ADU: Adaptive Detection of Unknown Categories in Black-Box Domain Adaptation

Supplementary Material

A. Proof of Eq. (3)

In Section 3.1, we introduce the SAKD loss as an enhancement to the traditional KD loss, specifically tailored for black box domain adaptation with unkonwn classes in the target domain. The SAKD loss, as defined in Eq. (15), aims to mitigate the impact of noisy pseudo-labels by preferentially amplifying the influence of high-confidence pseudo-labels:

$$\mathcal{L}_{SAKD} = -\mathbb{E}_{x_t \sim \mathcal{X}_t}[(1+p_s^{\hat{c}})^{\theta} \log p_t^{\hat{c}} + \sum_{\substack{c=1 \\ c \neq \hat{c}}}^{|L_s|} p_s^c \log p_t^c], \tag{15}$$

where $\theta \geq 1$.

For simplicity, we consider the case with a single sample, where the SAKD loss simplifies to:

$$\mathcal{L}_{\text{SAKD}} = -[(1+p_s^{\hat{c}})^{\theta} \log p_t^{\hat{c}} + \sum_{c=1, c \neq \hat{c}}^{|L_s|} p_s^c \log p_t^c]. \tag{16}$$

Next, we apply the generalized binomial theorem to expand $\left(1+p_s^{\hat{c}}\right)^{\theta}$ as follows:

$$\begin{aligned} \mathcal{L}_{\text{SAKD}} &= -\left[\sum_{k=0}^{\infty} \binom{\theta}{k} \left(p_s^{\hat{c}}\right)^k \log p_t^{\hat{c}} + \sum_{c=1, c \neq \hat{c}}^{|L_s|} p_s^c \log p_t^c\right] \\ &= \mathcal{L}_{\text{KD}} - \left[\sum_{k=1}^{\infty} \binom{\theta}{k} \left(p_s^{\hat{c}}\right)^k - p_s^{\hat{c}}\right] \log p_t^{\hat{c}} \\ &\approx \mathcal{L}_{\text{KD}} - \left[(\theta-1)\, p_s^{\hat{c}} + \frac{\theta\,(\theta-1)}{2} \left(p_s^{\hat{c}}\right)^2\right] \log p_t^{\hat{c}}. \end{aligned} \tag{17}$$

In Eq. (17), the first term $\mathcal{L}_{\text{KD}}$ represents the original KD loss, while the second term introduces an additional term, which is positive and solely depends on the target class $\hat{c}$.

We define the additional term as Δ, and we will prove its function in the following steps.

We assume that the weight of the loss function is denoted by ω, and the number of training iterations is represented by i. In our analysis, we adopt the following widely recognized assumptions:

Assumption 1 (L-smooth): The SAKD loss function $\mathcal{L}_{SAKD}$ is L-smooth, which means for all ω, the following inequality holds:

$$\begin{aligned} \mathcal{L}_{SAKD}(\omega_{i+1}) \leq \mathcal{L}_{SAKD}(\omega_i) + \langle \nabla \mathcal{L}_{SAKD}(\omega_i), \omega_{i+1} - \omega_i \rangle \\ + \frac{\beta}{2} \|\omega_{i+1} - w_i\|^2, \end{aligned} \tag{18}$$

where $\nabla \mathcal{L}_{SAKD}(\omega_i)$ represents the gradient of the loss function at iteration i, and $\beta > 0$ is the smoothness constant.

Assumption 2 (Bounded Gradients): The gradient of the SAKD loss function $\mathcal{L}_{SAKD}$ is bounded. Formally, there exists a constant $G > 0$ such that:

$$\|\nabla \mathcal{L}_{SAKD}(\omega_i)\| \leq G, \quad \forall i. \tag{19}$$

Assumption 3 (μ-smooth): The SAKD loss function $\mathcal{L}_{SAKD}$ is μ-smooth, for all ω:

$$\begin{aligned} \mathcal{L}_{SAKD}(\omega_{i+1}) \geq \mathcal{L}_{SAKD}(\omega_i) + \langle \nabla \mathcal{L}_{SAKD}(\omega_i), \omega_{i+1} - \omega_i \rangle \\ + \frac{\mu}{2} \|\omega_{i+1} - w_i\|^2, \end{aligned} \tag{20}$$

where $\mu > 0$ is a constant.

