

# A Prototype-oriented Clustering for Domain Shift with Source Privacy: Appendix

## A FULL EXPERIMENTAL RESULTS

### A.1 STANDARD SETTING

Table 6: Clustering accuracy (%) on different datasets for ResNet-18-based methods (supervised pre-training for all methods below the mid line) and (random initialization for all methods above the mid line).

Settings	Office-31				Office-Home					PACS				
	$\mathcal{R} \rightarrow A$	$\mathcal{R} \rightarrow W$	$\mathcal{R} \rightarrow D$	Avg	$\mathcal{R} \rightarrow Ar$	$\mathcal{R} \rightarrow Cl$	$\mathcal{R} \rightarrow Pr$	$\mathcal{R} \rightarrow Rw$	Avg	$\mathcal{R} \rightarrow P$	$\mathcal{R} \rightarrow A$	$\mathcal{R} \rightarrow C$	$\mathcal{R} \rightarrow S$	Avg
DeepCluster	19.6	18.9	18.7	19.1	8.9	11.1	16.9	13.3	12.6	27.9	22.2	24.4	27.1	25.4
IIC	31.9	37.0	34.0	34.4	12.0	15.2	22.5	15.9	16.4	70.6	39.8	39.6	46.6	49.2
ACIDS	33.4	37.5	36.1	35.7	12.0	16.2	23.9	15.7	17.0	64.4	42.1	44.5	51.1	50.5
PO	14.1 $\pm$ 1.6	17.9 $\pm$ 2.0	18.3 $\pm$ 2.9	16.8	11.4 $\pm$ 1.6	9.0 $\pm$ 1.6	12.9 $\pm$ 2.8	10.8 $\pm$ 1.7	11.0	30.5 $\pm$ 3.1	24.1 $\pm$ 0.6	19.8 $\pm$ 3.7	20.8 $\pm$ 1.7	23.8
SO	34.5 $\pm$ 0.5	46.7 $\pm$ 2.9	43.0 $\pm$ 2.9	41.4	23.6 $\pm$ 1.6	15.6 $\pm$ 1.9	23.1 $\pm$ 3.7	21.8 $\pm$ 2.9	21.0	30.8 $\pm$ 8.2	35.7 $\pm$ 3.9	27.6 $\pm$ 8.3	26.0 $\pm$ 3.7	30.0
TO	38.0 $\pm$ 3.2	46.6 $\pm$ 1.6	45.3 $\pm$ 1.5	43.3	21.3 $\pm$ 2.6	12.2 $\pm$ 0.7	30.6 $\pm$ 4.1	24.2 $\pm$ 0.7	22.1	88.4 $\pm$ 3.9	<b>56.5 <math>\pm</math> 4.1</b>	56.5 $\pm$ 11.1	49.1 $\pm$ 2.8	62.6
AO	42.8 $\pm$ 0.9	58.4 $\pm$ 3.7	55.8 $\pm$ 1.9	52.3	30.0 $\pm$ 1.7	22.7 $\pm$ 1.6	29.3 $\pm$ 4.1	24.4 $\pm$ 2.6	26.6	91.5 $\pm$ 5.9	47.7 $\pm$ 5.7	52.3 $\pm$ 1.2	49.1 $\pm$ 3.0	60.2
PCD	<b>46.8 <math>\pm</math> 1.7</b>	<b>60.0 <math>\pm</math> 2.6</b>	<b>57.8 <math>\pm</math> 5.9</b>	<b>54.9</b>	<b>33.3 <math>\pm</math> 1.0</b>	<b>24.4 <math>\pm</math> 1.5</b>	<b>31.4 <math>\pm</math> 4.7</b>	<b>28.1 <math>\pm</math> 2.5</b>	<b>29.3</b>	<b>92.6 <math>\pm</math> 2.4</b>	49.7 $\pm$ 5.0	<b>56.7 <math>\pm</math> 2.6</b>	<b>53.4 <math>\pm</math> 5.9</b>	<b>63.4</b>

Table 7: Clustering accuracy (%) on different datasets for ResNet-18-based methods (supervised pre-training).

