

LLMs were used solely for aiding and polishing the writing. The methodology, theoretical analysis, and experimental details are entirely described and developed by the authors in the paper.

A ERROR BOUND FOR IMBALANCED DOMAIN GENERALIZATION

We build upon the domain adaptation bound established in (Ben-David et al., 2010, Thm. 2), which relates the target error to the source error and a distributional discrepancy term. Specifically, for any hypothesis $h \in \mathcal{H}$, the target domain risk $\epsilon_T(h)$ is upper-bounded as:

$$\epsilon_T(h) \leq \epsilon_S(h) + d_{\mathcal{H}\Delta\mathcal{H}}(P_S(\mathbf{X}), P_T(\mathbf{X})) + \lambda_{\mathcal{H}} \quad (6)$$

where $\epsilon_S(h)$ is the source risk, $d_{\mathcal{H}\Delta\mathcal{H}}(P_S(\mathbf{X}), P_T(\mathbf{X}))$ is the $\mathcal{H}\Delta\mathcal{H}$ -divergence measuring the discrepancy between input distributions, and $\lambda_{\mathcal{H}} := \inf_{h \in \mathcal{H}} [\epsilon_S(h) + \epsilon_T(h)]$ denotes the error of the ideal joint hypothesis over both domains. While this bound has been influential in understanding domain shift, it primarily focuses on aligning marginal input distributions $P(\mathbf{X})$ and subsequent variants of this bound have been widely used in existing domain generalization theory analysis Albuquerque et al. (2019); Lu et al. (2024).

However, in IDG, aligning these marginal input distributions $P(\mathbf{X})$ may fail to guarantee effective transfer when label distributions $P(\mathbf{Y})$ vary significantly across domains. Even under invariant class-conditional distributions across domains $P(\mathbf{X}|\mathbf{Y})$, a shift in the label prior $P(\mathbf{Y})$ can lead to a substantial change in the posterior $P(\mathbf{Y}|\mathbf{X})$ via Bayesian rule. Consequently, the classifier trained on the source domain may fail to generalize to the target domain, despite exhibiting a low $\mathcal{H}\Delta\mathcal{H}$ -divergence. To illustrate this limitation, we present concrete examples where input alignment alone is insufficient for generalization. Consider a binary classification task with classes $\mathbf{Y} \in \{0, 1\}$, corresponding to two categories. Assume that the class-conditional distributions are identical across the source and target domains:

$$\begin{aligned} P_S(\mathbf{X}|\mathbf{Y}=0) &= P_T(\mathbf{X}|\mathbf{Y}=0) = \mathcal{N}([-1, 0], \mathbf{I}) \\ P_S(\mathbf{X}|\mathbf{Y}=1) &= P_T(\mathbf{X}|\mathbf{Y}=1) = \mathcal{N}([1, 0], \mathbf{I}) \end{aligned}$$

but the class priors differ significantly:

$$P_S(\mathbf{Y}=0) = 0.1, P_S(\mathbf{Y}=1) = 0.9; \quad P_T(\mathbf{Y}=0) = 0.9, P_T(\mathbf{Y}=1) = 0.1$$

Under these conditions, the resulting marginal input distributions are:

$$\begin{aligned} P_S(\mathbf{X}) &= 0.1 \cdot \mathcal{N}([-1, 0], \mathbf{I}) + 0.9 \cdot \mathcal{N}([1, 0], \mathbf{I}), \\ P_T(\mathbf{X}) &= 0.9 \cdot \mathcal{N}([-1, 0], \mathbf{I}) + 0.1 \cdot \mathcal{N}([1, 0], \mathbf{I}) \end{aligned}$$

which clearly diverge due to the shift in label priors, despite the shared class-conditional structure. This example highlights that marginal alignment of input features is insufficient for generalization when posterior distributions are influenced by label imbalance.

Consequently, it is more appropriate to measure the divergence between the joint distributions $P(\mathbf{X}, \mathbf{Y})$ across domains, rather than focusing solely on the marginal distributions $P(\mathbf{X})$.

Step 1: Refining the error bound in Eq. (6) with the joint distributions. To better capture prediction-induced discrepancies under distribution shift, we reformulate the classical discrepancy $d_{\mathcal{H}\Delta\mathcal{H}}(P_S(\mathbf{X}), P_T(\mathbf{X}))$ from Ben-David et al. (2010), where $h : \mathbf{X} \rightarrow \{0, 1\} \in \mathcal{H}$, into a form defined over the mappings $m = yh(x) : \mathbf{X} \times \mathbf{Y} \rightarrow \{-1, 1\} \in \mathcal{M}$, with $h : \mathbf{X} \rightarrow \{0, 1\}$ and $y \in \mathbf{Y} = \{-1, 1\}$. We then define a new discrepancy $d_{\mathcal{M}\Delta\mathcal{M}}$ over the auxiliary class \mathcal{M} , which acts directly on this composed formulation:

$$\Delta := \sup_{m, m'} |\mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim P_S(\mathbf{X}, \mathbf{Y})} [I[m(\mathbf{X}, \mathbf{Y}) \neq m'(\mathbf{X}, \mathbf{Y})]] - \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim P_T(\mathbf{X}, \mathbf{Y})} [I[m(\mathbf{X}, \mathbf{Y}) \neq m'(\mathbf{X}, \mathbf{Y})]]|$$

Notably, this transformation does not alter the underlying hypothesis \mathcal{H} class but instead reorganizes the evaluation space to emphasize decision-relevant characteristics.

In this way, we reformulate their error bound as follows:

$$|\epsilon_T(h) - \epsilon_S(h)| \leq d_{\mathcal{M}\Delta\mathcal{M}}(P_S(\mathbf{X}, \mathbf{Y}), P_T(\mathbf{X}, \mathbf{Y})) \quad (7)$$

Lemma 1 Let $P_S(\mathbf{X}, \mathbf{Y})$ and $P_T(\mathbf{X}, \mathbf{Y})$ be joint distributions from the source and target domains. Then the $\mathcal{M}\Delta\mathcal{M}$ -divergence between these joint distributions can be upper bounded by:

$$\begin{aligned} d_{\mathcal{M}\Delta\mathcal{M}}(P_S(\mathbf{X}, \mathbf{Y}), P_T(\mathbf{X}, \mathbf{Y})) &\leq d_{\mathcal{M}\Delta\mathcal{M}}^{P_S(\mathbf{Y}|\mathbf{X})}(P_S(\mathbf{X}), P_T(\mathbf{X})) \\ &\quad + \mathbb{E}_{x \sim P_T(\mathbf{X})} [d_{\mathcal{M}\Delta\mathcal{M}}(P_S(\mathbf{Y}|\mathbf{X}=x), P_T(\mathbf{Y}|\mathbf{X}=x))] \end{aligned}$$

where the non-standard $d_{\mathcal{M}\Delta\mathcal{M}}^{P_S(\mathbf{Y}|\mathbf{X})}(P_S(\mathbf{X}), P_T(\mathbf{X}))$ can be defined as:

$$\sup_{m, m'} \left| \int (P_S(\mathbf{X}) - P_T(\mathbf{X})) \mathbb{E}_{y \sim P_S(\mathbf{Y}|\mathbf{X})} [I[m(\mathbf{X}, \mathbf{Y}) \neq m'(\mathbf{X}, \mathbf{Y})]] dx \right|$$

which denotes a posterior-weighted symmetric difference divergence between the source and target marginal distributions and measures the discrepancy between $P_S(\mathbf{X})$ and $P_T(\mathbf{X})$ under the influence of a fixed source posterior $P_S(\mathbf{Y}|\mathbf{X})$.

Proof 1 By the chain rule of the joint distribution, we have:

$$P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{X}) \cdot P(\mathbf{Y}|\mathbf{X})$$

Consider the difference between two joint distributions, where $I[\cdot]$ denotes the indicator and the discrepancy between the hypothesis pairs $m, m' \in \mathcal{M}$ can be regarded as:

$$\Delta := \sup_{m, m'} \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim P_S(\mathbf{X}, \mathbf{Y})} [I[m(\mathbf{X}, \mathbf{Y}) \neq m'(\mathbf{X}, \mathbf{Y})]] - \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim P_T(\mathbf{X}, \mathbf{Y})} [I[m(\mathbf{X}, \mathbf{Y}) \neq m'(\mathbf{X}, \mathbf{Y})]]$$

Then, the chain rule of distributions can be applied as follows:

$$\Delta = \sup \left| \int P_S(\mathbf{X}) P_S(\mathbf{Y}|\mathbf{X}) I dx dy - \int P_T(\mathbf{X}) P_T(\mathbf{Y}|\mathbf{X}) I dx dy \right|$$

where I is short for $I[m(\mathbf{X}, \mathbf{Y}) \neq m'(\mathbf{X}, \mathbf{Y})]$. By introducing the term involving $P_T(\mathbf{X}) P_S(\mathbf{Y}|\mathbf{X})$, and applying the triangle inequality, we obtain:

$$\Delta \leq \left| \int (P_S(\mathbf{X}) - P_T(\mathbf{X})) \mathbb{E}_{y \sim P_S(\mathbf{Y}|\mathbf{X})} [I] dx \right| + \left| \int P_T(\mathbf{X}) (\mathbb{E}_{y \sim P_S(\mathbf{Y}|\mathbf{X})} [I] - \mathbb{E}_{y \sim P_T(\mathbf{Y}|\mathbf{X})} [I]) dx dy \right|$$

So that, we can obtain:

$$d_{\mathcal{M}\Delta\mathcal{M}}(P_S(\mathbf{X}, \mathbf{Y}), P_T(\mathbf{X}, \mathbf{Y})) \leq d_{\mathcal{M}\Delta\mathcal{M}}^{P_S(\mathbf{Y}|\mathbf{X})}(P_S(\mathbf{X}), P_T(\mathbf{X})) + \mathbb{E}_{x \sim P_T(\mathbf{X})} [d_{\mathcal{M}\Delta\mathcal{M}}(P_S(\mathbf{Y}|\mathbf{X} = x), P_T(\mathbf{Y}|\mathbf{X} = x))] \quad \square$$

Theorem 2 (Generalization Bound with the Joint Distributions) For any hypothesis $h \in \mathcal{H}$ with an auxiliary class \mathcal{M} where $h : \mathbf{X} \rightarrow \{-1, 1\} \in \mathcal{H}$ and $m = yh(x) : \mathbf{X} \times \mathbf{Y} \rightarrow \{-1, 1\} \in \mathcal{M}$, the target domain risk can be bounded by:

$$\epsilon_T(h) \leq \epsilon_S(h) + d_{\mathcal{M}\Delta\mathcal{M}}^{P_S(\mathbf{Y}|\mathbf{X})}(P_S(\mathbf{X}), P_T(\mathbf{X})) + \mathbb{E}_{x \sim P_T(\mathbf{X})} [d_{\mathcal{M}\Delta\mathcal{M}}(P_S(\mathbf{Y}|\mathbf{X} = x), P_T(\mathbf{Y}|\mathbf{X} = x))] + \lambda_{\mathcal{H}}$$

where $\lambda_{\mathcal{H}} := \inf_{h \in \mathcal{H}} [\epsilon_S(h) + \epsilon_T(h)]$. On the one hand, while retaining the original marginal discrepancy term $d_{\mathcal{M}\Delta\mathcal{M}}^{P_S(\mathbf{Y}|\mathbf{X})}(P_S(\mathbf{X}), P_T(\mathbf{X}))$, this extended bound enhances the classical formulation by incorporating a conditional divergence term that captures shifts in $P(\mathbf{Y}|\mathbf{X})$, which are typically overlooked by input alignment alone. Moreover, based on Bayes rule $P(\mathbf{Y}|\mathbf{X}) = \frac{P(\mathbf{X}|\mathbf{Y})P(\mathbf{Y})}{P(\mathbf{X})}$, aligning the posteriors $P(\mathbf{Y}|\mathbf{X})$ can effectively mitigate the influence of the label prior $P(\mathbf{Y})$ on the decision function. Even in scenarios where $P(\mathbf{Y})$ varies significantly across domains, maintaining consistency in $P(\mathbf{Y}|\mathbf{X})$ helps preserve a stable and generalizable decision boundary. On the other hand, $d_{\mathcal{M}\Delta\mathcal{M}}^{P_S(\mathbf{Y}|\mathbf{X})}(P_S(\mathbf{X}), P_T(\mathbf{X}))$ characterizes the alignment of the marginal distributions under the source conditional distribution $P_S(\mathbf{Y}|\mathbf{X})$. In other words, aligning only the marginals without taking the conditional distributions into account may distort the original representation structure, thereby weakening the discriminative information encoded in the feature space. Such distortions can in turn destabilize the decision boundary and hinder its ability to generalize across domains. Therefore, ensuring posterior consistency becomes crucial.

Step 2: Extending to multiple source domains

Theorem 3 (Generalization Bound with the Joint Distributions on multiple source domains)

For any hypothesis $h \in \mathcal{H}$ with an auxiliary class \mathcal{M} , the target domain risk can be defined based on a linear mixture across multiple source domains:

$$\epsilon_T(h) \leq \sum_{i=1}^N \pi_i \epsilon_S^i(h) + \underbrace{\zeta_{\text{in}} + \zeta_{\text{out}}}_{\text{prior distribution discrepancy}} + \underbrace{\eta_{\text{in}} + \eta_{\text{out}}}_{\text{posterior distribution discrepancy}} + \lambda_{\pi}$$

where $\sum_{i=1}^N \pi_i = 1$, $\lambda_\pi = \inf_{h \in \mathcal{H}} \left[\sum_{i=1}^N \pi_i \epsilon_S^i(h), \epsilon_T(h) \right]$. $\zeta_{\text{in}}/\eta_{\text{in}}$ denotes the maximum \mathcal{H} -divergence between any pair of the prior/posterior distributions $P(\mathbf{X})/P(\mathbf{Y}|\mathbf{X} = x)$. $\zeta_{\text{out}} = d_{\mathcal{M}\Delta\mathcal{M}}^{P_S(\mathbf{Y}|\mathbf{X})} \left(\sum_{i=1}^N \pi_i P_S^i(\mathbf{X}), P_T(\mathbf{X}) \right)$ and $\eta_{\text{out}} = d_{\mathcal{M}\Delta\mathcal{M}} \left(\sum_{i=1}^N \pi_i P_S^i(\mathbf{Y}|\mathbf{X} = x), P_T(\mathbf{Y}|\mathbf{X} = x) \right)$.

Proof 2

Remark 1 (Bounding the \mathcal{H} -divergence between domains in the convex hull Albuquerque et al. (2019))

Let a set S of source domains such that $|S| = N$ be denoted by $P_S^i, i \in [N]$. The convex hull Λ_S of S is defined as the set of mixture distributions given by: $\Lambda_S = \left\{ \bar{P} := \bar{P}(\cdot) = \sum_{i=1}^N \pi_i P_S^i(\cdot), \pi_i \in \Delta_N \right\}$. Let $d_{\mathcal{H}\Delta\mathcal{H}}(P_S^i(\mathbf{X}), P_S^j(\mathbf{X})) \leq \epsilon, \forall i, j \in [N]$. The following holds for the \mathcal{H} -divergence between any pair of domains $P', P'' \in \Lambda_S^2$:

$$d_{\mathcal{H}\Delta\mathcal{H}}(P', P'') \leq \epsilon$$

According to this remark, Albuquerque et al. (2019) further introduce unseen target \bar{P}_T , the element of Λ_S which is closest to P_T , i.e., \bar{P}_T is given by $\arg \min_{\pi_1, \dots, \pi_N} d_{\mathcal{H}\Delta\mathcal{H}}(P_T, \sum_{i=1}^N \pi_i P_S^i)$.

Along this line, we extend this divergence analysis to the joint distributions $P_i(\mathbf{X}, \mathbf{Y}), \forall i \in [N]$, in order to capture both covariate and label shifts. Specifically, the convex hull Λ_S can be defined as $\Lambda_S = \left\{ \bar{P} := \bar{P}(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^N \pi_i P_S^i(\mathbf{X}, \mathbf{Y}), \pi_i \in \Delta_N \right\}$. This definition is well-motivated, as under the two-stage sampling perspective Cao & Chen (2024), one first samples a joint distribution from the real-world domain mixture, and then draws specific input-label pairs from the selected joint distribution.

Consequently, Eq. (7) can be reformulated in terms of a linear mixture with Lemma 1:

$$\begin{aligned} \sum_{i=1}^N \pi_i \epsilon_T(h) &\leq \sum_{i=1}^N \pi_i \left[\epsilon_S^i(h) + d_{\mathcal{M}\Delta\mathcal{M}}(P_S^i(\mathbf{X}, \mathbf{Y}), P_T(\mathbf{X}, \mathbf{Y})) \right] + \lambda_\pi \\ &\leq \sum_{i=1}^N \pi_i \left[\epsilon_S^i(h) + d_{\mathcal{M}\Delta\mathcal{M}}(P_S^i(\mathbf{X}, \mathbf{Y}), \bar{P}_T(\mathbf{X}, \mathbf{Y})) \right. \\ &\quad \left. + d_{\mathcal{M}\Delta\mathcal{M}}(\bar{P}_T(\mathbf{X}, \mathbf{Y}), P_T(\mathbf{X}, \mathbf{Y})) \right] + \lambda_\pi \\ &\leq \sum_{i=1}^N \pi_i \left[\epsilon_S^i(h) + d_{\mathcal{M}\Delta\mathcal{M}}^{P_S^i(\mathbf{X})} + \mathbb{E}_{x \sim P_T(\mathbf{X})} d_{\mathcal{M}\Delta\mathcal{M}}^{P_S^i(\mathbf{Y}|\mathbf{X}=x)} \right. \\ &\quad \left. + d_{\mathcal{M}\Delta\mathcal{M}}^{P_T(\mathbf{X})} + \mathbb{E}_{x \sim P_T(\mathbf{X})} d_{\mathcal{M}\Delta\mathcal{M}}^{P_T(\mathbf{Y}|\mathbf{X}=x)} \right] + \lambda_\pi \\ &\leq \sum_{i=1}^N \pi_i \epsilon_S^i(h) + \zeta_{\text{in}} + \zeta_{\text{out}} + \eta_{\text{in}} + \eta_{\text{out}} + \lambda_\pi \end{aligned}$$

where the last inequality is based on Remark 1, and $d_{\mathcal{M}\Delta\mathcal{M}}^{P(\mathbf{X})} = d_{\mathcal{M}\Delta\mathcal{M}}^{P(\mathbf{Y}|\mathbf{X})}(P(\mathbf{X}), \bar{P}_T(\mathbf{X}))$, $d_{\mathcal{M}\Delta\mathcal{M}}^{P(\mathbf{Y}|\mathbf{X}=x)} = d_{\mathcal{M}\Delta\mathcal{M}}(P(\mathbf{Y}|\mathbf{X} = x), \bar{P}_T(\mathbf{Y}|\mathbf{X} = x))$. \square

Step 3: Introducing a structural regularization term via margin surrogate error. Note that the standard risk $\epsilon_S(h)$ measures the empirical error on the source domain. However, when $P(\mathbf{Y})$ is imbalanced, majority classes dominate the optimization of the empirical loss, which biases the decision boundary toward them. As a result, the margin around minority samples shrinks, reducing classification confidence and increasing error on rare classes. To address this, we introduce a convex surrogate error Yuan et al. (2021) on the source domain, denoted by $\gamma(h)$, defined as:

$$\gamma(h) := \phi \left(h_y(x) - \max_{k \neq y} h_k(x) \right)$$

where $\phi(\cdot)$ denotes a margin surrogate function, such as $\phi(z) = \max(0, -z)$ or $\log(1 + e^z)$. This term encourages the model to maintain a sufficiently large margin between classes.

Consequently, $\epsilon_S(h)$ can be decomposed through margin $\gamma(h)$ as follows:

$$\begin{aligned}\epsilon_S(h) &= \mathbb{E}_{(x,y) \sim P_S(\mathbf{X}, \mathbf{Y})} \ell(h(x), y) \\ &= \underbrace{\mathbb{E}_{(x,y) \sim P_S(\mathbf{X}, \mathbf{Y})} [\ell(h(x), y) \cdot 1_{\gamma(h) \leq \delta}]}_{\text{small-margin loss}} + \underbrace{\mathbb{E}_{(x,y) \sim P_S(\mathbf{X}, \mathbf{Y})} [\ell(h(x), y) \cdot 1_{\gamma(h) > \delta}]}_{\text{large-margin loss}} \\ &\leq \ell_{max} \cdot \Pr[\gamma(h) \leq \delta] + \mathbb{E}_{(x,y) \sim P_S(\mathbf{X}, \mathbf{Y})} [\ell(h(x), y) \cdot 1_{\gamma(h) > \delta}] \\ &\leq \ell_{max} \cdot \Pr[\gamma(h) \leq \delta] + \epsilon_S\end{aligned}$$

where δ is a margin threshold, ℓ_{max} denotes the upper bound of the loss function, and $\Pr[\gamma(h) \leq \delta]$ quantifies the proportion of samples with margin less than or equal to the threshold δ .

