A Theoretical Analysis and Complete Proofs

In this section, we explain the details of Theorem 1 in the main paper, and also formally describe Theorem 2. We start with giving additional definitions and providing a useful lemma and its proof, which invoked through the proof of the theorems. We then formally prove the arguments in Theorem 1 and 2.

A.1 Additional Definition, Lemma, and Theorem

Definition 4 ($(\widetilde{\alpha}, \widetilde{\beta}, \widetilde{\gamma})$ Calibrated Transferability Statistics). *The transferability graph can be further described by the following three components:*

$$\widetilde{\alpha} = \mathbb{E}_c \mathbb{E}_d \mathbb{E}_{d' \neq d} \left[\lambda_{d,c}^{d',c} \cdot \mathrm{trans}\big((d,c),(d',c)\big) \right],$$
$$\widetilde{\beta} = \mathbb{E}_d \mathbb{E}_c \mathbb{E}_{c' \neq c} \left[\lambda_{d,c}^{d,c'} \cdot \mathrm{trans}\big((d,c),(d,c')\big) \right],$$
$$\widetilde{\gamma} = \mathbb{E}_d \mathbb{E}_{d' \neq d} \mathbb{E}_c \mathbb{E}_{c' \neq c} \left[\lambda_{d,c}^{d',c'} \cdot \mathrm{trans}\big((d,c),(d',c')\big) \right],$$

where $\lambda_{d,c}^{d',c'} = \left(\frac{N_{d',c'}}{N_{d,c}}\right)^{\nu}$ *denotes the distance calibration coefficient.*

Lemma 1. *Let* $\eta, \pi > 0$ *and* $\varphi : \mathbb{R} \to \mathbb{R}$, $\varphi(x) = \log(\eta + \pi \exp(x))$. *Given a finite sequence* $x_1, x_2, \ldots, x_M \in \mathbb{R}$, *it holds that*

$$\frac{1}{M} \sum_{i=1}^{M} \varphi(x_i) \geq \varphi\left(\frac{1}{M} \sum_{i=1}^{M} x_i \right).$$

Proof. Note that φ is smooth and thus twice differentiable for all $x \in \mathbb{R}$. We obtain the second derivative of φ as

$$\varphi''(x) = \frac{\eta\pi \exp(x)}{(\eta + \pi \exp(x))^2} > 0, \quad \forall x \in \mathbb{R}.$$

Therefore, φ is convex. Thus, by Jensen's inequality, we obtain that $\frac{1}{M} \sum_{i=1}^{M} \varphi(x_i) \geq \varphi\left(\frac{1}{M} \sum_{i=1}^{M} x_i\right)$, which completes the proof.

Theorem 2 ($\widetilde{\mathcal{L}}_{\texttt{BoDA}}$ as an Upper Bound). *Given a multi-domain long-tailed dataset* $\mathcal{S}$ *with domain label space* $\mathcal{D}$ *and class label space* $\mathcal{C}$ *satisfying* $|\mathcal{D}| > 1$ *and* $|\mathcal{C}| > 1$, *let* $\mathcal{Z}$ *be the representation set of all training samples. It holds that*

$$\widetilde{\mathcal{L}}_{\texttt{BoDA}}(\mathcal{Z}, \{\boldsymbol{\mu}\}) \geq N \log\left(|\mathcal{D}| - 1 + |\mathcal{D}|(|\mathcal{C}| - 1) \exp\left(\tfrac{|\mathcal{C}||\mathcal{D}|}{N} \cdot \widetilde{\alpha} - \tfrac{|\mathcal{C}|}{N} \cdot \widetilde{\beta} - \tfrac{|\mathcal{C}|(|\mathcal{D}|-1)}{N} \cdot \widetilde{\gamma} \right) \right), \quad (5)$$

where $(\widetilde{\alpha}, \widetilde{\beta}, \widetilde{\gamma})$ *are the calibrated transferability statistics for* $\mathcal{S}$ *defined in Definition 4.*

A.2 Proof of Theorem 1

Recall that $\mathcal{M} = \mathcal{D} \times \mathcal{C} := \{(d, c) : d \in \mathcal{D}, c \in \mathcal{C}\}$ is the set of all domain-class pairs. $\mathcal{L}_{\texttt{BoDA}}$ is given by

$$\begin{aligned}\mathcal{L}_{\texttt{BoDA}}(\mathcal{Z}, \{\boldsymbol{\mu}\}) &= \sum_{\mathbf{z}_i \in \mathcal{Z}} \frac{-1}{|\mathcal{D}| - 1} \sum_{d \in \mathcal{D} \setminus \{d_i\}} \log \frac{\exp\left(-\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i})\right)}{\sum_{(d',c') \in \mathcal{M} \setminus \{(d_i,c_i)\}} \exp\left(-\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})\right)} \\ &= \sum_{\mathbf{z}_i \in \mathcal{Z}} \ell_{\texttt{BoDA}}(\mathbf{z}_i, \{\boldsymbol{\mu}\}),\end{aligned}$$

where $\ell_{\texttt{BoDA}}(\mathbf{z}_i, \{\boldsymbol{\mu}\})$ is the *sample-wise* BoDA loss. We rewrite $\ell_{\texttt{BoDA}}$ in the following format

$$\begin{aligned}\ell_{\texttt{BoDA}}(\mathbf{z}_i, \{\boldsymbol{\mu}\}) &= -\frac{1}{|\mathcal{D}| - 1} \sum_{d \in \mathcal{D} \setminus \{d_i\}} \log \frac{\exp\left(-\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i})\right)}{\sum_{(d',c') \in \mathcal{M} \setminus \{(d_i,c_i)\}} \exp\left(-\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})\right)} \\ &= \log \left(\frac{\sum_{(d',c') \in \mathcal{M} \setminus \{(d_i,c_i)\}} \exp\left(-\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})\right)}{\prod_{d \in \mathcal{D} \setminus \{d_i\}} \exp\left(-\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i})\right)^{\frac{1}{|\mathcal{D}|-1}}} \right) \\ &= \log \left(\frac{\sum_{(d',c') \in \mathcal{M} \setminus \{(d_i,c_i)\}} \exp\left(-\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})\right)}{\exp\left(-\frac{1}{|\mathcal{D}|-1} \sum_{d \in \mathcal{D} \setminus \{d_i\}} \widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i})\right)} \right). \end{aligned} \tag{6}$$

We will first focus on the term in the numerator of Eqn. (6). We can rewrite the sum into two terms

$$\begin{aligned}&\sum_{(d',c') \in \mathcal{M} \setminus \{(d_i,c_i)\}} \exp\left(-\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})\right) \\ &= \underbrace{\sum_{d' \in \mathcal{D} \setminus \{d_i\}} \sum_{c' \in \{c_i\}} \exp\left(-\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})\right)}_{T_1} + \underbrace{\sum_{d' \in \mathcal{D}} \sum_{c' \in \mathcal{C} \setminus \{c_i\}} \exp\left(-\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})\right)}_{T_2}.\end{aligned}$$

Since the exponential function $\exp(\cdot)$ is convex, we apply Jensen's inequality on both T_1 and T_2:

$$\begin{aligned}T_1 &\geq (|\mathcal{D}| - 1) \exp\left(-\frac{1}{|\mathcal{D}| - 1} \sum_{d' \in \mathcal{D} \setminus \{d_i\}} \sum_{c' \in \{c_i\}} \widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'}) \right) \\ &= (|\mathcal{D}| - 1) \exp\left(-\frac{1}{|\mathcal{D}| - 1} \sum_{d' \in \mathcal{D} \setminus \{d_i\}} \widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c_i}) \right), \\ T_2 &\geq |\mathcal{D}|(|\mathcal{C}| - 1) \exp\left(-\frac{1}{|\mathcal{D}|(|\mathcal{C}| - 1)} \sum_{d' \in \mathcal{D}} \sum_{c' \in \mathcal{C} \setminus \{c_i\}} \widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'}) \right).\end{aligned}$$

Thus, by using $\exp(x)/\exp(y) = \exp(x-y)$ and rearranging terms, we bound $\ell_{\texttt{BoDA}}$ by

$$\ell_{\texttt{BoDA}}(\mathbf{z}_i, \{\boldsymbol{\mu}\}) \geq \log\left(|\mathcal{D}| - 1 + |\mathcal{D}|(|\mathcal{C}|-1)\exp\Bigg(\underbrace{\frac{1}{|\mathcal{D}|-1}\sum_{d'\in\mathcal{D}\setminus\{d_i\}}\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c_i}) - \frac{1}{|\mathcal{D}|(|\mathcal{C}|-1)}\sum_{d'\in\mathcal{D}}\sum_{c'\in\mathcal{C}\setminus\{c_i\}}\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})}_{T(\mathbf{z}_i,\{\boldsymbol{\mu}\})}\Bigg)\right).$$

Leveraging Lemma 1, by setting $\eta = |\mathcal{D}|-1$, $\pi = |\mathcal{D}|(|\mathcal{C}|-1)$, and $x_i = T(\mathbf{z}_i, \{\boldsymbol{\mu}\})$, we further bound $\mathcal{L}_{\texttt{BoDA}}(\mathcal{Z}, \{\boldsymbol{\mu}\})$ by

$$\begin{aligned}\mathcal{L}_{\texttt{BoDA}}(\mathcal{Z}, \{\boldsymbol{\mu}\}) &= \sum_{\mathbf{z}_i\in\mathcal{Z}} \ell_{\texttt{BoDA}}(\mathbf{z}_i, \{\boldsymbol{\mu}\}) \\ &\geq \sum_{\mathbf{z}_i\in\mathcal{Z}} \log\left(|\mathcal{D}| - 1 + |\mathcal{D}|(|\mathcal{C}|-1)\exp\left(T(\mathbf{z}_i, \{\boldsymbol{\mu}\})\right)\right) \\ &\geq |\mathcal{Z}|\log\left(|\mathcal{D}| - 1 + |\mathcal{D}|(|\mathcal{C}|-1)\exp\left(\frac{1}{|\mathcal{Z}|}\sum_{\mathbf{z}_i\in\mathcal{Z}} T(\mathbf{z}_i, \{\boldsymbol{\mu}\})\right)\right).\end{aligned} \tag{7}$$

Note that the argument of the $\exp(\cdot)$ in Eqn. (7) can be expanded and further rearranged as

$$\begin{aligned}\frac{1}{|\mathcal{Z}|}\sum_{\mathbf{z}_i\in\mathcal{Z}} T(\mathbf{z}_i, \{\boldsymbol{\mu}\}) = {} & \frac{1}{|\mathcal{Z}|}\sum_{\mathbf{z}_i\in\mathcal{Z}}\frac{1}{|\mathcal{D}|-1}\sum_{d'\in\mathcal{D}\setminus\{d_i\}}\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c_i}) - \\ & \frac{1}{|\mathcal{Z}|}\sum_{\mathbf{z}_i\in\mathcal{Z}}\frac{1}{|\mathcal{D}|(|\mathcal{C}|-1)}\sum_{d'\in\mathcal{D}}\sum_{c'\in\mathcal{C}\setminus\{c_i\}}\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'}) \\ = {} & \underbrace{\frac{1}{|\mathcal{Z}|}\frac{1}{|\mathcal{D}|-1}\sum_{\mathbf{z}_i\in\mathcal{Z}}\sum_{d'\in\mathcal{D}\setminus\{d_i\}}\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c_i})}_{T_\alpha} - \\ & \underbrace{\frac{1}{|\mathcal{Z}|}\frac{1}{|\mathcal{D}|(|\mathcal{C}|-1)}\sum_{\mathbf{z}_i\in\mathcal{Z}}\sum_{c'\in\mathcal{C}\setminus\{c_i\}}\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d_i,c'})}_{T_\beta} - \\ & \underbrace{\frac{1}{|\mathcal{Z}|}\frac{1}{|\mathcal{D}|(|\mathcal{C}|-1)}\sum_{\mathbf{z}_i\in\mathcal{Z}}\sum_{d'\in\mathcal{D}\setminus\{d_i\}}\sum_{c'\in\mathcal{C}\setminus\{c_i\}}\widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})}_{T_\gamma}.\end{aligned} \tag{8}$$