Settings	PACS				
	$\mathcal{R} \rightarrow P$	$\mathcal{R} \rightarrow A$	$\mathcal{R} \rightarrow C$	$\mathcal{R} \rightarrow S$	Avg
ACIDS	80.9	48.2	50.5	<b>56.7</b>	59.1
PCD	<b>92.6</b>	<b>49.7</b>	<b>56.7</b>	53.4	<b>63.4</b>

Table 8: Clustering accuracy (%) on different initialization strategies for ResNet-50-based methods (supervised pre-training).

Settings	PACS				
	$\mathcal{R} \rightarrow P$	$\mathcal{R} \rightarrow A$	$\mathcal{R} \rightarrow C$	$\mathcal{R} \rightarrow S$	Avg
Self-supervised pre-training	82.1	53.4	50.8	43.6	57.5
Supervised pre-training	<b>93.3</b>	<b>54.6</b>	<b>59.1</b>	<b>56.8</b>	<b>66.0</b>

Table 9: Clustering accuracy (%) on different datasets for ResNet-50-based methods (self-supervised pre-training).

Settings	Office-31				Office-Home					PACS				
	$\mathcal{R} \rightarrow A$	$\mathcal{R} \rightarrow W$	$\mathcal{R} \rightarrow D$	Avg	$\mathcal{R} \rightarrow Ar$	$\mathcal{R} \rightarrow Cl$	$\mathcal{R} \rightarrow Pr$	$\mathcal{R} \rightarrow Rw$	Avg	$\mathcal{R} \rightarrow P$	$\mathcal{R} \rightarrow A$	$\mathcal{R} \rightarrow C$	$\mathcal{R} \rightarrow S$	Avg
PO	13.5 $\pm$ 0.9	16.7 $\pm$ 0.3	19.3 $\pm$ 2.4	16.5	10.5 $\pm$ 0.6	8.4 $\pm$ 0.2	10.5 $\pm$ 0.7	9.1 $\pm$ 0.9	9.6	28.3 $\pm$ 7.4	22.9 $\pm$ 2.6	24.0 $\pm$ 1.2	29.1 $\pm$ 2.3	26.1
SO	19.0 $\pm$ 5.0	26.3 $\pm$ 2.0	27.5 $\pm$ 3.0	24.3	18.3 $\pm$ 1.2	11.2 $\pm$ 0.3	16.4 $\pm$ 1.3	16.7 $\pm$ 1.7	15.7	40.7 $\pm$ 12.2	25.0 $\pm$ 1.4	29.7 $\pm$ 5.7	35.0 $\pm$ 5.0	32.6
TO	31.6 $\pm$ 1.8	34.3 $\pm$ 4.3	33.7 $\pm$ 2.8	33.2	17.9 $\pm$ 2.0	10.1 $\pm$ 0.1	20.7 $\pm$ 1.9	16.6 $\pm$ 1.8	16.3	80.4 $\pm$ 6.8	51.9 $\pm$ 2.4	44.8 $\pm$ 1.3	32.8 $\pm$ 1.3	52.5
AO	33.3 $\pm$ 0.6	37.6 $\pm$ 5.3	41.9 $\pm$ 2.7	37.6	21.7 $\pm$ 2.2	17.9 $\pm$ 0.8	20.8 $\pm$ 3.7	27.5 $\pm$ 3.4	22.0	80.0 $\pm$ 3.8	38.9 $\pm$ 3.6	55.6 $\pm$ 3.4	43.5 $\pm$ 3.7	54.5
PCD	<b>37.8 <math>\pm</math> 1.5</b>	<b>48.2 <math>\pm</math> 5.4</b>	<b>51.0 <math>\pm</math> 4.8</b>	<b>45.6</b>	<b>23.8 <math>\pm</math> 0.9</b>	<b>18.4 <math>\pm</math> 0.4</b>	<b>30.6 <math>\pm</math> 1.5</b>	<b>27.6 <math>\pm</math> 1.2</b>	<b>25.1</b>	<b>82.1 <math>\pm</math> 4.0</b>	<b>53.4 <math>\pm</math> 4.4</b>	<b>50.8 <math>\pm</math> 4.5</b>	<b>43.6 <math>\pm</math> 4.7</b>	<b>57.5</b>

### A.2 MODEL TRANSFER SETTING

Table 10: Clustering accuracy (%) on Office-31 for different model transfer methods.

