Prior Knowledge Guided Unsupervised Domain Adaptation Supplementary Material

Tao Sun[1], Cheng Lu[2], and Haibin Ling[1]

[1]Stony Brook University [2]XPeng Motors
{tao,hling}@cs.stonybrook.edu, luc@xiaopeng.com

A Combining Two Types of Prior Knowledge

In the paper, we considered two types of prior knowledge, *Unary Bound* and *Binary Relationship*. The two knowledge may have some overlapping. For example, it is possible to infer BR from UB when the unary bounds are tight, and vice versa. Nevertheless, when the bounds are not tight, one knowledge may provide complementary information for the other. Table A.1 lists results when combining UB and BR together in kSHOT on VisDA-2017. UB($\sigma = 0.5$)+BR performs slightly better than both UB($\sigma = 0.5$) and BR. UB($\sigma = 1.0$)+BR is on par with BR as UB($\sigma = 1.0$) is not informative.

Table A.1. Classification accuracies (%) on **VisDA-2017**.

Method	$\mathcal{K}$	σ	aero.	bike	bus	car	horse	knife	moto.	pers.	plant	sktb.	train	truck	Avg.
SHOT	–	–	94.3	88.5	80.1	57.3	93.1	94.9	80.7	80.3	91.5	89.1	86.3	58.2	82.9
kSHOT	UB	0.0	95.7	88.7	**81.4**	**73.4**	94.7	94.2	**88.1**	82.5	93.4	91.1	87.2	**63.1**	**86.1**
	UB	0.1	96.1	**90.2**	80.7	71.5	**96.0**	91.3	85.7	83.5	**94.5**	**91.3**	87.1	61.5	85.8
	UB	0.5	95.2	89.6	79.7	59.6	94.8	90.7	82.0	**86.2**	92.7	90.2	86.8	59.8	83.9
	UB	1.0	94.8	88.3	79.1	56.8	93.8	92.8	80.6	82.7	91.0	90.9	86.2	59.0	83.0
	UB	2.0	94.6	87.7	78.9	55.9	93.4	94.8	80.2	81.4	89.3	89.9	86.1	58.6	82.6
	BR	–	**96.3**	89.2	79.7	58.0	94.2	92.7	81.1	81.1	92.2	90.9	**88.7**	59.2	83.6
	UB+BR	0.5	95.8	89.1	81.1	60.2	95.1	91.5	84.3	82.7	93.4	91.4	88.7	59.8	84.4
	UB+BR	1.0	96.2	89.1	79.7	58.0	94.2	92.6	81.1	81.1	92.2	90.9	88.7	59.3	83.6

B Visualization of Standard Deviations

For all tables and figures in the paper, we report the mean evaluation results of three repeated experiments with different random seeds. Figure A.1 visualizes the standard deviations on Office-Home of comparison methods. As can be seen, the performances are stable to different initializations.

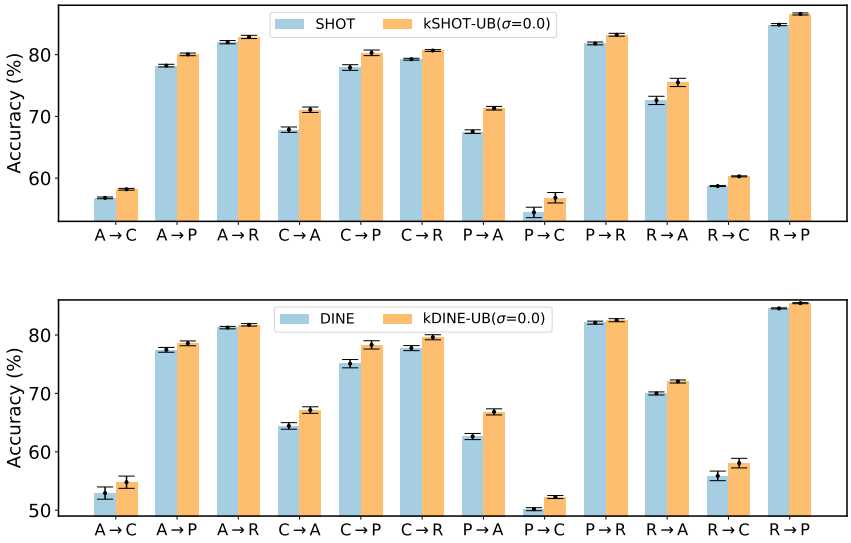

Fig. A.1. Visualization of standard deviations on Office-Home.