Recall that each $\mathbf{z}_i \in \mathcal{Z}$ belongs to a domain-class pair (d_i, c_i), and $\mathcal{Z}_{d,c}$ denotes the representation set of $\mathcal{S}_{d,c}$ with size $N_{d,c}$. For simplicity, we remove the

subscript i in the following derivation. We can further rewrite $T_\alpha, T_\beta, T_\gamma$ as

$$\begin{aligned}
T_\alpha &= \frac{1}{|\mathcal{Z}|}\frac{1}{|\mathcal{D}|-1}\sum_{c\in\mathcal{C}}\sum_{d\in\mathcal{D}}\sum_{d'\in\mathcal{D}\setminus\{d\}}\sum_{\mathbf{z}\in\mathcal{Z}_{d,c}}\widetilde{\mathsf{d}}(\mathbf{z},\boldsymbol{\mu}_{d',c}) \\
&= \frac{1}{|\mathcal{Z}|}\frac{1}{|\mathcal{D}|-1}|\mathcal{C}||\mathcal{D}|(|\mathcal{D}|-1)\mathbb{E}_c\mathbb{E}_d\mathbb{E}_{d'\neq d}\mathbb{E}_{\mathbf{z}\in\mathcal{Z}_{d,c}}\big[\underbrace{N_{d,c}\cdot\widetilde{\mathsf{d}}(\mathbf{z},\boldsymbol{\mu}_{d',c})}_{\mathsf{d}(\mathbf{z},\boldsymbol{\mu}_{d',c})}\big] \\
&= \frac{|\mathcal{C}||\mathcal{D}|}{|\mathcal{Z}|}\underbrace{\mathbb{E}_c\mathbb{E}_d\mathbb{E}_{d'\neq d}\mathbb{E}_{\mathbf{z}\in\mathcal{Z}_{d,c}}\big[\mathsf{d}(\mathbf{z},\boldsymbol{\mu}_{d',c})\big]}_{\alpha}, \qquad (9) \\
T_\beta &= \frac{1}{|\mathcal{Z}|}\frac{1}{|\mathcal{D}|(|\mathcal{C}|-1)}\sum_{c\in\mathcal{C}}\sum_{d\in\mathcal{D}}\sum_{c'\in\mathcal{C}\setminus\{c\}}\sum_{\mathbf{z}\in\mathcal{Z}_{d,c}}\widetilde{\mathsf{d}}(\mathbf{z},\boldsymbol{\mu}_{d,c'}) \\
&= \frac{1}{|\mathcal{Z}|}\frac{1}{|\mathcal{D}|(|\mathcal{C}|-1)}|\mathcal{C}||\mathcal{D}|(|\mathcal{C}|-1)\mathbb{E}_d\mathbb{E}_c\mathbb{E}_{c'\neq c}\mathbb{E}_{\mathbf{z}\in\mathcal{Z}_{d,c}}\big[\underbrace{N_{d,c}\cdot\widetilde{\mathsf{d}}(\mathbf{z},\boldsymbol{\mu}_{d,c'})}_{\mathsf{d}(\mathbf{z},\boldsymbol{\mu}_{d,c'})}\big] \\
&= \frac{|\mathcal{C}|}{|\mathcal{Z}|}\underbrace{\mathbb{E}_d\mathbb{E}_c\mathbb{E}_{c'\neq c}\mathbb{E}_{\mathbf{z}\in\mathcal{Z}_{d,c}}\big[\mathsf{d}(\mathbf{z},\boldsymbol{\mu}_{d,c'})\big]}_{\beta}, \qquad (10) \\
T_\gamma &= \frac{1}{|\mathcal{Z}|}\frac{1}{|\mathcal{D}|(|\mathcal{C}|-1)}\sum_{c\in\mathcal{C}}\sum_{d\in\mathcal{D}}\sum_{d'\in\mathcal{D}\setminus\{d\}}\sum_{c'\in\mathcal{C}\setminus\{c\}}\sum_{\mathbf{z}\in\mathcal{Z}_{d,c}}\widetilde{\mathsf{d}}(\mathbf{z},\boldsymbol{\mu}_{d',c'}) \\
&= \frac{1}{|\mathcal{Z}|}\frac{|\mathcal{C}||\mathcal{D}|(|\mathcal{D}|-1)(|\mathcal{C}|-1)}{|\mathcal{D}|(|\mathcal{C}|-1)}\mathbb{E}_d\mathbb{E}_{d'\neq d}\mathbb{E}_c\mathbb{E}_{c'\neq c}\mathbb{E}_{\mathbf{z}\in\mathcal{Z}_{d,c}}\big[\underbrace{N_{d,c}\cdot\widetilde{\mathsf{d}}(\mathbf{z},\boldsymbol{\mu}_{d',c'})}_{\mathsf{d}(\mathbf{z},\boldsymbol{\mu}_{d',c'})}\big] \\
&= \frac{|\mathcal{C}|(|\mathcal{D}|-1)}{|\mathcal{Z}|}\underbrace{\mathbb{E}_d\mathbb{E}_{d'\neq d}\mathbb{E}_c\mathbb{E}_{c'\neq c}\mathbb{E}_{\mathbf{z}\in\mathcal{Z}_{d,c}}\big[\mathsf{d}(\mathbf{z},\boldsymbol{\mu}_{d',c'})\big]}_{\gamma}, \qquad (11)
\end{aligned}$$

where (α,β,γ) are the transferability statistics for $\mathcal{S}$ as in Definition 3. Finally, replace $|\mathcal{Z}| = N$ and combine Eqn. (7), (8), (9), (10), and (11), we have

$$\mathcal{L}_{\texttt{BoDA}}(\mathcal{Z},\{\boldsymbol{\mu}\}) \geq N\log\Big(|\mathcal{D}|-1+|\mathcal{D}|(|\mathcal{C}|-1)\exp\Big(\tfrac{|\mathcal{C}||\mathcal{D}|}{N}\cdot\alpha-\tfrac{|\mathcal{C}|}{N}\cdot\beta-\tfrac{|\mathcal{C}|(|\mathcal{D}|-1)}{N}\cdot\gamma\Big)\Big).$$

This completes the proof.

A.3 Proof of Theorem 2

We first define a notion of *calibrated distance* $\widehat{\mathsf{d}}$. Let $\mathbf{z}\in\mathcal{Z}_{d,c}$, we have

$$\widehat{\mathsf{d}}(\mathbf{z},\boldsymbol{\mu}_{d',c'}) \triangleq \lambda_{d,c}^{d',c'}\cdot\widetilde{\mathsf{d}}(\mathbf{z},\boldsymbol{\mu}_{d',c'}) = \left(\frac{N_{d',c'}}{N_{d,c}}\right)^{\nu}\cdot\widetilde{\mathsf{d}}(\mathbf{z},\boldsymbol{\mu}_{d',c'}).$$

From Theorem 1, by substituting $\widetilde{\mathsf{d}}$ with $\widehat{\mathsf{d}}$, it holds that

$$\begin{aligned}
\widetilde{\mathcal{L}}_{\texttt{BoDA}}(\mathcal{Z},\{\boldsymbol{\mu}\}) &= \mathcal{L}_{\texttt{BoDA}}(\mathcal{Z},\{\boldsymbol{\mu}\})\Big|_{\widetilde{\mathsf{d}}\to\widehat{\mathsf{d}}} \\
&\geq N\log\left(|\mathcal{D}|-1+|\mathcal{D}|(|\mathcal{C}|-1)\exp\left(T'_\alpha-T'_\beta-T'_\gamma\right)\right),
\end{aligned} \qquad (12)$$

where T'_α, T'_β, and T'_γ can be expressed as

$$\begin{aligned} T'_\alpha &= \frac{|\mathcal{C}||\mathcal{D}|}{N}\mathbb{E}_c\mathbb{E}_d\mathbb{E}_{d'\neq d}\mathbb{E}_{\mathbf{z}\in\mathcal{Z}_{d,c}}\left[N_{d,c}\cdot\widehat{\mathsf{d}}(\mathbf{z},\boldsymbol{\mu}_{d',c})\right] \\ &= \frac{|\mathcal{C}||\mathcal{D}|}{N}\mathbb{E}_c\mathbb{E}_d\mathbb{E}_{d'\neq d}\mathbb{E}_{\mathbf{z}\in\mathcal{Z}_{d,c}}\Big[\lambda_{d,c}^{d',c}\cdot\underbrace{N_{d,c}\cdot\widetilde{\mathsf{d}}(\mathbf{z},\boldsymbol{\mu}_{d',c})}_{\mathsf{d}(\mathbf{z},\boldsymbol{\mu}_{d',c})}\Big] \\ &= \frac{|\mathcal{C}||\mathcal{D}|}{N}\underbrace{\mathbb{E}_c\mathbb{E}_d\mathbb{E}_{d'\neq d}\left[\lambda_{d,c}^{d',c}\cdot\mathbb{E}_{\mathbf{z}\in\mathcal{Z}_{d,c}}\left[\mathsf{d}(\mathbf{z},\boldsymbol{\mu}_{d',c})\right]\right]}_{\widetilde{\alpha}}, \end{aligned} \tag{13}$$

$$\begin{aligned} T'_\beta &= \frac{|\mathcal{C}|}{N}\mathbb{E}_d\mathbb{E}_c\mathbb{E}_{c'\neq c}\mathbb{E}_{\mathbf{z}\in\mathcal{Z}_{d,c}}\left[N_{d,c}\cdot\widehat{\mathsf{d}}(\mathbf{z},\boldsymbol{\mu}_{d,c'})\right] \\ &= \frac{|\mathcal{C}|}{N}\mathbb{E}_d\mathbb{E}_c\mathbb{E}_{c'\neq c}\mathbb{E}_{\mathbf{z}\in\mathcal{Z}_{d,c}}\Big[\lambda_{d,c}^{d,c'}\cdot\underbrace{N_{d,c}\cdot\widetilde{\mathsf{d}}(\mathbf{z},\boldsymbol{\mu}_{d,c'})}_{\mathsf{d}(\mathbf{z},\boldsymbol{\mu}_{d,c'})}\Big] \\ &= \frac{|\mathcal{C}|}{N}\underbrace{\mathbb{E}_d\mathbb{E}_c\mathbb{E}_{c'\neq c}\left[\lambda_{d,c}^{d,c'}\cdot\mathbb{E}_{\mathbf{z}\in\mathcal{Z}_{d,c}}\left[\mathsf{d}(\mathbf{z},\boldsymbol{\mu}_{d,c'})\right]\right]}_{\widetilde{\beta}}, \end{aligned} \tag{14}$$

$$\begin{aligned} T'_\gamma &= \frac{|\mathcal{C}|(|\mathcal{D}|-1)}{N}\mathbb{E}_d\mathbb{E}_{d'\neq d}\mathbb{E}_c\mathbb{E}_{c'\neq c}\mathbb{E}_{\mathbf{z}\in\mathcal{Z}_{d,c}}\left[N_{d,c}\cdot\widehat{\mathsf{d}}(\mathbf{z},\boldsymbol{\mu}_{d',c'})\right] \\ &= \frac{|\mathcal{C}|(|\mathcal{D}|-1)}{N}\mathbb{E}_d\mathbb{E}_{d'\neq d}\mathbb{E}_c\mathbb{E}_{c'\neq c}\mathbb{E}_{\mathbf{z}\in\mathcal{Z}_{d,c}}\Big[\lambda_{d,c}^{d',c'}\cdot\underbrace{N_{d,c}\cdot\widetilde{\mathsf{d}}(\mathbf{z},\boldsymbol{\mu}_{d',c'})}_{\mathsf{d}(\mathbf{z},\boldsymbol{\mu}_{d',c'})}\Big] \\ &= \frac{|\mathcal{C}|(|\mathcal{D}|-1)}{N}\underbrace{\mathbb{E}_d\mathbb{E}_{d'\neq d}\mathbb{E}_c\mathbb{E}_{c'\neq c}\left[\lambda_{d,c}^{d',c'}\cdot\mathbb{E}_{\mathbf{z}\in\mathcal{Z}_{d,c}}\left[\mathsf{d}(\mathbf{z},\boldsymbol{\mu}_{d',c'})\right]\right]}_{\widetilde{\gamma}}, \end{aligned} \tag{15}$$

where $(\widetilde{\alpha},\widetilde{\beta},\widetilde{\gamma})$ are formally defined in Definition 4. Combine Eqn. (12), (13), (14), and (15), we have

$$\widetilde{\mathcal{L}}_{\texttt{BoDA}}(\mathcal{Z},\{\boldsymbol{\mu}\}) \geq N\log\left(|\mathcal{D}|-1+|\mathcal{D}|(|\mathcal{C}|-1)\exp\left(\tfrac{|\mathcal{C}||\mathcal{D}|}{N}\cdot\widetilde{\alpha}-\tfrac{|\mathcal{C}|}{N}\cdot\widetilde{\beta}-\tfrac{|\mathcal{C}|(|\mathcal{D}|-1)}{N}\cdot\widetilde{\gamma}\right)\right),$$

which completes the proof.

B Additional Discussions, Properties, and Interpretations

B.1 Unified Interpretation for Single- and Multi-Domain Imbalance

In the main paper we show that, in the multi-domain setting, label imbalance implicitly brings *label divergence* across domains, which brings additional challenges and potentially harms MDLT performance. Here we provide a unified viewpoint from the *label divergence* perspective to explain single- and multi-domain data imbalance.