Settings	ViT-B/16 (ssl) $\rightarrow$ ResNet-50 (ssl)				ViT-B/16 (ssl) $\rightarrow$ ResNet-50 (sup)				ViT-B/16 (ssl) $\rightarrow$ ResNet-18 (sup)			
	$\mathcal{R} \rightarrow A$	$\mathcal{R} \rightarrow W$	$\mathcal{R} \rightarrow D$	Avg	$\mathcal{R} \rightarrow A$	$\mathcal{R} \rightarrow W$	$\mathcal{R} \rightarrow D$	Avg	$\mathcal{R} \rightarrow A$	$\mathcal{R} \rightarrow W$	$\mathcal{R} \rightarrow D$	Avg
PO	20.1 $\pm$ 0.2/13.5 $\pm$ 0.9	26.7 $\pm$ 0.8/16.7 $\pm$ 0.3	27.2 $\pm$ 0.3/19.3 $\pm$ 2.4	24.7/16.5	20.1 $\pm$ 0.2/15.7 $\pm$ 0.7	26.7 $\pm$ 0.8/24.2 $\pm$ 3.5	27.2 $\pm$ 0.3/18.8 $\pm$ 2.2	24.7/16.6	20.1 $\pm$ 0.2/14.1 $\pm$ 1.6	26.7 $\pm$ 0.8/17.9 $\pm$ 2.0	27.2 $\pm$ 0.3/18.3 $\pm$ 2.9	24.7/16.8
SO	43.2 $\pm$ 5.0	46.4 $\pm$ 4.5	37.2 $\pm$ 9.0	42.3	43.2 $\pm$ 5.0	46.4 $\pm$ 4.5	37.2 $\pm$ 9.0	42.3	43.2 $\pm$ 5.0	46.4 $\pm$ 4.5	37.2 $\pm$ 9.0	42.3 $\pm$ 5.0
TO	32.6 $\pm$ 1.8	34.3 $\pm$ 4.3	33.7 $\pm$ 2.8	33.5	43.7 $\pm$ 1.6	55.8 $\pm$ 2.1	52.0 $\pm$ 3.8	50.5	38.0 $\pm$ 3.2	45.3 $\pm$ 1.6	46.6 $\pm$ 1.5	43.3
AO	50.6 $\pm$ 3.7	49.7 $\pm$ 4.0	36.4 $\pm$ 5.0	45.6	52.5 $\pm$ 3.5	53.7 $\pm$ 1.9	44.2 $\pm$ 4.1	50.1	53.0 $\pm$ 3.8	47.2 $\pm$ 3.3	43.9 $\pm$ 4.9	48.0
PCD	<b>51.7 <math>\pm</math> 2.9</b>	<b>51.7 <math>\pm</math> 2.0</b>	<b>41.8 <math>\pm</math> 3.2</b>	<b>48.4</b>	<b>54.4 <math>\pm</math> 2.4</b>	<b>60.8 <math>\pm</math> 2.1</b>	<b>49.2 <math>\pm</math> 3.3</b>	<b>54.8</b>	<b>54.6 <math>\pm</math> 2.5</b>	<b>53.6 <math>\pm</math> 5.4</b>	<b>46.7 <math>\pm</math> 4.8</b>	<b>51.6</b>

### A.3 LIMITED-DATA AND CLUSTER-IMBALANCED AND SETTING

Due to space constraints, we provide additional results for the sub-sampled versions of all three datasets in the appendix. PCD again outperforms other alternative methods consistently. AO, on

Table 11: Clustering accuracy (%) on sub-sampled versions of different datasets for ResNet-50-based methods (self-supervised pre-training).