During the training process, all values in Δ are fixed, except for $p_t^{\hat{c}}$. Additionally, since Assumption 3 is holds, the derivative of $p_t^{\hat{c}}$ with respect to i is positive, i.e.,

$$\frac{d(p_t^{\hat{c}})}{di} > 0. \tag{21}$$

Now, we proceed to compute the derivative of Δ with respect to i. By applying the chain rule, we obtain:

$$\frac{d\Delta}{di} = -\left[(\theta-1)\, p_s^{\hat{c}} + \frac{\theta\,(\theta-1)}{2} \left(p_s^{\hat{c}}\right)^2\right] \frac{1}{p_t^{\hat{c}}} \frac{d(p_t^{\hat{c}})}{di}. \tag{22}$$

Given that $\theta > 1$, it follows that: $\delta = \frac{d\Delta}{di} < 0$.

Next, we compute the derivative of δ with respect to $p_s^{\hat{c}}$:

$$\frac{d\delta}{dp_s^{\hat{c}}} = -\left[(\theta-1) + \theta\,(\theta-1)\, p_s^{\hat{c}}\right] \frac{1}{p_t^{\hat{c}}} \frac{d(p_t^{\hat{c}})}{di} < 0 \tag{23}$$

.

Thus, as $p_s^{\hat{c}}$ increases, δ decreases, and the absolute value of δ increases, which means that the gradient of Δ decreases at a faster rate. In other words, as the confidence of the pseudo-label increases, the gradient descent speed of Δ becomes faster.

Table 6. Ablation Study on OfficeHome. H-score (%) of different variants in OPBDA scenarios. $\mathcal{L}^1_{HQ}$, $\mathcal{L}^2_{HQ}$, $\mathcal{L}_{LQ}$ refer to the objectives corresponding to the negative loss in $\mathcal{L}_{HQ}$, entropy loss in $\mathcal{L}_{HQ}$, and loss associated with low-quality labels, respectively.

$\mathcal{L}^1_{HQ}$	$\mathcal{L}^2_{HQ}$	$\mathcal{L}_{LQ}$	Ar→Cl	Ar→Pr	Ar→Re	Cl→Ar	Cl→Pr	Cl→Re	Pr→Ar	Pr→Cl	Pr→Re	Re→Ar	Re→Cl	Re→Pr	Avg.
-	-	-	61.2	68.9	73.1	67.2	67.0	72.9	71.5	57.9	81.0	70.2	61.0	70.2	68.5
✓	-	-	61.2	70.4	74.5	69.5	68.7	73.9	72.0	60.8	82.2	71.4	62.7	71.1	69.9
-	✓	-	61.7	70.8	74.8	69.8	69.6	74.3	72.1	61.7	82.0	71.0	62.5	71.4	70.1
-	-	✓	61.4	71.2	74.9	69.9	69.4	74.1	72.6	60.7	82.8	71.2	62.6	71.9	70.2
✓	✓	-	62.7	69.7	74.8	69.4	68.7	73.7	71.7	60.8	81.5	71.4	61.3	71.2	69.7
✓	-	✓	**63.3**	71.4	74.6	69.9	69.8	74.7	72.3	61.9	82.1	71.5	62.6	72.2	70.5
-	✓	✓	62.6	71.5	75.7	70.0	70.6	74.6	72.1	61.8	82.6	71.9	63.1	72.3	70.7
✓	✓	✓	61.2	**72.7**	**77.9**	**70.3**	**72.5**	**77.3**	**75.9**	**62.0**	**84.7**	**73.2**	**64.1**	**74.9**	**72.2**

Table 7. Ablation Study on Office31. H-score (%) of different variants in OPBDA scenarios.