To elaborate, in single domain imbalanced learning, we essentially cope with the divergence between the imbalanced training label distribution and the uniform test label distribution:

$$\mathrm{div}(p(y) \parallel \mathcal{U}),$$

where $\mathrm{div}(\cdot\|\cdot)$ indicates certain divergence measure. In contrast, when extending to the multi-domain scenario, given $|\mathcal{D}|$ domains with (different) imbalanced label distributions, the target divergence becomes

$$\underbrace{\sum_{d} \mathrm{div}\big(p_d(y) \parallel \mathcal{U}\big)}_{\text{imbalanced training}} + \text{const} \cdot \underbrace{\sum_{d \neq d'} \mathrm{div}\big(p_d(y) \parallel p_{d'}(y)\big)}_{\text{divergence across domains}},$$

where one not only needs to tackle the imbalanced training data for each domain $d \in \mathcal{D}$ in order to generalize to the balanced test set, but also takes into consideration the *label divergence* across domains.

Such interpretation echoes our `BoDA` objective: We design the `DA` loss for cross-domain distribution alignment to tackle the latter term, and further adapt it to `BoDA` via balanced distance to address the former term.

B.2 A Probabilistic Perspective of $\mathcal{L}_{\texttt{DA}}$ Derivation

Recall $\mathcal{M} = \mathcal{D} \times \mathcal{C}$ the set of all (d, c) pairs. Let $(\mathbf{x}_i, c_i, d_i)$ denote a sample with feature $\mathbf{z}_i$. Following the metric learning setting [17], we model the likelihood of $\boldsymbol{\mu}_{d,c}$ given $\mathbf{z}_i$ to decay exponentially with respect to their distance in the representation space. Such modeling can be viewed as performing a random walk with transition probability inversely related to distance [16]. For domain-class pairs that share the same class label but different domain labels with $\mathbf{x}_i$ (i.e., $(d, c_i), d \neq d_i$), the normalized likelihood of $\boldsymbol{\mu}_{d,c_i}$ given $\mathbf{z}_i$ can be written as

$$\mathbb{P}((d, c_i)|\mathbf{z}_i) = \frac{\exp\left(-\mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i})\right)}{\sum_{(d',c') \in \mathcal{M} \setminus \{(d_i,c_i)\}} \exp\left(-\mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})\right)},$$

where the denominator is a sum over all domain-class pairs except (d_i, c_i). As motivated, we want to concentrate all $\mathbf{z}_i$ from the same class across different domains (i.e., smaller α), while separating $\mathbf{z}_i$ from different classes within and across domains (i.e., larger β, γ). Therefore, the positive domain-class pairs with $\mathbf{x}_i$ are those share the same class labels but different domain labels. As a result, we define the per-sample loss as the average negative log-likelihood over all positive domain-class pairs:

$$\ell_{\texttt{DA}}(\mathbf{z}_i, \{\boldsymbol{\mu}\}) = -\frac{1}{|\mathcal{D}| - 1} \sum_{d \in \mathcal{D} \setminus \{d_i\}} \log \frac{\exp\left(-\mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i})\right)}{\sum_{(d',c') \in \mathcal{M} \setminus \{(d_i,c_i)\}} \exp\left(-\mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})\right)}.$$

Given a set of all training samples with representation set as $\mathcal{Z}$, the total loss can then be derived as

$$\mathcal{L}_{\texttt{DA}}(\mathcal{Z}, \{\boldsymbol{\mu}\}) = \sum_{\mathbf{z}_i \in \mathcal{Z}} \frac{-1}{|\mathcal{D}| - 1} \sum_{d \in \mathcal{D} \setminus \{d_i\}} \log \frac{\exp\left(-\mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i})\right)}{\sum_{(d',c') \in \mathcal{M} \setminus \{(d_i,c_i)\}} \exp\left(-\mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})\right)}.$$

B.3 Intrinsic Hardness-Aware Property of BoDA

Below, we demonstrate an additional property of BoDA: the intrinsic *hardness-aware* property. Specifically, we analyze the gradients of BoDA loss with respect to positive (d, c) pairs and different negative (d, c) pairs. We observe that the gradient contributions from *hard* positives/negatives are larger than that from the *easy* ones, indicating that BoDA automatically concentrates on the *hard* (d, c) pairs, where penalties are given according to their hardness.

Recall that the sample-wise calibrated BoDA loss $\widetilde{\ell}_{\texttt{BoDA}}$ can be written as

$$\begin{aligned}
&\widetilde{\ell}_{\texttt{BoDA}}(\mathbf{z}_i, \{\boldsymbol{\mu}\}) \\
&= -\frac{1}{|\mathcal{D}|-1} \sum_{d \in \mathcal{D} \setminus \{d_i\}} \log \frac{\exp\left(-\lambda_{d_i,c_i}^{d,c_i} \widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i})\right)}{\sum_{(d',c') \in \mathcal{M} \setminus \{(d_i,c_i)\}} \exp\left(-\lambda_{d_i,c_i}^{d',c'} \widetilde{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})\right)} \\
&= -\frac{1}{|\mathcal{D}|-1} \sum_{d \in \mathcal{D} \setminus \{d_i\}} \log \frac{\exp\left(-\frac{\lambda_{d_i,c_i}^{d,c_i}}{N_{d_i,c_i}} \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i})\right)}{\sum_{(d',c') \in \mathcal{M} \setminus \{(d_i,c_i)\}} \exp\left(-\frac{\lambda_{d_i,c_i}^{d',c'}}{N_{d_i,c_i}} \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})\right)},
\end{aligned} \tag{16}$$

where $\mathbf{z}_i \in \mathcal{Z}_{d_i,c_i}$. For convenience, we further define the probability of $\mathbf{z}_i$ being recognized as belonging to $\boldsymbol{\mu}_{d,c}$ as

$$P_{d,c}^i \triangleq \frac{\exp\left(-\frac{\lambda_{d_i,c_i}^{d,c}}{N_{d_i,c_i}} \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c})\right)}{\sum_{(d',c') \in \mathcal{M} \setminus \{(d_i,c_i)\}} \exp\left(-\frac{\lambda_{d_i,c_i}^{d',c'}}{N_{d_i,c_i}} \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})\right)}, \quad (d,c) \in \mathcal{M} \setminus \{(d_i,c_i)\}.$$

Note that the essential goal of Eqn. (16) is to align (minimize) *positive* distances $\mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i})$ and to separate (maximize) *negative* distances $\mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})$. Therefore, we analyze the gradients with respect to positive distance and different negative distances to explore the properties of $\widetilde{\ell}_{\texttt{BoDA}}$. Specifically, we have

$$\begin{aligned}
&\frac{\partial \widetilde{\ell}_{\texttt{BoDA}}(\mathbf{z}_i, \{\boldsymbol{\mu}\})}{\partial \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i})} \\
&= \frac{-1}{|\mathcal{D}|-1} \sum_{d \in \mathcal{D} \setminus \{d_i\}} \frac{\partial}{\partial \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i})} \left\{ -\frac{\lambda_{d_i,c_i}^{d,c_i}}{N_{d_i,c_i}} \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i}) - \log \sum_{(d',c') \in \mathcal{M} \setminus \{(d_i,c_i)\}} \exp\left(-\frac{\lambda_{d_i,c_i}^{d',c'}}{N_{d_i,c_i}} \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})\right) \right\} \\
&= \frac{1}{|\mathcal{D}|-1} \sum_{d \in \mathcal{D} \setminus \{d_i\}} \frac{\lambda_{d_i,c_i}^{d,c_i}}{N_{d_i,c_i}} \left(1 - \frac{\exp\left(-\frac{\lambda_{d_i,c_i}^{d,c_i}}{N_{d_i,c_i}} \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i})\right)}{\sum_{(d',c') \in \mathcal{M} \setminus \{(d_i,c_i)\}} \exp\left(-\frac{\lambda_{d_i,c_i}^{d',c'}}{N_{d_i,c_i}} \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})\right)}\right) \\
&= \frac{1}{|\mathcal{D}|-1} \sum_{d \in \mathcal{D} \setminus \{d_i\}} \frac{N_{d,c_i}^{\nu}}{N_{d_i,c_i}^{(1+\nu)}} \left(1 - P_{d,c_i}^i\right) \\
&\propto \sum_{d \in \mathcal{D} \setminus \{d_i\}} N_{d,c_i}^{\nu} \left(1 - P_{d,c_i}^i\right),
\end{aligned}$$

$$
\begin{aligned}
&\frac{\partial \widetilde{\ell}_{\texttt{BoDA}}(\mathbf{z}_i, \{\boldsymbol{\mu}\})}{\partial \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})} \\
&= \frac{-1}{|\mathcal{D}|-1} \sum_{d \in \mathcal{D}\setminus\{d_i\}} \frac{\partial}{\partial \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})} \left\{ -\frac{\lambda_{d_i,c_i}^{d,c_i}}{N_{d_i,c_i}} \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i}) - \log \sum_{(d',c') \in \mathcal{M}\setminus\{(d_i,c_i)\}} \exp\left(-\frac{\lambda_{d_i,c_i}^{d',c'}}{N_{d_i,c_i}} \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'}) \right) \right\} \\
&= -\frac{1}{|\mathcal{D}|-1} \sum_{d \in \mathcal{D}\setminus\{d_i\}} \frac{\lambda_{d_i,c_i}^{d',c'}}{N_{d_i,c_i}} \frac{\exp\left(-\frac{\lambda_{d_i,c_i}^{d',c'}}{N_{d_i,c_i}} \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i}) \right)}{\sum_{(d',c') \in \mathcal{M}\setminus\{(d_i,c_i)\}} \exp\left(-\frac{\lambda_{d_i,c_i}^{d',c'}}{N_{d_i,c_i}} \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'}) \right)} \\
&= -\frac{1}{|\mathcal{D}|-1} \sum_{d \in \mathcal{D}\setminus\{d_i\}} \frac{N_{d',c'}^{\nu}}{N_{d_i,c_i}^{(1+\nu)}} P_{d',c'}^{i} \\
&\propto -N_{d',c'}^{\nu} P_{d',c'}^{i}.
\end{aligned}
$$

Combine the above results, we have

$$
\text{positive:} \quad \frac{\partial \widetilde{\ell}_{\texttt{BoDA}}(\mathbf{z}_i, \{\boldsymbol{\mu}\})}{\partial \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d,c_i})} \propto \sum_{d \in \mathcal{D}\setminus\{d_i\}} N_{d,c_i}^{\nu} \left(1 - P_{d,c_i}^{i}\right), \tag{17}
$$

$$
\text{negative:} \quad \frac{\partial \widetilde{\ell}_{\texttt{BoDA}}(\mathbf{z}_i, \{\boldsymbol{\mu}\})}{\partial \mathsf{d}(\mathbf{z}_i, \boldsymbol{\mu}_{d',c'})} \propto -N_{d',c'}^{\nu} P_{d',c'}^{i}. \tag{18}
$$

Interpretation. Eqn. (17) and (18) illustrate several interesting and important properties of `BoDA`:

1. *Intrinsic hard positive and negative mining.* For positive pairs, we observe that the gradient magnitudes are proportional to $(1-P_{d,c_i}^{i})$, where for an easy (d, c_i) pair, $P_{d,c_i}^{i} \approx 1$ and $(1 - P_{d,c_i}^{i}) \approx 0$, and for a hard (d, c_i) pair, $P_{d,c_i}^{i} \approx 0$ and $(1 - P_{d,c_i}^{i}) \approx 1$, indicating that the gradient contributions from *hard* positives are larger than *easy* ones. Similarly, for negative pairs, the gradient magnitudes are proportional to $P_{d',c'}^{i}$, where an easy (d', c') pair has $P_{d',c'}^{i} \approx 0$ and a hard (d, c_i) pair induces $P_{d',c'}^{i} \approx 1$, showing that the gradient contribution is large for hard negatives and small for easy negatives. Therefore, `BoDA` is a hardness-aware loss with intrinsic hard positive/negative mining property.
2. *Scaling gradients according to the number of samples of each* (d, c). Furthermore, as we have shown in Fig. 5, when data are imbalanced across different (d, c) pairs, minority pairs with smaller number of samples would induce worse $\boldsymbol{\mu}_{d,c}$ estimates. We further observe that the gradients for both positive and negative pairs are proportional to their number of samples (i.e., N_{d,c_i}^{ν} and $N_{d',c'}^{\nu}$). This suggests that `BoDA` automatically adjusts the gradient scale for each (d, c) according to how accurate the estimation of $\boldsymbol{\mu}_{d,c}$ is. The appealing property highlights that `BoDA` also implicitly calibrates the gradient scale, emphasizing gradients from majority pairs (which are more reliable) while suppressing gradients from minority pairs (which are less reliable). Such behavior is essential for better statistics transfer as we demonstrated in the main paper.

C Pseudo Code for `BoDA`

We provide the pseudo code of `BoDA` in Algorithm 1.