Settings	Office-31				Office-Home					PACS				
	$\mathcal{R} \rightarrow \text{sub-A}$	$\mathcal{R} \rightarrow \text{sub-W}$	$\mathcal{R} \rightarrow \text{sub-D}$	Avg	$\mathcal{R} \rightarrow \text{sub-Ar}$	$\mathcal{R} \rightarrow \text{sub-Cl}$	$\mathcal{R} \rightarrow \text{sub-Pr}$	$\mathcal{R} \rightarrow \text{sub-Rw}$	Avg	$\mathcal{R} \rightarrow \text{sub-P}$	$\mathcal{R} \rightarrow \text{sub-A}$	$\mathcal{R} \rightarrow \text{sub-C}$	$\mathcal{R} \rightarrow \text{sub-S}$	Avg
PO	14.6 $\pm$ 1.2	16.7 $\pm$ 1.4	21.5 $\pm$ 1.5	17.6	12.5 $\pm$ 0.8	9.3 $\pm$ 0.8	14.4 $\pm$ 1.1	12.2 $\pm$ 1.3	12.1	43.0 $\pm$ 8.9	32.3 $\pm$ 5.9	29.1 $\pm$ 1.1	39.8 $\pm$ 3.7	36.1
SO	21.1 $\pm$ 3.0	32.5 $\pm$ 2.6	36.0 $\pm$ 3.2	29.9	26.2 $\pm$ 1.3	19.5 $\pm$ 0.3	23.9 $\pm$ 1.0	26.9 $\pm$ 1.1	24.1	54.8 $\pm$ 4.4	37.8 $\pm$ 4.2	41.0 $\pm$ 5.5	47.2 $\pm$ 6.8	45.2
TO	31.4 $\pm$ 3.0	41.9 $\pm$ 3.6	45.1 $\pm$ 3.1	39.5	21.6 $\pm$ 1.8	11.9 $\pm$ 1.0	28.7 $\pm$ 1.3	22.8 $\pm$ 5.3	21.2	65.1 $\pm$ 2.8	46.4 $\pm$ 2.3	47.8 $\pm$ 5.4	40.9 $\pm$ 1.2	50.0
AO	34.7 $\pm$ 3.5	40.8 $\pm$ 3.9	43.9 $\pm$ 3.6	39.8	28.1 $\pm$ 0.6	21.8 $\pm$ 0.6	30.3 $\pm$ 2.6	29.4 $\pm$ 2.1	27.4	65.4 $\pm$ 5.4	43.2 $\pm$ 6.7	51.1 $\pm$ 1.8	43.4 $\pm$ 2.3	50.7
PCD	37.8 $\pm$ 3.6	46.4 $\pm$ 3.3	47.0 $\pm$ 3.7	43.7	28.7 $\pm$ 1.0	22.3 $\pm$ 0.4	32.7 $\pm$ 2.6	31.2 $\pm$ 3.2	28.7	66.1 $\pm$ 4.7	48.0 $\pm$ 1.5	51.8 $\pm$ 2.0	43.4 $\pm$ 1.9	52.4

average, performs better than TO, meaning that knowledge from the source can benefit target training. Similarly, SO improves upon PO in all cases. PCD achieves higher clustering accuracy than AO (1 – 4%), illustrating that target model refinement is crucial for PCD’s success.

#### A.4 ABLATION STUDY

Table 12: Full ablation study on Office-31 dataset.

lightgraylightgray						
Settings	$\mathcal{R} \rightarrow \text{W}$	diff	$\mathcal{R} \rightarrow \text{A}$	diff	$\mathcal{R} \rightarrow \text{D}$	diff
Full	60.0	0.0	46.8	0.0	57.8	0.0
w/o prototype clustering	49.7	−10.3	43.2	−3.6	53.2	−4.6
w/o MI	52.7	−7.3	39.1	−7.7	54.79	−3.01
w/o CutMix	54.5	−5.5	46.1	−0.7	54.6	−3.2
w/o Temporal Ensemble	58.3	−1.7	46.6	−0.2	57.3	−0.5
w/o model privacy	63.1	3.1	49.2	2.4	61.4	3.6
pooled source	57.8	−2.2	45.6	−1.2	57.0	−0.8

Table 13: Clustering accuracy (%) on the task  $\mathcal{R} \rightarrow \text{W}$  (Office-31) under different variants (ResNet-18).