Table A.2. Classification accuracies (%) on **Office-Home** for partial-set DA.

Method	$\mathcal{K}$	σ	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg.
SHOT	–	–	64.8	85.2	92.7	76.3	77.6	88.8	79.7	64.3	89.5	80.6	66.4	85.8	79.3
kSHOT	UB	0.0	**74.1**	**94.4**	**94.3**	**84.3**	**93.1**	**93.0**	**85.3**	**73.4**	**93.5**	**86.7**	**74.7**	**95.0**	**86.8**
	BR	–	72.2	92.9	92.8	82.3	89.8	90.9	83.6	69.6	92.6	86.0	71.7	93.3	84.8
DINE	–	–	58.1	83.4	89.2	78.0	80.0	80.6	74.2	56.6	85.9	80.6	62.9	84.8	76.2
DINE*	–	–	54.9	80.8	87.3	70.3	75.2	78.8	70.9	51.2	85.7	78.1	58.3	84.1	73.0
kDINE	UB	0.0	**65.5**	**91.4**	**92.3**	**80.2**	**89.3**	**91.2**	**81.6**	**64.4**	**91.6**	**84.5**	**69.3**	**93.2**	**82.9**
	BR	–	62.5	89.2	91.1	77.3	85.0	87.2	78.5	60.3	90.3	83.4	67.1	90.3	80.2

C More Detailed Results

Table A.2 presents the detailed results on Office-Home for partial-set DA. Accuracies per class on VisDA-2017 are listed in Tab. A.1.

D Effects of Label Smoothing in kDINE

DINE [1] distillates knowledge from the source predictor to a target model. As mentioned in [1], using label smoothing for the teacher probability is superior to using one-hot encoding. To show its effect in kDINE, we compare a variant of

Table A.3. Classification accuracies (%) on **Office-Home**.

Method	$\mathcal{K}$	σ	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg.
DINE	–	–	52.2	78.4	81.3	65.3	76.6	78.7	62.7	49.6	82.2	69.8	55.8	84.2	69.7
DINE*	–	–	51.8	76.0	79.6	63.1	75.1	76.5	60.4	48.5	80.7	69.4	55.9	83.5	68.4
DINE**	–	–	51.3	75.4	79.2	62.7	74.6	75.8	59.8	48.1	80.2	68.9	55.7	83.1	67.9
kDINE	UB	0.0	54.8	78.6	**81.7**	**67.1**	78.3	**79.6**	**66.8**	**52.3**	82.5	**72.0**	**58.1**	**85.4**	**71.4**
	UB	0.1	**55.0**	78.8	81.1	66.4	77.7	79.2	66.4	51.8	82.3	71.5	58.0	84.9	71.1
	UB	0.5	52.9	76.7	79.9	64.5	76.3	77.8	63.8	51.0	80.9	70.5	57.1	84.2	69.6
	UB	1.0	52.3	76.0	79.6	63.5	75.2	76.5	62.1	49.0	80.7	69.9	56.4	83.4	68.7
	UB	2.0	51.8	76.0	79.6	63.0	75.1	76.5	60.8	49.2	80.7	69.6	55.5	83.5	68.4
	BR	–	54.2	**79.4**	81.5	66.8	**78.6**	79.2	65.6	50.9	**82.6**	71.4	**58.1**	85.3	71.1
kDINE*	UB	0.0	54.8	78.4	81.4	66.8	77.6	79.2	67.0	51.6	82.4	71.8	58.0	85.2	71.2
	UB	0.1	54.4	78.0	80.9	66.5	76.9	78.9	66.0	51.1	82.1	71.4	57.7	84.7	70.7
	UB	0.5	52.9	76.2	79.5	64.4	75.7	77.3	63.2	50.6	80.5	70.3	56.7	84.0	69.3
	UB	1.0	52.1	75.4	79.2	63.2	74.9	75.9	61.6	48.9	80.3	69.8	56.0	83.2	68.4
	UB	2.0	51.7	75.4	79.2	62.7	74.6	75.8	60.6	48.9	80.2	69.4	55.5	83.1	68.1
	BR	–	54.2	78.7	81.4	66.2	78.1	78.6	65.0	50.9	82.3	71.1	57.5	85.0	70.7