Algorithm 1 Balanced Domain-Class Distribution Alignment (`BoDA`)

Input: Training set $\mathcal{D} = \{(\mathbf{x}_i, c_i, d_i)\}_{i=1}^{N}$, all domain-class pairs $\mathcal{M} = \{(d, c)\}$, encoder f, classifier g, total training epochs E, calibration parameter ν, loss weight ω, momentum α
for all $(d, c) \in \mathcal{M}$ **do**
  Initialize the feature statistics $\{\boldsymbol{\mu}_{d,c}^{(0)}, \boldsymbol{\Sigma}_{d,c}^{(0)}\}$
end for
for $e = 0$ **to** E **do**
  repeat
    Sample a mini-batch $\{(\mathbf{x}_i, c_i, d_i)\}_{i=1}^{m}$ from $\mathcal{D}$
    for $i = 1$ **to** m (in parallel) **do**
      $\mathbf{z}_i = f(\mathbf{x}_i)$
      $\widehat{c}_i = g(\mathbf{z}_i)$
    end for
    Calculate $\widetilde{\mathcal{L}}_{\texttt{BoDA}}$ using $\{\mathbf{z}_i\}$ based on Eqn. (4)
    Calculate $\mathcal{L}_{\texttt{CE}}$ using $\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(\widehat{c}_i, c_i)$
    Do one training step with loss $\mathcal{L}_{\texttt{CE}} + \omega\widetilde{\mathcal{L}}_{\texttt{BoDA}}$
  until iterate over all training samples at current epoch e
  `/* Update feature statistics with momentum updating */`
  for all $(d, c) \in \mathcal{M}$ **do**
    Estimate current feature statistics $\{\boldsymbol{\mu}_{d,c}, \boldsymbol{\Sigma}_{d,c}\}$
    $\boldsymbol{\mu}_{d,c}^{(e+1)} \leftarrow \alpha \times \boldsymbol{\mu}_{d,c}^{(e)} + (1-\alpha) \times \boldsymbol{\mu}_{d,c}$
    $\boldsymbol{\Sigma}_{d,c}^{(e+1)} \leftarrow \alpha \times \boldsymbol{\Sigma}_{d,c}^{(e)} + (1-\alpha) \times \boldsymbol{\Sigma}_{d,c}$
  end for
end for

D Details of MDLT Datasets

In this section, we provide the detailed information of the curated MDLT datasets we used in our experiments. Table 10 provides an overview of the datasets. Table 11 provides the image examples across domains for each MDLT dataset.

`Digits-MLT`. We construct `Digits-MLT` by combining two digit datasets: (1) MNIST-M [15], a variant of the original MNIST handwritten digit classification dataset [26] with colorful background, and (2) SVHN [36]. The original MNIST-M dataset contains 60,000 training samples and 10,000 testing examples, and the original SVHN dataset contains 73,257 images for training and 26,032 images for testing. Both datasets have examples of dimension $(3, 32, 32)$ and 10 classes. We create `Digits-MLT` with controllable degrees of data imbalance, where we keep the maximum number of samples each (d, c) to be 1,000, and manually vary the imbalance degree to adjust the number of samples for minority (d, c). For validation and test set, we use the original test set of the two datasets, but keep the number of samples each (d, c) to be 800.

`VLCS-MLT`. The original `VLCS` dataset [14] is an object recognition dataset that comprises photographic domains $d \in \{$ Caltech101, LabelMe, SUN09, VOC2007 $\}$, with scenes captured from urban to rural. The dataset contains 5 classes with

Table 10. Detailed statistics of the curated MDLT datasets used in our experiments. For the synthetic `Digits-MLT` dataset, we manually vary the minimum (d, c) size to simulate different degrees of imbalance.

Dataset	# Domains	# Classes	Max (d, c) size	Min (d, c) size	# Training set	# Val. set	# Test set
`Digits-MLT`	2	10	1,000	$10 \sim 1{,}000$	$20{,}000 \sim 4{,}956$	16,000	16,000
`VLCS-MLT`	4	5	1,454	0	9,872	285	572
`PACS-MLT`	4	7	741	5	7,891	700	1,400
`OfficeHome-MLT`	4	65	84	0	11,688	1,300	2,600
`TerraInc-MLT`	4	10	4,455	0	23,269	353	708
`DomainNet-MLT`	6	345	778	0	468,574	39,240	78,761

Table 11. Overview of images from different domains in all MDLT datasets. For each dataset, we pick a single class and show illustrative images from each domain.

Dataset	Domains					
`Digits-MLT`	MNIST-M 	SVHN 				
`VLCS-MLT`	Caltech101 	LabelMe 	SUN09 	VOC2007 		
`PACS-MLT`	Art 	Cartoon 	Photo 	Sketch 		
`OfficeHome-MLT`	Art 	Clipart 	Product 	Photo 		
`TerraInc-MLT`	L100 	L38 	L43 	L46 		
	(camera trap location)					
`DomainNet-MLT`	Clipart 	Infographic 	Painting 	QuickDraw 	Photo 	Sketch 

10,729 examples of dimension $(3, 224, 224)$. To construct `VLCS-MLT`, for each (d, c) we split out a validation set of size 15 and a test set of size 30, and leave the rest for training.

`PACS-MLT`. The original `PACS` dataset [28] is an object recognition dataset that comprises four domains $d \in \{$ art, cartoons, photos, sketches $\}$ with image style changes. It contains 7 classes with 9,991 examples of dimension $(3, 224, 224)$. We construct `PACS-MLT` in a simialr manner as `VLCS-MLT`, where we split out a

validation set of size 25 and a test set of size 50 for each (d, c), and leave the rest for training.

`OfficeHome-MLT`. The original `OfficeHome` dataset [47] includes domains $d \in \{$ art, clipart, product, real $\}$, containing 15,588 examples of dimension $(3, 224, 224)$ and 65 classes. We make `OfficeHome-MLT` by splitting out a validation set of size 5 and a test set of size 10 for each (d, c), leaving the rest for training.

`TerraInc-MLT`. `TerraInc-MLT` is constructed from `TerraIncognita` dataset [2], a species classification dataset that contains photographs of wild animals taken by camera traps at locations $d \in \{\text{L100}, \text{L38}, \text{L43}, \text{L46}\}$. The dataset contains 10 classes with 24,788 examples of dimension $(3, 224, 224)$. For each (d, c), we split out a validation set of size 10 and a test set of size 20, and use all remaining samples for training.

`DomainNet-MLT`. We construct `DomainNet-MLT` using `DomainNet` dataset [38], a large-scale multi-domain dataset for object recognition that consists of six domains $d \in \{$ clipart, infograph, painting, quickdraw, real, sketch $\}$, 345 classes, and 586,575 examples of size $(3, 224, 224)$. To construct `DomainNet-MLT`, for each (d, c) we split out a validation set of size 20 and a test set of size 40, and leave the rest for training.

E Experimental Settings

E.1 Implementation Details

For the synthetic `Digits-MLT` dataset, we fix the network architecture as a small MNIST CNN [19] for all algorithms, and use no data augmentation. For all other MDLT datasets, following [19], we use the pretrained ResNet-50 model [21] as the backbone network for all algorithms, and use the same data augmentation protocol as [19]: random crop and resize to 224×224 pixels, random horizontal flips, random color jitter, grayscaling the image with 10% probability, and normalization using the ImageNet channel statistics. We train all models using the Adam optimizer [24] for 5,000 steps on all MDLT datasets except `DomainNet-MLT`, on which we train longer for 15,000 steps to ensure convergence. We fix a batch size of 64 per domain for `Digits-MLT` experiments, a batch size of 32 per domain for `DomainNet-MLT` experiments, and a batch size of 24 per domain for experiments on all other datasets.

For all MDLT datasets except `OfficeHome-MLT` and `TerraInc-MLT`, we define *many-shot* (d, c) pairs as with over 100 training samples, *medium-shot* as with 20∼100 training samples, and *few-shot* as with under 20 training samples. For `OfficeHome-MLT`, we define *many-shot* as (d, c) pairs with over 60 training samples, *medium-shot* as with 20∼60 training samples, and *few-shot* as with under 20 training samples. For `TerraInc-MLT`, we define *many-shot* as (d, c) pairs with over 100 training samples, *medium-shot* as with 25∼100 training samples, and *few-shot* as with under 25 training samples.

E.2 Competing Algorithms

We compare `BoDA` to a large number of algorithms that span different learning strategies. We group them according to their categories, and provide detailed descriptions for each algorithm below.

- *Vanilla:* The empirical risk minimization (**ERM**) [46] minimizes the sum of errors across all domains and samples.
- *Distributionally robust optimization:* Group distributionally robust optimization (**GroupDRO**) [40] performs ERM while increasing the importance of domains with larger errors.
- *Cross-domain data augmentation:* Inter-domain mixup (**Mixup**) [50] performs ERM on linear interpolations of examples from random pairs of domains and their labels. Style-agnostic network (**SagNet**) [35] disentangles style encodings from image content by randomizing and augmenting styles.
- *Meta-learning:* Meta-learning for domain generalization (**MLDG**) [27] leverages meta-learning to learn how to generalize across domains.
- *Domain-invariant representation learning:* Invariant risk minimization (**IRM**) [1] learns a feature representation such that the optimal linear classifier on top of that representation matches across domains. Domain adversarial neural networks (**DANN**) [15] employ an adversarial network to match feature distributions. Class-conditional DANN (**CDANN**) [31] builds upon DANN but further matches the conditional distributions across domains for all labels. Deep correlation alignment (**CORAL**) [45] matches the mean and covariance of feature distributions. Maximum mean discrepancy (**MMD**) [29] matches the MMD [18] of feature distributions.
- *Transfer learning:* Marginal transfer learning (**MTL**) [4] estimates a mean embedding per domain, passed as a second argument to the classifier.
- *Multi-task learning:* Gradient matching for domain generalization (**Fish**) [42] maximizes the inner product between gradients from different domains through a multi-task objective.
- *Imbalanced learning:* Focal loss (**Focal**) [32] reduces the relative loss for well-classified samples and focuses on difficult samples. Class-balanced loss (**CBLoss**) [10] proposes re-weighting by the inverse effective number of samples. The LDAM loss (**LDAM**) [6] employs a modified marginal loss that favors minority samples more. Balanced-Softmax (**BSoftmax**) [39] extends Softmax to an unbiased estimation that considers the number of samples of each class. Self-supervised pre-training (**SSP**) [52] uses self-supervised learning as a first-stage pre-training to alleviate the network dependence on imbalanced labels. Classifier re-training (**CRT**) [23] decomposes the representation and classifier learning into two stages, where it fine-tunes the classifier using class-balanced sampling with representation fixed in the second stage.

E.3 Hyperparameters Search Protocol

For a fair evaluation across different algorithms, following the training protocol in [19], for each algorithm we conduct a random search of 20 trials over a joint

Table 12. Hyperparameters search space for all experiments.

Condition	Parameter	Default value	Random distribution
General:			
ResNet	learning rate	0.00005	$10^{\text{Uniform}(-5,-3.5)}$
	dropout	0	RandomChoice($[0, 0.1, 0.5]$)
	generator learning rate	0.00005	$10^{\text{Uniform}(-5,-3.5)}$
	discriminator learning rate	0.00005	$10^{\text{Uniform}(-5,-3.5)}$
not ResNet	learning rate	0.001	$10^{\text{Uniform}(-4.5,-3.5)}$
	generator learning rate	0.001	$10^{\text{Uniform}(-4.5,-2.5)}$
	discriminator learning rate	0.001	$10^{\text{Uniform}(-4.5,-2.5)}$
`Digits-MLT`	weight decay	0	0
	generator weight decay	0	0
not `Digits-MLT`	weight decay	0	$10^{\text{Uniform}(-6,-2)}$
	generator weight decay	0	$10^{\text{Uniform}(-6,-2)}$
Algorithm-specific:			
IRM	lambda	100	$10^{\text{Uniform}(-1,5)}$
	iterations of penalty annealing	500	$10^{\text{Uniform}(0,4)}$
GroupDRO	eta	0.01	$10^{\text{Uniform}(-3,-1)}$
Mixup	alpha	0.2	$10^{\text{Uniform}(0,4)}$
MLDG	beta	1	$10^{\text{Uniform}(-1,1)}$
CORAL, MMD	gamma	1	$10^{\text{Uniform}(-1,1)}$
DANN, CDANN	lambda	1.0	$10^{\text{Uniform}(-2,2)}$
	discriminator weight decay	0	$10^{\text{Uniform}(-6,-2)}$
	discriminator steps	1	$2^{\text{Uniform}(0,3)}$
	gradient penalty	0	$10^{\text{Uniform}(-2,1)}$
	adam β_1	0.5	RandomChoice($[0, 0.5]$)
MTL	ema	0.99	RandomChoice($[.5, .9, .99, 1]$)
SagNet	adversary weight	0.1	$10^{\text{Uniform}(-2,1)}$
Fish	meta learning rate	0.5	RandomChoice($[.05, .1, .5]$)
Focal	gamma	1	$0.5 * 10^{\text{Uniform}(0,1)}$
CBLoss	beta	0.9999	$1 - 10^{\text{Uniform}(-5,-2)}$
LDAM	max_m	0.5	$10^{\text{Uniform}(-1,-0.1)}$
	scale	30	RandomChoice($[10, 30]$)
BoDA	nu	1	$10^{\text{Uniform}(-0.5,0)}$
	`BoDA` loss weight	0.1	$10^{\text{Uniform}(-2,-0.5)}$

distribution of its all hyperparameters. We then use the validation set to select the best hyperparameters for each algorithm, fix them and rerun the experiments under 3 different random seeds to report the final average accuracy (and standard deviation). Such process ensures the comparison is best-versus-best, and the hyperparameters are optimized for all algorithms.