Full	w/o prototype clustering	w/o MI	w/o CutMix	w/o Temporal Ensemble	w/o model privacy	pooled source
60.0 $\pm$ 2.6	49.7 $\pm$ 2.8	52.7 $\pm$ 3.0	54.5 $\pm$ 5.2	58.3 $\pm$ 2.4	63.1 $\pm$ 2.1	57.8 $\pm$ 2.5

## B SENSITIVITY PLOT

In Figure 4, we plot the sensitivity of the target clustering accuracy when we vary the coefficient in front of the loss. We can see that our method is not sensitive to different values of the coefficients except for when the  $\lambda_{mix}$  coefficient is set to 5. This result is expected since the  $\lambda_{mix}$  is used as a regularization term and should not be set too high. We also observe that the performance can get even better via oracle validation by setting the  $\lambda_{mi}$  to 2 or 5. However, we set the coefficient to 1 for all three losses for all experiments.

## C RUNNING TIME AND PARAMETER SIZE

Table 14: Number of parameters and average running time per step for different clustering approaches for ResNet-18-based models.

Methods	Parameter size (millions)	Running time (s/step)
ACIDS	11.94 M	head1 - 0.52 s/step / head2 - 0.44 s/step
PCD	11.32 M	0.16 s/step

## D CONNECTION WITH DEEPCUSTER AND SeLA

Caron et al. (2018) propose DeepCluster to perform clustering and representation learning simultaneously. This method alternates between K-means for clustering and cross-entropy minimization

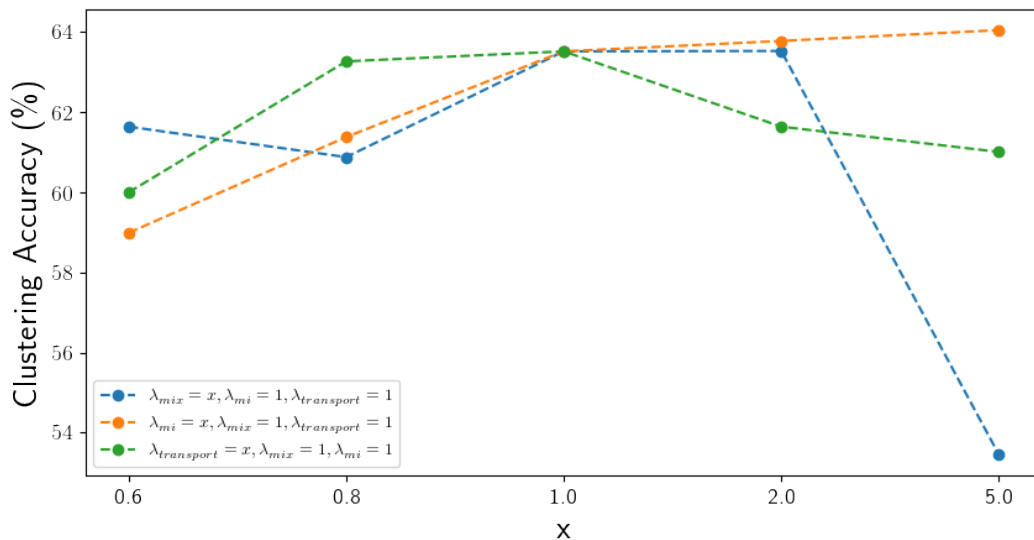


Figure 4: Sensitivity plot for the coefficient of the losses. We fix the coefficient of the two losses to 1.0 while varying the third loss from 0.6 to 5.0 and plot the clustering accuracy on the target data.

Table 15: Average running time per step for different clustering approaches for ResNet-18-based models.

