Deep Co-Training with Task Decomposition for Semi-Supervised Domain Adaptation
(Supplementary Material)

Luyu Yang[1], **Yan Wang**[2], **Mingfei Gao**[3], **Abhinav Shrivastava**[1], **Kilian Q. Weinberger**[2], **Wei-Lun Chao**[4], **Ser-Nam Lim**[5]

[1]University of Maryland [2]Cornell University
[3]Salesforce Research [4]Ohio State University [5]Facebook AI

We provide details omitted in the previous sections.

- Appendix A: additional details on experimental setups (cf. section 4 of the main paper).
- Appendix B: additional details on experimental results (cf. section 4 of the main paper).

A. Experimental Setups

In section 5 of the main paper, we compare variants of our approach, including **MIST**, two-view **MIST**, and **DECOTA**. Here we give some more discussions. These three methods are different by 1) how many classifiers they train; 2) what labeled data they use; 3) which classifier provides the pseudo-labels. Fig. 5 gives an illustrative comparison. Fig. 4 illustrates the framework pipeline of **DECOTA**.

- **MIST** learns a single model $\boldsymbol{w}$, using both labeled source data $D_{\mathcal{S}}$ and labeled target data $D_{\mathcal{T}}$. **MIST** also updates $\boldsymbol{w}$ using pseudo-labels on the unlabeled target data $D_{\mathcal{U}}$, where the pseudo-labels are predicted by the current $\boldsymbol{w}$.
- **Two-view MIST** (*i.e.*, two-task **MIST**) learns two models, $\boldsymbol{w}_{\mathcal{T}}$ and $\boldsymbol{w}_{\mathcal{S}}$ (cf. subsection 3.2 of the main paper). $\boldsymbol{w}_{\mathcal{T}}$ is updated using $D_{\mathcal{T}}$ and pseudo-labeled data on $D_{\mathcal{U}}$, where the pseudo-labels are predicted by the current $\boldsymbol{w}_{\mathcal{T}}$. $\boldsymbol{w}_{\mathcal{S}}$ is updated using $D_{\mathcal{S}}$ and pseudo-labeled data on $D_{\mathcal{U}}$, where the pseudo-labels are predicted by the current $\boldsymbol{w}_{\mathcal{S}}$.
- **DECOTA** learns two models, $\boldsymbol{w}_f$ and $\boldsymbol{w}_g$. $\boldsymbol{w}_f$ is updated using $D_{\mathcal{T}}$ and pseudo-labeled data on $D_{\mathcal{U}}$, where the pseudo-labels are predicted by the current $\boldsymbol{w}_g$. $\boldsymbol{w}_g$ is updated using $D_{\mathcal{S}}$ and pseudo-labeled data on $D_{\mathcal{U}}$, where the pseudo-labels are predicted by the current $\boldsymbol{w}_f$.

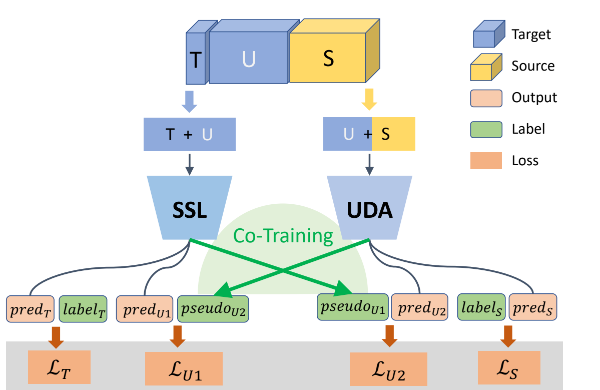

Figure 4: The overall framework of **DECOTA**. It decomposes the SSDA task into SSL and UDA tasks that exchange pseudo-labels for unlabeled target U.