We detail the hyperparameter choices for each algorithm in Table 12.

E.4 Settings for DG Experiments

For DG experiments, we strictly follow the training protocols described in [19]. Across all benchmark DG datasets, we keep the same hyperparameter search space for `BoDA` as in Table 12. We fix all other training parameters unchanged so that the results of `BoDA` are directly comparable to the results in [19].

For model selection, we use the *training-domain validation set* protocol in [19] with $80\% - 20\%$ training-validation split, and the average out-domain test performance is reported across all runs for each domain.

F Complete Results for MDLT

We provide complete evaluation results on the five MDLT datasets. In addition to the reported results in the main paper, for each dataset we also include the accuracy on each domain together with the averaged and the worst accuracy.

F.1 `VLCS-MLT`

Table 13. Complete evaluation results on `VLCS-MLT`.

	Accuracy (by domain)						Accuracy (by shot)			
Algorithm	**C**	**L**	**S**	**V**	**Average**	**Worst**	**Many**	**Medium**	**Few**	**Zero**
ERM	99.3 ±0.3	53.6 ±1.1	65.9 ±1.2	86.4 ±0.7	76.3 ±0.4	53.6 ±1.1	84.6 ±0.5	76.6 ±0.4	–	32.9 ±0.4
IRM	99.1 ±0.4	52.3 ±0.7	68.8 ±1.4	86.0 ±0.3	76.5 ±0.2	52.3 ±0.7	85.3 ±0.6	75.5 ±1.0	–	33.5 ±1.0
GroupDRO	98.7 ±0.3	54.1 ±1.3	67.5 ±1.5	86.7 ±0.3	76.7 ±0.4	54.1 ±1.3	85.3 ±0.9	76.2 ±1.0	–	34.5 ±2.0
Mixup	99.3 ±0.3	52.7 ±1.3	66.1 ±0.0	85.3 ±1.1	75.9 ±0.1	52.7 ±1.3	84.4 ±0.2	77.1 ±0.6	–	29.2 ±1.4
MLDG	99.3 ±0.3	53.6 ±0.5	68.3 ±0.4	86.4 ±0.5	76.9 ±0.2	53.6 ±0.5	84.9 ±0.3	77.5 ±1.0	–	34.4 ±0.9
CORAL	99.3 ±0.3	51.6 ±0.7	67.5 ±1.8	85.3 ±0.9	75.9 ±0.5	51.6 ±0.7	84.3 ±0.6	75.5 ±0.5	–	34.5 ±0.8
MMD	99.6 ±0.2	53.4 ±0.3	65.6 ±0.8	86.7 ±1.1	76.3 ±0.6	53.4 ±0.3	84.5 ±0.8	77.1 ±0.5	–	32.7 ±0.3
DANN	99.6 ±0.2	54.1 ±0.3	69.9 ±0.2	86.7 ±0.0	77.5 ±0.1	54.1 ±0.3	85.9 ±0.5	76.0 ±0.4	–	38.0 ±2.3
CDANN	99.6 ±0.4	53.6 ±0.4	67.5 ±0.6	85.8 ±0.8	76.6 ±0.4	53.6 ±0.4	84.4 ±0.7	77.3 ±0.8	–	35.0 ±0.8
MTL	99.1 ±0.2	52.9 ±0.5	66.7 ±0.4	86.7 ±0.6	76.3 ±0.3	52.9 ±0.5	84.8 ±0.9	76.2 ±0.6	–	33.3 ±1.4
SagNet	99.6 ±0.4	52.3 ±0.2	67.2 ±0.2	86.2 ±1.0	76.3 ±0.2	52.3 ±0.2	85.3 ±0.3	75.1 ±0.2	–	32.9 ±0.3
Fish	98.7 ±0.3	54.3 ±0.4	69.4 ±0.8	87.6 ±0.4	77.5 ±0.3	54.3 ±0.4	86.2 ±0.5	76.0 ±0.4	–	35.6 ±2.2
Focal	99.1 ±0.4	52.3 ±0.2	66.1 ±0.8	84.9 ±0.2	75.6 ±0.4	52.3 ±0.2	84.0 ±0.2	75.5 ±0.6	–	32.7 ±0.9
CBLoss	99.1 ±0.2	52.5 ±0.5	68.5 ±1.0	87.1 ±1.0	76.8 ±0.3	52.5 ±0.5	84.8 ±0.7	77.5 ±1.4	–	33.2 ±1.6
LDAM	98.9 ±0.2	52.9 ±0.2	69.4 ±1.4	**88.0** ±1.3	77.5 ±0.1	52.9 ±0.2	**86.5** ±0.4	75.5 ±0.5	–	35.2 ±0.6
BSoftmax	99.3 ±0.3	52.9 ±0.9	68.0 ±0.2	86.7 ±0.8	76.7 ±0.5	52.9 ±0.9	84.4 ±0.9	78.2 ±0.6	–	34.3 ±0.9
SSP	99.1 ±0.2	52.3 ±1.0	68.0 ±0.2	85.1 ±0.4	76.1 ±0.3	52.3 ±1.0	83.8 ±0.3	76.0 ±1.2	–	37.1 ±0.7
CRT	99.6 ±0.3	51.4 ±0.3	66.9 ±0.8	86.9 ±0.4	76.3 ±0.2	51.4 ±0.3	84.5 ±0.1	77.3 ±0.0	–	31.7 ±1.0
BoDA$_r$	99.3 ±0.3	51.4 ±0.3	70.2 ±0.4	86.7 ±0.3	76.9 ±0.5	51.4 ±0.3	85.3 ±0.3	77.3 ±0.2	–	33.3 ±0.5
BoDA-M$_r$	**100.0** ±0.0	53.4 ±0.3	68.5 ±0.4	**88.0** ±0.8	77.5 ±0.3	53.4 ±0.3	85.8 ±0.2	77.3 ±0.2	–	35.7 ±0.7
BoDA$_{r,c}$	99.3 ±0.3	53.4 ±0.3	68.5 ±0.4	**88.0** ±0.4	77.3 ±0.2	53.4 ±0.3	85.3 ±0.3	78.0 ±0.2	–	38.6 ±0.7
BoDA-M$_{r,c}$	**100.0** ±0.0	**55.4** ±0.5	**72.6** ±0.3	84.7 ±0.5	**78.2** ±0.4	**55.4** ±0.5	85.3 ±0.3	**79.3** ±0.6	–	**43.3** ±1.1
BoDA *vs.* ERM	**+0.7**	**+1.8**	**+6.7**	**+1.6**	**+1.9**	**+1.8**	**+0.7**	**+2.7**	–	**+10.4**

F.2 PACS-MLT

Table 14. Complete evaluation results on PACS-MLT.

	Accuracy (by domain)						Accuracy (by shot)			
Algorithm	**A**	**C**	**P**	**S**	**Average**	**Worst**	**Many**	**Medium**	**Few**	**Zero**
ERM	96.8 ±0.1	97.0 ±0.3	98.9 ±0.3	95.8 ±0.2	97.1 ±0.1	95.8 ±0.2	97.1 ±0.0	97.0 ±0.0	98.0 ±0.9	–
IRM	96.8 ±0.1	96.3 ±0.7	98.7 ±0.2	95.2 ±0.4	96.7 ±0.2	95.2 ±0.4	96.8 ±0.2	96.7 ±0.7	94.7 ±1.4	–
GroupDRO	96.9 ±0.2	97.0 ±0.4	99.0 ±0.1	95.3 ±0.4	97.0 ±0.1	95.3 ±0.4	97.3 ±0.1	95.3 ±1.2	94.7 ±3.6	–
Mixup	96.5 ±0.3	96.9 ±0.7	98.5 ±0.2	95.1 ±0.2	96.7 ±0.2	95.1 ±0.2	97.0 ±0.1	96.7 ±0.3	91.3 ±2.7	–
MLDG	96.6 ±0.2	97.2 ±0.3	98.5 ±0.1	94.1 ±0.3	96.6 ±0.1	94.1 ±0.3	96.8 ±0.1	96.3 ±0.7	92.7 ±0.5	–
CORAL	96.9 ±0.4	97.0 ±0.5	98.3 ±0.3	94.3 ±0.7	96.6 ±0.5	94.3 ±0.7	96.6 ±0.5	97.0 ±0.8	94.7 ±0.5	–
MMD	96.8 ±0.2	97.1 ±0.4	97.4 ±0.3	96.3 ±0.3	96.9 ±0.1	96.2 ±0.2	96.9 ±0.2	97.0 ±0.0	96.7 ±0.5	–
DANN	95.7 ±0.3	97.2 ±0.4	98.9 ±0.1	94.3 ±0.1	96.5 ±0.0	94.3 ±0.1	96.5 ±0.1	98.0 ±0.0	94.7 ±2.4	–
CDANN	95.5 ±0.5	96.7 ±0.2	97.2 ±0.3	94.9 ±0.5	96.1 ±0.1	94.5 ±0.2	96.1 ±0.1	96.3 ±0.5	94.0 ±0.9	–
MTL	96.3 ±0.4	97.9 ±0.3	98.2 ±0.3	94.6 ±0.7	96.7 ±0.2	94.5 ±0.6	96.8 ±0.1	95.3 ±1.7	97.3 ±1.1	–
SagNet	97.0 ±0.2	97.8 ±0.4	98.9 ±0.1	95.2 ±0.3	**97.2** ±0.1	95.2 ±0.3	**97.4** ±0.1	96.7 ±0.5	95.3 ±0.5	–
Fish	95.5 ±0.2	97.9 ±0.4	98.2 ±0.3	95.9 ±0.5	96.9 ±0.2	95.2 ±0.2	97.0 ±0.1	97.0 ±0.5	94.7 ±1.1	–
Focal	96.6 ±0.4	96.6 ±0.8	98.1 ±0.2	94.6 ±0.7	96.5 ±0.2	94.6 ±0.7	96.6 ±0.1	95.0 ±1.7	96.7 ±0.5	–
CBLoss	**97.3** ±0.1	97.4 ±0.5	97.8 ±0.6	95.1 ±0.4	96.9 ±0.1	95.1 ±0.4	96.8 ±0.2	97.0 ±1.2	**100.0** ±0.0	–
LDAM	96.9 ±0.1	96.6 ±0.6	97.9 ±0.1	94.7 ±0.2	96.5 ±0.2	94.7 ±0.2	96.6 ±0.1	95.7 ±1.4	96.0 ±0.0	–
BSoftmax	96.0 ±0.5	96.9 ±0.6	98.8 ±0.6	95.9 ±0.1	96.9 ±0.3	95.6 ±0.3	96.6 ±0.4	**98.7** ±0.7	99.3 ±0.5	–
SSP	96.2 ±0.5	96.8 ±0.2	98.9 ±0.1	95.7 ±0.3	96.9 ±0.2	95.4 ±0.4	96.7 ±0.2	98.3 ±0.5	98.0 ±0.9	–
CRT	95.3 ±0.2	96.7 ±0.1	98.5 ±0.1	94.9 ±0.1	96.3 ±0.1	94.9 ±0.1	96.3 ±0.1	97.3 ±0.3	94.0 ±0.9	–
BoDA$_r$	96.9 ±0.4	97.4 ±0.2	98.6 ±0.2	95.1 ±0.4	97.0 ±0.1	95.1 ±0.4	97.0 ±0.1	96.3 ±0.5	98.0 ±0.9	–
BoDA-M$_r$	96.6 ±0.2	**98.0** ±0.2	99.1 ±0.2	94.9 ±0.1	97.1 ±0.1	94.9 ±0.1	97.3 ±0.1	96.3 ±0.5	96.0 ±0.0	–
BoDA$_{r,c}$	96.3 ±0.1	97.4 ±0.5	**99.4** ±0.3	95.7 ±0.3	**97.2** ±0.1	95.7 ±0.3	**97.4** ±0.1	97.0 ±0.0	94.7 ±1.1	–
BoDA-M$_{r,c}$	96.3 ±0.4	97.7 ±0.2	98.1 ±0.4	**96.4** ±0.2	97.1 ±0.2	**96.3** ±0.1	97.1 ±0.0	97.0 ±0.8	96.0 ±0.0	–
BoDA *vs.* ERM	**−0.5**	**+0.7**	**+0.5**	**+0.6**	**+0.1**	**+0.5**	**+0.3**	**+0.0**	**−2.0**	–

F.3 OfficeHome-MLT

Table 15. Complete evaluation results on OfficeHome-MLT.