ACIDS	PCD
head1 - 0.52 s/step / head2 - 0.44 s/step	0.16 s/step

for representation learning. While compatible with deep learning frameworks, the approach does have an obvious degenerate solution where all the samples get assigned to one cluster, yielding a constant representation. To overcome this, Asano et al. (2019) invent SeLa, which is similar to DeepCluster in the cross-entropy minimization step but differs from it in the pseudo-label assignment step. The authors explain that solving the K-means problem with equal partitioning constraints can avoid the degenerate solution. Asano et al. (2019) further recognize this as an instance of an optimal transport problem. Our clustering method is similar to SeLa in that we also solve the optimal transport problem during the pseudo-label assignment step. Unlike SeLa, we do not use the simplistic assumption that each cluster contains an equal number of data points. Instead, we dynamically update the cluster proportions using the predicted cluster probabilities. We also offer the interpretation of our method from the distribution alignment perspective. Moreover, our method is designed specifically for multi-domain data, and we also explore the use of our framework under the domain shift scenario.

## E PSEUDO-CODE

## F FULL IMPLEMENTATION DETAILS

We follow the standard protocols for source-free unsupervised domain adaptation (Liang et al., 2020). Specifically, we use mini-batch SGD with a momentum of 0.9 and weight decay of 0.001. Both source and target encoders are initialized with ImageNet pre-trained networks (Russakovsky et al., 2015), but the prototypes are initialized with a random linear layer. The initial learning rates are set to 0.001 for the pre-trained encoders and 0.01 for the randomly initialized layer. The learning rates,  $\eta$ , follows the following schedule:  $\eta = \eta_0(1 + 10p)^{-0.75}$  where  $\eta_0$  is the initial learning rate. We use the batch size of 64 in both source and target learning. The initial value  $\beta_0$  to learn domain-specific proportions is set to 0.9999 for source clustering and 0.99 for target clustering in all settings. We set the entropic regularization parameter,  $\epsilon$ , to 0.01. The concentration parameter,  $\alpha$ , in the CutMix

---

**Algorithm 1:** Pseudo code for our framework.

---

**1. Source model training**

**Input:** source data -  $\mathcal{X}^s = \{\mathcal{X}_d^s\}_{d=1}^D$ , source model -  $G^s = C_{\mu^s}(F_{\theta^s}(\cdot))$  (a randomly initialized  $C_{\mu^s}$  and a pre-trained  $F_{\theta^s}$ )

**Output:** updated  $\theta^s, \mu^s$

**for**  $t = 1$  **to**  $T$  **do**

- Sample a mini-batch of source data
- Update the proportions  $B$  with Eq. (2)
- Solve the optimal transport problem in Eq. (1) to obtain the transport map for each domain
- Update the encoder and prototypes using Eq. (4) with the transport map from the previous step

**end for**

**2. Target model clustering**

**Input:** target data -  $\mathcal{X}^t = \{x_j^t\}_{j=1}^{n_t}$ , cluster labels from the source model -  $G^s(x^t)$ , target model -  $G^t = C_{\mu^t}(F_{\theta^t}(\cdot))$  (a randomly initialized  $C_{\mu^t}$  and a pre-trained  $F_{\theta^t}$ )

**Output:** updated  $\theta^t, \mu^t$

**for**  $t = 1$  **to**  $T$  **do**

- Sample a mini-batch of target data
- Refine the hard-label with label smoothing and temporal ensemble
- Update the the target model with the loss in Eq. (2.3.1)

**end for**

**3. Target model refinement**

**Input:**  $\mathcal{X}^t = \{x_j^t\}_{j=1}^{n_t}$ , target model -  $G^t$  ( $C_{\mu^t}$  and  $F_{\theta^t}$  from step 2's output)

**Output:** updated  $\theta^t, \mu^t$

**for**  $t = 1$  **to**  $T$  **do**

- Sample a mini-batch of target data
- Update the proportions  $B$  with Eq. (2)
- Solve the optimal transport problem in Eq. (1) to obtain the transport map for the target domain
- Update the encoder and prototype using Eq. (2.3.2) with the transport map from the previous step

**end for**

---

loss is set to 0.3. The temporal ensemble coefficient,  $\tau$ , is equal to 0.6. The source model and hyper-parameters are selected using the validation set of the source domain. The target model is trained using all the target data. We run our method with three different random seeds to calculate the standard deviation. We implement our method in PyTorch.