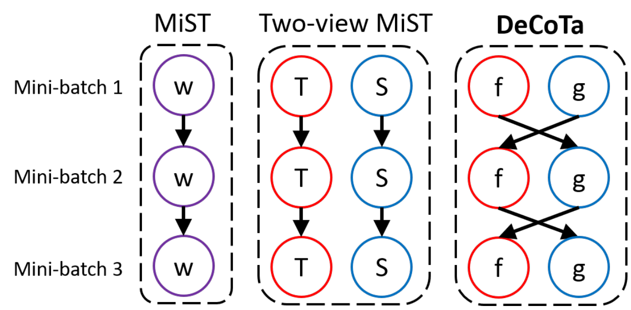

Figure 5: Comparison among **MIST**, two-view **MIST** (*i.e.*, two-task **MIST**), and **DECOTA**. The color on the circles means the labeled data: red for $D_{\mathcal{T}}$, blue for $D_{\mathcal{S}}$, and purple for both. The arrows indicate which model provides the pseudo-labels for which model to learn from.

DECOTA has two hyper-parameters: the confidence threshold τ (cf. Equation 1 of the main paper) and α in MIXUP (cf. Equation 2 of the main paper). We follow [51] to select these hyper-parameters using three other labeled examples per class in the target domain. Specifically, we only select hyper-parameters based on DomainNet three-shot setting, Real to Clipart. We then fix the selected hyper-parameters, $\tau = 0.5$ and $\alpha = 1.0$, for all other experiments.

B. Experimental Results

B.1. Main results on the one-shot setting

We report the comparison with baselines in the one-shot setting on DomainNet in Table 5 and Office-Home in Table 6. **DECOTA** outperforms the state-of-the-art methods by 4.9% on DomainNet (ResNet-34), while performs slightly worse than [45] by 0.6% on Office-Home (VGG-16). Neverthe-

Table 5: Accuracy on DomainNet (%) for the one-shot setting with four domains, using ResNet-34.

Method	R to C	R to P	P to C	C to S	S to P	R to S	P to R	Mean
S+T	58.1	61.8	57.7	51.5	55.4	49.1	73.1	58.1
DANN [13]	61.2	62.3	56.4	54.0	57.9	55.9	65.6	59.0
ENT [51]	60.0	60.2	54.9	48.3	55.8	49.4	74.4	57.6
MME [51]	69.5	68.1	64.4	56.7	62.0	59.2	76.9	65.3
UODA [45]	72.7	70.3	69.8	60.5	66.4	62.7	77.3	68.5
APE [26]	70.4	70.8	72.9	56.7	64.5	63.0	76.6	67.6
ELP [22]	72.8	70.8	72.0	59.6	66.7	63.3	77.8	69.0
DECOTA	**79.1**	**74.9**	**76.9**	**65.1**	**72.0**	**69.7**	**79.6**	**73.9**

Table 6: Accuracy on Office-Home (%) for the one-shot setting with four domains, using VGG-16.

Method	R to C	R to P	R to A	P to R	P to C	P to A	A to P	A to C	A to R	C to R	C to A	C to P	Mean
S+T	39.5	75.3	61.2	71.6	37.0	52.0	63.6	37.5	69.5	64.5	51.4	65.9	57.4
DANN [13]	**52.0**	75.7	62.7	72.7	45.9	51.3	64.3	44.4	68.9	64.2	52.3	65.3	60.0
ENT [51]	23.7	77.5	64.0	74.6	21.3	44.6	66.0	22.4	70.6	62.1	25.1	67.7	51.6
MME [51]	49.1	78.7	65.1	74.4	46.2	56.0	68.6	45.8	72.2	68.0	57.5	71.3	62.7
UODA [45]	49.6	79.8	**66.1**	75.4	45.5	**58.8**	**72.5**	43.3	**73.3**	70.5	**59.3**	72.1	**63.9**
ELP [22]	49.2	79.7	65.5	75.3	46.7	56.3	69.0	**46.1**	72.4	68.2	67.4	71.6	63.1
DECOTA	47.2	**80.3**	64.6	**75.5**	**47.2**	56.6	71.1	42.5	73.1	**71.0**	57.8	**72.9**	63.3

less, **DECOTA** attains the highest accuracy on 5 adaptation scenarios of Office-Home in the one-shot setting.

B.2. Office-Home results on other backbones

We report the comparison with baselines on Office-Home using a ResNet-34 backbone in Table 7, following [26][3]. **DECOTA** attains the state-of-the-art result.