Notably, this inequality holds under the assumption that the loss function is monotonically decreasing with respect to the margin $\gamma(h)$, whose property can be satisfied by commonly used surrogate losses, such as cross-entropy loss.

Combined with the above, the final upper bound can be defined as:

$$\begin{aligned}\epsilon_T(h) &\leq \sum_{i=1}^N \pi_i (\epsilon_S^i(h) + \ell_{max}^i \cdot \Pr[\gamma^i(h) \leq \delta]) + \zeta_{in} + \zeta_{out} + \eta_{in} + \eta_{out} + \lambda_\pi \\ &\leq \sum_{i=1}^N \pi_i \epsilon_S^i(h) + \ell_{max} \cdot \Pr[\gamma(h) \leq \delta] + \zeta_{in} + \zeta_{out} + \eta_{in} + \eta_{out} + \lambda_\pi\end{aligned}$$

where ℓ_{max} denotes the upper bound of the loss function across all source domains, and $\gamma(h)$ denotes the expected margin over the source domains.

B GRADIENT DERIVATION

Our proposed InfoNCE-like contrastive loss can be formulated as follows:

$$\mathcal{L}_{con}^{\text{InfoNCE-ND}} = \sum_{i \in \mathcal{I}} -\log \left\{ \frac{1}{|\mathcal{N}(i)|} \sum_{n \in \mathcal{N}(i)} \frac{1 - s(\mathbf{p}_i, \mathbf{p}_n)}{\sum_{a \in \mathcal{A}(i)} (1 - s(\mathbf{p}_i, \mathbf{p}_a))} \right\}$$

In analogy, SupCon-like contrastive loss can be formulated as follows:

$$\mathcal{L}_{con}^{\text{SupCon-ND}} = \sum_{i \in \mathcal{I}} -\frac{1}{|\mathcal{N}(i)|} \sum_{n \in \mathcal{N}(i)} \log \left\{ \frac{1 - s(\mathbf{p}_i, \mathbf{p}_n)}{\sum_{a \in \mathcal{A}(i)} (1 - s(\mathbf{p}_i, \mathbf{p}_a))} \right\}$$

And, classical InfoNCE contrastive loss dominated by positives can be formulated as follows:

$$\mathcal{L}_{con}^{\text{InfoNCE}} = \sum_{i \in \mathcal{I}} -\log \left\{ \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \frac{s(\mathbf{p}_i, \mathbf{p}_p)}{\sum_{a \in \mathcal{A}(i)} s(\mathbf{p}_i, \mathbf{p}_a)} \right\}$$

Let $f_{ip} = s(\mathbf{p}_i, \mathbf{p}_p)$, $Z_i = \sum_{a \in \mathcal{A}(i)} s(\mathbf{p}_i, \mathbf{p}_a)$, and $Q_i = \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \frac{f_{ip}}{Z_i}$. Then, the derivation of $\mathcal{L}_{con}^{\text{InfoNCE}}$ w.r.t. \mathbf{p}_i can be expressed as:

$$\begin{aligned}\frac{\partial \mathcal{L}_{con}^{\text{InfoNCE}}}{\partial \mathbf{p}_i} &= \nabla_{\mathbf{p}_i} \mathcal{L}_{con}^{\text{InfoNCE}} = -\frac{1}{Q_i} \cdot \nabla_{\mathbf{p}_i} Q_i \\ &= -\frac{1}{Q_i} \cdot \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} \left(\frac{f_{ip}}{Z_i} \right) \\ &\stackrel{\text{Quotient rule}}{=} -\frac{1}{Q_i} \cdot \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \frac{\nabla_{\mathbf{p}_i} f_{ip} \cdot Z_i - f_{ip} \cdot \nabla_{\mathbf{p}_i} Z_i}{Z_i^2} \\ &= -\frac{1}{Q_i} \cdot \frac{1}{|\mathcal{P}(i)|} \cdot \frac{1}{Z_i^2} \sum_{p \in \mathcal{P}(i)} [\nabla_{\mathbf{p}_i} f_{ip} \cdot Z_i - f_{ip} \cdot \nabla_{\mathbf{p}_i} Z_i]\end{aligned}$$

Next, we organize the final summation terms, and let $\Psi = \sum_{p \in \mathcal{P}(i)} [\nabla_{\mathbf{p}_i} f_{ip} \cdot Z_i - f_{ip} \cdot \nabla_{\mathbf{p}_i} Z_i]$ and $f_{in} = s(\mathbf{p}_i, \mathbf{p}_n)$:

$$\begin{aligned}
\Psi &= \sum_{p \in \mathcal{P}(i)} \left[\nabla_{\mathbf{p}_i} f_{ip} \cdot Z_i - f_{ip} \cdot \left(\sum_{p' \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} f_{ip'} + \sum_{n \in \mathcal{N}(i)} \nabla_{\mathbf{p}_i} f_{in} \right) \right] \\
&= \sum_{p \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} f_{ip} \cdot Z_i - \sum_{p \in \mathcal{P}(i)} \sum_{p' \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} f_{ip'} \cdot f_{ip} - \sum_{p \in \mathcal{P}(i)} \sum_{n \in \mathcal{N}(i)} \nabla_{\mathbf{p}_i} f_{in} \cdot f_{ip} \\
&= \sum_{p \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} f_{ip} \cdot Z_i - \sum_{p' \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} f_{ip'} \cdot \sum_{p \in \mathcal{P}(i)} f_{ip} - \sum_{n \in \mathcal{N}(i)} \nabla_{\mathbf{p}_i} f_{in} \cdot \sum_{p \in \mathcal{P}(i)} f_{ip} \\
&= \sum_{p \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} f_{ip} \cdot Z_i - \sum_{p \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} f_{ip} \cdot \sum_{p \in \mathcal{P}(i)} f_{ip} - \sum_{n \in \mathcal{N}(i)} \nabla_{\mathbf{p}_i} f_{in} \cdot \sum_{p \in \mathcal{P}(i)} f_{ip} \\
&= - \left(\sum_{p \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} f_{ip} \cdot \left(\sum_{p \in \mathcal{P}(i)} f_{ip} - Z_i \right) + \sum_{n \in \mathcal{N}(i)} \nabla_{\mathbf{p}_i} f_{in} \cdot \sum_{p \in \mathcal{P}(i)} f_{ip} \right) \\
&= - \left(- \sum_{p \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} f_{ip} \cdot \sum_{n \in \mathcal{N}(i)} f_{in} + \sum_{n \in \mathcal{N}(i)} \nabla_{\mathbf{p}_i} f_{in} \cdot \sum_{p \in \mathcal{P}(i)} f_{ip} \right)
\end{aligned}$$

Therefore, we can obtain:

$$\begin{aligned}
\frac{\partial \mathcal{L}_{con}^{\text{InfoNCE}}}{\partial \mathbf{p}_i} &= -\frac{1}{Q_i} \cdot \frac{1}{|\mathcal{P}(i)|} \cdot \frac{1}{Z_i^2} \left[- \left(- \sum_{p \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} f_{ip} \cdot \sum_{n \in \mathcal{N}(i)} f_{in} + \sum_{n \in \mathcal{N}(i)} \nabla_{\mathbf{p}_i} f_{in} \cdot \sum_{p \in \mathcal{P}(i)} f_{ip} \right) \right] \\
&= \frac{1}{Q_i} \cdot \frac{1}{|\mathcal{P}(i)|} \cdot \frac{1}{Z_i} \left[- \sum_{p \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} f_{ip} \cdot \frac{\sum_{n \in \mathcal{N}(i)} f_{in}}{Z_i} + \sum_{n \in \mathcal{N}(i)} \nabla_{\mathbf{p}_i} f_{in} \cdot \frac{\sum_{p \in \mathcal{P}(i)} f_{ip}}{Z_i} \right] \\
&= \frac{1}{Z_i} \left[- \sum_{p \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} f_{ip} \cdot \frac{\sum_{n \in \mathcal{N}(i)} f_{in}}{\sum_{p \in \mathcal{P}(i)} f_{ip}} + \sum_{n \in \mathcal{N}(i)} \nabla_{\mathbf{p}_i} f_{in} \right] \\
&= \frac{1}{Z_i} \left[- \sum_{p \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} s(\mathbf{p}_i, \mathbf{p}_p) \cdot \frac{\sum_{n \in \mathcal{N}(i)} f_{in}}{\sum_{p \in \mathcal{P}(i)} f_{ip}} + \sum_{n \in \mathcal{N}(i)} \nabla_{\mathbf{p}_i} s(\mathbf{p}_i, \mathbf{p}_n) \right]
\end{aligned}$$

Similar to $\mathcal{L}_{con}^{\text{InfoNCE}}$, the gradient of $\mathcal{L}_{con}^{\text{InfoNCE-ND}}$ can be expressed in terms of the notations $f'_{ip} = 1 - s(\mathbf{p}_i, \mathbf{p}_p)$, $f'_{in} = 1 - s(\mathbf{p}_i, \mathbf{p}_n)$, $Z'_i = \sum_{a \in \mathcal{A}(i)} (1 - s(\mathbf{p}_i, \mathbf{p}_a))$, and $Q'_i = \frac{1}{|\mathcal{N}(i)|} \sum_{n \in \mathcal{N}(i)} \frac{f'_{in}}{Z'_i}$:

$$\begin{aligned}
\frac{\partial \mathcal{L}_{con}^{\text{InfoNCE-ND}}}{\partial \mathbf{p}_i} &= \frac{1}{Z'_i} \left[\sum_{p \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} f'_{ip} - \sum_{n \in \mathcal{N}(i)} \nabla_{\mathbf{p}_i} f'_{in} \cdot \frac{\sum_{p \in \mathcal{P}(i)} f'_{ip}}{\sum_{n \in \mathcal{N}(i)} f'_{in}} \right] \\
&= \frac{1}{Z'_i} \left[- \sum_{p \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} s(\mathbf{p}_i, \mathbf{p}_p) + \sum_{n \in \mathcal{N}(i)} \nabla_{\mathbf{p}_i} s(\mathbf{p}_i, \mathbf{p}_n) \cdot \frac{\sum_{p \in \mathcal{P}(i)} f'_{ip}}{\sum_{n \in \mathcal{N}(i)} f'_{in}} \right]
\end{aligned}$$

These two gradient formulations exhibit fundamentally different optimization focuses, each expressed as a term modulated by a weighted coefficient. $\partial \mathcal{L}_{con}^{\text{InfoNCE}} / \partial \mathbf{p}_i$ adaptively emphasizes positive pair tightening by amplifying gradients for hard positives with low similarity or surrounded by similar negatives, while applying uniform repulsion to all negatives. In contrast, $\mathcal{L}_{con}^{\text{InfoNCE-ND}}$ prioritizes hard negative repulsion by adaptively scaling negative gradients according to overall similarity, while applying uniform attraction to all positives. This contrast reveals a fundamental shift: **from enforcing positive compactness to enhancing negative separation**. This distinction is particularly relevant in the context of class imbalance. Under label shift, minority-class

samples tend to be absorbed by majority-class clusters, requiring stronger emphasis on negative separation. $\mathcal{L}_{con}^{\text{InfoNCE-ND}}$ addresses this issue by amplifying the repulsion of confusing majority-class negatives when $\sum_{n \in \mathcal{N}(i)} f_{in}$ is small, thereby preserving minority-class decision boundaries. In contrast, the gradient of $\mathcal{L}_{con}^{\text{InfoNCE}}$ can explode when $\sum_{p \in \mathcal{P}(i)} f_{ip}$ becomes vanishingly small, which is exacerbated by the scarcity of positive samples in minority classes. Consequently, $\mathcal{L}_{con}^{\text{InfoNCE-ND}}$ appears more suitable for IDG.

Finally, the gradient of $\mathcal{L}_{con}^{\text{SupCon-ND}}$ can also be expressed in terms of $f'_{ip} = 1 - s(\mathbf{p}_i, \mathbf{p}_p)$, $f'_{in} = 1 - s(\mathbf{p}_i, \mathbf{p}_n)$, and $Z'_i = \sum_{a \in \mathcal{A}(i)} (1 - s(\mathbf{p}_i, \mathbf{p}_a))$:

$$\begin{aligned} \frac{\partial \mathcal{L}_{con}^{\text{SupCon-ND}}}{\partial \mathbf{p}_i} &= \frac{1}{Z'_i} \left[\sum_{p \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} f'_{ip} - \sum_{n \in \mathcal{N}(i)} \nabla_{\mathbf{p}_i} f'_{in} \cdot \left(\frac{Z_i}{|\mathcal{N}(i)| \sum_{n \in \mathcal{N}(i)} f_{in}} - 1 \right) \right] \\ &= \frac{1}{Z'_i} \left[- \sum_{p \in \mathcal{P}(i)} \nabla_{\mathbf{p}_i} s(\mathbf{p}_i, \mathbf{p}_p) + \sum_{n \in \mathcal{N}(i)} \nabla_{\mathbf{p}_i} s(\mathbf{p}_i, \mathbf{p}_n) \cdot \left(\frac{Z'_i}{|\mathcal{N}(i)| \sum_{n \in \mathcal{N}(i)} f'_{in}} - 1 \right) \right] \end{aligned}$$

Similar to the conclusion in SupCon Khosla et al. (2020), this gradient enforces equal treatment across all negatives, which can suppress the gradient signal from rare but critical minority-class negatives. In contrast, our proposed $\mathcal{L}_{con}^{\text{InfoNCE-ND}}$ naturally emphasizes closer negatives, thereby more confusing, leading to stronger repulsion for hard negatives and attenuated updates for already-separated ones. This dynamic aligns well with the goal of maintaining class separation in IDG. In summary, while SupCon emphasizes balanced treatment of positives, we shift focus by dynamically weighting negatives based on similarity. This inversion transforms SupCon’s design into a complementary strength under imbalance, enabling targeted repulsion of hard negatives and improved minority-class separation. Meanwhile, an empirical experiment has been provided in E.5.

C ALGORITHM

Algorithm 1 NDCL

Input: Training set, batch size B , number of source domains N , number of class K , maximum iteration T , the trade-offs α and β , and Beta distribution parameter ρ

Parameters: A learnable network f_θ

Output: f_θ

```

1: for  $t \leftarrow 1$  to  $T$  do
2:   Randomly sample a batch  $\mathcal{S} = \{(\mathbf{x}_b, y_b, d_b)\}_{b=1}^B$ 
3:    $\mathbf{p}_b \leftarrow f_\theta(\mathbf{x}_b)$ 
4:   # calculate sub-objective function  $\mathcal{L}_{ce}$ 
    $\omega_b \leftarrow \frac{\exp(\ell_{ce}(\mathbf{p}_b, y_b))}{\sum_{\mathbf{x}_i \in \mathcal{X}_k} \exp(\ell_{ce}(\mathbf{p}_i, k))}$ 
5:    $\mathcal{L}_{ce} \leftarrow$  Eq. (4) with  $\{(\mathbf{p}_b, y_b, \omega_b)\}_{b=1}^B$ 
6:   # calculate sub-objective function  $\mathcal{L}_{const}$ 
    $\mu_k^d \leftarrow \frac{1}{n_k^d} \sum_{\mathbf{x}_i \in \mathcal{X}_k^d} \mathbf{p}_i$ 
7:    $\mathcal{L}_{const} \leftarrow$  Eq. (3) with  $\{\mu_k^d\}_{d=1, \dots, N; k=1, \dots, K}$ 
8:   # calculate sub-objective function  $\mathcal{L}_{con}$ 
   for a do
9:     resort all instances by  $p_k$  in batch  $\mathcal{S}$ 
10:     $\hat{\mathcal{N}}_k \leftarrow$  Eq. (2)
11:     $(\hat{\mathbf{p}}_k)_i \leftarrow f_\theta((\mathbf{x}_k)_i), \forall (\mathbf{x}_k)_i \in \hat{\mathcal{N}}_k$ 
12:   end for
13:    $\mathcal{L}_{con} \leftarrow$  Eq. (1) with  $\bigcup_{k=1}^K \{(\hat{\mathbf{p}}_k)_i\}_{i=1}^{\hat{n}_k}$ 
14:   # total loss
    $\mathcal{L}_{total} \leftarrow \mathcal{L}_{ce} + \alpha \mathcal{L}_{con} + \beta \mathcal{L}_{const}$ 
15:   Update network parameters
16: end for

```

Note that the number of the augmented negative set $\hat{\mathcal{N}}_k$ is \hat{n}_k , which is inversely proportional to the number of observed samples of the k -th class across all domains. The underlying intuition is that the classifier is easier to learn on the major classes.

Thus, p_k in Eq. (2) is *not a tunable hyper-parameter*, which would be dynamically determined in the current mini-batch. For example, when mining hard negatives for class 1, we first aim to augment $\hat{n}_k = 100$ instances. Then, we sort all instances by their predicted probability p_1 , and subsequently select the *bottom quarter of those* labeled as class 1, e.g., 10 instances, as X_k^{low} . Accordingly, X_k^{high} consists of the top 10 non-class-1 instances with the highest predicted p_1 scores, since $100/10 = 10$. This adaptive mechanism enables the dynamic selection of boundary samples per class and per iteration, ensuring robustness to both class imbalance and evolving model predictions.

D RELATED WORKS

Existing methods in conventional Domain Generalization can be divided into four taxonomies. 1) *Domain-invariant representation (DIR)*. MMD Li et al. (2018b) leveraged the kernel mean embedding to align arbitrary order moments of two distributions. CORAL Sun & Saenko (2016) achieved good performance by transferring only second-order moment. DANN Ganin et al. (2016) introduced an additional domain classifier to interfere with the representations, so that they do not contain domain information. IRM Arjovsky et al. (2019) constructed a causal diagram from a causal perspective to obtain invariant representations of causality. RDM Nguyen et al. (2024) aligned risk distribution to indirectly eliminate domain information in the representation space. TCRI Salaudeen & Koyejo (2024) proposed a causal strategy to learn domain general representation, which is similar to DIR. 2) *Data augmentation*. Mixup Yan et al. (2020) provided more interpolated instances to enhance the smoothness, thereby improving generalizability. EDM Cao & Chen (2024) enlarged the supported region spanned by source domains through generated extrapolated domains. 3) *Optimization-based*. Fish Shi et al. (2022) proposed aligning gradients across domains by minimizing domain-specific gradient discrepancies. PGrad Wang et al. (2023) leveraged principal gradients to align feature distributions across multiple source domains, enabling better transferability. 4) *Meta learning*. MLDG Li et al. (2018a) utilized the meta-learning to simulate the marginal shift between domains. Despite their remarkable success, these methods appear insufficient to fully address the complex challenges posed by the entangled domain and label shifts in IDG.

Recent studies have begun tackling IDG through a divide-and-conquer strategy. Xia et al. (2023) propose a two-stage framework focused on synthesizing reliable samples to compensate for minority domains or classes, while incorporating domain-invariant representation learning via adversarial training. Su et al. (2024) seek flat local minima while amplifying gradients for low-confidence samples, leveraging domain-invariant representations to address long-tailed data across domains. While intuitive, the former incurs substantial computational overhead, and the latter does not effectively address those coupling effects. Moreover, the inherent limitation of domain-invariant representations, where target domains may lie outside the support of source domains Cao & Chen (2024), is often exacerbated under such coupling. In contrast, Yang et al. (2022) propose a unified method that adaptively re-weights each domain-class pair through a designed transferability graph in the representation space to jointly address both shifts. this method remains biased toward majority classes, and its overemphasis on underrepresented minority classes may lead to overfitting and degraded generalization performance. In this paper, we first present the generalization bound for IDG to reveal the key factors that influence the generalization of the target domain. Building on this theoretical insight, our designed NDCL leverages negative signals to naturally repel samples from neighboring classes, thereby enhancing generalization. In addition, **this design cleverly incorporates the structure of InfoNCE objective**, without requiring any modifications to the original data distribution.

E ADDITIONAL EXPERIMENTS

In this section, we provide a more detailed experimental analysis.



Figure 5: TotalHeavyTail setting on three Benchmark.

E.1 DATASETS AND SETTINGS

VLCS is a widely used domain generalization benchmark that combines four distinct datasets: VOC2007 (V), LabelMe (L), Caltech (C), and SUN09 (S). It contains 5 object categories shared across domains, with significant variations in scene style and object depiction, making it suitable for evaluating cross-domain recognition performance.

PACS is a more challenging DG dataset consisting of four visually diverse domains: Photo (P), Art painting (A), Cartoon (C), and Sketch (S). It includes 7 object categories and features large domain shifts due to differences in texture, abstraction, and artistic style.

OfficeHome is another more challenging benchmark for domain generalization, consisting of 65 categories across four diverse domains: Art, Clipart, Product, and Real-World. It exhibits substantial domain discrepancies and category imbalance, thus providing a comprehensive evaluation setting for generalization algorithms.