kDINE without label smoothing. The objective function is

$$\mathcal{L}^*_{\text{kdine}} = \mathbb{E}_{\boldsymbol{x}_i^t} \mathcal{D}_{\text{kl}} \left(\frac{P^{\text{tch}}(\boldsymbol{x}_i^t) + \boldsymbol{l}_i^{(\text{pk}_1)}}{2} \middle\| f_t(\boldsymbol{x}_i^t) \right) + \beta \mathcal{L}_{\text{mix}} - \mathcal{L}_{\text{im}} \tag{A.1}$$

Compared with Eq. 10 of the paper, the smoothed label $\tilde{\boldsymbol{l}}_i^{(\text{pk}_1)}$ is replaced with the one-hot label $\boldsymbol{l}_i^{(\text{pk}_1)}$. We term this variant as kDINE*. For a fair comparison with DINE, we also replace $\boldsymbol{l}_i^{(\text{pk}_1)}$ with $\boldsymbol{l}_i^{(\text{dine})}$ in Eq. A.1 and term it as DINE**. Table A.3 shows that using one-hot labels indeed degrades the performance in both DINE** and kDINE*. Nevertheless, with class prior knowledge, kDINE* still achieves much better accuracies than DINE**. This verifies the effectiveness of considering prior knowledge and our proposed rectification module.

E Analysis of Prior Knowledge in kDINE

In Fig. 5 of the paper, we show that prior knowledge rectifies distribution of pseudo labels in kSHOT. To see how prior knowledge is helpful in kDINE, let us consider the K-L divergences between the mean teacher probability and the ground-truth target class distribution. The divergences for DINE and kDINE are $\mathcal{D}_{\text{kl}} \left(\mathbb{E}_{\boldsymbol{x}_i^t} \left[P^{\text{tch}}(\boldsymbol{x}_i^t) \right] \middle\| p_t(y) \right)$ and $\mathcal{D}_{\text{kl}} \left(\mathbb{E}_{\boldsymbol{x}_i^t} \left[\frac{P^{\text{tch}}(\boldsymbol{x}_i^t) + \tilde{\boldsymbol{l}}_i^{(\text{pk}_1)}}{2} \right] \middle\| p_t(y) \right)$, respectively. Divergences for DINE*, DINE** and kDINE* can be similarly defined. Figure A.2 plots these divergences on two domain adaption tasks at a step of generating teacher probability. Clearly using prior knowledge leads to much smaller divergences, which may benefit the distillation stage.

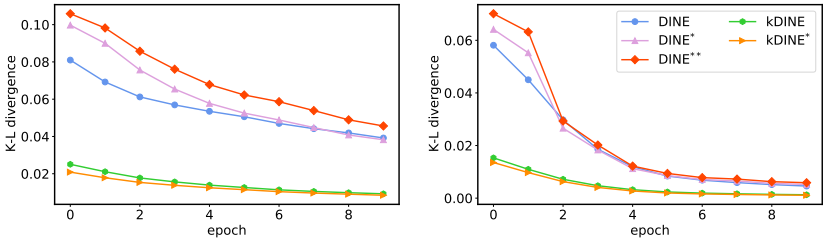

Fig. A.2. K-L divergences between ground-truth target class distribution and mean teacher probabilities during training on (left) Office-Home P→A and (right) Office A→W. (See text for details.)

References

1. Liang, J., Hu, D., Feng, J., He, R.: Dine: Domain adaptation from single and multiple black-box predictors. In: CVPR. pp. 8003–8013 (2022)