B.3. Results on Office-31

We report the comparison with available baseline results on Office-31 [50] in Table 8, using ResNet-34 backbone. Following [51], two adaptation scenarios are compared (Webcam to Amazon, DSLR to Amazon). Our approach DECOTA consistently outperforms the compared methods.

B.4. Larger-shot results

We provide *10,20,50*-shot SSDA results on DomainNet in Table 9. We randomly select and add additional samples per class from the target domain to the target labeled pool. As a semi-supervised setting, we compared with both domain adaptation (DA) and semi-supervised learning (SSL) baselines [59]. The implementation details are the same as those of *1,3*-shot. DECOTA improves along with more shots and can outperform baselines.

[3] Most existing papers only reported Office-Home results using VGG-16. We followed [26] to further report ResNet-34. Some algorithms reported in Table 3 are missing in Table 7 since they do not release code.

B.5. Numbers and accuracy of pseudo-labels

We showed the number of total and correct pseudo-labels by the two classifiers of **DECOTA** along the training iterations in Figure 3 (c) of the main paper. The analysis is on DomainNet three-shot setting, from Real to Clipart. Concretely, for every $1K$ iterations (*i.e.*, 24K unlabeled data), we accumulated the number of unlabeled data that have confident (with confidence $> \tau = 0.5$) and correct predictions by at least one classifier. We further plot them independently for each classifier (*i.e.*, $\boldsymbol{w}_f$ and $\boldsymbol{w}_g$) in Fig. 6. The accuracy of pseudo-labels remains stable (*i.e.*, the number of confident and correct predictions divided by the number of confident predictions) but the number increases along training.

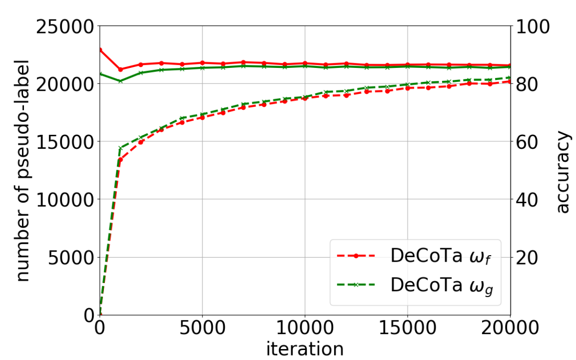

Figure 6: Number (dashed, left) and accuracy (solid, right) of pseudo-labels on DomainNet three-shot setting, Real to Clipart.

Table 7: Accuracy on Office-Home (%) for the three-shot setting with four domains, using ResNet-34.

Method	R to C	R to P	R to A	P to R	P to C	P to A	A to P	A to C	A to R	C to R	C to A	C to P	Mean
S+T	55.7	80.8	67.8	73.1	53.8	63.5	73.1	54.0	74.2	68.3	57.6	72.3	66.2
DANN [13]	57.3	75.5	65.2	69.2	51.8	56.6	68.3	54.7	73.8	67.1	55.1	67.5	63.5
ENT [51]	62.6	85.7	70.2	79.9	60.5	63.9	79.5	61.3	79.1	76.4	64.7	79.1	71.9
MME [51]	64.6	85.5	71.3	80.1	64.6	65.5	79.0	63.6	79.7	76.6	67.2	79.3	73.1
APE [26]	66.4	86.2	73.4	82.0	65.2	66.1	81.1	63.9	80.2	76.8	66.6	79.9	74.0
DECOTA	**70.4**	**87.7**	**74.0**	**82.1**	**68.0**	**69.9**	**81.8**	**64.0**	**80.5**	**79.0**	**68.0**	**83.2**	**75.7**

Table 8: SSDA results on Office-31, on two scenarios (following [51]).

Method	Webcam (W) to Amazon (A)		DSLR (D) to Amazon (A)	
	1-shot	3-shot	1-shot	3-shot
S+T	69.2	73.2	68.2	73.3
DANN [13]	69.3	75.4	70.4	74.6
ENT [51]	69.1	75.4	72.1	75.1
MME [51]	73.1	76.3	73.6	77.6
Ours	**76.0**	**76.8**	**74.2**	**78.3**

B.6. Task decomposition

We report the comparison of **DECOTA** and **MIST** on DomainNet and Office-Home in all the adaptation scenarios. As shown in Table 10, **DECOTA** outperform **MIST** on all the setting by $1 \sim 2\%$ on DomainNet and $3 \sim 5\%$ on Office-Home, which further confirms the effectiveness of task decomposition — explicitly considering the discrepancy between the two sources of supervision — in **DECOTA**.