	Accuracy (by domain)						Accuracy (by shot)			
Algorithm	**A**	**C**	**P**	**R**	**Average**	**Worst**	**Many**	**Medium**	**Few**	**Zero**
ERM	71.3 ±0.1	78.4 ±0.2	89.6 ±0.3	83.3 ±0.2	80.7 ±0.0	71.3 ±0.1	87.8 ±0.2	81.0 ±0.2	63.1 ±0.1	63.3 ±7.2
IRM	70.7 ±0.2	78.5 ±0.8	89.4 ±0.5	83.8 ±0.6	80.6 ±0.4	70.7 ±0.2	87.6 ±0.4	81.5 ±0.4	61.1 ±0.9	56.7 ±1.4
GroupDRO	68.7 ±0.9	79.0 ±0.2	89.4 ±0.4	83.3 ±0.5	80.1 ±0.3	68.7 ±0.9	88.1 ±0.2	80.8 ±0.4	59.8 ±1.2	51.7 ±3.6
Mixup	72.3 ±0.6	79.1 ±0.4	89.7 ±0.1	83.9 ±0.2	81.2 ±0.2	72.3 ±0.6	87.9 ±0.4	81.8 ±0.1	64.1 ±0.4	60.0 ±4.1
MLDG	70.2 ±0.6	78.2 ±0.5	89.4 ±0.4	83.7 ±0.3	80.4 ±0.2	70.2 ±0.6	87.1 ±0.1	81.3 ±0.3	61.3 ±1.0	61.7 ±1.4
CORAL	**72.7** ±0.6	80.9 ±0.3	89.9 ±0.2	84.2 ±0.4	81.9 ±0.1	**72.7** ±0.6	87.9 ±0.1	83.0 ±0.1	63.5 ±0.7	65.0 ±2.4
MMD	67.7 ±0.8	77.8 ±0.2	87.4 ±0.5	80.6 ±0.4	78.4 ±0.4	67.7 ±0.8	85.2 ±0.2	79.4 ±0.7	58.8 ±0.4	56.7 ±3.6
DANN	70.2 ±0.9	77.3 ±0.3	87.3 ±0.5	82.1 ±0.4	79.2 ±0.2	70.2 ±0.9	86.2 ±0.1	80.0 ±0.1	60.3 ±1.1	61.7 ±5.9
CDANN	69.4 ±0.3	77.2 ±0.3	87.7 ±0.2	81.5 ±0.3	79.0 ±0.2	69.4 ±0.3	86.4 ±0.6	79.8 ±0.1	58.9 ±0.8	50.0 ±4.7
MTL	69.8 ±0.6	77.6 ±0.3	87.9 ±0.1	82.4 ±0.3	79.5 ±0.2	69.8 ±0.6	87.3 ±0.3	79.8 ±0.2	61.1 ±0.2	51.7 ±2.7
SagNet	70.5 ±0.5	79.6 ±0.5	89.3 ±0.4	83.9 ±0.1	80.9 ±0.1	70.5 ±0.5	87.8 ±0.4	81.9 ±0.1	61.2 ±0.9	56.7 ±3.6
Fish	71.3 ±0.7	79.1 ±0.1	**90.2** ±0.6	84.7 ±0.4	81.3 ±0.3	71.3 ±0.7	**88.2** ±0.2	81.9 ±0.3	63.2 ±0.8	61.7 ±1.4
Focal	67.6 ±0.4	76.6 ±0.8	87.1 ±0.5	80.2 ±0.3	77.9 ±0.0	67.6 ±0.4	86.5 ±0.3	78.3 ±0.1	57.4 ±0.3	46.7 ±3.6
CBLoss	69.5 ±0.7	78.7 ±0.3	88.9 ±0.4	82.2 ±0.1	79.8 ±0.2	69.5 ±0.7	86.6 ±0.4	80.6 ±0.2	61.1 ±1.4	65.0 ±2.4
LDAM	69.9 ±0.5	78.9 ±0.4	89.4 ±0.3	83.0 ±0.4	80.3 ±0.2	69.9 ±0.5	87.1 ±0.2	81.3 ±0.3	61.1 ±0.2	51.7 ±2.7
BSoftmax	70.9 ±0.5	78.7 ±0.2	89.0 ±0.8	83.0 ±0.3	80.4 ±0.2	70.9 ±0.5	86.7 ±0.5	81.3 ±0.3	62.4 ±1.0	60.0 ±4.1
SSP	71.1 ±0.3	79.6 ±0.8	89.4 ±0.3	84.2 ±0.2	81.1 ±0.3	71.1 ±0.3	87.3 ±0.6	82.3 ±0.3	61.6 ±0.7	63.3 ±1.4
CRT	72.5 ±0.2	79.6 ±0.2	88.9 ±0.1	83.6 ±0.2	81.2 ±0.0	72.5 ±0.2	87.7 ±0.1	81.8 ±0.1	64.0 ±0.1	65.0 ±2.4
BoDA$_r$	71.8 ±0.1	80.3 ±0.3	89.1 ±0.4	84.6 ±0.2	81.5 ±0.1	71.8 ±0.1	87.7 ±0.2	82.3 ±0.1	**64.2** ±0.3	63.3 ±1.4
BoDA-M$_r$	71.6 ±0.2	80.5 ±0.3	89.2 ±0.2	85.7 ±0.4	81.9 ±0.2	71.6 ±0.2	87.3 ±0.3	83.4 ±0.2	62.3 ±0.3	65.0 ±2.4
BoDA$_{r,c}$	72.3 ±0.3	80.8 ±0.2	89.4 ±0.4	**86.3** ±0.3	82.3 ±0.1	72.3 ±0.3	87.1 ±0.2	**83.9** ±0.3	63.2 ±0.2	65.0 ±2.4
BoDA-M$_{r,c}$	72.3 ±0.3	**81.5** ±0.4	89.5 ±0.3	85.8 ±0.2	**82.4** ±0.2	72.3 ±0.3	87.7 ±0.1	**83.9** ±0.6	**64.2** ±0.3	**66.7** ±2.7
BoDA *vs.* ERM	**+1.0**	**+3.1**	**−0.1**	**+3.0**	**+1.7**	**+1.0**	**−0.1**	**+2.9**	**+1.1**	**+3.4**

F.4 TerraInc-MLT

Table 16. Complete evaluation results on TerraInc-MLT.

	Accuracy (by domain)						Accuracy (by shot)			
Algorithm	**L100**	**L38**	**L43**	**L46**	**Average**	**Worst**	**Many**	**Medium**	**Few**	**Zero**
ERM	80.3 ±1.3	71.2 ±0.7	82.2 ±0.3	67.4 ±0.3	75.3 ±0.3	67.4 ±0.3	85.6 ±0.8	69.6 ±3.2	66.1 ±2.4	14.4 ±2.8
IRM	78.2 ±0.9	69.6 ±2.0	81.1 ±0.7	64.3 ±1.3	73.3 ±0.7	64.3 ±1.3	83.5 ±0.6	70.0 ±1.8	58.3 ±3.4	20.1 ±1.4
GroupDRO	68.3 ±1.0	68.8 ±1.3	82.6 ±0.2	68.1 ±0.8	72.0 ±0.4	66.6 ±0.2	84.7 ±1.1	64.6 ±4.7	38.9 ±1.2	13.5 ±1.1
Mixup	75.4 ±1.4	70.2 ±1.3	78.3 ±0.6	60.4 ±1.1	71.1 ±0.7	60.4 ±1.1	83.2 ±0.7	60.0 ±0.6	56.1 ±3.0	12.2 ±2.1
MLDG	82.3 ±0.9	73.5 ±2.0	83.8 ±1.4	66.9 ±0.5	76.6 ±0.2	66.9 ±0.5	86.1 ±0.6	73.8 ±3.9	70.6 ±3.7	18.8 ±2.4
CORAL	81.6 ±1.0	72.0 ±0.6	84.2 ±0.2	67.8 ±0.9	76.4 ±0.5	67.8 ±0.9	86.3 ±0.3	77.5 ±3.1	66.1 ±2.0	11.0 ±1.4
MMD	78.9 ±0.6	68.8 ±1.0	81.9 ±0.9	63.7 ±1.1	73.3 ±0.4	63.7 ±1.1	84.0 ±0.4	67.9 ±2.7	60.6 ±1.6	13.6 ±2.6
DANN	74.1 ±0.8	63.1 ±1.9	75.9 ±0.2	61.5 ±0.9	68.7 ±0.9	61.1 ±1.0	79.6 ±1.2	62.5 ±8.1	48.9 ±2.8	13.3 ±1.1
CDANN	73.0 ±1.3	67.8 ±2.0	75.0 ±0.6	65.2 ±1.1	70.3 ±0.5	63.9 ±1.0	83.5 ±0.8	50.0 ±4.2	43.9 ±4.7	20.4 ±3.1
MTL	79.4 ±0.8	70.8 ±0.6	81.9 ±0.8	67.8 ±1.4	75.0 ±0.7	67.7 ±1.4	85.2 ±0.7	73.8 ±1.6	61.1 ±2.8	12.4 ±4.0
SagNet	79.4 ±1.8	71.2 ±0.7	83.4 ±2.4	66.5 ±2.1	75.1 ±1.6	66.5 ±2.1	85.5 ±0.9	77.1 ±5.0	57.8 ±4.3	13.0 ±3.4
Fish	80.1 ±1.9	70.2 ±0.2	84.4 ±0.9	66.3 ±0.5	75.3 ±0.5	66.3 ±0.5	85.8 ±0.2	73.3 ±3.9	61.1 ±3.0	13.7 ±3.3
Focal	80.9 ±0.7	71.6 ±1.6	84.4 ±1.3	66.1 ±1.7	75.7 ±0.4	65.3 ±1.1	85.7 ±0.3	76.2 ±3.9	68.9 ±3.2	12.6 ±1.9
CBLoss	84.9 ±0.6	78.0 ±1.2	80.7 ±0.3	68.3 ±2.0	78.0 ±0.4	68.3 ±2.0	85.0 ±0.1	89.2 ±1.2	83.9 ±2.5	9.3 ±3.9
LDAM	83.0 ±0.9	70.6 ±0.6	81.3 ±1.1	64.1 ±1.4	74.7 ±0.9	64.1 ±1.4	85.1 ±0.6	70.8 ±3.5	67.8 ±1.2	11.1 ±2.4
BSoftmax	83.5 ±2.1	75.5 ±0.4	82.1 ±0.7	65.6 ±1.3	76.7 ±1.0	65.6 ±1.3	83.4 ±0.8	90.8 ±0.9	78.3 ±3.9	12.6 ±2.4
SSP	82.6 ±1.3	80.7 ±1.8	83.2 ±0.6	67.3 ±0.4	78.5 ±0.7	67.3 ±0.4	85.5 ±1.0	87.8 ±0.9	82.6 ±1.2	13.2 ±2.8
CRT	89.0 ±0.1	81.8 ±0.3	85.8 ±0.3	70.0 ±0.4	81.6 ±0.1	70.0 ±0.4	**89.7** ±0.2	90.4 ±0.3	83.9 ±0.5	12.9 ±0.0
BoDA$_r$	86.7 ±0.7	74.1 ±1.1	85.2 ±0.7	68.5 ±0.3	78.6 ±0.4	68.5 ±0.3	86.4 ±0.1	85.0 ±1.0	80.0 ±0.9	13.7 ±2.1
BoDA-M$_r$	87.8 ±0.9	76.5 ±0.9	82.2 ±0.3	71.3 ±0.4	79.4 ±0.6	71.3 ±0.4	88.4 ±0.3	76.2 ±2.7	88.3 ±1.6	14.4 ±1.4
BoDA$_{r,c}$	88.3 ±0.6	**82.9** ±0.5	**89.3** ±0.9	68.5 ±0.6	82.3 ±0.3	68.5 ±0.6	89.2 ±0.2	**92.5** ±0.9	88.3 ±1.2	21.3 ±0.7
BoDA-M$_{r,c}$	**90.4** ±0.3	81.2 ±0.7	85.8 ±0.4	**74.6** ±0.7	**83.0** ±0.4	**74.6** ±0.7	89.2 ±0.2	91.2 ±0.6	**91.7** ±2.0	**21.7** ±1.4
BoDA *vs.* ERM	**+10.1**	**+11.7**	**+7.1**	**+7.2**	**+7.7**	**+7.2**	**+3.6**	**+22.9**	**+25.6**	**+7.3**

F.5 DomainNet-MLT

Table 17. Complete evaluation results on DomainNet-MLT.