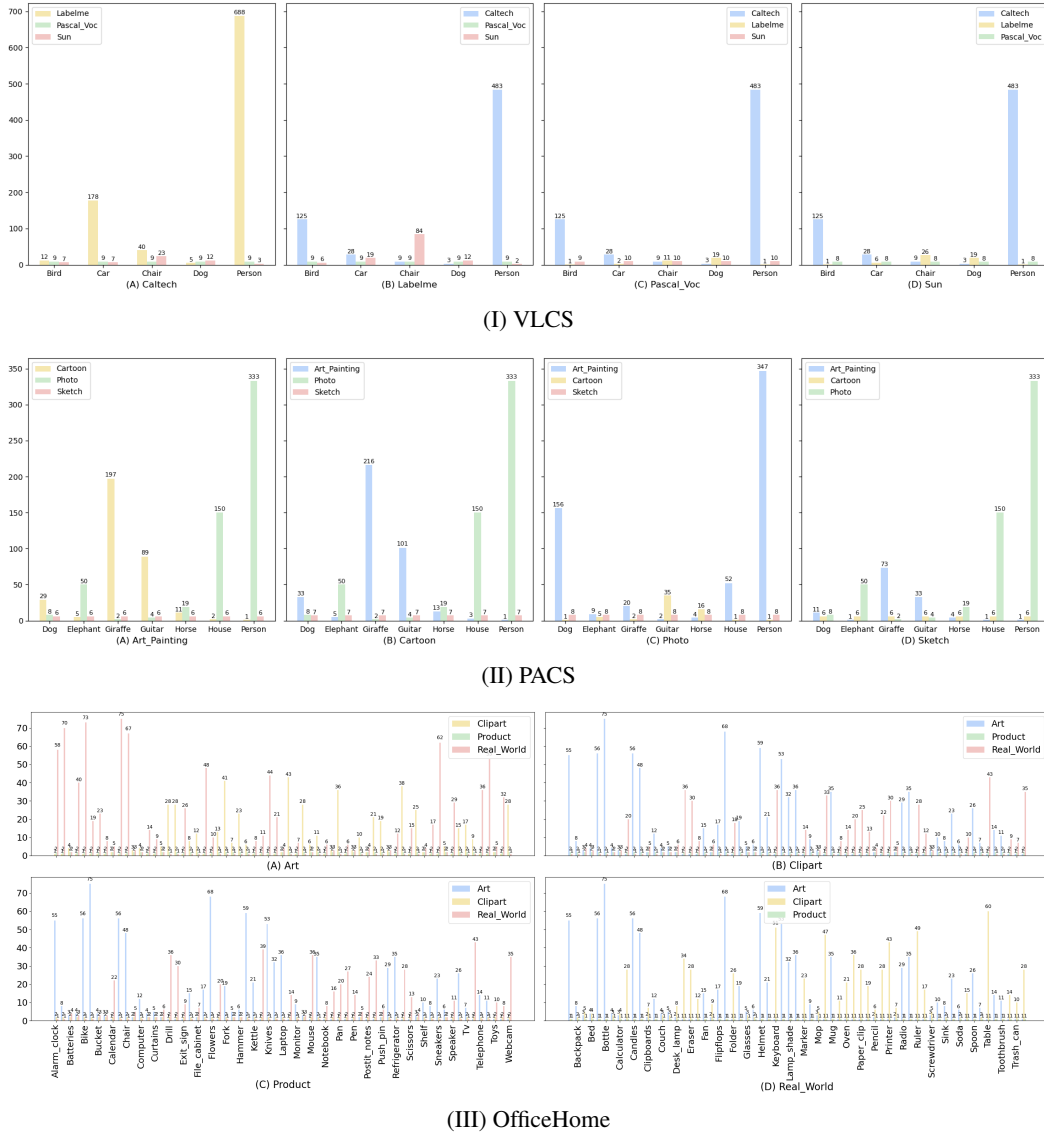


Figure 6: Duality setting on three Benchmark.

Since existing works do not provide standardized data splitting protocols, particularly on the OfficeHome Benchmark, and typically omit performance breakdowns across both coarse-and-fine-grained class levels Su et al. (2024); Xia et al. (2023), it becomes challenging to ensure systematic evaluation and reproducibility. To fill this gap, we provide a script, namely `idg_generate.py`, in the code folder of the supplementary materials, which enables the generation of diverse and controllable dataset configurations.

```

1  # class TotalHeavyTail() and main() are defined in idg_generate.py
2  generator = TotalHeavyTail(num_of_validation, percent_of_test, heavytail_distribution_parameter)
3  stats = main('PACS', num_of_validation, thred_many, thred_few, generator, file=sys.stdout)

```

Listing 1: Example of generating the TotalHeavyTail setting on PACS

TotalHeavyTail and Duality are two representative settings generated by this script, where Fig. 5 and Fig. 6 visualize the number of training samples per class in each setting, respectively. Specifically, the TotalHeavyTail setting represents a cross-domain consistent long-tailed distribution, with relatively balanced domain sampling. In contrast, the Duality setting introduces a symmetric long-tailed distribution with noticeable domain-level sampling discrepancies. Tab. 5 compares the imbalance

Table 5: Statistical information about imbalance ratios across all benchmarks under three different settings. CR denotes the overall class ratio, i.e., $\frac{\max(\{n_k\}_{k=1}^K)}{\min(\{n_k\}_{k=1}^K)}$. DR denotes the sampling ratio across domains, i.e., $\frac{\max(\{N^d\}_{d=1}^N)}{\min(\{N^d\}_{d=1}^N)}$. ECR denotes the class ratio in each source domain, i.e., $\frac{\max(\{n_k^d\}_{k=1}^K)}{\min(\{n_k^d\}_{k=1}^K)}$.

| | | GINIDG | | | TotalHeavyTail | | | Duality | | |
|------------|---|--------|------|------------------------|----------------|------|--------------------------|---------|-------|-----------------------|
| | | CR | DR | ECR | CR | DR | ECR | CR | DR | ECR |
| VLCS | C | 10.15 | 1.38 | [30.95, 90.28, 4.90] | 131.00 | 1.21 | [229.33, 64.23, 704.00] | 26.92 | 20.51 | [137.60, 1.00, 7.67] |
| | L | 10.00 | 2.45 | [14.08, 99.42, 7.36] | 144.42 | 2.04 | [483.00, 69.58, 704.00] | 20.58 | 14.40 | [161.00, 1.00, 42.00] |
| | S | 10.00 | 2.25 | [19.32, 42.87, 5.22] | 143.28 | 1.97 | [483.00, 688.00, 69.58] | 16.40 | 16.20 | [161.00, 26.00, 1.00] |
| | V | 26.66 | 2.08 | [13.23, 33.36, 107.14] | 144.23 | 1.81 | [69.00, 172.00, 704.00] | 15.43 | 16.20 | [161.00, 22.00, 1.11] |
| PACS | A | 12.50 | 3.35 | [8.54, 3.65, 46.44] | 154.57 | 3.88 | [176.50, 74.50, 580.00] | 26.92 | 20.51 | [137.60, 1.00, 7.67] |
| | C | 12.36 | 3.35 | [7.98, 4.12, 48.27] | 147.86 | 4.16 | [145.50, 69.50, 605.00] | 20.58 | 14.40 | [161.00, 1.00, 42.00] |
| | P | 13.00 | 2.34 | [7.27, 8.20, 40.75] | 132.78 | 2.31 | [97.00, 149.50, 605.00] | 16.40 | 16.20 | [161.00, 26.00, 1.00] |
| | S | 16.50 | 1.80 | [19.47, 25.00, 9.00] | 198.40 | 1.27 | [173.50, 312.00, 166.50] | 13.60 | 13.47 | [73.00, 1.00, 166.50] |
| OfficeHome | A | - | - | - | 74.00 | 1.16 | [78.00, 78.00, 72.00] | 9.75 | 7.42 | [43.00, 1.00, 75.00] |
| | C | - | - | - | 66.00 | 1.54 | [78.00, 75.00, 74.00] | 9.75 | 7.04 | [75.00, 1.00, 43.00] |
| | P | - | - | - | 69.33 | 1.63 | [75.00, 75.00, 74.00] | 9.75 | 7.04 | [75.00, 1.00, 43.00] |
| | R | - | - | - | 70.33 | 1.80 | [76.00, 77.00, 77.00] | 9.62 | 14.08 | [75.00, 60.00, 1.00] |

ratios across the three settings used in the main experiments. It can be observed that, compared to the GINIDG setting, our proposed TotalHeavyTail and Duality settings place increased demands on model robustness and generalization capacity.

We adopt the DomainBed protocol Gulrajani & Lopez-Paz (2021) to reproduce all methods, aiming to systematically evaluate which approaches are effective under which settings. ResNet-50 architecture has been adopted as the backbone, and each input image is resized and cropped to 224×224 . Adam is utilized as the optimizer. Training data is randomly sampled from the full source domains under each setting. Unlike the original DomainBed training-domain validation protocol, which utilizes the remaining data for model selection, we utilize a **cross-domain balanced validation set** for model selection. And, the remaining data from the training domains serves as the in-distribution test set, while the held-out domains are used as target domains. Batch size is set to 32 for each source domain and 64 for test, respectively. Maximum iteration T is set to 5,000. The default learning rate is set to $5e-5$. The range of the trade-offs involved in each method can be referred to the supplementary material. All codes are run on Python 3.9, PyTorch 1.13 on Arch Linux with NVIDIA GeForce RTX 4090 GPUs.

E.2 ILLUSTRATING IDG FAILURE MODES ON TOY DATA

In this subsection, we construct additional experiments to demonstrate the challenges of IDG on toy PACS, mentioned in section 1.

To demonstrate **how heterogeneous long-tailed distributions can lead to misalignment**, we design a toy experiment on the PACS dataset with a symmetric cross-domain long-tail distribution. Specifically, an imbalance ratio of 400:1 is used, as illustrated in Fig. 7 (C), where person is the dominant class in Domain A and a minority class in Domain C, and vice versa for the other classes. Several key observations emerge: 1) Under this imbalanced setting, the ROC-AUC on the validation set decreases significantly. 2) Employing domain alignment strategies further reduces the ROC-AUC, suggesting that majority classes dominate the alignment process and likely cause mismatches. 3) The dominant class consistently lies closest to its domain centroid, as illustrated by the bar plots in Fig. 7 (C). And, the Wasserstein distance between domains is comparable to the Wasserstein distance Cao & Chen (2024) between their dominant classes. These findings further imply that majority classes play a dominant role in the domain alignment process.

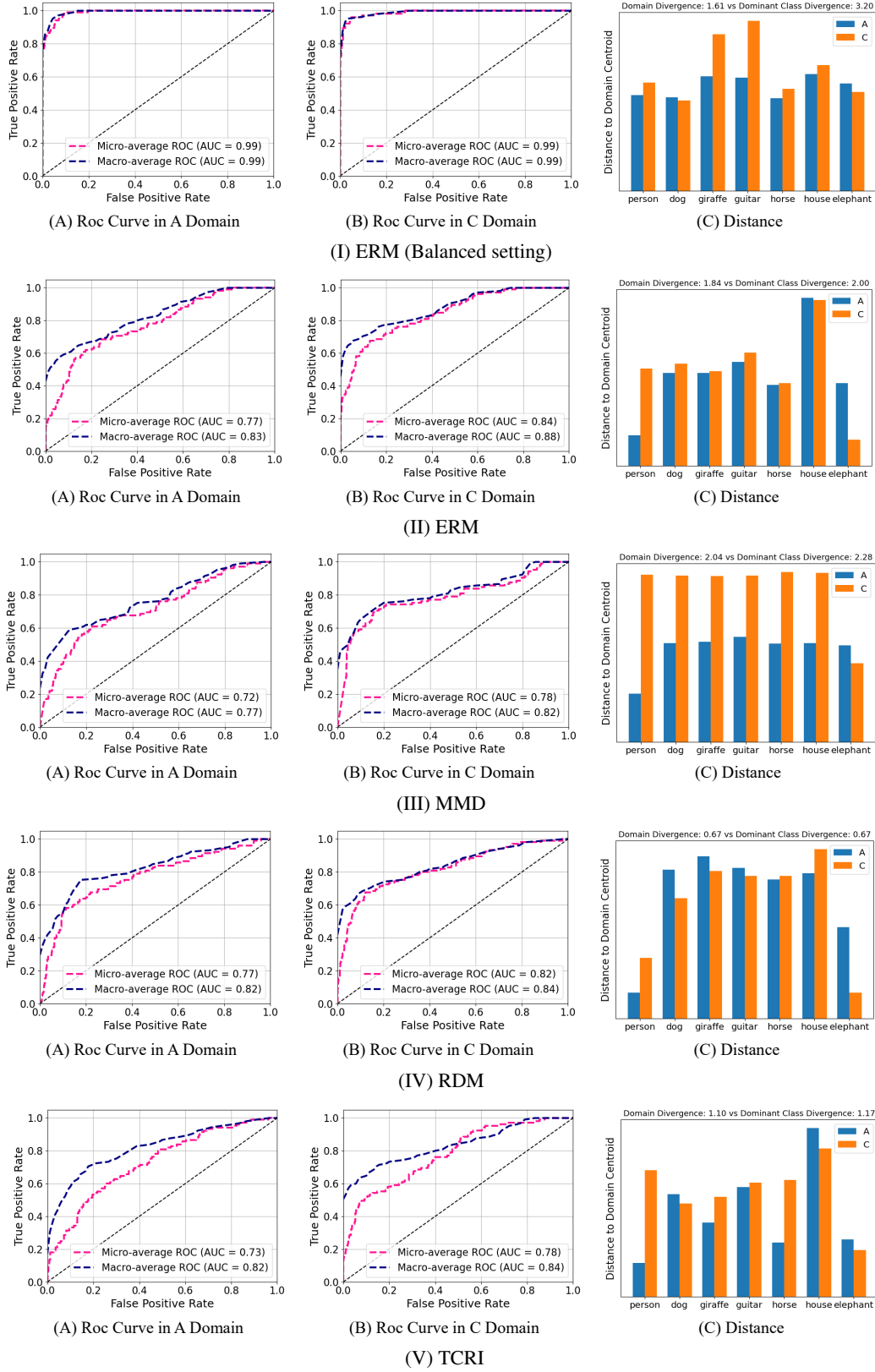


Figure 7: An illustration of missalignment caused by heterogeneous long-tailed distributions. (A) and (B) illustrates the Roc curves on the validation set for each source domain. (C) illustrates the distance between each class and its corresponding domain centroid. In Domain A, class sample sizes decrease from left to right along the x-axis, whereas in Domain C, they increase accordingly.

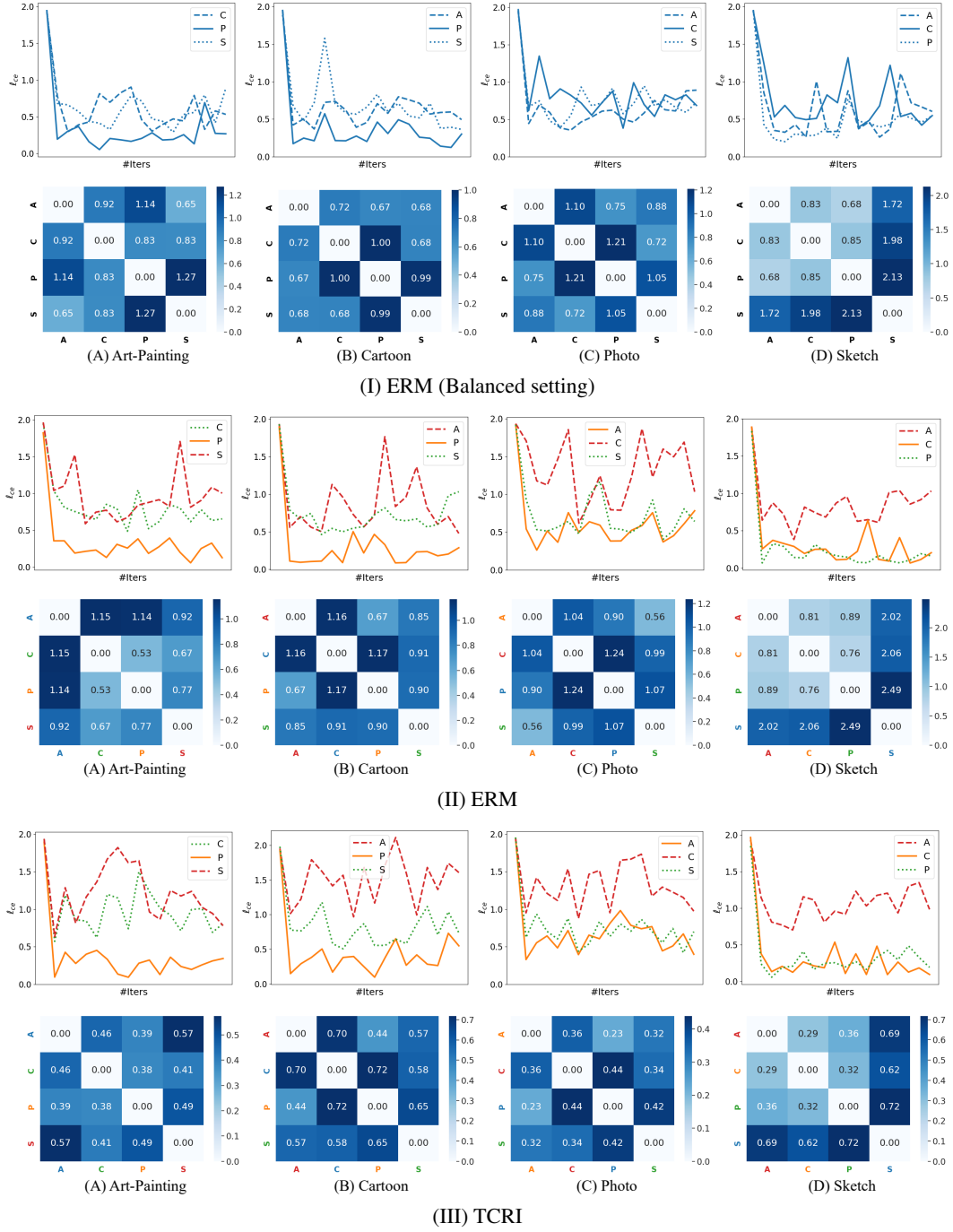


Figure 8: An illustration of underrepresented domains being absorbed by more populous ones. Minority, majority, and medium domains are color-coded in red, orange, and green, respectively. The line plot tracks the validation cross-entropy loss over training epochs for each source domain, while the heatmap quantifies inter-domain discrepancies under the optimal model, measured by the Wasserstein distanceCao & Chen (2024).

To demonstrate **how inter-domain sampling imbalance can cause underrepresented domains to be absorbed by more populous ones**, thereby suppressing inter-domain support set and impairing generalization, we design a toy experiment with imbalanced domain sampling but balanced class sampling within each domain. Specifically, we sample 100 examples per class for the major do-

Table 6: Impact of domain-level resampling on BSoftmax under TotalHeavyTail and Duality settings.

| | TotalHeavyTail | | | | Duality | | | |
|----------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| | Average | Many | Medium | Few | Average | Many | Medium | Few |
| Applying | 48.6 \pm 0.5 | 69.5 \pm 0.9 | 61.9 \pm 0.5 | 33.0 \pm 0.5 | 52.8 \pm 0.5 | 67.4 \pm 1.4 | 62.1 \pm 0.6 | 44.4 \pm 1.0 |
| Omitting | 48.4 \pm 0.5 | 69.1 \pm 0.7 | 61.8 \pm 1.0 | 32.4 \pm 0.5 | 51.9 \pm 0.3 | 65.4 \pm 1.2 | 59.1 \pm 0.8 | 43.7 \pm 0.3 |

Table 7: Discrepancy of prior distributions in learned embeddings with on OfficeHome under TotalHeavyTail and Duality settings across several representative methods. SS indicates source-source divergence. ST indicates source-target divergence.

| | | ERM | MMD | RDM | PGrad | TCRI | BSoftmax | GINIDG | BoDA | NDCL (<i>ours</i>) |
|----------------|----|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|----------------------|
| TotalHeavyTail | SS | 0.45 \pm 0.02 | 0.43 \pm 0.05 | 0.55 \pm 0.09 | 0.36 \pm 0.06 | 0.17 \pm 0.03 | 0.38 \pm 0.06 | 0.77 \pm 0.16 | 0.13 \pm 0.00 | 0.30 \pm 0.03 |
| | ST | 0.99 \pm 0.08 | 1.03 \pm 0.15 | 1.37 \pm 0.13 | 0.78 \pm 0.13 | 0.38 \pm 0.07 | 0.99 \pm 0.14 | 2.32 \pm 0.19 | 0.29 \pm 0.01 | 0.69 \pm 0.08 |
| Duality | SS | 0.84 \pm 0.10 | 0.70 \pm 0.07 | 0.82 \pm 0.16 | 0.65 \pm 0.05 | 0.42 \pm 0.04 | 0.63 \pm 0.06 | 1.38 \pm 0.14 | 0.36 \pm 0.02 | 0.53 \pm 0.05 |
| | ST | 1.10 \pm 0.09 | 1.06 \pm 0.09 | 0.90 \pm 0.17 | 0.72 \pm 0.07 | 0.55 \pm 0.06 | 0.81 \pm 0.11 | 2.36 \pm 0.19 | 0.43 \pm 0.05 | 0.70 \pm 0.08 |

main, 20 per class for the medium domain, and 4 per class for the minority domain. This setting is compared against a fully balanced baseline in which each domain contributes only 10 samples per class. As illustrated in Fig. 8, several key observations emerge: 1) Under the balanced setting, losses across domains remain similar. In contrast, the imbalanced setting leads to substantial loss discrepancies, with the minority domain exhibiting the highest and most unstable validation loss. This indicates that domain-level sampling imbalance negatively affects the generalization ability of each source domain. 2) Comparing inter-domain discrepancies in SubFig. 8 (I) and (II), we observe that the divergence between majority and minority domains under the imbalanced setting is significantly smaller than in the balanced case, suggesting that the minority domain has been absorbed by the majority. 3) Regarding source-to-target domain shift, the imbalanced setting exhibits notably larger discrepancies compared to the balanced case, confirming that domain-level imbalance undermines cross-domain generalization. 4) Even with the use of a recent distribution alignment method, i.e., TRCI Salaudeen & Koyejo (2024), the loss gap between majority and minority domains persists, although the overall inter-domain distributional divergence is noticeably reduced.