B.7. One-direction training

We further consider another variant of **DECOTA** named **one-direction teaching**, in which only one task teaches the other. Instead of co-training, we use either $\boldsymbol{w}_f$ or $\boldsymbol{w}_g$ to generate pseudo-labels for both tasks[4], while keeping the other setups the same as **DECOTA**. This study is designed to measure the complementary specialties of the two tasks. As shown in Table 11, the performance drops notably by using one-direction teaching. The results suggest that the two tasks provide unique expertise and complement each other, instead of one dominating the other.

B.8. Results on the source domain

We report the results on the source domain test set using $\boldsymbol{w}_f$ and $\boldsymbol{w}_g$ of **DECOTA** on DomainNet (three-shot) in Table 12. While $\boldsymbol{w}_f$ and $\boldsymbol{w}_g$ have similar accuracy on the target domain test set, the fact that $\boldsymbol{w}_f$ does not learn from $D_\mathcal{S}$ suggests their difference in classifying source domain data. Table 12 confirms this: we see that $\boldsymbol{w}_g$ clearly dominates $\boldsymbol{w}_f$. Its accuracy is even on a par with a model trained only on $D_\mathcal{S}$, showing one advantage of DECOTA— the model can keep its discriminative ability on the source domain.

[4]That is, **one-direction teaching** constructs both pseudo-label sets, *i.e.*, $U^{(f)}$ and $U^{(g)}$ in Equation 1 of the main text, by the same model (we hence have two versions, $\boldsymbol{w}_f$ teaching or $\boldsymbol{w}_g$ teaching).

B.9. Sensitivity to the confidence threshold τ

We investigate **DECOTA**'s sensitivity to the confidence threshold τ for assigning pseudo-labels (cf. Equation 1 and Equation 4 of the main paper). As shown in Fig. 7, the variance in accuracy is small when $\tau \leq 0.7$. The accuracy drops notably when $\tau \geq 0.9$. We surmise that it is due to too few pseudo-labeled data are picked under a high threshold.

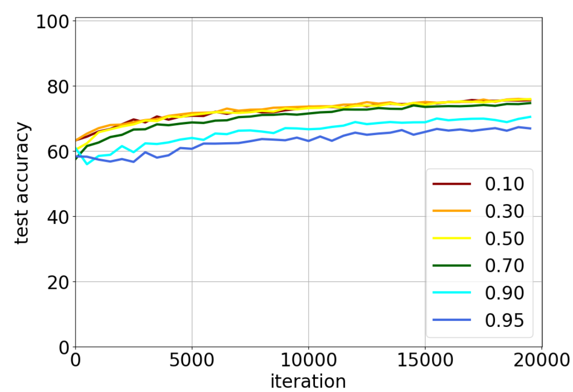

Figure 7: **DECOTA**'s sensitivity to pseudo-label threshold τ on DomainNet three-shot setting, Real to Clipart.

B.10. Analysis on the Beta distribution coefficient α

Fig. 8 shows **DECOTA**'s sensitivity to the MIXUP hyper-parameter α in Equation 2 of the main paper: α is the coefficient of the Beta distribution, which influences the sampled value of λ, an indicator of the "propotion" in the

MIXUP algorithm. We report **DECOTA**'s result on DomainNet three-shot setting, adapting from Real to Clipart. The best performance is achieved by $\alpha = 1.0$, equivalent to a uniform distribution of $\lambda \in [0, 1]$. This result is consistent with our hypothesis that MIXUP connects the source and target domains with interpolated feature spaces in-between.