$\mathcal{L}^1_{HQ}$	$\mathcal{L}^2_{HQ}$	$\mathcal{L}_{LQ}$	A → D	A → W	D→A	D →W	W → A	W →D	Avg.
-	-	-	82.1	82.9	84.0	92.6	80.2	85.4	84.5
✓	-	-	83.2	84.3	85.6	93.5	82.6	89.4	86.4
-	✓	-	82.8	83.5	86.5	92.7	83.0	87.9	86.1
-	-	✓	83.5	84.1	86.0	93.8	83.6	85.8	86.1
✓	✓	-	80.3	84.3	86.0	94.1	83.3	88.7	86.1
✓	-	✓	84.6	83.1	86.3	94.2	83.3	89.2	86.8
-	✓	✓	85.8	85.1	86.1	94.0	83.5	88.8	87.2
✓	✓	✓	**87.5**	**85.2**	**87.0**	**94.4**	**83.8**	**90.5**	**88.1**

In conclusion, we demonstrate that this additional term enables the SAKD loss to prioritize pseudo-labels with high confidence, while mitigating the influence of noisy pseudo-labels. This mechanism improves the robustness of the model during training, ensuring that more reliable information is captured from pseudo-labels with higher confidence.

B. Additional experiment results

Ablation study. In our study, we perform ablation study on two datasets and measure H-score to illustrate the impact of each component of our method. However, due to space constraints, we only present results of six tasks in the paper. Hence, in this section, we display additional ablation study results in Table 6 and 7. It is important to emphasize that in all ablation experiments, we consistently employ the SAKD loss, which is a critical component of the ADU framework. The same as the ablation study results of six tasks, we can still draw the following conclusions: (i) The introduction of any component alongside the SAKD loss leads to performance improvements, underscoring the vital role of the EDLD module. (ii) The full EDLD loss, which includes the negative loss term, yields better performance compared to its version without the negative loss, demonstrating the effectiveness of incorporating this term. (iii) The integration of all components results in the highest H-scores, providing clear evidence of the synergy and efficacy of the combined modules.

Table 8. Sensitivity analysis of λ.

λ	A → W	D → W	Ar → Cl	Cl → Re	Pr →Ar	Re → Cl	Avg.
0.0	82.9	92.6	61.2	72.9	71.5	61.0	73.7
0.2	84.4	93.5	62.5	73.4	71.5	62.8	74.7
0.5	83.2	93.7	**62.8**	73.7	71.7	63.4	74.8
1.0	**85.2**	**94.4**	61.2	**77.3**	**75.9**	**64.1**	**76.3**
2.0	83.8	93.1	61.9	73.3	71.0	61.6	74.1
5.0	85.8	93.5	60.5	71.6	70.6	60.4	73.7

Table 9. Sensitivity analysis of θ.

θ	A → W	D → W	Ar → Cl	Cl → Re	Pr →Ar	Re → Cl	Avg.
1.00	83.6	93.5	**63.7**	76.9	73.3	63.6	75.8
1.05	84.6	93.6	63.0	**77.3**	73.6	63.6	76.0
1.10	**85.2**	**94.4**	61.2	**77.3**	**75.9**	**64.1**	**76.3**
1.15	84.2	92.2	61.0	75.7	72.4	63.0	74.8
1.20	82.2	91.0	59.4	70.6	67.3	58.5	71.5
1.25	80.2	90.5	58.2	69.5	66.0	58.4	70.5

Table 10. Sensitivity analysis of γ.

γ	A → W	D → W	Ar → Cl	Cl → Re	Pr →Ar	Re → Cl	Avg.
0.0	84.7	91.0	63.2	73.5	70.9	62.5	74.3
0.2	**85.2**	91.9	62.4	73.4	70.2	62.2	74.2
0.4	84.9	93.4	**62.5**	73.9	72.7	62.9	75.1
0.6	**85.2**	94.4	61.2	**77.3**	**75.9**	**64.1**	**76.3**
0.8	82.8	**95.1**	60.6	74.1	71.7	61.5	74.3
1.0	79.0	94.3	59.8	73.9	69.8	59.7	72.8

Parameters sensitivity analysis. To illustrate the sensitivity of our method to the hyperparameters λ, θ and γ, we conduct experiments on two tasks from the Office31 dataset (A→W and D→W) and four tasks of the Office-Home dataset (Ar→Cl, Cl→Re, Pr→Ar and Re→Cl). Due to space constraints, we only present the average values in the paper. Hence, in this section, we display additional results in Table 8, 9 and 10. It is evident that the results are stable around the selected values of $\lambda = 1.0$, $\theta = 1.1$, and $\gamma = 0.6$.