E.3 IMPACT OF DOMAIN-LEVEL RESAMPLING

To ensure a fair comparison, a domain-level resampling strategy is implicitly applied to all long-tailed methods. Specifically, each training iteration consists of samples from all domains. This strategy is commonly adopted in domain generalization, particularly by alignment-based approaches. Tab. 6 reports the impact of applying or omitting this resampling strategy on the BSoftmax method. These results indicate that under the TotalHeavyTail setting, which is dominated by label imbalance, this strategy has minimal effect, suggesting that class discriminability should be prioritized in this case. In contrast, under the Duality setting, where heterogeneous label shift reduces inter-class competition and amplifies domain shift, this resampling strategy helps to mitigate domain divergence during training, thereby improving the robustness of long-tailed methods.

E.4 PRIOR DISTRIBUTION DISCREPANCIES IN LEARNED EMBEDDINGS

In Tab. 7, we investigate the marginal distribution discrepancies, i.e., prior distribution discrepancies, of learned representations across several representative methods. Despite not explicitly aligning prior distributions, our proposed NDCL achieves lower prior discrepancies both among source domains (SS) and between source and target domains (ST), compared to existing domain alignment approaches such as MMD and RDM. This suggests that our posterior-alignment strategy implicitly mitigates prior mismatches to some extent. Another potential explanation, discussed in Subsection E.2, is that the methods, such as MMD and RDM, may suffer from misalignment due to label shift, particularly under the TotalHeavyTail setting. TCRI achieves lower prior discrepancy than ours, possibly due to its causal approach, which aligns support sets rather than full representations and thus avoids misalignment when sampling is sufficient for each domain. BoDA, which directly manipulates the representations, yields the lowest prior discrepancy overall.

Table 8: Comparison of Contrastive Loss Variants for our NDCL. $\text{NDCL}^{\text{InfoNCE-ND}}$ denotes our proposed method in the main manuscript, dominated by negatives with InfoNCE-like objective. $\text{NDCL}^{\text{SupCon-ND}}$ denotes NDCL with SupCon-like objective dominated by negatives, while $\text{NDCL}^{\text{SupCon}}$ denotes NDCL with SupCon objective dominated by positives.

| | Average | TotalHeavyTail Many | Medium | Few | Average | Duality Many | Medium | Few |
|-----------------------------------|----------------|------------------------|----------------|----------------|----------------|-----------------|----------------|----------------|
| $\text{NDCL}^{\text{SupCon}}$ | 47.3 ± 0.2 | 76.1 ± 0.7 | 64.8 ± 0.9 | 26.7 ± 0.4 | 53.1 ± 0.3 | 70.2 ± 1.4 | 62.6 ± 0.5 | 43.7 ± 0.7 |
| $\text{NDCL}^{\text{SupCon-ND}}$ | 48.0 ± 0.3 | 75.3 ± 1.3 | 65.3 ± 0.4 | 27.7 ± 0.5 | 53.2 ± 0.2 | 69.7 ± 1.4 | 63.1 ± 0.6 | 43.6 ± 0.4 |
| $\text{NDCL}^{\text{InfoNCE-ND}}$ | 49.0 ± 0.2 | 71.6 ± 0.8 | 66.0 ± 0.2 | 30.5 ± 0.3 | 55.7 ± 0.2 | 71.4 ± 1.0 | 65.8 ± 0.5 | 47.6 ± 0.6 |

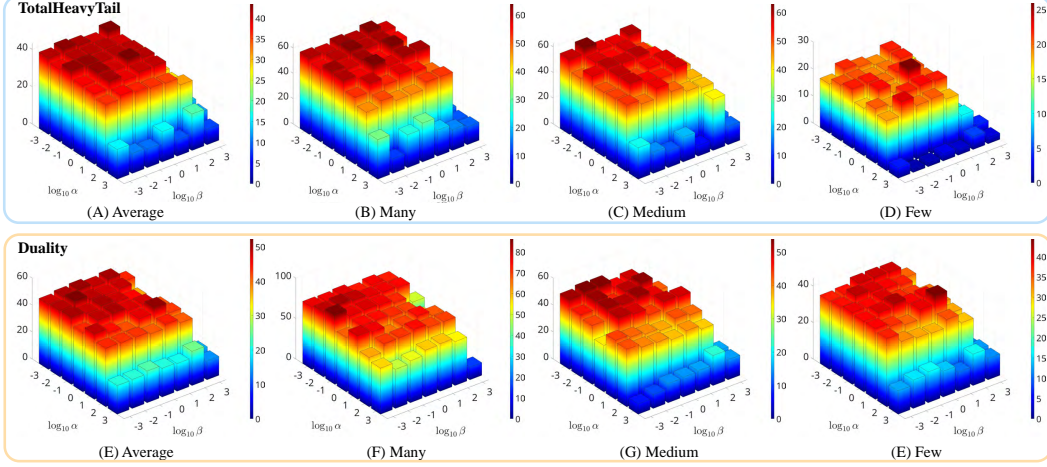


Figure 9: Joint influence of hyperparameters α and β on OfficeHome under two different settings. The upper part corresponds to the TotalHeavyTail setting, and the lower part to the Duality setting. Log-scale axes are used, e.g., $0 = \log_{10} 1$.

E.5 EMPIRICAL EVIDENCE OF PROPOSED CONTRASTIVE OBJECTIVE

The results in Tab. 8 demonstrate the performance differences among three contrastive loss variants integrated into our NDCL framework. Both $\text{NDCL}^{\text{SupCon-ND}}$ and $\text{NDCL}^{\text{InfoNCE-ND}}$ show clear improvements over $\text{NDCL}^{\text{SupCon}}$, particularly on minority classes, confirming that negative-dominated separation alleviates the absorption of small clusters into majority ones. Among them, $\text{NDCL}^{\text{InfoNCE-ND}}$, which is our proposed method in the main manuscript, achieves the best balance, yielding the highest overall accuracy while **substantially boosting Few class performance** across both datasets. These findings highlight the importance of posterior-aligned negative separation in maintaining stable and generalizable decision boundaries under imbalanced domain shifts, which is consistent with our gradient analysis.

E.6 PARAMETER STUDIES

Fig. 9 reports the joint influence of the two hyperparameters involved in our NDCL. These results indicate that both of them are effective within the range of $10^{[-3,2]}$, with optimal average performance typically around 10^{-1} and 10^{-2} , respectively. Notably, under the severely imbalanced TotalHeavyTail setting, larger values lead to improved performance on minority classes, indicating that our design can effectively enhance class separability for underrepresented categories.

E.7 COMPUTATIONAL COMPLEXITY ANALYSIS

First, we would like to clarify that the number of hyper-parameters involved in our NDCL is comparable to those in other IDG methods. NDCL has 3, GINIDG has 1 but involves 3 sub-objective functions, SAMALTDG has 4, BoDA has 7. **Second**, in Tab. 9, we have empirically investigated the computational complexity of our NDCL. From this table, we can observe that: 1) Overall,

Table 9: Average per-batch training cost (in seconds).

| | RDM | PGrad | TRCI | BSoftMax | GINIDG | BoDA | SAMALTDG | NDCL | NDCL-NoCon | NDCL-NoAug | NDCL-NoConst |
|------------|-------|-------|-------|----------|--------|-------|----------|-------|------------|------------|--------------|
| VLCS | 0.224 | 0.679 | 0.433 | 0.248 | 0.617 | 0.315 | 0.489 | 0.500 | 0.229 | 0.263 | 0.483 |
| PACS | 0.208 | 0.597 | 0.415 | 0.219 | 0.606 | 0.267 | 0.478 | 0.468 | 0.215 | 0.260 | 0.453 |
| OfficeHome | 0.228 | 0.654 | 0.466 | 0.254 | 0.613 | 0.298 | 0.493 | 0.505 | 0.221 | 0.296 | 0.494 |

Table 10: Worst-domain *in-distribution* accuracy per target under the Duality setting on three benchmarks. The **bold**, underline, and dashline items are the best, the second-best, and the third-best results, respectively. Column V on VLCS denotes the in-distribution accuracy of the worst source domain when V is used as the target domain.

| | VLCS | | | | PACS | | | | OfficeHome | | | |
|----------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| | C | L | S | V | A | C | P | S | A | C | P | R |
| ERM | 68.1 ± 0.4 | 62.9 ± 1.7 | 62.0 ± 2.9 | 33.7 ± 2.5 | 77.8 ± 0.6 | 79.6 ± 0.5 | 76.6 ± 1.5 | 75.8 ± 0.7 | 67.0 ± 0.3 | 66.5 ± 0.6 | 49.4 ± 0.3 | 55.9 ± 0.8 |
| IRM | 67.1 ± 4.6 | 64.1 ± 1.1 | 60.3 ± 0.9 | 31.3 ± 2.8 | 75.1 ± 1.2 | 75.1 ± 2.9 | 75.8 ± 2.0 | 74.9 ± 1.9 | 64.9 ± 0.6 | 65.5 ± 0.7 | 46.9 ± 0.9 | 54.4 ± 0.9 |
| GroupDRO | 67.6 ± 1.5 | 64.7 ± 2.1 | 60.8 ± 0.4 | 41.7 ± 5.9 | 75.7 ± 0.2 | 78.4 ± 0.9 | 78.8 ± 1.0 | 76.9 ± 1.7 | 66.1 ± 0.8 | 66.7 ± 0.9 | 48.5 ± 1.1 | 56.9 ± 0.6 |
| Mixup | 71.1 ± 2.5 | 64.4 ± 2.5 | 57.8 ± 1.6 | 36.7 ± 1.8 | 76.2 ± 1.1 | 77.3 ± 1.8 | 78.5 ± 0.7 | 76.2 ± 1.2 | 67.0 ± 1.3 | 67.1 ± 0.2 | 48.5 ± 0.2 | 59.5 ± 0.2 |
| MLDG | 70.2 ± 2.8 | 63.9 ± 2.0 | 60.3 ± 1.5 | 35.7 ± 1.8 | 76.9 ± 1.2 | 77.9 ± 1.2 | 76.3 ± 1.6 | 75.3 ± 2.4 | 66.9 ± 0.7 | 67.7 ± 0.4 | 48.3 ± 1.0 | 56.5 ± 0.8 |
| CORAL | 69.8 ± 2.1 | 65.3 ± 0.6 | 63.7 ± 0.2 | 38.4 ± 0.8 | 76.1 ± 1.1 | 75.5 ± 0.9 | 74.8 ± 1.3 | 73.6 ± 0.7 | 68.4 ± 0.8 | 69.2 ± 0.9 | 51.8 ± 0.4 | 59.7 ± 0.5 |
| MMD | 63.3 ± 2.8 | 64.5 ± 2.6 | 59.9 ± 2.1 | 31.4 ± 0.7 | 79.7 ± 1.4 | 79.9 ± 0.7 | 76.9 ± 2.2 | 74.9 ± 1.1 | 66.5 ± 1.1 | 66.6 ± 0.1 | 48.4 ± 0.4 | 53.7 ± 0.5 |
| DANN | 71.9 ± 2.7 | 62.3 ± 3.0 | 61.1 ± 3.5 | 41.0 ± 5.1 | 78.4 ± 0.3 | 80.5 ± 0.6 | 75.4 ± 3.1 | 74.9 ± 1.2 | 66.8 ± 0.5 | 66.9 ± 0.8 | 49.1 ± 0.5 | 55.8 ± 0.7 |
| SagNet | 71.7 ± 0.3 | 66.4 ± 0.8 | 64.8 ± 1.9 | 32.0 ± 1.6 | 77.6 ± 1.0 | 78.0 ± 0.6 | 76.7 ± 2.3 | 75.3 ± 2.3 | 68.6 ± 0.1 | 69.5 ± 0.3 | 52.3 ± 0.7 | 59.4 ± 0.8 |
| VREx | 65.7 ± 3.9 | 63.8 ± 0.6 | 64.9 ± 0.9 | 32.0 ± 2.5 | 74.6 ± 2.1 | 77.2 ± 0.8 | 75.4 ± 1.1 | 78.2 ± 1.4 | 66.1 ± 0.7 | 66.4 ± 0.5 | 49.7 ± 0.8 | 57.8 ± 0.4 |
| Fish | 72.6 ± 2.3 | 66.1 ± 2.4 | 62.0 ± 1.4 | 34.5 ± 2.3 | 76.1 ± 1.2 | 77.6 ± 0.7 | 77.3 ± 1.8 | 77.6 ± 0.8 | 67.6 ± 0.3 | 68.8 ± 0.2 | 50.2 ± 0.8 | 59.7 ± 1.2 |
| Fishr | 68.8 ± 2.3 | 66.3 ± 2.5 | 64.0 ± 1.5 | 36.8 ± 1.8 | 77.3 ± 0.7 | 77.8 ± 1.1 | 76.4 ± 1.6 | 76.8 ± 1.1 | 67.6 ± 0.7 | 66.4 ± 0.2 | 49.9 ± 0.6 | 58.4 ± 0.5 |
| EQRM | 69.7 ± 3.8 | 61.0 ± 1.5 | 62.6 ± 1.1 | 34.2 ± 1.9 | 79.0 ± 1.3 | 80.9 ± 0.4 | 76.9 ± 1.3 | 77.3 ± 1.4 | 66.1 ± 1.1 | 67.1 ± 0.1 | 50.0 ± 0.4 | 56.7 ± 0.3 |
| RDM | 72.5 ± 1.8 | 65.1 ± 1.8 | 59.5 ± 0.4 | 35.5 ± 2.4 | 76.9 ± 0.2 | 78.6 ± 0.1 | 79.4 ± 1.2 | 75.6 ± 1.6 | 67.8 ± 0.7 | 67.6 ± 1.2 | 49.3 ± 0.9 | 57.2 ± 0.6 |
| PGrad | 74.8 ± 1.4 | 65.5 ± 2.3 | 63.5 ± 0.9 | 33.1 ± 1.5 | 79.8 ± 0.5 | 81.1 ± 0.5 | <u>80.4 ± 1.7</u> | 78.3 ± 2.5 | 69.2 ± 0.4 | 69.9 ± 0.5 | 52.1 ± 0.9 | 62.0 ± 0.8 |
| TRCI | 65.9 ± 3.0 | <u>67.0 ± 1.6</u> | 64.7 ± 3.1 | 31.5 ± 1.5 | 76.2 ± 1.1 | 77.8 ± 1.1 | 77.5 ± 0.9 | 79.0 ± 1.0 | 69.2 ± 0.4 | 69.5 ± 0.5 | 50.1 ± 1.2 | 59.9 ± 0.5 |
| Focal | 68.0 ± 0.7 | 63.9 ± 2.2 | 61.1 ± 1.3 | 32.5 ± 0.3 | 79.6 ± 0.6 | 75.6 ± 0.6 | 77.4 ± 1.4 | 77.0 ± 0.4 | 66.7 ± 0.4 | 66.6 ± 0.4 | 48.1 ± 0.8 | 56.0 ± 0.2 |
| ReWeight | 64.7 ± 1.1 | 64.4 ± 0.9 | 63.3 ± 1.2 | 33.0 ± 0.9 | 79.2 ± 0.5 | 79.4 ± 0.3 | 78.2 ± 0.5 | 76.0 ± 1.8 | 68.2 ± 0.7 | 67.6 ± 0.3 | 48.8 ± 0.5 | 57.1 ± 0.8 |
| BSoftmax | 66.8 ± 0.9 | 67.3 ± 1.9 | 61.4 ± 1.3 | 38.7 ± 2.4 | 80.2 ± 0.7 | 78.6 ± 1.7 | 77.4 ± 1.9 | 77.0 ± 0.6 | 68.0 ± 0.6 | 67.4 ± 0.5 | 49.2 ± 0.8 | 57.6 ± 0.5 |
| LDAM | 68.9 ± 1.1 | 63.8 ± 1.3 | 60.5 ± 1.1 | 37.3 ± 1.0 | 78.7 ± 1.5 | 78.6 ± 2.1 | 78.5 ± 0.2 | 74.9 ± 0.7 | 65.8 ± 0.4 | 66.5 ± 0.3 | 47.0 ± 0.5 | 55.5 ± 1.0 |
| BoDA | 71.6 ± 3.0 | 65.1 ± 2.8 | 64.3 ± 1.6 | 31.4 ± 2.4 | 74.9 ± 0.3 | 76.2 ± 0.9 | 76.3 ± 0.5 | 74.2 ± 2.2 | 68.2 ± 0.9 | 68.5 ± 0.6 | 51.8 ± 1.1 | 58.8 ± 0.2 |
| GINIDG | 72.9 ± 1.8 | 62.6 ± 0.4 | 63.0 ± 3.4 | 31.7 ± 3.2 | 69.2 ± 3.2 | 73.7 ± 1.1 | 73.9 ± 1.6 | 72.6 ± 1.7 | 65.2 ± 1.3 | 65.8 ± 0.8 | 49.5 ± 1.5 | 54.1 ± 0.7 |
| SAMALTDG | <u>70.7 ± 0.7</u> | 63.0 ± 2.6 | 63.4 ± 1.7 | 34.3 ± 0.8 | <u>79.9 ± 0.3</u> | 79.3 ± 0.3 | 77.8 ± 1.3 | 75.7 ± 1.6 | 66.3 ± 0.8 | 66.9 ± 0.2 | 49.2 ± 0.5 | 58.0 ± 0.6 |
| NDCL | 74.9 ± 0.5 | 68.1 ± 1.2 | 62.9 ± 1.5 | 41.0 ± 1.8 | <u>80.0 ± 0.4</u> | 81.9 ± 0.2 | 81.2 ± 1.7 | <u>78.3 ± 1.3</u> | 69.4 ± 0.7 | <u>69.8 ± 0.3</u> | 51.3 ± 0.5 | <u>60.0 ± 0.4</u> |

NDCL exhibits moderate training cost, significantly lower than PGrad with gradient manipulation and GINIDG with adversarial generation. 2) Compared to the IDG methods, NDCL achieves comparable efficiency to SAMALTDG, and although its cost is marginally higher than BoDA, NDCL consistently outperforms these methods across all cases. We believe this constitutes a reasonable trade-off between efficiency and effectiveness. 3) Compared with NDCL and NDCL-NoConst, the cost of predictive center alignment is relatively minor. 4) Comparing NDCL-NoCon and NDCL-NoAug, the cost of hard negative mining is also relatively modest, and both variants are among the more efficient in terms of runtime. These findings indicate that the primary computational cost of NDCL arises from training on augmented data.

According to Tab. 4 in the main manuscript, NDCL-NoAug, which does not employ hard negative mining, exhibits a training cost comparable to that of BoDA, while still achieving consistently better performance.

E.8 COMPREHENSIVE RESULTS

In this subsection, we report more comprehensive results of our designed experiments.

Tab. 10 reports the worst-domain in-distribution accuracy under the Duality setting, that is, the performance on the minority source domain in each experiment. These results highlight several key findings: 1) Our proposed NDCL consistently improves performance on the worst-case domain, demonstrating its effectiveness in mitigating the absorption of minority domains by majority ones, as discussed in Subsection E.2. 2) Domain generalization methods generally perform well, underscoring domain shift as the primary bottleneck in this setting. 3) Some other methods achieve competitive results sporadically, suggesting that enhancing class discriminability alone can offer limited gains, though less reliably than addressing domain shift directly under this setting for IDG.

Tab. 11 and 12 report detailed results on VLCS and PACS benchmarks under the GINIDG setting, which are the performance on the target domains. Tab. 13, 14, and 15 report detailed results on three benchmarks under the TotalHeavyTail setting. Tab. 16, 17, and 18 report detailed results on three benchmarks under the Duality setting. From the detailed tables, we observe some variability in performance across different domain combinations and settings. Nevertheless, our NDCL consistently ranks among the top three in most cases. *Notably, when there is a large discrepancy between source and target domains, such as when domain S is selected as the target domain on PACS, our NDCL achieves consistently strong results across all settings.* This demonstrates the effectiveness and robustness of our NDCL under various forms of label shift.