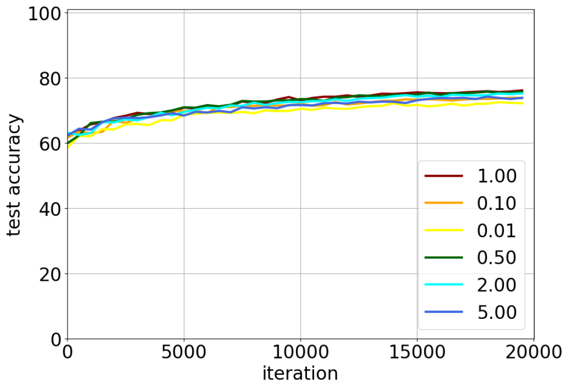

Figure 8: **DECOTA**'s sensitivity to the Beta distribution coefficient α on DomainNet three-shot setting, Real to Clipart.

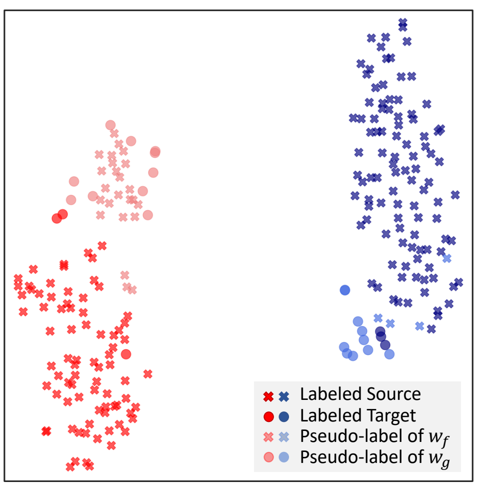

Figure 9: t-SNE visualization of pseudo-labels assigned by $\boldsymbol{w}_f$ and $\boldsymbol{w}_g$ in **DECOTA** (see text for details).

B.11. Training time

DECOTA does not increase the training time much for two reasons. First, at each iteration (*i.e.*, mini-batch), it only updates and learns from the pseudo-labels of *the current mini-batch* of unlabeled data, not the entire unlabeled data. Second, assigning pseudo-labels only requires a forward pass of the mini-batch, just like most domain adaptation algorithms normally do to compute training losses. The only difference is that DECOTA trains two classifiers and needs to perform the forward pass of unlabeled data twice.

B.12. t-SNE visualizations on DECOTA tasks

We visualize $D_{\mathcal{S}}$, $D_{\mathcal{T}}$, and the $D_{\mathcal{U}}$ pseudo-labels by each task of **DECOTA** in Fig. 9. For clarity, we select two classes for illustration. The colors blue and red represent the two classes; the shapes circle and cross represent data from $D_{\mathcal{T}}$ (labeled target data) and $D_{\mathcal{S}}$ (labeled source data), respectively. The colors light blue and light red represent the pseudo-labels of each class on $D_{\mathcal{U}}$, in which the shape circle indicates that the pseudo-labels are provided by $\boldsymbol{w}_f$ (learned with $D_{\mathcal{T}}$) and the shape cross indicates that the pseudo-labels are provided by $\boldsymbol{w}_g$ (learned with $D_{\mathcal{S}}$). The visualization is based on DomainNet three-shot setting, from Real to Clipart, trained for $10,000$ iterations. We see that $\boldsymbol{w}_f$ tends to assign pseudo-labels to unlabeled data whose features are closer to $D_{\mathcal{T}}$; $\boldsymbol{w}_g$ tends to assign pseudo-labels to unlabeled data whose features are closer to $D_{\mathcal{S}}$. Such a behavior is aligned with the seminal work of semi-supervised learning by [77].

Table 9: Results on DomainNet at *10, 20, 50*-shot, using ResNet-34. We tune hyper-parameters for SSL methods similarly to DA methods.