	Accuracy (by domain)								Accuracy (by shot)			
Algorithm	**clip**	**info**	**paint**	**quick**	**real**	**sketch**	**Average**	**Worst**	**Many**	**Medium**	**Few**	**Zero**
ERM	68.6 ±0.1	29.4 ±0.3	57.1 ±0.2	62.8 ±0.3	72.1 ±0.2	61.7 ±0.2	58.6 ±0.2	29.4 ±0.3	66.0 ±0.1	56.1 ±0.1	35.9 ±0.5	27.6 ±0.3
IRM	66.7 ±0.2	27.6 ±0.1	56.0 ±0.2	60.1 ±0.1	72.0 ±0.0	60.2 ±0.2	57.1 ±0.1	27.6 ±0.1	64.7 ±0.1	54.3 ±0.3	33.5 ±0.3	25.8 ±0.3
GroupDRO	60.1 ±0.2	25.9 ±0.2	50.3 ±0.1	63.9 ±0.2	64.9 ±0.2	56.7 ±0.3	53.6 ±0.1	25.9 ±0.2	61.8 ±0.1	49.1 ±0.3	30.7 ±0.7	22.0 ±0.1
Mixup	67.6 ±0.2	28.7 ±0.0	56.4 ±0.2	60.0 ±0.4	72.1 ±0.1	60.9 ±0.1	57.6 ±0.1	28.7 ±0.0	64.9 ±0.2	54.5 ±0.1	35.6 ±0.2	27.3 ±0.3
MLDG	68.0 ±0.2	28.7 ±0.1	57.2 ±0.1	61.6 ±0.2	73.3 ±0.1	61.9 ±0.2	58.5 ±0.0	28.7 ±0.1	66.0 ±0.1	55.7 ±0.1	35.3 ±0.2	26.9 ±0.3
CORAL	69.1 ±0.3	30.1 ±0.4	57.8 ±0.2	63.4 ±0.2	72.8 ±0.2	63.3 ±0.3	59.4 ±0.1	30.1 ±0.4	66.4 ±0.1	57.1 ±0.0	37.7 ±0.6	29.9 ±0.2
MMD	66.1 ±0.1	27.2 ±0.2	55.9 ±0.1	59.3 ±0.2	71.9 ±0.1	60.0 ±0.2	56.7 ±0.0	27.2 ±0.2	64.2 ±0.1	54.0 ±0.0	33.9 ±0.2	25.4 ±0.2
DANN	65.5 ±0.3	26.9 ±0.4	55.2 ±0.1	57.4 ±0.2	70.6 ±0.1	59.0 ±0.2	55.8 ±0.1	26.9 ±0.4	63.0 ±0.1	52.7 ±0.1	34.2 ±0.4	26.8 ±0.4
CDANN	65.9 ±0.1	27.7 ±0.1	55.3 ±0.1	57.6 ±0.2	70.9 ±0.2	58.7 ±0.1	56.0 ±0.1	27.7 ±0.1	63.2 ±0.0	52.7 ±0.2	34.3 ±0.5	27.6 ±0.1
MTL	68.2 ±0.2	29.3 ±0.2	57.3 ±0.1	62.1 ±0.1	72.9 ±0.1	61.8 ±0.2	58.6 ±0.1	29.3 ±0.2	65.9 ±0.1	56.0 ±0.4	35.4 ±0.1	28.2 ±0.3
SagNet	68.5 ±0.1	29.4 ±0.2	57.8 ±0.2	62.1 ±0.2	73.3 ±0.1	62.4 ±0.1	58.9 ±0.0	29.4 ±0.2	66.3 ±0.1	56.4 ±0.0	36.2 ±0.3	27.2 ±0.4
Fish	68.7 ±0.1	29.1 ±0.1	58.4 ±0.1	64.1 ±0.1	73.9 ±0.1	63.7 ±0.1	59.6 ±0.1	29.1 ±0.1	67.1 ±0.1	57.2 ±0.1	36.8 ±0.4	27.8 ±0.3
Focal	67.6 ±0.1	27.5 ±0.1	56.5 ±0.3	62.3 ±0.3	71.7 ±0.3	61.4 ±0.3	57.8 ±0.2	27.5 ±0.1	65.2 ±0.2	55.1 ±0.2	35.8 ±0.1	26.3 ±0.1
CBLoss	68.3 ±0.2	30.1 ±0.1	57.8 ±0.1	60.8 ±0.1	73.3 ±0.2	63.3 ±0.1	58.9 ±0.1	30.1 ±0.1	64.3 ±0.0	61.0 ±0.3	42.5 ±0.4	28.1 ±0.2
LDAM	68.8 ±0.2	29.2 ±0.2	57.1 ±0.1	65.0 ±0.0	72.3 ±0.1	63.1 ±0.1	59.2 ±0.0	29.2 ±0.2	66.6 ±0.0	57.0 ±0.0	37.1 ±0.2	27.8 ±0.3
BSoftmax	68.5 ±0.1	29.9 ±0.1	57.8 ±0.1	60.5 ±0.3	73.4 ±0.1	63.3 ±0.0	58.9 ±0.1	29.9 ±0.1	64.3 ±0.1	60.9 ±0.3	42.4 ±0.6	28.2 ±0.1
SSP	69.7 ±0.1	31.6 ±0.2	58.8 ±0.1	59.7 ±0.3	73.9 ±0.1	64.2 ±0.1	59.7 ±0.0	31.6 ±0.2	64.3 ±0.1	62.6 ±0.1	45.0 ±0.3	30.5 ±0.0
CRT	70.0 ±0.1	31.6 ±0.1	59.2 ±0.2	64.0 ±0.1	73.4 ±0.1	64.4 ±0.1	60.4 ±0.2	31.6 ±0.1	66.8 ±0.0	61.6 ±0.1	45.7 ±0.1	29.7 ±0.1
BoDA$_r$	70.0 ±0.1	32.6 ±0.1	59.1 ±0.1	61.2 ±0.4	73.3 ±0.1	64.1 ±0.1	60.1 ±0.2	32.6 ±0.1	65.7 ±0.2	60.6 ±0.1	42.6 ±0.3	30.5 ±0.2
BoDA-M$_r$	70.6 ±0.1	32.2 ±0.2	57.7 ±0.3	**65.5** ±0.3	70.2 ±0.1	64.5 ±0.1	60.1 ±0.2	32.2 ±0.2	65.9 ±0.2	60.7 ±0.1	42.9 ±0.3	30.0 ±0.1
BoDA$_{r,c}$	**72.0** ±0.2	**33.4** ±0.1	60.7 ±0.2	63.6 ±0.2	**74.6** ±0.1	65.5 ±0.2	**61.7** ±0.1	**33.4** ±0.1	**67.0** ±0.1	62.7 ±0.1	46.0 ±0.2	**32.2** ±0.3
BoDA-M$_{r,c}$	71.8 ±0.1	33.3 ±0.1	**60.8** ±0.1	63.7 ±0.3	**74.6** ±0.1	**65.8** ±0.2	**61.7** ±0.2	33.3 ±0.1	**67.0** ±0.1	**63.0** ±0.3	**46.6** ±0.4	31.8 ±0.2
BoDA *vs.* ERM	**+3.4**	**+4.0**	**+3.7**	**+0.9**	**+2.5**	**+4.1**	**+3.1**	**+4.0**	**+1.0**	**+6.9**	**+10.7**	**+4.6**

G Complete Results for DG

We provide detailed results of Table 9 across five DG benchmarks [19]. Results for all algorithms except `BoDA` are directly copied from [19].

G.1 `VLCS`

Table 18. Complete domain generalization results on `VLCS`.

Algorithm	C	L	S	V	Avg
ERM	97.7 ±0.4	64.3 ±0.9	73.4 ±0.5	74.6 ±1.3	77.5
IRM	98.6 ±0.1	64.9 ±0.9	73.4 ±0.6	77.3 ±0.9	78.5
GroupDRO	97.3 ±0.3	63.4 ±0.9	69.5 ±0.8	76.7 ±0.7	76.7
Mixup	98.3 ±0.6	64.8 ±1.0	72.1 ±0.5	74.3 ±0.8	77.4
MLDG	97.4 ±0.2	65.2 ±0.7	71.0 ±1.4	75.3 ±1.0	77.2
CORAL	98.3 ±0.1	**66.1** ±1.2	73.4 ±0.3	77.5 ±1.2	**78.8**
MMD	97.7 ±0.1	64.0 ±1.1	72.8 ±0.2	75.3 ±3.3	77.5
DANN	**99.0** ±0.3	65.1 ±1.4	73.1 ±0.3	77.2 ±0.6	78.6
CDANN	97.1 ±0.3	65.1 ±1.2	70.7 ±0.8	77.1 ±1.5	77.5
MTL	97.8 ±0.4	64.3 ±0.3	71.5 ±0.7	75.3 ±1.7	77.2
SagNet	97.9 ±0.4	64.5 ±0.5	71.4 ±1.3	77.5 ±0.5	77.8
ARM	98.7 ±0.2	63.6 ±0.7	71.3 ±1.2	76.7 ±0.6	77.6
VREx	98.4 ±0.3	64.4 ±1.4	74.1 ±0.4	76.2 ±1.3	78.3
RSC	97.9 ±0.1	62.5 ±0.7	72.3 ±1.2	75.6 ±0.8	77.1
BoDA	98.1 ±0.3	64.5 ±0.4	**74.3** ±0.3	**78.0** ±0.6	78.5

G.2 `PACS`

Table 19. Complete domain generalization results on `PACS`.

Algorithm	A	C	P	S	Avg
ERM	84.7 ±0.4	80.8 ±0.6	97.2 ±0.3	79.3 ±1.0	85.5
IRM	84.8 ±1.3	76.4 ±1.1	96.7 ±0.6	76.1 ±1.0	83.5
GroupDRO	83.5 ±0.9	79.1 ±0.6	96.7 ±0.3	78.3 ±2.0	84.4
Mixup	86.1 ±0.5	78.9 ±0.8	97.6 ±0.1	75.8 ±1.8	84.6
MLDG	85.5 ±1.4	80.1 ±1.7	97.4 ±0.3	76.6 ±1.1	84.9
CORAL	**88.3** ±0.2	80.0 ±0.5	97.5 ±0.3	78.8 ±1.3	86.2
MMD	86.1 ±1.4	79.4 ±0.9	96.6 ±0.2	76.5 ±0.5	84.6
DANN	86.4 ±0.8	77.4 ±0.8	97.3 ±0.4	73.5 ±2.3	83.7
CDANN	84.6 ±1.8	75.5 ±0.9	96.8 ±0.3	73.5 ±0.6	82.6
MTL	87.5 ±0.8	77.1 ±0.5	96.4 ±0.8	77.3 ±1.8	84.6
SagNet	87.4 ±1.0	80.7 ±0.6	97.1 ±0.1	80.0 ±0.4	86.3
ARM	86.8 ±0.6	76.8 ±0.5	97.4 ±0.3	79.3 ±1.2	85.1
VREx	86.0 ±1.6	79.1 ±0.6	96.9 ±0.5	77.7 ±1.7	84.9
RSC	85.4 ±0.8	79.7 ±1.8	97.6 ±0.3	78.2 ±1.2	85.2
BoDA	88.2 ±0.2	**81.7** ±0.3	**97.8** ±0.2	**80.2** ±0.3	**86.9**

G.3 OfficeHome

Table 20. Complete domain generalization results on OfficeHome.

Algorithm	A	C	P	R	Avg
ERM	61.3 ±0.7	52.4 ±0.3	75.8 ±0.1	76.6 ±0.3	66.5
IRM	58.9 ±2.3	52.2 ±1.6	72.1 ±2.9	74.0 ±2.5	64.3
GroupDRO	60.4 ±0.7	52.7 ±1.0	75.0 ±0.7	76.0 ±0.7	66.0
Mixup	62.4 ±0.8	54.8 ±0.6	76.9 ±0.3	78.3 ±0.2	68.1
MLDG	61.5 ±0.9	53.2 ±0.6	75.0 ±1.2	77.5 ±0.4	66.8
CORAL	65.3 ±0.4	54.4 ±0.5	76.5 ±0.1	78.4 ±0.5	68.7
MMD	60.4 ±0.2	53.3 ±0.3	74.3 ±0.1	77.4 ±0.6	66.3
DANN	59.9 ±1.3	53.0 ±0.3	73.6 ±0.7	76.9 ±0.5	65.9
CDANN	61.5 ±1.4	50.4 ±2.4	74.4 ±0.9	76.6 ±0.8	65.8
MTL	61.5 ±0.7	52.4 ±0.6	74.9 ±0.4	76.8 ±0.4	66.4
SagNet	63.4 ±0.2	54.8 ±0.4	75.8 ±0.4	78.3 ±0.3	68.1
ARM	58.9 ±0.8	51.0 ±0.5	74.1 ±0.1	75.2 ±0.3	64.8
VREx	60.7 ±0.9	53.0 ±0.9	75.3 ±0.1	76.6 ±0.5	66.4
RSC	60.7 ±1.4	51.4 ±0.3	74.8 ±1.1	75.1 ±1.3	65.5
BoDA	**65.4** ±0.1	**55.4** ±0.3	**77.1** ±0.1	**79.5** ±0.3	**69.3**

G.4 TerraInc

Table 21. Complete domain generalization results on TerraInc.