Table 11: Detailed results of the target domain on VLCS benchmark under the GINIDG setting.

| | C | | | L | | | S | | | Y | | |
|----------|------------|-------------|-------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| | Average | Many | Few | Average | Many | Few | Average | Many | Few | Average | Many | Few |
| ERM | 97.4 ± 0.7 | 98.5 ± 0.9 | 96.7 ± 0.7 | 62.4 ± 0.4 | 64.3 ± 0.6 | 50.0 ± 1.5 | 70.1 ± 0.6 | 65.2 ± 1.7 | 72.8 ± 5.7 | 67.9 ± 4.4 | 66.6 ± 3.1 | 64.6 ± 8.7 |
| IRM | 97.2 ± 0.6 | 99.4 ± 0.2 | 91.5 ± 2.1 | 64.5 ± 0.5 | 67.0 ± 0.6 | 21.7 ± 3.1 | 73.0 ± 0.6 | 70.0 ± 2.3 | 65.2 ± 0.4 | 71.7 ± 0.4 | 70.9 ± 1.6 | 69.1 ± 7.9 |
| GroupDRO | 97.5 ± 0.4 | 99.3 ± 0.3 | 95.9 ± 4.2 | 63.2 ± 1.1 | 65.3 ± 1.2 | 3.5 ± 48.8 | 68.8 ± 1.8 | 60.4 ± 3.4 | 74.0 ± 2.2 | 74.9 ± 1.8 | 75.5 ± 1.2 | 71.6 ± 5.0 |
| Mixup | 95.6 ± 1.5 | 99.7 ± 0.1 | 96.4 ± 1.6 | 60.1 ± 0.9 | 62.2 ± 1.1 | 18.4 ± 3.5 | 67.5 ± 1.3 | 64.7 ± 3.6 | 69.5 ± 3.4 | 72.7 ± 0.7 | 76.1 ± 0.9 | 70.7 ± 2.2 |
| MLDG | 97.5 ± 0.5 | 99.4 ± 0.1 | 97.5 ± 0.9 | 65.4 ± 1.3 | 67.6 ± 1.6 | 26.1 ± 1.8 | 68.5 ± 1.0 | 62.3 ± 1.5 | 66.7 ± 3.4 | 71.3 ± 1.2 | 68.3 ± 2.0 | 74.5 ± 3.1 |
| CORAL | 96.7 ± 0.5 | 98.4 ± 0.2 | 96.3 ± 0.8 | 63.3 ± 0.7 | 65.5 ± 0.5 | 25.0 ± 0.7 | 68.6 ± 1.9 | 67.7 ± 1.6 | 62.0 ± 3.7 | 74.3 ± 0.8 | 75.2 ± 0.5 | 69.2 ± 2.3 |
| MMD | 97.5 ± 0.3 | 98.7 ± 0.2 | 95.5 ± 1.3 | 63.3 ± 1.3 | 65.6 ± 1.3 | 24.7 ± 2.2 | 68.3 ± 1.2 | 64.6 ± 2.7 | 67.7 ± 1.7 | 75.8 ± 1.8 | 77.1 ± 0.6 | 68.3 ± 1.4 |
| DANN | 97.7 ± 0.1 | 99.5 ± 0.2 | 94.6 ± 1.8 | 64.6 ± 0.4 | 65.5 ± 0.6 | 22.2 ± 1.8 | 71.0 ± 0.1 | 66.5 ± 1.4 | 66.8 ± 3.6 | 74.5 ± 1.3 | 72.0 ± 2.3 | 76.2 ± 3.5 |
| SagNet | 95.3 ± 0.2 | 98.6 ± 0.4 | 90.2 ± 3.1 | 63.3 ± 1.6 | 67.7 ± 1.6 | 21.5 ± 0.9 | 69.9 ± 1.1 | 66.9 ± 2.8 | 61.5 ± 4.9 | 72.8 ± 0.7 | 74.5 ± 0.7 | 72.0 ± 0.9 |
| VREx | 93.8 ± 2.1 | 96.2 ± 1.6 | 86.6 ± 3.4 | 63.2 ± 0.4 | 65.3 ± 0.2 | 24.1 ± 3.4 | 66.5 ± 0.5 | 61.6 ± 2.3 | 68.7 ± 5.5 | 72.1 ± 4.1 | 72.1 ± 4.4 | 73.8 ± 6.7 |
| Fish | 97.8 ± 0.2 | 98.8 ± 0.7 | 90.6 ± 2.4 | 64.3 ± 1.2 | 66.4 ± 1.3 | 23.0 ± 0.9 | 71.9 ± 0.4 | 70.1 ± 0.4 | 64.2 ± 3.5 | 75.3 ± 2.5 | 78.0 ± 1.3 | 73.5 ± 1.5 |
| Fishr | 96.9 ± 0.5 | 99.6 ± 0.3 | 95.7 ± 0.6 | 62.9 ± 0.5 | 64.9 ± 0.4 | 24.9 ± 4.3 | 70.4 ± 0.9 | 65.6 ± 3.1 | 68.6 ± 1.6 | 73.7 ± 1.0 | 71.9 ± 3.1 | 76.6 ± 3.4 |
| EQRM | 97.8 ± 0.7 | 98.8 ± 0.6 | 95.8 ± 4.3 | 68.3 ± 4.3 | 69.8 ± 0.8 | 28.4 ± 1.0 | 70.2 ± 0.5 | 68.8 ± 1.7 | 60.2 ± 7.4 | 70.2 ± 1.5 | 69.4 ± 3.5 | 81.1 ± 2.4 |
| EQRM | 96.6 ± 0.8 | 99.7 ± 0.2 | 93.7 ± 2.4 | 62.9 ± 0.7 | 65.2 ± 0.8 | 20.8 ± 2.6 | 68.4 ± 0.6 | 62.6 ± 1.0 | 73.3 ± 3.2 | 72.1 ± 0.8 | 71.2 ± 1.3 | 76.4 ± 0.8 |
| RDM | 98.7 ± 0.1 | 99.6 ± 0.2 | 95.3 ± 1.8 | 97.6 ± 0.8 | 63.8 ± 1.0 | 66.0 ± 0.9 | 71.9 ± 0.5 | 69.2 ± 1.0 | 64.3 ± 3.6 | 75.8 ± 0.5 | 78.1 ± 0.6 | 70.8 ± 2.2 |
| PGrad | 98.7 ± 0.1 | 99.6 ± 0.2 | 95.3 ± 1.8 | 97.6 ± 0.8 | 63.8 ± 1.0 | 66.0 ± 0.9 | 71.9 ± 0.5 | 69.2 ± 1.0 | 64.3 ± 3.6 | 75.8 ± 0.5 | 78.1 ± 0.6 | 70.8 ± 2.2 |
| TCRI | 97.0 ± 0.9 | 95.6 ± 3.5 | 92.0 ± 2.2 | 63.0 ± 1.7 | 65.1 ± 2.0 | 24.5 ± 3.7 | 67.4 ± 1.2 | 59.5 ± 1.5 | 67.2 ± 3.1 | 75.2 ± 1.5 | 76.8 ± 3.2 | 69.3 ± 1.3 |
| Focal | 97.5 ± 0.4 | 99.2 ± 0.4 | 95.4 ± 1.5 | 63.1 ± 0.9 | 65.3 ± 0.9 | 22.1 ± 1.7 | 69.0 ± 0.8 | 67.5 ± 3.7 | 64.7 ± 7.6 | 71.7 ± 3.2 | 72.4 ± 3.6 | 67.5 ± 1.4 |
| ReWeight | 95.7 ± 0.4 | 97.8 ± 0.4 | 87.9 ± 6.1 | 63.0 ± 0.6 | 64.9 ± 0.5 | 25.2 ± 1.7 | 66.0 ± 1.8 | 57.7 ± 3.1 | 75.4 ± 1.5 | 73.1 ± 2.4 | 70.3 ± 4.2 | 63.9 ± 4.6 |
| BSoftmax | 97.4 ± 0.8 | 99.1 ± 0.4 | 81.5 ± 12.2 | 97.1 ± 0.8 | 62.5 ± 1.3 | 64.2 ± 1.1 | 65.0 ± 1.6 | 54.4 ± 3.1 | 81.7 ± 3.4 | 67.0 ± 2.1 | 61.9 ± 2.6 | 75.5 ± 2.0 |
| LDAM | 96.8 ± 1.4 | 99.1 ± 0.5 | 71.5 ± 10.0 | 97.3 ± 0.6 | 62.6 ± 1.1 | 29.9 ± 1.0 | 68.4 ± 1.6 | 62.3 ± 3.1 | 68.4 ± 4.9 | 73.4 ± 0.8 | 71.7 ± 2.7 | 74.1 ± 3.7 |
| GINIDG | 96.4 ± 0.6 | 97.3 ± 0.8 | 82.8 ± 6.0 | 95.3 ± 1.5 | 65.2 ± 0.3 | 67.4 ± 0.4 | 66.0 ± 1.3 | 60.9 ± 3.1 | 65.0 ± 4.1 | 69.3 ± 0.7 | 69.3 ± 2.1 | 72.2 ± 3.6 |
| BoDA | 96.9 ± 1.0 | 98.8 ± 0.1 | 80.8 ± 8.0 | 97.0 ± 0.6 | 64.6 ± 1.0 | 66.8 ± 0.9 | 68.6 ± 0.8 | 66.9 ± 1.6 | 62.2 ± 3.7 | 75.2 ± 0.7 | 75.0 ± 0.7 | 76.8 ± 1.8 |
| SAMALTDG | 98.3 ± 0.3 | 99.6 ± 0.1 | 89.4 ± 2.4 | 97.2 ± 1.0 | 60.4 ± 1.9 | 62.6 ± 2.1 | 69.3 ± 1.1 | 62.0 ± 1.9 | 73.7 ± 5.4 | 69.8 ± 1.8 | 66.0 ± 2.8 | 84.6 ± 1.9 |
| NDCL | 99.1 ± 0.4 | 100.0 ± 0.0 | 94.7 ± 2.9 | 98.3 ± 0.7 | 64.8 ± 0.7 | 66.9 ± 1.0 | 72.2 ± 0.4 | 67.6 ± 1.4 | 65.9 ± 3.1 | 76.1 ± 0.1 | 77.8 ± 0.9 | 65.1 ± 2.5 |

Table 12: Detailed results of the target domain on PACS benchmark under the GINIDG setting.

| | A | | | C | | | P | | | S | | |
|----------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|-------------|-------------|
| | Average | Many | Few | Average | Many | Few | Average | Many | Few | Average | Many | Few |
| ERM | 82.5 ± 1.0 | 89.0 ± 0.3 | 62.1 ± 3.0 | 73.5 ± 1.0 | 78.6 ± 1.3 | 83.0 ± 0.3 | 96.0 ± 0.1 | 95.9 ± 0.5 | 92.8 ± 0.4 | 76.8 ± 0.7 | 72.0 ± 4.9 | 69.6 ± 5.4 |
| IRM | 82.0 ± 1.8 | 88.6 ± 1.8 | 67.6 ± 0.3 | 71.7 ± 2.8 | 76.4 ± 2.7 | 80.1 ± 1.1 | 96.3 ± 0.5 | 95.6 ± 1.0 | 94.0 ± 0.9 | 74.7 ± 2.4 | 74.8 ± 4.3 | 54.9 ± 2.4 |
| GroupDRO | 80.8 ± 1.1 | 86.0 ± 1.0 | 60.0 ± 6.2 | 73.2 ± 1.4 | 79.8 ± 1.0 | 82.6 ± 1.6 | 95.3 ± 0.6 | 93.6 ± 0.9 | 92.9 ± 1.6 | 69.3 ± 1.7 | 61.9 ± 6.0 | 64.7 ± 6.2 |
| Mixup | 82.4 ± 0.8 | 89.3 ± 0.3 | 57.7 ± 3.7 | 78.5 ± 0.7 | 84.5 ± 1.6 | 85.5 ± 1.1 | 95.8 ± 1.0 | 95.0 ± 1.0 | 93.5 ± 1.3 | 71.9 ± 1.2 | 74.1 ± 1.5 | 42.3 ± 17.1 |
| MLDG | 83.9 ± 1.1 | 89.0 ± 0.2 | 66.1 ± 3.0 | 75.8 ± 1.0 | 82.4 ± 0.8 | 82.9 ± 1.6 | 95.3 ± 0.3 | 93.7 ± 0.7 | 92.5 ± 1.5 | 73.1 ± 1.8 | 72.9 ± 3.5 | 47.6 ± 5.6 |
| CORAL | 83.6 ± 1.8 | 89.5 ± 0.6 | 66.2 ± 7.2 | 75.8 ± 1.4 | 84.0 ± 1.9 | 81.5 ± 1.1 | 96.6 ± 0.4 | 95.6 ± 0.4 | 94.5 ± 1.0 | 75.4 ± 1.7 | 69.1 ± 4.8 | 70.4 ± 6.6 |
| MMD | 82.8 ± 1.3 | 87.8 ± 1.0 | 61.5 ± 3.7 | 73.6 ± 2.6 | 79.6 ± 2.0 | 82.6 ± 0.5 | 95.8 ± 0.5 | 95.9 ± 0.8 | 91.9 ± 0.5 | 76.6 ± 0.5 | 69.3 ± 1.5 | 69.7 ± 4.3 |
| DANN | 84.6 ± 1.5 | 88.3 ± 1.4 | 71.1 ± 4.5 | 76.4 ± 0.4 | 80.8 ± 1.3 | 81.0 ± 0.3 | 96.5 ± 0.6 | 95.5 ± 0.8 | 94.5 ± 1.5 | 73.7 ± 1.0 | 66.4 ± 3.4 | 65.6 ± 6.8 |
| SagNet | 79.3 ± 2.3 | 84.5 ± 1.4 | 58.0 ± 8.1 | 75.9 ± 0.3 | 84.5 ± 0.9 | 82.7 ± 1.2 | 95.1 ± 0.3 | 93.4 ± 0.6 | 92.5 ± 1.4 | 74.7 ± 3.4 | 64.2 ± 10.1 | 67.5 ± 5.5 |
| VREx | 85.4 ± 0.5 | 89.5 ± 1.2 | 71.0 ± 5.1 | 75.9 ± 0.1 | 82.8 ± 0.6 | 83.6 ± 1.2 | 95.2 ± 0.4 | 94.4 ± 0.4 | 93.0 ± 0.5 | 74.8 ± 1.8 | 72.7 ± 3.9 | 59.4 ± 5.2 |
| Fish | 83.1 ± 1.0 | 89.2 ± 1.5 | 61.5 ± 4.8 | 72.2 ± 3.2 | 77.9 ± 3.8 | 82.6 ± 2.3 | 97.4 ± 0.3 | 96.8 ± 0.4 | 95.5 ± 0.2 | 69.9 ± 1.9 | 61.3 ± 4.1 | 79.6 ± 1.1 |
| Fishr | 84.1 ± 1.4 | 89.7 ± 0.4 | 64.0 ± 6.0 | 78.1 ± 0.5 | 81.7 ± 1.1 | 85.4 ± 0.9 | 96.4 ± 0.3 | 96.3 ± 0.9 | 93.2 ± 0.2 | 70.0 ± 1.8 | 56.2 ± 7.5 | 81.9 ± 1.8 |
| EQRM | 82.6 ± 1.6 | 87.3 ± 1.1 | 68.1 ± 7.0 | 77.3 ± 0.8 | 81.0 ± 1.0 | 85.0 ± 1.2 | 97.0 ± 0.1 | 96.8 ± 0.4 | 94.5 ± 0.6 | 72.5 ± 2.8 | 64.0 ± 5.2 | 80.5 ± 1.4 |
| EQRM | 82.3 ± 2.0 | 90.2 ± 0.9 | 82.1 ± 1.7 | 64.9 ± 6.4 | 81.7 ± 1.9 | 87.0 ± 1.3 | 96.1 ± 0.4 | 95.7 ± 0.9 | 93.9 ± 0.2 | 76.5 ± 2.4 | 70.2 ± 5.4 | 83.6 ± 0.9 |
| RDM | 87.8 ± 0.3 | 93.9 ± 0.4 | 72.5 ± 0.9 | 76.5 ± 0.9 | 83.5 ± 0.7 | 83.4 ± 1.0 | 96.9 ± 0.3 | 97.1 ± 0.2 | 94.2 ± 0.8 | 76.0 ± 2.0 | 76.0 ± 1.7 | 58.8 ± 1.1 |
| PGrad | 84.5 ± 0.7 | 88.5 ± 1.3 | 69.6 ± 6.4 | 76.8 ± 0.9 | 81.8 ± 0.3 | 86.5 ± 1.4 | 96.7 ± 0.6 | 95.6 ± 1.2 | 95.0 ± 1.2 | 75.0 ± 1.7 | 63.3 ± 5.2 | 54.5 ± 5.8 |
| TCRI | 82.4 ± 0.6 | 89.9 ± 1.0 | 60.7 ± 2.1 | 77.6 ± 1.0 | 83.1 ± 1.1 | 84.4 ± 1.4 | 95.8 ± 0.8 | 95.0 ± 1.2 | 92.8 ± 1.2 | 68.4 ± 1.5 | 64.7 ± 7.0 | 52.3 ± 0.9 |
| Focal | 83.9 ± 1.9 | 90.1 ± 0.9 | 61.2 ± 8.0 | 76.4 ± 0.4 | 82.6 ± 0.5 | 85.4 ± 0.9 | 96.3 ± 0.5 | 96.1 ± 0.5 | 93.4 ± 1.1 | 71.8 ± 2.2 | 62.0 ± 2.1 | 80.6 ± 5.1 |
| ReWeight | 84.0 ± 1.1 | 87.9 ± 2.2 | 70.6 ± 3.3 | 77.3 ± 1.8 | 77.9 ± 3.8 | 84.5 ± 0.9 | 96.0 ± 0.5 | 94.3 ± 1.1 | 94.2 ± 0.4 | 75.7 ± 0.9 | 63.3 ± 2.7 | 83.3 ± 0.6 |
| BSoftmax | 85.6 ± 1.4 | 90.4 ± 0.6 | 68.8 ± 5.1 | 78.3 ± 0.3 | 83.9 ± 0.8 | 85.2 ± 1.0 | 96.0 ± 0.5 | 94.9 ± 0.5 | 93.6 ± 0.9 | 72.0 ± 0.4 | 67.8 ± 0.8 | 76.5 ± 2.4 |
| LDAM | 85.6 ± 1.4 | 90.4 ± 0.6 | 68.8 ± 5.1 | 78.3 ± 0.3 | 83.9 ± 0.8 | 85.2 ± 1.0 | 96.0 ± 0.5 | 94.9 ± 0.5 | 93.6 ± 0.9 | 72.0 ± 0.4 | 67.8 ± 0.8 | 76.5 ± 2.4 |
| GINIDG | 75.3 ± 1.2 | 80.0 ± 4.9 | 54.9 ± 7.3 | 76.2 ± 1.0 | 83.6 ± 1.4 | 82.1 ± 1.8 | 96.0 ± 0.6 | 93.7 ± 0.9 | 94.8 ± 0.8 | 76.9 ± 3.0 | 69.9 ± 4.9 | 84.0 ± 1.2 |
| BoDA | 83.8 ± 1.3 | 90.2 ± 0.2 | 84.4 ± 1.5 | 74.3 ± 0.5 | 82.1 ± 1.3 | 82.0 ± 0.9 | 95.5 ± 0.7 | 93.2 ± 1.2 | 93.9 ± 1.1 | 74.7 ± 1.7 | 76.8 ± 1.2 | 75.9 ± 1.6 |
| SAMALTDG | 81.3 ± 2.1 | 87.9 ± 0.6 | 87.4 ± 1.8 | 77.3 ± 0.7 | 83.7 ± 0.9 | 83.7 ± 1.4 | 96.7 ± 0.3 | 94.9 ± 0.3 | 95.6 ± 1.0 | 74.2 ± 2.5 | 71.5 ± 1.2 | 78.0 ± 4.0 |
| NDCL | 87.7 ± 0.6 | 90.3 ± 0.6 | 89.0 ± 1.1 | 80.4 ± 4.8 | 84.4 ± 0.4 | 85.2 ± 0.6 | 97.3 ± 0.3 | 96.6 ± 0.3 | 95.5 ± 1.2 | 79.3 ± 0.4 | 74.8 ± 1.8 | 83.8 ± 0.9 |
| | | | | | | | | | | | | 76.9 ± 7.0 |

Table 13: Detailed results of the target domain on VLCS benchmark under the TotalHeavyTail setting.