	R to C			R to P			P to C			C to S			S to P			R to S			P to R			Mean		
n-shot $\rightarrow$	10	20	50	10	20	50	10	20	50	10	20	50	10	20	50	10	20	50	10	20	50	10	20	50
S+T	69.1	72.4	77.5	67.3	70.2	73.4	68.2	72.5	77.7	62.9	67.3	71.8	64.8	67.9	72.6	61.3	65.5	70.2	78.0	79.3	82.2	67.4	70.7	75.1
DANN [13]	66.2	68.0	71.1	65.1	67.1	69.0	62.4	64.5	68.2	60.0	62.4	66.8	61.3	63.8	67.6	61.4	63.2	66.9	71.6	74.7	78.1	64.0	66.2	69.7
ENT [51]	77.9	80.0	83.0	72.3	74.9	77.7	77.5	79.1	82.3	66.3	70.1	75.0	66.3	71.0	75.7	63.9	68.3	74.6	**81.2**	**82.9**	84.5	72.2	75.2	79.0
MME [51]	77.0	78.5	80.9	71.9	74.0	76.4	75.6	76.9	80.4	65.9	68.6	72.5	68.6	70.9	74.4	66.7	69.7	72.7	80.8	82.2	83.3	72.4	74.4	77.2
Mixup [76]	73.4	79.5	83.1	68.3	72.2	75.4	75.0	79.5	83.1	63.7	69.4	75.0	68.5	72.4	76.2	62.9	69.9	75.0	78.8	82.3	**84.7**	70.1	75.0	78.9
FixMatch [59]	76.6	79.5	82.3	73.0	74.7	76.4	75.8	79.4	83.3	70.1	73.1	76.9	71.3	73.3	77.0	68.7	71.6	74.2	79.7	81.9	84.2	73.6	76.2	79.2
DeCoTa	**81.8**	**82.6**	**85.0**	**75.1**	**76.6**	**78.7**	**81.3**	**81.7**	**84.5**	**73.7**	**75.3**	**78.0**	**73.4**	**75.7**	**77.7**	**73.7**	**75.5**	**77.8**	80.7	80.1	83.9	**77.1**	**78.2**	**80.8**

Table 10: Comparison between **DeCoTa** and **MiST**: test accuracy on DomainNet and Office-Home dataset (%).

(a) DomainNet

Setting	Method	R to C	R to P	P to C	C to S	S to P	R to S	P to R	Mean
1-shot	**MiST**	74.8	73.6	74.5	65.0	72.0	67.0	77.6	72.1
	DeCoTa	79.1	74.9	76.9	65.1	72.0	69.7	79.6	73.9
3-shot	**MiST**	78.1	75.2	76.7	68.3	72.6	71.5	79.8	74.6
	DeCoTa	80.4	75.2	78.7	68.6	72.7	71.9	81.5	75.6

(b) Office-Home

Setting	Method	R to C	R to P	R to A	P to R	P to C	P to A	A to P	A to C	A to R	C to R	C to A	C to P	Mean
1-shot	**MiST**	42.7	77.5	62.9	73.1	39.4	54.8	67.1	40.0	66.9	67.9	56.8	69.4	59.9
	DeCoTa	47.2	80.3	64.6	75.5	47.2	56.6	71.1	42.5	73.1	71.0	57.8	72.9	63.3
3-shot	**MiST**	54.7	81.2	64.0	69.4	51.7	58.8	69.1	47.6	70.6	65.3	60.8	73.8	63.9
	DeCoTa	59.9	83.9	67.7	77.3	57.7	60.7	78.0	54.9	76.0	74.3	63.2	78.4	69.3

Table 11: Comparison between **DeCoTa** and **one-direction teaching**: accuracy on DomainNet (%) three-shot setting.

Method	R to C	R to P	P to C	C to S	S to P	R to S	P to R	Mean
$\boldsymbol{w}_f$ teaching	73.8	67.2	73.7	63.1	65.9	61.7	78.2	69.1
$\boldsymbol{w}_g$ teaching	77.5	74.5	74.2	64.8	71.6	69.0	79.0	72.9
DeCoTa	80.4	75.2	78.7	68.6	72.7	71.9	81.5	75.6

Table 12: Comparison on the source domain test data of DomainNet (%). Here we compare the two-task models of **DeCoTa** in the three-shot setting to the source-only model (S).

Method	R to C	R to P	P to C	C to S	S to P	R to S	P to R	Mean
$\boldsymbol{w}_f$	55.2	68.2	43.8	59.5	50.8	56.9	61.0	56.3
$\boldsymbol{w}_g$	97.2	97.1	99.3	98.7	98.9	96.8	99.4	98.2
S	98.1	98.2	99.5	98.9	99.2	98.2	99.6	98.8