-M* is the target, transferring from only *USPS*, leads to an extremely poor performance of **21.3**% - here, USPS is a strong outlier. Despite the presence of such a poor source, our framework is mostly able to correctly negate the transfer from USPS, achieving a performance of **93**%, close to the best source performance of **94**%. On removing *USPS* as a source, DECISION outperforms the best source by achieving an accuracy of **94.5**%. In the future, we plan to actively use the weights to simultaneously remove poor sources while adaptation in order to boost the performance.

SOURCE	METHOD	C,P,I,S,R → Q	Q,P,I,S,R → C	Q,C,I,S,R → P	Q,C,P,S,R → I	Q,C,P,I,R → S	Q,C,P,I,S → R	AVG.
Multiple(w)	DAN[25]	16.2	39.1	33.3	11.4	29.7	42.1	28.6
	DCTN[46]	7.2	48.6	48.8	23.4	47.3	53.5	38.1
	MCD[37]	7.6	54.3	45.7	22.1	43.5	58.4	38.6
	M^3SDA-β[32]	6.3	58.6	52.3	26	49.5	62.7	42.5
Single(w/o)	Source-best	11.9	49.9	47.5	20	41.1	57.7	38
	Source-worst	2.3	12.2	2.2	1.1	8.7	4.8	5.2
	SHOT[22]-best	18.7	58.3	53	22.7	48.4	65.9	44.5
	SHOT[22]-worst	3.8	14.8	3.5	1	11.9	6.6	7
Multiple(w/o)	SHOT[22]-Ens	15.3	58.6	**55.3**	**25.2**	**52.4**	**70.5**	**46.2**
	DECISION(Ours)	**18.9**	**61.5**	54.6	21.6	51	67.5	45.9

Table 1: **Results on DomainNet**:Q,C,P,I,S and R are abbreviations of *quickdraw*, *clipart*, *painting*, *infograph*, *sketch* and *real*.

DomainNet [32]: This is a relatively new and large dataset where there are six domains under the common object categories, namely quickdraw (Q), clipart (C), painting (P), infograph (I), sketch (S) and real (R) with a total of 345 object classes in each domain. Experimental results on this dataset are shown in Table 1. Our method consistently outperforms the best adapted source baselines (SHOT-best) except for *infograph* as a target. However the average performance over all the domains as target is slightly less than the SHOT-Ens. Note that for *quickdraw* and *clipart* as target, our method outperforms all the state of the art methods including source free and with source data single and multi source state-of-the-art DA methods.

Distillation. Our results on using the distillation strategy outlined in Section 5.4 of the main paper are shown in Table 2. Despite the model compression, the performance remains consistent.

METHOD	OFFICE-HOME				OFFICE-CALTECH				OFFICE		
	RW	PR	CL	AR	A	C	D	W	A	D	W
DECISION (original)	83.6	84.4	59.4	74.5	95.9	95.9	100	99.6	75.4	99.6	98.4
DECISION (distillation)	83.7	84.4	59.1	74.4	96.0	95.7	99.4	99.6	75.4	99.6	98.1

Table 2: **Distillation results on object recognition tasks.** Performance remains consistent across all datasets despite distilling into a single target model.