| | C | | | L | | | S | | | Y | | |
|----------|------------|-------------|------------|-------------|------------|------------|------------|------------|------------|------------|------------|-------------|
| | Average | Many | Few | Average | Many | Few | Average | Many | Few | Average | Many | Few |
| ERM | 91.0 ± 2.1 | 99.8 ± 0.1 | 81.5 ± 6.6 | 66.3 ± 0.3 | 45.5 ± 1.8 | 54.7 ± 0.9 | 37.5 ± 0.8 | 66.8 ± 0.8 | 81.8 ± 3.1 | 49.7 ± 5.4 | 36.8 ± 1.7 | 156.6 ± 3.1 |
| IRM | 89.3 ± 0.9 | 98.7 ± 1.1 | 78.9 ± 2.8 | 68.7 ± 10.1 | 65.1 ± 0.2 | 42.3 ± 1.7 | 53.3 ± 2.0 | 32.9 ± 4.5 | 60.8 ± 1.3 | 91.7 ± 0.5 | 38.6 ± 2.0 | 156.6 ± 3.1 |
| GroupDRO | 92.7 ± 1.7 | 98.7 ± 0.9 | 84.5 ± 6.0 | 77.5 ± 5.9 | 64.2 ± 0.6 | 38.0 ± 2.0 | 57.6 ± 1.2 | 37.4 ± 1.5 | 68.8 ± 1.8 | 87.1 ± 2.7 | 49.9 ± 9.8 | 156.6 ± 3.1 |
| Mixup | 87.7 ± 1.2 | 99.8 ± 0.1 | 70.9 ± 6.2 | 70.1 ± 3.8 | 65.3 ± 0.4 | 40.9 ± 1.2 | 58.2 ± 1.2 | 37.7 ± 1.2 | 62.2 ± 2.0 | 91.4 ± 1.6 | 39.2 ± 3.3 | 156.6 ± 3.1 |
| MLDG | 92.2 ± 1.0 | 99.9 ± 0.1 | 78.8 ± 8.7 | 82.3 ± 5.7 | 65.2 ± 0.7 | 40.3 ± 1.4 | 58.3 ± 0.9 | 36.9 ± 3.6 | 65.5 ± 0.4 | 86.9 ± 1.6 | 44.1 ± 3.4 | 156.6 ± 3.1 |
| CORAL | 92.5 ± 0.4 | 99.9 ± 0.1 | 86.0 ± 2.3 | 76.4 ± 0.6 | 67.2 ± 0.8 | 46.6 ± 1.6 | 54.7 ± 0.7 | 36.9 ± 1.2 | 63.9 ± 1.1 | 90.1 ± 1.7 | 41.4 ± 3.0 | 156.6 ± 3.1 |
| MMD | 91.6 ± 1.2 | 99.5 ± 0.1 | 82.6 ± 4.4 | 78.5 ± 6.1 | 64.2 ± 1.8 | 40.0 ± 5.5 | 55.3 ± 1.2 | 36.4 ± 1.3 | 64.1 ± 0.1 | 91.7 ± 0.5 | 37.3 ± 2.5 | 156.6 ± 3.1 |
| DANN | 91.7 ± 2.6 | 98.9 ± 0.9 | 83.0 ± 5.5 | 82.4 ± 4.5 | 62.3 ± 1.9 | 31.8 ± 4.5 | 58.2 ± 1.4 | 41.3 ± 2.0 | 64.6 ± 2.7 | 84.7 ± 1.0 | 49.7 ± 3.0 | 156.6 ± 3.1 |
| SagNet | 95.1 ± 1.5 | 99.9 ± 0.1 | 86.7 ± 7.7 | 85.5 ± 7.3 | 67.4 ± 0.5 | 38.3 ± 1.4 | 57.0 ± 1.9 | 37.0 ± 2.1 | 61.2 ± 2.1 | 88.7 ± 2.3 | 42.2 ± 6.1 | 156.6 ± 3.1 |
| VREx | 91.4 ± 1.1 | 96.8 ± 2.3 | 85.2 ± 4.6 | 80.6 ± 3.6 | 64.4 ± 0.3 | 47.4 ± 3.7 | 57.0 ± 1.9 | 34.8 ± 2.5 | 62.2 ± 2.1 | 84.8 ± 2.5 | 49.1 ± 5.7 | 156.6 ± 3.1 |
| Fish | 94.8 ± 0.6 | 100.0 ± 0.0 | 81.7 ± 4.9 | 90.3 ± 3.1 | 65.6 ± 0.5 | 42.6 ± 2.0 | 55.0 ± 1.7 | 36.2 ± 1.8 | 65.4 ± 0.1 | 91.3 ± 1.6 | 39.5 ± 5.0 | 156.6 ± 3.1 |
| Fishr | 92.0 ± 1.1 | 100.0 ± 0.0 | 77.0 ± 3.8 | 84.1 ± 3.1 | 65.7 ± 0.6 | 44.9 ± 0.1 | 55.1 ± 1.9 | 29.5 ± 0.9 | 65.6 ± 1.9 | 83.2 ± 1.6 | 49.3 ± 2.7 | 156.6 ± 3.1 |
| EORM | 92.1 ± 0.6 | 99.9 ± 0.1 | 74.3 ± 2.0 | 80.5 ± 5.2 | 64.6 ± 0.3 | 39.2 ± 1.3 | 57.4 ± 1.1 | 38.0 ± 0.9 | 63.7 ± 1.4 | 83.4 ± 4.2 | 46.2 ± 7.0 | 156.6 ± 3.1 |
| RDM | 91.7 ± 0.4 | 100.0 ± 0.0 | 82.6 ± 3.5 | 73.0 ± 6.5 | 64.0 ± 1.8 | 38.2 ± 4.9 | 57.2 ± 0.6 | 35.8 ± 0.9 | 60.6 ± 2.3 | 88.0 ± 2.0 | 39.3 ± 6.3 | 156.6 ± 3.1 |
| PGrad | 92.5 ± 0.9 | 100.0 ± 0.0 | 83.8 ± 3.3 | 81.5 ± 3.7 | 65.6 ± 0.3 | 42.2 ± 0.7 | 56.8 ± 0.3 | 35.6 ± 0.5 | 60.1 ± 0.5 | 94.0 ± 0.6 | 36.0 ± 2.4 | 156.6 ± 3.1 |
| TCRI | 91.0 ± 0.8 | 98.1 ± 1.5 | 79.3 ± 5.6 | 83.7 ± 1.5 | 63.5 ± 0.9 | 36.0 ± 2.2 | 58.6 ± 3.1 | 40.8 ± 0.5 | 61.4 ± 1.9 | 90.4 ± 2.7 | 35.0 ± 2.5 | 156.6 ± 3.1 |
| Focal | 89.5 ± 1.3 | 98.7 ± 1.1 | 82.3 ± 3.0 | 76.6 ± 3.8 | 65.5 ± 0.6 | 43.2 ± 1.2 | 53.7 ± 0.4 | 37.2 ± 3.1 | 62.2 ± 3.4 | 88.6 ± 3.2 | 41.1 ± 4.8 | 156.6 ± 3.1 |
| ReWeight | 94.8 ± 1.1 | 99.9 ± 0.1 | 80.3 ± 7.8 | 93.5 ± 0.7 | 58.5 ± 0.9 | 23.6 ± 2.2 | 60.6 ± 1.5 | 45.7 ± 3.3 | 63.4 ± 2.9 | 75.5 ± 7.2 | 49.8 ± 5.7 | 156.6 ± 3.1 |
| BSoftmax | 94.8 ± 0.1 | 99.0 ± 0.6 | 76.2 ± 0.9 | 95.0 ± 1.3 | 65.9 ± 0.7 | 42.2 ± 2.1 | 57.9 ± 0.6 | 42.1 ± 1.7 | 65.4 ± 2.1 | 72.0 ± 5.5 | 56.8 ± 7.6 | 156.6 ± 3.1 |
| LDAM | 93.0 ± 1.4 | 99.8 ± 0.2 | 77.7 ± 5.8 | 86.1 ± 2.1 | 65.9 ± 0.8 | 43.6 ± 2.2 | 55.7 ± 2.0 | 41.1 ± 2.6 | 60.8 ± 3.4 | 88.8 ± 3.4 | 42.4 ± 2.8 | 156.6 ± 3.1 |
| GINIDG | 88.9 ± 1.1 | 100.0 ± 0.0 | 65.1 ± 2.9 | 74.7 ± 8.2 | 66.7 ± 0.5 | 49.0 ± 1.4 | 54.5 ± 0.4 | 29.0 ± 3.9 | 59.1 ± 3.7 | 87.3 ± 1.6 | 42.8 ± 2.1 | 156.6 ± 3.1 |
| BoDA | 91.3 ± 0.5 | 99.9 ± 0.0 | 79.8 ± 4.8 | 74.1 ± 3.1 | 65.9 ± 1.2 | 44.0 ± 3.5 | 53.6 ± 0.6 | 37.5 ± 1.4 | 59.5 ± 2.5 | 89.4 ± 3.4 | 41.9 ± 3.2 | 156.6 ± 3.1 |
| SAMALTDG | 88.8 ± 1.2 | 95.2 ± 2.0 | 77.4 ± 3.2 | 86.9 ± 4.4 | 61.8 ± 1.8 | 31.6 ± 3.9 | 56.5 ± 0.4 | 42.2 ± 1.7 | 66.7 ± 2.2 | 73.3 ± 7.5 | 61.0 ± 3.6 | 156.6 ± 3.1 |
| NDCL | 95.7 ± 0.7 | 100.0 ± 0.0 | 78.9 ± 4.6 | 94.6 ± 0.9 | 65.2 ± 1.0 | 39.9 ± 3.4 | 59.5 ± 2.8 | 41.7 ± 2.5 | 65.3 ± 1.5 | 52.4 ± 4.1 | 58.4 ± 1.7 | 156.6 ± 3.1 |

Table 14: Detailed results of the target domain on PACS benchmark under the TotalHeavyTail setting.

| | A | | | C | | | P | | | S | | |
|----------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|-------------|
| | Average | Many | Few | Average | Many | Few | Average | Many | Few | Average | Many | Few |
| ERM | 68.6 ± 2.6 | 90.8 ± 1.6 | 64.8 ± 6.2 | 64.2 ± 2.0 | 93.6 ± 1.0 | 36.1 ± 5.0 | 92.9 ± 0.5 | 92.9 ± 1.5 | 94.7 ± 2.2 | 91.3 ± 0.9 | 87.5 ± 5.3 | 39.7 ± 15.0 |
| IRM | 72.8 ± 1.6 | 91.3 ± 1.4 | 64.9 ± 1.6 | 67.9 ± 4.5 | 67.1 ± 1.0 | 90.4 ± 1.2 | 45.1 ± 1.5 | 89.5 ± 3.7 | 89.5 ± 2.6 | 92.4 ± 1.7 | 84.3 ± 3.6 | 49.7 ± 8.1 |
| GroupDRO | 75.4 ± 2.2 | 86.1 ± 0.2 | 74.4 ± 1.8 | 65.4 ± 5.2 | 66.9 ± 1.8 | 93.0 ± 0.5 | 44.0 ± 3.0 | 85.6 ± 2.4 | 91.8 ± 1.5 | 96.1 ± 0.4 | 86.6 ± 2.4 | 38.3 ± 11.1 |
| Mixup | 71.7 ± 0.6 | 90.4 ± 1.1 | 67.7 ± 1.0 | 60.6 ± 1.2 | 63.6 ± 0.6 | 88.3 ± 3.1 | 42.4 ± 2.7 | 85.1 ± 1.3 | 88.7 ± 2.3 | 95.2 ± 2.0 | 79.2 ± 3.1 | 13.2 ± 3.7 |
| MLDG | 69.8 ± 2.3 | 89.0 ± 1.6 | 66.4 ± 3.3 | 58.6 ± 5.5 | 68.3 ± 1.6 | 92.1 ± 2.2 | 47.7 ± 3.8 | 82.8 ± 4.0 | 92.2 ± 0.8 | 94.4 ± 0.2 | 89.7 ± 2.5 | 30.3 ± 2.1 |
| CORAL | 65.7 ± 1.8 | 92.3 ± 0.4 | 56.7 ± 4.2 | 52.1 ± 3.8 | 66.7 ± 2.0 | 89.9 ± 2.5 | 45.5 ± 4.8 | 87.6 ± 1.2 | 90.7 ± 1.5 | 92.4 ± 2.2 | 87.5 ± 0.5 | 22.0 ± 8.2 |
| MMD | 69.5 ± 2.2 | 85.3 ± 1.7 | 64.8 ± 1.6 | 63.8 ± 2.5 | 65.4 ± 1.2 | 93.2 ± 0.9 | 41.2 ± 2.2 | 86.0 ± 0.9 | 90.7 ± 1.6 | 95.3 ± 1.4 | 83.1 ± 1.3 | 30.6 ± 7.8 |
| DANN | 76.3 ± 2.1 | 92.3 ± 1.0 | 71.9 ± 2.1 | 66.6 ± 3.9 | 69.0 ± 1.3 | 90.9 ± 2.5 | 49.5 ± 3.2 | 85.5 ± 0.9 | 93.8 ± 1.4 | 94.8 ± 2.0 | 91.3 ± 2.7 | 41.0 ± 6.8 |
| SagNet | 65.1 ± 0.4 | 85.8 ± 2.3 | 60.1 ± 0.8 | 53.6 ± 5.5 | 71.2 ± 1.8 | 87.0 ± 1.2 | 54.6 ± 3.8 | 88.2 ± 2.3 | 91.4 ± 1.1 | 95.7 ± 0.9 | 86.2 ± 2.5 | 52.4 ± 11.5 |
| VREx | 67.1 ± 1.4 | 85.4 ± 3.8 | 65.1 ± 2.9 | 49.8 ± 5.4 | 64.7 ± 0.8 | 91.0 ± 1.4 | 40.1 ± 2.4 | 88.1 ± 1.2 | 90.8 ± 1.3 | 92.4 ± 2.3 | 89.2 ± 2.0 | 45.9 ± 11.0 |
| Fish | 74.5 ± 1.5 | 91.7 ± 1.1 | 69.2 ± 3.2 | 67.6 ± 2.5 | 63.4 ± 3.1 | 94.1 ± 1.5 | 35.8 ± 4.5 | 90.5 ± 2.0 | 90.5 ± 1.2 | 94.9 ± 1.4 | 84.2 ± 1.1 | 34.1 ± 8.6 |
| Fishr | 69.2 ± 0.8 | 90.9 ± 1.0 | 62.2 ± 2.5 | 61.3 ± 3.1 | 64.5 ± 0.9 | 89.1 ± 2.6 | 40.4 ± 2.0 | 91.1 ± 2.2 | 92.7 ± 0.5 | 94.4 ± 0.3 | 89.0 ± 0.8 | 41.4 ± 10.0 |
| EORM | 73.7 ± 1.0 | 88.9 ± 1.9 | 68.4 ± 0.9 | 68.8 ± 5.4 | 68.8 ± 1.8 | 91.4 ± 0.9 | 48.2 ± 4.2 | 89.6 ± 1.4 | 91.8 ± 0.5 | 93.7 ± 1.4 | 88.5 ± 1.9 | 41.7 ± 8.4 |
| PDG | 72.7 ± 1.2 | 90.5 ± 1.0 | 66.6 ± 1.8 | 64.9 ± 3.6 | 66.2 ± 1.4 | 91.5 ± 1.3 | 42.7 ± 2.7 | 88.5 ± 1.7 | 92.6 ± 1.3 | 92.9 ± 1.2 | 87.1 ± 1.1 | 47.9 ± 4.8 |
| RDM | 72.4 ± 1.2 | 92.8 ± 0.9 | 66.8 ± 0.6 | 60.9 ± 5.2 | 64.0 ± 0.4 | 93.5 ± 1.9 | 36.6 ± 0.9 | 89.7 ± 1.3 | 92.6 ± 0.6 | 98.1 ± 0.3 | 86.1 ± 0.9 | 33.3 ± 10.9 |
| TCRI | 71.3 ± 3.6 | 90.5 ± 0.9 | 63.2 ± 4.8 | 63.8 ± 5.8 | 67.7 ± 1.2 | 89.0 ± 0.5 | 47.2 ± 2.3 | 90.8 ± 1.8 | 92.3 ± 0.2 | 95.3 ± 0.2 | 89.1 ± 1.0 | 45.7 ± 6.3 |
| Focal | 69.3 ± 3.7 | 89.3 ± 1.9 | 65.8 ± 5.1 | 53.4 ± 4.9 | 64.7 ± 1.3 | 92.9 ± 1.2 | 39.3 ± 2.2 | 88.8 ± 0.9 | 92.2 ± 0.6 | 94.5 ± 1.3 | 89.2 ± 1.3 | 37.1 ± 10.1 |
| ReWeight | 73.8 ± 2.2 | 87.5 ± 3.3 | 71.7 ± 4.5 | 65.1 ± 3.8 | 66.1 ± 1.9 | 94.6 ± 0.5 | 49.1 ± 3.1 | 89.9 ± 0.5 | 94.8 ± 0.6 | 92.3 ± 2.1 | 93.2 ± 1.5 | 52.1 ± 7.4 |
| BSoftmax | 74.0 ± 0.7 | 88.8 ± 0.5 | 69.7 ± 0.5 | 69.3 ± 4.8 | 68.3 ± 2.9 | 94.7 ± 5.1 | 49.5 ± 7.0 | 95.5 ± 1.3 | 93.5 ± 1.0 | 90.0 ± 3.6 | 91.2 ± 2.3 | 63.6 ± 5.9 |
| LDAM | 70.6 ± 1.6 | 87.1 ± 0.3 | 65.4 ± 2.0 | 60.9 ± 2.8 | 68.2 ± 2.4 | 90.0 ± 1.0 | 46.3 ± 3.2 | 90.7 ± 1.9 | 93.7 ± 1.1 | 92.8 ± 2.0 | 89.7 ± 1.9 | 45.8 ± 8.1 |
| GINIDG | 66.8 ± 0.7 | 89.5 ± 0.9 | 57.8 ± 0.8 | 60.3 ± 2.0 | 63.0 ± 1.0 | 91.5 ± 0.6 | 39.2 ± 3.0 | 82.6 ± 1.5 | 91.3 ± 1.7 | 93.6 ± 1.2 | 86.8 ± 3.0 | 29.7 ± 3.1 |
| BoDA | 69.8 ± 2.8 | 91.0 ± 0.4 | 66.6 ± 3.4 | 50.7 ± 2.7 | 65.3 ± 1.1 | 86.1 ± 1.6 | 45.1 ± 3.0 | 88.9 ± 2.1 | 90.9 ± 0.8 | 94.2 ± 2.3 | 87.4 ± 1.3 | 32.4 ± 10.0 |
| SAMALTDG | 70.4 ± 1.6 | 89.7 ± 1.2 | 65.0 ± 0.8 | 63.1 ± 5.1 | 66.4 ± 2.2 | 88.3 ± 3.6 | 43.9 ± 5.4 | 92.7 ± 3.2 | 92.9 ± 1.3 | 92.5 ± 2.8 | 91.5 ± 1.6 | 53.7 ± 10.5 |
| NDCL | 74.6 ± 1.9 | 84.1 ± 1.7 | 74.8 ± 2.3 | 64.9 ± 1.1 | 70.3 ± 1.4 | 86.1 ± 3.5 | 51.9 ± 4.3 | 94.0 ± 0.8 | 93.6 ± 0.7 | 95.7 ± 0.5 | 89.9 ± 1.9 | 66.4 ± 4.1 |

Table 15: Detailed results of the target domain on OfficeHome benchmark under the TotalHeavyTail setting.

| | A | | | C | | | P | | | R | | |
|----------|-------------------|-------------------|------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| | Average | Many | Few | Average | Many | Few | Average | Many | Few | Average | Many | Few |
| ERM | 41.9 ± 0.3 | 62.5 ± 0.7 | 61.7 ± 1.1 | 20.0 ± 0.9 | 38.5 ± 0.7 | 58.4 ± 1.2 | 19.9 ± 1.8 | 49.8 ± 0.8 | 84.3 ± 0.9 | 63.8 ± 1.6 | 30.4 ± 1.4 | 75.4 ± 1.1 |
| IRM | 39.3 ± 0.9 | 62.9 ± 1.1 | 58.1 ± 1.9 | 18.6 ± 1.2 | 37.0 ± 0.3 | 56.6 ± 2.1 | 18.9 ± 0.2 | 46.3 ± 2.0 | 82.9 ± 0.6 | 58.7 ± 4.0 | 27.2 ± 2.5 | 71.3 ± 3.1 |
| GroupDRO | 40.5 ± 0.3 | 61.4 ± 1.1 | 58.9 ± 1.3 | 18.7 ± 0.7 | 39.9 ± 0.4 | 58.5 ± 1.4 | 21.7 ± 0.9 | 48.1 ± 0.7 | 84.1 ± 0.6 | 61.1 ± 1.6 | 29.7 ± 1.3 | 77.6 ± 0.5 |
| Mixup | 40.6 ± 0.6 | 65.3 ± 1.7 | 58.3 ± 0.5 | 17.9 ± 1.6 | 37.1 ± 1.0 | 58.5 ± 1.4 | 18.5 ± 1.4 | 48.6 ± 0.7 | 83.9 ± 0.4 | 63.1 ± 1.6 | 28.3 ± 0.9 | 77.0 ± 0.5 |
| MLDG | 40.6 ± 1.2 | 62.2 ± 1.7 | 59.4 ± 0.8 | 18.6 ± 0.9 | 37.6 ± 0.8 | 55.3 ± 1.8 | 20.7 ± 1.0 | 49.2 ± 0.9 | 81.5 ± 0.7 | 61.5 ± 1.8 | 30.5 ± 1.0 | 77.1 ± 1.1 |
| CORAL | 40.3 ± 0.8 | 64.6 ± 1.5 | 57.5 ± 1.6 | 17.1 ± 0.6 | 39.1 ± 0.7 | 61.8 ± 0.4 | 17.3 ± 0.9 | 49.3 ± 1.4 | 87.0 ± 0.6 | 61.3 ± 2.4 | 29.2 ± 1.3 | 79.0 ± 0.5 |
| MMD | 38.8 ± 1.3 | 59.2 ± 1.4 | 59.7 ± 1.6 | 17.3 ± 1.1 | 39.7 ± 1.1 | 59.7 ± 1.0 | 21.0 ± 1.2 | 50.3 ± 1.2 | 83.8 ± 1.3 | 62.2 ± 0.8 | 31.7 ± 0.7 | 75.9 ± 1.3 |
| DANN | 40.1 ± 0.5 | 64.6 ± 2.6 | 58.9 ± 1.6 | 16.6 ± 0.3 | 37.2 ± 1.2 | 68.4 ± 3.1 | 19.1 ± 1.4 | 47.7 ± 1.2 | 82.3 ± 1.4 | 62.0 ± 1.2 | 26.9 ± 1.5 | 75.7 ± 1.0 |
| SagNet | 42.3 ± 0.2 | 65.2 ± 0.7 | 61.1 ± 1.9 | 19.2 ± 0.5 | 38.9 ± 0.8 | 60.4 ± 0.2 | 17.9 ± 0.8 | 49.2 ± 1.0 | 84.8 ± 0.2 | 62.9 ± 1.3 | 29.4 ± 1.0 | 79.7 ± 1.6 |
| VRex | 40.5 ± 0.3 | 62.0 ± 0.3 | 58.6 ± 2.5 | 19.7 ± 0.6 | 38.1 ± 1.1 | 68.1 ± 1.2 | 20.1 ± 2.3 | 49.1 ± 0.2 | 84.9 ± 0.8 | 60.6 ± 0.3 | 29.5 ± 0.3 | 76.9 ± 1.6 |
| Fish | 41.8 ± 0.6 | 64.7 ± 1.0 | 59.7 ± 0.5 | 20.3 ± 0.4 | 38.8 ± 0.3 | 72.0 ± 1.9 | 18.1 ± 0.7 | 50.2 ± 0.3 | 86.1 ± 1.0 | 63.1 ± 1.9 | 30.1 ± 0.5 | 78.9 ± 2.1 |
| Fishr | 41.5 ± 0.8 | 64.7 ± 0.5 | 60.7 ± 0.4 | 19.3 ± 1.3 | 38.3 ± 0.7 | 57.3 ± 0.4 | 20.1 ± 1.2 | 50.1 ± 0.9 | 84.3 ± 0.5 | 62.2 ± 0.7 | 31.4 ± 1.2 | 78.2 ± 1.2 |
| EQRm | 40.5 ± 0.3 | 66.4 ± 1.5 | 56.7 ± 0.7 | 18.3 ± 0.5 | 39.0 ± 0.7 | 71.4 ± 0.6 | 19.7 ± 1.2 | 50.5 ± 0.5 | 84.2 ± 2.1 | 62.2 ± 1.3 | 32.3 ± 1.2 | 76.8 ± 0.5 |
| RDM | 42.1 ± 0.8 | 62.7 ± 1.6 | 59.5 ± 1.5 | 20.7 ± 0.8 | 39.4 ± 0.5 | 69.3 ± 1.8 | 21.8 ± 1.0 | 51.3 ± 0.4 | 82.8 ± 1.4 | 63.2 ± 1.5 | 32.4 ± 0.8 | 76.3 ± 0.9 |
| PGrad | 42.4 ± 1.0 | 65.4 ± 0.8 | 60.8 ± 1.1 | 19.8 ± 0.6 | 41.5 ± 0.3 | 75.3 ± 0.4 | 21.5 ± 1.1 | 52.1 ± 0.4 | 85.6 ± 0.7 | 64.4 ± 1.5 | 32.8 ± 0.2 | 78.8 ± 0.8 |
| TCRI | 42.4 ± 0.3 | 60.7 ± 1.3 | 59.3 ± 0.2 | 24.3 ± 0.7 | 39.4 ± 0.3 | 66.7 ± 1.6 | 21.4 ± 1.0 | 52.7 ± 0.5 | 81.6 ± 0.3 | 62.3 ± 0.8 | 36.7 ± 0.2 | 78.5 ± 1.5 |
| Focal | 40.3 ± 0.9 | 65.4 ± 1.4 | 57.7 ± 1.3 | 17.6 ± 0.5 | 36.6 ± 0.5 | 66.3 ± 2.6 | 18.6 ± 0.4 | 49.1 ± 0.8 | 83.7 ± 1.9 | 59.0 ± 0.9 | 30.8 ± 1.5 | 78.2 ± 0.2 |
| ReWeight | 41.8 ± 0.8 | 65.6 ± 1.0 | 61.3 ± 1.0 | 18.8 ± 1.2 | 37.8 ± 1.3 | 64.2 ± 3.7 | 22.4 ± 0.6 | 52.3 ± 0.7 | 83.7 ± 1.3 | 63.1 ± 1.2 | 34.5 ± 0.3 | 79.7 ± 0.8 |
| BSofmax | 42.2 ± 0.2 | 55.9 ± 0.1 | 56.0 ± 0.8 | 26.5 ± 0.5 | 39.7 ± 1.1 | 50.2 ± 2.2 | 26.8 ± 1.6 | 53.6 ± 0.6 | 79.2 ± 1.8 | 60.2 ± 0.5 | 39.3 ± 1.3 | 58.8 ± 0.9 |
| LDAM | 41.7 ± 0.8 | 61.2 ± 1.1 | 59.5 ± 0.9 | 21.3 ± 0.1 | 38.6 ± 0.1 | 70.2 ± 1.3 | 19.7 ± 0.6 | 51.3 ± 0.6 | 82.1 ± 2.1 | 63.9 ± 1.3 | 32.8 ± 1.3 | 76.0 ± 0.4 |
| GINiDG | 39.9 ± 0.2 | 60.3 ± 1.8 | 58.5 ± 1.0 | 19.9 ± 0.5 | 39.0 ± 0.3 | 64.4 ± 1.4 | 22.2 ± 1.0 | 48.5 ± 1.0 | 81.8 ± 1.2 | 59.0 ± 0.4 | 30.4 ± 1.7 | 74.9 ± 2.0 |
| BoDA | 39.6 ± 0.6 | 62.7 ± 0.7 | 59.3 ± 1.6 | 15.5 ± 0.6 | 39.3 ± 0.5 | 76.2 ± 1.0 | 17.8 ± 0.9 | 48.0 ± 0.3 | 85.8 ± 1.0 | 61.9 ± 0.5 | 26.9 ± 0.1 | 74.4 ± 1.7 |
| SAMALTDG | 42.4 ± 1.5 | 63.7 ± 1.8 | 59.2 ± 1.3 | 20.5 ± 1.6 | 37.8 ± 0.9 | 63.7 ± 3.4 | 20.5 ± 0.6 | 49.2 ± 0.9 | 84.6 ± 0.7 | 60.5 ± 0.5 | 30.3 ± 1.2 | 77.8 ± 0.5 |
| NDCL | 43.0 ± 0.0 | 59.2 ± 0.4 | 59.0 ± 1.5 | 22.8 ± 0.6 | 41.0 ± 1.1 | 65.4 ± 4.6 | 24.9 ± 0.8 | 54.8 ± 0.2 | 78.1 ± 0.3 | 64.3 ± 0.7 | 39.9 ± 0.6 | 80.4 ± 0.6 |

31

Table 16: Detailed results of the target domain on VLCS benchmark under the Duality setting.

| | C | | | L | | | S | | | V | | |
|----------|-------------------|-------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|--------------------|-------------------|
| | Average | Many | Few | Average | Many | Few | Average | Many | Few | Average | Many | Few |
| ERM | 74.2 ± 11.2 | 67.2 ± 20.0 | 82.0 ± 4.3 | 81.7 ± 8.5 | 48.6 ± 0.3 | 66.8 ± 0.5 | 43.6 ± 1.1 | 11.4 ± 1.1 | 61.5 ± 6.9 | 49.5 ± 0.6 | 38.3 ± 3.9 | 83.7 ± 0.9 |
| IRM | 67.5 ± 12.7 | 52.5 ± 21.0 | 88.1 ± 1.7 | 92.8 ± 3.6 | 47.7 ± 1.2 | 63.6 ± 2.1 | 42.1 ± 0.2 | 10.6 ± 1.8 | 57.5 ± 3.7 | 47.8 ± 2.4 | 42.2 ± 2.2 | 71.6 ± 4.1 |
| GroupDRO | 83.4 ± 4.9 | 81.1 ± 8.1 | 77.5 ± 2.3 | 94.8 ± 0.9 | 48.1 ± 1.1 | 69.6 ± 1.7 | 40.6 ± 1.4 | 8.5 ± 1.2 | 67.3 ± 4.2 | 50.3 ± 1.0 | 53.6 ± 10.8 | 87.6 ± 8.3 |
| Mixup | 75.4 ± 7.2 | 67.7 ± 13.7 | 78.4 ± 8.7 | 94.0 ± 2.0 | 48.5 ± 1.0 | 57.3 ± 3.8 | 41.1 ± 0.5 | 11.6 ± 2.9 | 61.2 ± 10.7 | 50.1 ± 1.8 | 52.5 ± 2.4 | 80.1 ± 2.3 |
| MLDG | 65.5 ± 11.9 | 52.1 ± 17.2 | 86.3 ± 7.5 | 88.7 ± 3.9 | 47.2 ± 1.8 | 68.0 ± 2.8 | 41.8 ± 0.6 | 11.9 ± 0.8 | 55.8 ± 9.7 | 49.5 ± 1.7 | 50.9 ± 2.1 | 77.7 ± 2.5 |
| CORAL | 69.6 ± 8.8 | 62.5 ± 14.3 | 67.2 ± 5.3 | 92.1 ± 3.9 | 50.2 ± 2.0 | 63.4 ± 4.0 | 47.0 ± 0.3 | 24.6 ± 2.1 | 61.7 ± 3.7 | 50.5 ± 1.7 | 54.5 ± 1.6 | 78.8 ± 1.8 |
| MMD | 64.5 ± 10.2 | 53.6 ± 16.4 | 69.4 ± 2.8 | 89.6 ± 1.8 | 48.2 ± 2.4 | 58.0 ± 3.5 | 40.9 ± 0.6 | 8.6 ± 1.0 | 67.2 ± 3.4 | 46.8 ± 1.2 | 38.1 ± 2.7 | 81.4 ± 2.1 |
| DANN | 87.2 ± 4.4 | 84.6 ± 7.7 | 91.2 ± 3.1 | 87.8 ± 3.9 | 52.0 ± 2.2 | 90.0 ± 3.5 | 48.8 ± 4.7 | 27.3 ± 8.6 | 44.7 ± 13.3 | 48.5 ± 1.7 | 55.7 ± 6.1 | 83.9 ± 3.4 |
| SagNet | 69.7 ± 9.8 | 54.5 ± 16.8 | 90.4 ± 2.6 | 96.2 ± 0.7 | 47.4 ± 0.3 | 65.7 ± 3.4 | 44.6 ± 0.7 | 13.3 ± 2.3 | 55.4 ± 8.8 | 51.5 ± 0.7 | 52.9 ± 3.2 | 82.0 ± 2.1 |
| VRex | 62.0 ± 8.4 | 43.4 ± 13.1 | 87.1 ± 3.7 | 95.1 ± 1.1 | 48.0 ± 0.9 | 61.6 ± 4.1 | 42.4 ± 1.3 | 11.4 ± 3.0 | 55.9 ± 8.7 | 49.4 ± 2.0 | 53.6 ± 2.9 | 83.3 ± 2.7 |
| Fish | 80.9 ± 10.1 | 75.4 ± 18.1 | 80.8 ± 13.2 | 95.1 ± 1.2 | 49.0 ± 0.5 | 61.8 ± 0.3 | 43.8 ± 0.8 | 12.1 ± 0.5 | 59.9 ± 6.5 | 50.7 ± 2.2 | 52.8 ± 1.7 | 82.3 ± 1.1 |
| Fishr | 63.4 ± 13.6 | 47.0 ± 22.6 | 86.7 ± 4.2 | 93.4 ± 0.5 | 48.6 ± 1.2 | 59.4 ± 0.8 | 44.6 ± 1.1 | 13.4 ± 0.7 | 55.7 ± 6.0 | 51.6 ± 2.0 | 39.8 ± 2.4 | 80.5 ± 1.8 |
| EQRm | 67.3 ± 12.2 | 56.7 ± 17.0 | 74.7 ± 7.8 | 86.3 ± 4.4 | 47.9 ± 0.5 | 61.6 ± 1.3 | 42.4 ± 1.9 | 9.7 ± 1.5 | 65.4 ± 0.5 | 50.5 ± 2.7 | 53.9 ± 2.5 | 80.9 ± 1.7 |
| RDM | 73.9 ± 15.5 | 64.8 ± 26.1 | 80.7 ± 3.3 | 94.3 ± 0.6 | 46.8 ± 1.1 | 61.5 ± 3.6 | 42.3 ± 1.2 | 10.3 ± 1.3 | 59.6 ± 5.6 | 48.8 ± 0.4 | 55.9 ± 2.7 | 84.5 ± 1.4 |
| PGrad | 76.1 ± 6.1 | 66.4 ± 26.0 | 87.7 ± 2.4 | 96.1 ± 0.6 | 49.0 ± 1.0 | 63.4 ± 0.4 | 43.1 ± 0.7 | 11.1 ± 0.2 | 44.3 ± 3.6 | 51.2 ± 0.7 | 54.6 ± 2.4 | 79.1 ± 2.9 |
| TCRI | 60.6 ± 11.0 | 40.0 ± 17.9 | 87.8 ± 1.8 | 98.3 ± 0.7 | 49.5 ± 0.2 | 58.7 ± 1.4 | 43.2 ± 1.1 | 13.5 ± 3.8 | 50.0 ± 4.2 | 52.1 ± 2.9 | 40.7 ± 5.0 | 80.7 ± 1.6 |
| Focal | 66.1 ± 12.3 | 53.7 ± 18.8 | 80.7 ± 6.3 | 93.4 ± 1.0 | 48.0 ± 0.9 | 62.0 ± 1.9 | 41.1 ± 1.2 | 11.1 ± 1.3 | 73.3 ± 5.2 | 48.3 ± 0.7 | 50.8 ± 1.2 | 80.5 ± 3.1 |
| ReWeight | 64.0 ± 11.6 | 50.4 ± 19.8 | 72.7 ± 5.9 | 96.5 ± 1.1 | 47.3 ± 1.3 | 61.7 ± 2.7 | 43.0 ± 0.2 | 12.2 ± 1.3 | 67.2 ± 3.4 | 51.4 ± 1.8 | 51.6 ± 2.8 | 69.6 ± 10.8 |
| BSofmax | 57.0 ± 5.7 | 34.2 ± 9.4 | 90.5 ± 1.9 | 96.5 ± 0.4 | 50.2 ± 1.9 | 74.4 ± 3.3 | 43.5 ± 0.9 | 14.4 ± 1.2 | 80.6 ± 4.4 | 48.8 ± 1.4 | 50.4 ± 3.0 | 78.6 ± 3.7 |
| LDAM | 59.7 ± 9.3 | 43.8 ± 17.5 | 75.4 ± 12.1 | 93.4 ± 1.0 | 49.8 ± 1.3 | 62.6 ± 4.2 | 40.1 ± 1.0 | 7.9 ± 0.6 | 73.1 ± 4.4 | 51.4 ± 2.3 | 51.1 ± 3.2 | 81.0 ± 5.0 |
| BoDA | 89.6 ± 1.8 | 90.5 ± 3.7 | 79.9 ± 5.8 | 95.8 ± 0.8 | 49.0 ± 2.4 | 64.6 ± 3.7 | 45.3 ± 1.2 | 20.8 ± 1.4 | 69.4 ± 3.7 | 50.3 ± 1.5 | 49.0 ± 3.0 | 41.1 ± 1.2 |
| GINiDG | 94.6 ± 1.4 | 97.4 ± 0.7 | 84.4 ± 11.0 | 93.6 ± 1.8 | 46.6 ± 1.9 | 61.4 ± 2.6 | 47.5 ± 2.3 | 33.3 ± 9.7 | 53.7 ± 15.5 | 45.2 ± 2.5 | 45.5 ± 3.8 | 74.2 ± 4.4 |
| SAMALTDG | 86.7 ± 6.4 | 82.7 ± 8.7 | 93.7 ± 3.7 | 93.2 ± 1.0 | 48.5 ± 0.5 | 64.8 ± 0.2 | 43.9 ± 0.6 | 9.1 ± 1.0 | 69.3 ± 3.0 | 53.7 ± 0.5 | 50.8 ± 2.7 | 79.5 ± 1.3 |
| NDCL | 94.0 ± 0.9 | 92.8 ± 1.1 | 96.9 ± 1.1 | 96.5 ± 0.5 | 49.4 ± 0.9 | 59.6 ± 0.8 | 43.8 ± 0.7 | 11.6 ± 2.0 | 69.2 ± 7.0 | 48.8 ± 1.7 | 59.1 ± 3.7 | 79.8 ± 1.9 |

67.8 ± 5.7

Table 17: Detailed results of the target domain on PACS benchmark under the Duality setting.

| | Average | Many | Median | Few | C | Average | Many | Median | Few | P | Average | Many | Median | Few | S | Average | Many | Median | Few |
|----------|------------|-------------|------------|------------|------------|-------------|------------|------------|------------|-------------|------------|------------|------------|-------------|-------------|-------------|-------------|-------------|-------------|
| ERM | 67.3 ± 1.3 | 27.7 ± 3.5 | 80.4 ± 0.3 | 77.3 ± 1.0 | 62.0 ± 1.8 | 53.1 ± 1.6 | 95.5 ± 0.1 | 43.3 ± 3.9 | 86.8 ± 0.6 | 100.0 ± 0.0 | 94.2 ± 0.6 | 67.5 ± 2.3 | 55.5 ± 2.5 | 54.2 ± 4.5 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| IRM | 70.7 ± 1.3 | 39.3 ± 5.5 | 82.0 ± 0.9 | 79.6 ± 0.7 | 57.8 ± 0.8 | 37.1 ± 8.9 | 95.7 ± 0.5 | 40.6 ± 3.9 | 84.4 ± 1.8 | 100.0 ± 0.0 | 94.9 ± 0.5 | 60.0 ± 4.8 | 47.9 ± 2.3 | 27.6 ± 14.8 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| GroupDRO | 67.4 ± 1.1 | 31.4 ± 1.7 | 80.2 ± 0.7 | 75.5 ± 2.1 | 60.4 ± 2.1 | 38.5 ± 5.9 | 96.6 ± 0.4 | 44.1 ± 3.1 | 82.0 ± 0.4 | 100.0 ± 0.0 | 93.2 ± 1.5 | 55.0 ± 1.7 | 56.9 ± 2.8 | 38.5 ± 5.2 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| Mixup | 68.3 ± 0.4 | 35.7 ± 5.7 | 83.0 ± 1.9 | 73.8 ± 3.3 | 56.4 ± 1.9 | 25.3 ± 4.2 | 94.8 ± 0.6 | 42.0 ± 2.2 | 84.6 ± 1.3 | 100.0 ± 0.0 | 95.6 ± 0.7 | 60.3 ± 3.2 | 46.3 ± 1.6 | 51.0 ± 9.1 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| MLDG | 70.0 ± 0.8 | 42.8 ± 2.5 | 84.3 ± 0.6 | 74.9 ± 2.7 | 59.3 ± 2.6 | 44.2 ± 1.6 | 96.4 ± 1.0 | 40.4 ± 4.7 | 86.3 ± 2.9 | 100.0 ± 0.0 | 94.5 ± 0.3 | 65.7 ± 8.3 | 48.8 ± 0.2 | 54.0 ± 20.8 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| CORAL | 67.8 ± 1.3 | 45.1 ± 5.6 | 83.6 ± 0.5 | 68.6 ± 4.2 | 55.3 ± 2.0 | 46.4 ± 5.4 | 97.0 ± 1.5 | 31.1 ± 3.0 | 84.2 ± 1.2 | 100.0 ± 0.0 | 95.9 ± 0.7 | 58.6 ± 4.5 | 52.8 ± 6.2 | 49.0 ± 4.1 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| MMD | 73.5 ± 2.1 | 47.6 ± 6.9 | 85.1 ± 1.2 | 77.2 ± 3.2 | 63.5 ± 1.6 | 44.2 ± 3.4 | 96.9 ± 0.3 | 47.9 ± 3.6 | 85.0 ± 0.8 | 100.0 ± 0.0 | 93.5 ± 1.5 | 62.8 ± 3.9 | 50.6 ± 4.2 | 45.4 ± 11.4 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| DANN | 72.3 ± 0.9 | 47.2 ± 4.8 | 85.0 ± 2.0 | 75.4 ± 2.2 | 65.3 ± 1.6 | 59.2 ± 7.1 | 97.2 ± 0.4 | 45.8 ± 5.4 | 87.7 ± 0.8 | 100.0 ± 0.0 | 92.8 ± 1.4 | 72.0 ± 1.1 | 48.0 ± 3.8 | 57.0 ± 11.0 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| SagNet | 70.6 ± 1.4 | 37.4 ± 8.6 | 86.3 ± 1.3 | 75.7 ± 1.7 | 59.2 ± 1.4 | 52.0 ± 7.4 | 95.9 ± 0.8 | 38.1 ± 4.3 | 84.8 ± 2.6 | 100.0 ± 0.0 | 94.7 ± 1.6 | 62.2 ± 5.5 | 49.5 ± 4.3 | 24.0 ± 3.5 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| VREx | 67.8 ± 2.2 | 40.0 ± 5.5 | 79.7 ± 2.1 | 75.1 ± 2.3 | 56.7 ± 0.1 | 39.5 ± 0.2 | 97.3 ± 0.2 | 36.6 ± 0.8 | 86.5 ± 1.0 | 100.0 ± 0.0 | 94.0 ± 1.0 | 67.0 ± 2.8 | 51.5 ± 5.0 | 46.5 ± 6.1 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| Fishr | 69.8 ± 1.4 | 39.0 ± 1.3 | 82.0 ± 2.1 | 75.3 ± 2.8 | 60.0 ± 2.0 | 48.2 ± 7.7 | 95.3 ± 0.6 | 40.7 ± 1.6 | 84.7 ± 1.0 | 100.0 ± 0.0 | 94.9 ± 0.7 | 60.8 ± 2.9 | 49.9 ± 2.6 | 58.5 ± 5.5 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| Fishr | 69.1 ± 1.3 | 45.7 ± 1.1 | 85.4 ± 1.8 | 69.6 ± 5.0 | 58.4 ± 1.4 | 39.7 ± 2.1 | 97.0 ± 0.4 | 39.9 ± 2.6 | 86.1 ± 2.4 | 100.0 ± 0.0 | 94.8 ± 1.0 | 65.2 ± 6.6 | 52.2 ± 3.9 | 75.2 ± 10.4 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| EQRM | 68.1 ± 4.1 | 32.7 ± 11.7 | 84.2 ± 1.5 | 74.6 ± 3.0 | 66.8 ± 3.4 | 58.0 ± 9.5 | 96.6 ± 0.6 | 49.7 ± 2.6 | 86.5 ± 0.7 | 100.0 ± 0.0 | 92.9 ± 2.4 | 68.3 ± 0.7 | 45.3 ± 2.4 | 21.0 ± 7.6 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| RDM | 72.5 ± 1.8 | 42.4 ± 9.3 | 83.6 ± 1.4 | 80.2 ± 1.1 | 58.9 ± 1.1 | 43.3 ± 3.8 | 97.1 ± 0.3 | 38.8 ± 3.1 | 82.7 ± 1.5 | 100.0 ± 0.0 | 92.8 ± 0.1 | 57.3 ± 4.2 | 54.4 ± 4.3 | 41.5 ± 2.6 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| PGrad | 75.3 ± 0.5 | 53.5 ± 3.8 | 85.5 ± 0.9 | 79.1 ± 3.4 | 62.2 ± 0.9 | 44.4 ± 6.9 | 97.7 ± 0.2 | 44.1 ± 3.5 | 87.9 ± 1.4 | 100.0 ± 0.0 | 95.8 ± 0.8 | 69.2 ± 4.6 | 56.9 ± 3.6 | 30.3 ± 10.1 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| TCRI | 68.8 ± 1.4 | 53.6 ± 3.7 | 81.7 ± 4.0 | 67.2 ± 4.0 | 61.9 ± 3.1 | 56.6 ± 9.1 | 96.5 ± 0.1 | 40.3 ± 4.1 | 87.1 ± 1.1 | 100.0 ± 0.0 | 95.6 ± 1.1 | 66.8 ± 1.4 | 58.3 ± 1.3 | 67.8 ± 4.3 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| Focal | 69.1 ± 1.2 | 33.5 ± 6.6 | 82.5 ± 1.4 | 78.0 ± 0.7 | 63.3 ± 1.6 | 61.0 ± 10.6 | 95.4 ± 1.1 | 42.8 ± 4.7 | 85.9 ± 2.2 | 100.0 ± 0.0 | 92.9 ± 1.8 | 66.4 ± 4.5 | 52.0 ± 4.3 | 49.4 ± 5.3 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| ReWeight | 70.1 ± 1.6 | 42.1 ± 5.0 | 83.2 ± 1.9 | 75.2 ± 1.8 | 58.0 ± 1.6 | 31.6 ± 3.5 | 96.7 ± 1.1 | 41.7 ± 4.0 | 86.3 ± 1.7 | 100.0 ± 0.0 | 92.3 ± 2.5 | 68.2 ± 2.3 | 48.4 ± 5.9 | 53.4 ± 10.4 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| BSofmax | 68.2 ± 1.5 | 22.4 ± 4.7 | 84.2 ± 1.5 | 78.6 ± 1.7 | 60.7 ± 1.4 | 36.2 ± 5.2 | 96.3 ± 0.2 | 45.0 ± 2.1 | 88.7 ± 1.3 | 100.0 ± 0.0 | 95.0 ± 1.6 | 72.4 ± 3.8 | 53.1 ± 4.7 | 50.9 ± 8.7 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| LDAM | 67.9 ± 0.1 | 33.5 ± 5.4 | 82.2 ± 0.8 | 74.9 ± 1.8 | 59.4 ± 2.8 | 30.4 ± 1.6 | 96.0 ± 0.5 | 45.5 ± 4.9 | 84.0 ± 0.9 | 100.0 ± 0.0 | 92.3 ± 1.7 | 61.8 ± 1.8 | 55.1 ± 1.3 | 39.3 ± 12.7 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| BoDA | 68.4 ± 0.5 | 30.2 ± 4.8 | 83.6 ± 0.6 | 76.3 ± 2.4 | 56.1 ± 2.1 | 45.4 ± 2.7 | 97.4 ± 0.3 | 32.6 ± 3.3 | 85.0 ± 0.5 | 100.0 ± 0.0 | 94.6 ± 1.0 | 62.1 ± 2.0 | 48.4 ± 6.3 | 45.3 ± 6.8 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| GINIDG | 64.5 ± 1.7 | 39.4 ± 3.5 | 78.5 ± 2.0 | 68.2 ± 0.2 | 52.4 ± 2.5 | 42.2 ± 13.4 | 91.5 ± 2.0 | 31.0 ± 1.0 | 85.5 ± 1.4 | 100.0 ± 0.0 | 93.2 ± 2.0 | 65.4 ± 2.3 | 50.4 ± 5.4 | 60.0 ± 19.7 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| SAMALTDG | 67.7 ± 0.6 | 27.4 ± 1.2 | 81.0 ± 2.2 | 77.3 ± 1.2 | 61.0 ± 1.3 | 40.5 ± 5.1 | 96.5 ± 0.4 | 44.9 ± 1.6 | 86.1 ± 0.5 | 100.0 ± 0.0 | 95.4 ± 0.2 | 64.8 ± 1.6 | 49.7 ± 2.7 | 30.7 ± 11.2 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |
| NDCL | 70.7 ± 1.8 | 39.8 ± 2.5 | 83.5 ± 3.2 | 78.7 ± 0.2 | 62.8 ± 0.6 | 46.1 ± 9.6 | 94.2 ± 0.8 | 48.0 ± 3.3 | 86.8 ± 1.0 | 100.0 ± 0.0 | 94.6 ± 0.3 | 67.1 ± 2.9 | 65.3 ± 2.2 | 32.0 ± 5.1 | 167.7 ± 1.1 | 168.5 ± 1.3 | 168.8 ± 1.5 | 167.8 ± 1.7 | 167.4 ± 1.9 |

Table 18: Detailed results of the target domain on OfficeHome benchmark under the Duality setting.

| | A | | | | C | | | | P | | | | R | | | |
|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| | Average | Many | Median | Few | Average | Many | Median | Few | Average | Many | Median | Few | Average | Many | Median | Few |
| ERM | 48.6 ± 0.7 | 76.9 ± 0.3 | 49.3 ± 0.0 | 38.3 ± 1.2 | 42.5 ± 0.4 | 48.0 ± 2.6 | 55.1 ± 0.2 | 36.7 ± 0.7 | 59.6 ± 0.5 | 63.2 ± 4.3 | 75.2 ± 0.4 | 50.1 ± 0.6 | 58.7 ± 0.9 | 87.3 ± 0.5 | 67.2 ± 0.4 | 46.6 ± 1.3 |
| IRM | 44.1 ± 1.4 | 72.4 ± 1.3 | 42.8 ± 2.3 | 34.7 ± 1.5 | 39.6 ± 1.3 | 47.6 ± 0.8 | 54.7 ± 2.3 | 33.3 ± 1.1 | 56.5 ± 1.4 | 61.7 ± 6.2 | 73.3 ± 1.7 | 46.5 ± 1.1 | 57.5 ± 0.6 | 86.5 ± 1.2 | 69.6 ± 0.7 | 44.8 ± 1.4 |
| GroupDRO | 45.2 ± 0.2 | 73.3 ± 2.8 | 44.0 ± 2.6 | 35.3 ± 0.5 | 42.9 ± 0.6 | 53.3 ± 0.8 | 55.5 ± 0.9 | 36.7 ± 0.7 | 59.1 ± 0.5 | 64.9 ± 3.3 | 77.4 ± 1.0 | 48.4 ± 0.8 | 59.6 ± 0.3 | 90.6 ± 0.7 | 71.8 ± 0.6 | 46.5 ± 0.1 |
| Mixup | 46.1 ± 1.4 | 73.9 ± 1.6 | 45.6 ± 1.8 | 36.2 ± 2.3 | 42.8 ± 0.7 | 51.8 ± 0.4 | 55.0 ± 1.0 | 36.9 ± 1.3 | 59.5 ± 0.8 | 70.0 ± 3.3 | 76.9 ± 0.8 | 48.0 ± 0.8 | 60.6 ± 0.2 | 86.7 ± 0.8 | 72.8 ± 0.6 | 47.9 ± 0.5 |
| MLDG | 49.1 ± 0.6 | 76.9 ± 1.2 | 47.9 ± 2.2 | 39.2 ± 1.0 | 42.1 ± 1.4 | 46.5 ± 1.7 | 56.2 ± 2.7 | 36.6 ± 1.0 | 58.4 ± 1.1 | 61.9 ± 4.8 | 76.6 ± 0.7 | 48.2 ± 1.3 | 60.5 ± 0.2 | 86.5 ± 0.7 | 71.0 ± 1.6 | 48.5 ± 0.4 |
| CORAL | 50.6 ± 0.8 | 77.9 ± 0.5 | 52.5 ± 0.9 | 39.6 ± 0.9 | 45.0 ± 0.2 | 57.4 ± 2.5 | 58.8 ± 0.5 | 38.5 ± 0.6 | 60.4 ± 0.7 | 70.6 ± 4.3 | 78.8 ± 0.2 | 48.9 ± 0.5 | 61.9 ± 0.4 | 91.2 ± 0.5 | 76.3 ± 0.1 | 48.1 ± 0.4 |
| MMD | 45.7 ± 0.3 | 73.9 ± 1.3 | 43.1 ± 2.3 | 36.6 ± 0.4 | 41.2 ± 1.0 | 51.0 ± 2.1 | 54.1 ± 1.4 | 35.7 ± 1.9 | 59.3 ± 0.9 | 61.5 ± 3.6 | 75.9 ± 0.7 | 49.1 ± 1.1 | 57.4 ± 0.2 | 89.6 ± 1.6 | 68.2 ± 0.4 | 44.1 ± 0.7 |
| DANN | 49.3 ± 0.6 | 75.7 ± 0.8 | 47.2 ± 1.9 | 40.7 ± 1.4 | 40.6 ± 0.3 | 48.6 ± 1.3 | 54.0 ± 1.3 | 35.5 ± 1.4 | 58.0 ± 1.4 | 66.2 ± 6.4 | 74.4 ± 2.3 | 47.0 ± 1.1 | 59.2 ± 0.8 | 89.7 ± 1.1 | 72.1 ± 1.9 | 45.7 ± 0.8 |
| SagNet | 48.7 ± 0.0 | 80.2 ± 1.3 | 47.3 ± 1.1 | 38.2 ± 0.1 | 44.3 ± 0.8 | 55.1 ± 2.2 | 56.7 ± 0.9 | 38.5 ± 0.9 | 60.6 ± 0.3 | 66.5 ± 4.2 | 77.7 ± 0.0 | 49.9 ± 1.1 | 61.8 ± 0.9 | 90.5 ± 1.7 | 73.6 ± 0.9 | 48.9 ± 1.3 |
| VREx | 48.6 ± 1.0 | 77.7 ± 0.9 | 46.3 ± 1.7 | 39.0 ± 1.2 | 43.4 ± 0.9 | 50.6 ± 1.4 | 57.2 ± 0.7 | 36.9 ± 1.3 | 58.5 ± 0.4 | 65.5 ± 3.1 | 75.6 ± 0.6 | 48.3 ± 0.8 | 60.9 ± 0.2 | 86.5 ± 3.0 | 71.0 ± 1.1 | 49.1 ± 0.6 |
| Fish | 47.5 ± 0.4 | 77.5 ± 1.4 | 46.6 ± 1.0 | 37.4 ± 0.1 | 43.3 ± 1.1 | 48.4 ± 1.3 | 56.9 ± 2.2 | 37.4 ± 1.0 | 60.2 ± 0.6 | 62.3 ± 4.0 | 76.8 ± 0.6 | 50.5 ± 0.4 | 60.5 ± 0.2 | 89.8 ± 0.9 | 69.9 ± 0.7 | 47.8 ± 0.5 |
| Fishr | 47.5 ± 0.4 | 79.6 ± 0.8 | 46.3 ± 2.3 | 37.5 ± 0.8 | 41.8 ± 0.2 | 51.7 ± 1.7 | 53.2 ± 1.2 | 36.5 ± 1.0 | 60.3 ± 0.3 | 63.5 ± 3.5 | 78.4 ± 0.2 | 49.8 ± 1.1 | 60.3 ± 0.7 | 89.8 ± 1.2 | 70.2 ± 1.8 | 48.4 ± 0.9 |
| EQRM | 48.3 ± 0.4 | 74.7 ± 1.7 | 47.2 ± 1.2 | 38.2 ± 0.7 | 42.5 ± 0.6 | 48.2 ± 1.0 | 55.5 ± 0.5 | 37.7 ± 0.4 | 59.4 ± 0.9 | 62.4 ± 3.9 | 78.2 ± 0.5 | 48.7 ± 1.3 | 59.5 ± 1.0 | 86.7 ± 0.5 | 71.2 ± 0.5 | 46.9 ± 1.3 |
| RDM | 48.2 ± 0.6 | 78.1 ± 1.4 | 46.4 ± 1.3 | 38.5 ± 1.0 | 42.8 ± 0.7 | 51.5 ± 1.0 | 55.2 ± 1.4 | 37.6 ± 1.1 | 60.5 ± 0.8 | 65.5 ± 4.0 | 75.2 ± 1.1 | 51.1 ± 1.5 | 60.1 ± 0.6 | 89.5 ± 1.3 | 69.4 ± 1.0 | 47.6 ± 0.9 |
| PGrad | 51.4 ± 0.3 | 78.8 ± 0.2 | 50.9 ± 0.9 | 41.8 ± 0.6 | 45.6 ± 1.0 | 56.4 ± 1.2 | 58.9 ± 1.5 | 39.6 ± 0.4 | 61.2 ± 0.3 | 65.7 ± 1.9 | 79.9 ± 0.8 | 50.6 ± 0.7 | 63.5 ± 1.2 | 90.9 ± 0.9 | 73.0 ± 1.4 | 52.1 ± 1.3 |
| TCRI, HSIC | 50.8 ± 1.3 | 75.5 ± 1.6 | 48.5 ± 4.4 | 42.2 ± 1.3 | 43.3 ± 0.5 | 52.1 ± 1.1 | 57.4 ± 2.4 | 37.3 ± 1.4 | 61.5 ± 0.4 | 67.0 ± 3.4 | 78.1 ± 0.4 | 51.4 ± 0.7 | 62.7 ± 0.8 | 88.6 ± 1.8 | 73.9 ± 1.4 | 50.8 ± 0.9 |
| Focal | 45.9 ± 1.4 | 73.9 ± 2.4 | 42.5 ± 2.7 | 37.3 ± 1.4 | 41.9 ± 0.8 | 49.3 ± 2.9 | 54.3 ± 1.6 | 36.6 ± 0.7 | 57.1 ± 1.1 | 58.7 ± 2.5 | 75.3 ± 1.5 | 47.3 ± 1.1 | 58.1 ± 0.2 | 88.2 ± 1.7 | 69.0 ± 1.7 | 44.9 ± 0.2 |
| ReWeight | 48.2 ± 0.7 | 74.0 ± 1.0 | 46.8 ± 1.6 | 40.3 ± 1.6 | 42.2 ± 0.7 | 51.0 ± 2.2 | 52.9 ± 0.9 | 37.3 ± 0.8 | 59.7 ± 0.5 | 64.3 ± 3.5 | 74.9 ± 0.5 | 49.8 ± 0.5 | 59.7 ± 0.5 | 86.3 ± 0.6 | 70.6 ± 1.1 | 47.2 ± 0.7 |
| BSofmax | 48.2 ± 0.8 | 68.8 ± 3.1 | 46.1 ± 0.9 | 41.3 ± 1.3 | 42.6 ± 1.7 | 47.4 ± 2.0 | 53.1 ± 1.1 | 38.0 ± 2.1 | 60.5 ± 0.7 | 65.4 ± 3.7 | 78.4 ± 0.6 | 50.3 ± 1.6 | 59.8 ± 0.8 | 88.0 ± 1.5 | 70.8 ± 1.6 | 48.0 ± 0.6 |
| LDAM | 46.2 ± 0.9 | 75.4 ± 2.4 | 44.0 ± 1.4 | 37.2 ± 0.9 | 41.5 ± 0.5 | 48.4 ± 2.1 | 56.1 ± 1.3 | 35.3 ± 0.6 | 55.3 ± 0.9 | 59.8 ± 3.8 | 70.8 ± 0.6 | 45.4 ± 0.9 | 57.5 ± 0.5 | 88.7 ± 1.5 | 67.6 ± 2.0 | 45.2 ± 1.6 |
| BoDA | 49.3 ± 1.0 | 78.3 ± 0.3 | 51.6 ± 0.8 | 38.4 ± 1.5 | 44.0 ± 1.1 | 61.4 ± 0.3 | 57.5 ± 0.7 | 36.8 ± 1.8 | 59.5 ± 0.3 | 67.2 ± 2.6 | 78.0 ± 0.6 | 48.4 ± 0.8 | 61.1 ± 0.6 | 91.5 ± 1.4 | 74.9 ± 0.4 | 47.3 ± 0.5 |
| GINDG | 45.7 ± 0.9 | 72.7 ± 1.9 | 42.5 ± 1.8 | 36.5 ± 1.6 | 41.4 ± 1.0 | 52.2 ± 1.7 | 54.9 ± 0.9 | 35.2 ± 0.9 | 57.1 ± 0.1 | 63.3 ± 2.8 | 74.0 ± 1.0 | 46.4 ± 0.3 | 56.6 ± 0.8 | 89.4 ± 1.2 | 67.9 ± 0.4 | 43.8 ± 1.2 |
| SAMALTDG | 42.4 ± 1.5 | 63.7 ± 1.8 | 59.2 ± 1.3 | 20.5 ± 1.6 | 37.8 ± 0.9 | 63.7 ± 3.4 | 58.5 ± 0.6 | 20.5 ± 0.6 | 49.2 ± 0.9 | 84.6 ± 0.7 | 60.5 ± 0.5 | 30.3 ± 1.2 | 54.8 ± 0.5 | 84.3 ± 0.5 | 77.8 ± 0.5 | 31.2 ± 0.5 |
| NDCL | 50.0 ± 0.5 | 75.6 ± 1.9 | 51.5 ± 2.7 | 41.4 ± 0.8 | 45.7 ± 0.6 | 52.3 ± 1.5 | 58.2 ± 1.6 | 41.5 ± 0.8 | 63.5 ± 0.5 | 66.7 ± 4.7 | 80.2 ± 1.4 | 54.4 ± 0.8 | 63.6 ± 0.3 | 90.9 ± 1.8 | 73.2 ± 0.9 | 53.2 ± 1.0 |