Algorithm	L100	L38	L43	L46	Avg
ERM	49.8 ±4.4	42.1 ±1.4	56.9 ±1.8	35.7 ±3.9	46.1
IRM	54.6 ±1.3	39.8 ±1.9	56.2 ±1.8	39.6 ±0.8	47.6
GroupDRO	41.2 ±0.7	38.6 ±2.1	56.7 ±0.9	36.4 ±2.1	43.2
Mixup	**59.6** ±2.0	42.2 ±1.4	55.9 ±0.8	33.9 ±1.4	47.9
MLDG	54.2 ±3.0	44.3 ±1.1	55.6 ±0.3	36.9 ±2.2	47.7
CORAL	51.6 ±2.4	42.2 ±1.0	57.0 ±1.0	39.8 ±2.9	47.6
MMD	41.9 ±3.0	34.8 ±1.0	57.0 ±1.9	35.2 ±1.8	42.2
DANN	51.1 ±3.5	40.6 ±0.6	57.4 ±0.5	37.7 ±1.8	46.7
CDANN	47.0 ±1.9	41.3 ±4.8	54.9 ±1.7	39.8 ±2.3	45.8
MTL	49.3 ±1.2	39.6 ±6.3	55.6 ±1.1	37.8 ±0.8	45.6
SagNet	53.0 ±2.9	43.0 ±2.5	57.9 ±0.6	40.4 ±1.3	48.6
ARM	49.3 ±0.7	38.3 ±2.4	55.8 ±0.8	38.7 ±1.3	45.5
VREx	48.2 ±4.3	41.7 ±1.3	56.8 ±0.8	38.7 ±3.1	46.4
RSC	50.2 ±2.2	39.2 ±1.4	56.3 ±1.4	40.8 ±0.6	46.6
BoDA	54.0 ±0.3	**46.5** ±0.2	**59.5** ±0.3	**41.0** ±0.4	**50.2**

G.5 DomainNet

Table 22. Complete domain generalization results on DomainNet.

Algorithm	clip	info	paint	quick	real	sketch	Avg
ERM	58.1 ±0.3	18.8 ±0.3	46.7 ±0.3	12.2 ±0.4	59.6 ±0.1	49.8 ±0.4	40.9
IRM	48.5 ±2.8	15.0 ±1.5	38.3 ±4.3	10.9 ±0.5	48.2 ±5.2	42.3 ±3.1	33.9
GroupDRO	47.2 ±0.5	17.5 ±0.4	33.8 ±0.5	9.3 ±0.3	51.6 ±0.4	40.1 ±0.6	33.3
Mixup	55.7 ±0.3	18.5 ±0.5	44.3 ±0.5	12.5 ±0.4	55.8 ±0.3	48.2 ±0.5	39.2
MLDG	59.1 ±0.2	19.1 ±0.3	45.8 ±0.7	13.4 ±0.3	59.6 ±0.2	50.2 ±0.4	41.2
CORAL	59.2 ±0.1	19.7 ±0.2	46.6 ±0.3	13.4 ±0.4	59.8 ±0.2	50.1 ±0.6	41.5
MMD	32.1 ±13.3	11.0 ±4.6	26.8 ±11.3	8.7 ±2.1	32.7 ±13.8	28.9 ±11.9	23.4
DANN	53.1 ±0.2	18.3 ±0.1	44.2 ±0.7	11.8 ±0.1	55.5 ±0.4	46.8 ±0.6	38.3
CDANN	54.6 ±0.4	17.3 ±0.1	43.7 ±0.9	12.1 ±0.7	56.2 ±0.4	45.9 ±0.5	38.3
MTL	57.9 ±0.5	18.5 ±0.4	46.0 ±0.1	12.5 ±0.1	59.5 ±0.3	49.2 ±0.1	40.6
SagNet	57.7 ±0.3	19.0 ±0.2	45.3 ±0.3	12.7 ±0.5	58.1 ±0.5	48.8 ±0.2	40.3
ARM	49.7 ±0.3	16.3 ±0.5	40.9 ±1.1	9.4 ±0.1	53.4 ±0.4	43.5 ±0.4	35.5
VREx	47.3 ±3.5	16.0 ±1.5	35.8 ±4.6	10.9 ±0.3	49.6 ±4.9	42.0 ±3.0	33.6
RSC	55.0 ±1.2	18.3 ±0.5	44.4 ±0.6	12.2 ±0.2	55.7 ±0.7	47.8 ±0.9	38.9
BoDA	**62.1** ±0.4	**20.5** ±0.7	**48.0** ±0.1	**13.8** ±0.6	**60.6** ±0.4	**51.4** ±0.3	**42.7**

G.6 Averages

Table 23. Complete domain generalization results over all DG benchmarks.

Algorithm	VLCS	PACS	OfficeHome	TerraInc	DomainNet	Avg
ERM	77.5 ±0.4	85.5 ±0.2	66.5 ±0.3	46.1 ±1.8	40.9 ±0.1	63.3
IRM	78.5 ±0.5	83.5 ±0.8	64.3 ±2.2	47.6 ±0.8	33.9 ±2.8	61.6
GroupDRO	76.7 ±0.6	84.4 ±0.8	66.0 ±0.7	43.2 ±1.1	33.3 ±0.2	60.7
Mixup	77.4 ±0.6	84.6 ±0.6	68.1 ±0.3	47.9 ±0.8	39.2 ±0.1	63.4
MLDG	77.2 ±0.4	84.9 ±1.0	66.8 ±0.6	47.7 ±0.9	41.2 ±0.1	63.6
CORAL	**78.8** ±0.6	86.2 ±0.3	68.7 ±0.3	47.6 ±1.0	41.5 ±0.1	64.5
MMD	77.5 ±0.9	84.6 ±0.5	66.3 ±0.1	42.2 ±1.6	23.4 ±9.5	58.8
DANN	78.6 ±0.4	83.6 ±0.4	65.9 ±0.6	46.7 ±0.5	38.3 ±0.1	62.6
CDANN	77.5 ±0.1	82.6 ±0.9	65.8 ±1.3	45.8 ±1.6	38.3 ±0.3	62.0
MTL	77.2 ±0.4	84.6 ±0.5	66.4 ±0.5	45.6 ±1.2	40.6 ±0.1	62.9
SagNet	77.8 ±0.5	86.3 ±0.2	68.1 ±0.1	48.6 ±1.0	40.3 ±0.1	64.2
ARM	77.6 ±0.3	85.1 ±0.4	64.8 ±0.3	45.5 ±0.3	35.5 ±0.2	61.7
VREx	78.3 ±0.2	84.9 ±0.6	66.4 ±0.6	46.4 ±0.6	33.6 ±2.9	61.9
RSC	77.1 ±0.5	85.2 ±0.9	65.5 ±0.9	46.6 ±1.0	38.9 ±0.5	62.7
BoDA	78.5 ±0.3	**86.9** ±0.4	**69.3** ±0.1	**50.2** ±0.4	**42.7** ±0.1	**65.5**

H Additional Analysis and Studies

H.1 Ablation Studies for BoDA

Effect of Balanced Distance. We study the effect of adding balanced distance in BoDA compared to the vanilla DA loss. As Table 24 demonstrates, incorporating balanced distance in BoDA is essential for addressing MDLT: we observe that BoDA improves over DA by a large margin, resulting in an averaged improvements of 2.3% over all MDLT benchmarks. The improvements are especially large on datasets with severe data imbalance across domains (e.g., TerraInc-MLT).

Table 24. Ablation study on effect of adding balanced distance in BoDA.

	VLCS-MLT	PACS-MLT	OfficeHome-MLT	TerraInc-MLT	DomainNet-MLT	**Avg**
DA	76.6 ±0.4	96.8 ±0.2	80.7 ±0.3	76.4 ±0.5	58.9 ±0.2	77.9
BoDA	**77.3** ±0.2	**97.2** ±0.1	**82.3** ±0.1	**82.3** ±0.3	**61.7** ±0.1	**80.2**
Gains	**+0.7**	**+0.4**	**+1.6**	**+5.9**	**+2.8**	**+2.3**

Effect of Different Distance Calibration Coefficient $\lambda_{d,c}^{d',c'}$. We further investigate the effect of different distance calibration coefficients in BoDA. Recall that $\lambda_{d,c}^{d',c'} = (N_{d',c'}/N_{d,c})^{\nu}$ indicates how much we would like to transfer (d, c) to (d', c'), based on their relative sample sizes. We vary the value of ν, and study its effect on BoDA performance across all MDLT datasets. Table 25 reveals several interesting findings. First, when $\nu = 0$ (i.e., no calibration is used as the coefficient is always equal to 1), BoDA performance is lower than those with a positive ν, confirming the effectiveness of the calibrated distance. Moreover, when we vary ν between $0.5 - 1.5$, the overall performance gains are similar across different choices, where ν around 0.9 seems to achieve the best results. Finally, when compared to ERM, we demonstrate that BoDA consistently obtains notable gains across different ν.

Table 25. Ablation study on effect of distance calibration coefficient $\lambda_{d,c}^{d',c'}$ in BoDA. We vary the value of ν and report the averaged results over all five MDLT datasets.

ν	0	0.5	0.7	0.9	1	1.1	1.2	1.5	ERM
BoDA	78.9	80.1	80.0	80.2	80.1	79.8	79.6	79.2	77.6

H.2 Absolute Accuracy Gains on All MDLT Benchmarks

We provide additional results for understanding how BoDA performs across *all* domain-class pair when cross-domain imbalance occurs. Similar to Fig. 7 in the main text, we plot the absolute gains of BoDA over ERM on all five MDLT

datasets, shown in Figs. 9, 10, 11, 12, and 13. Across all datasets, we observe that BoDA establishes large improvements w.r.t. all regions, especially for the few-shot and zero-shot ones.

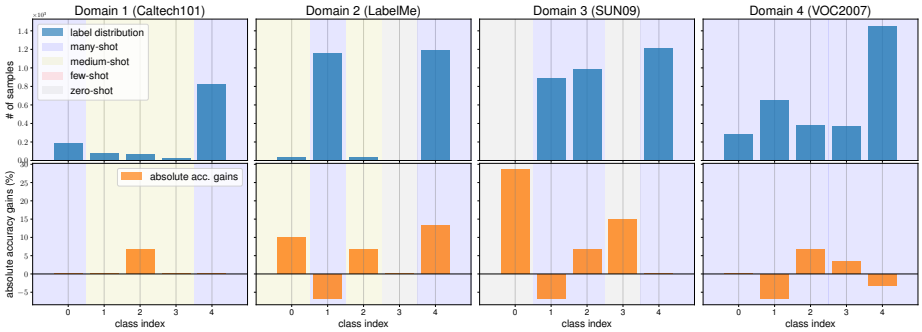

Fig. 9. The absolute accuracy gains of BoDA *vs.* ERM over all domain-class pairs on VLCS-MLT.

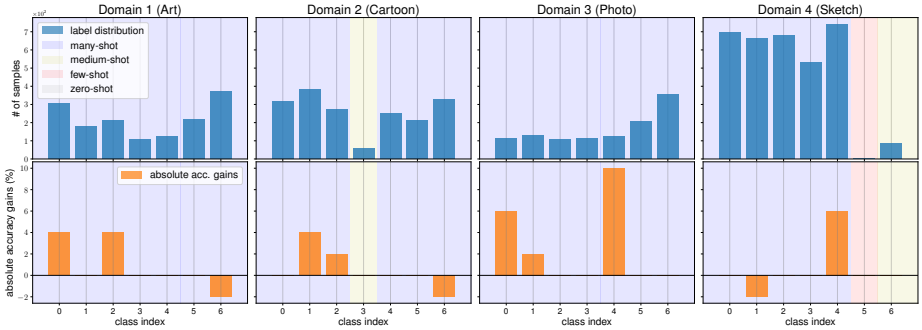

Fig. 10. The absolute accuracy gains of BoDA *vs.* ERM over all domain-class pairs on PACS-MLT.

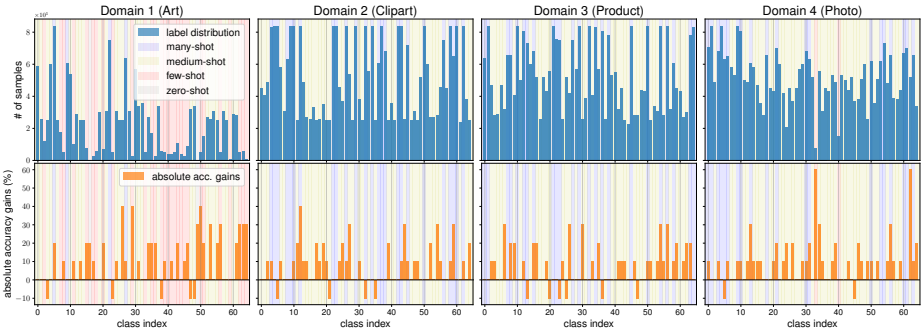

Fig. 11. The absolute accuracy gains of BoDA *vs.* ERM over all domain-class pairs on OfficeHome-MLT.

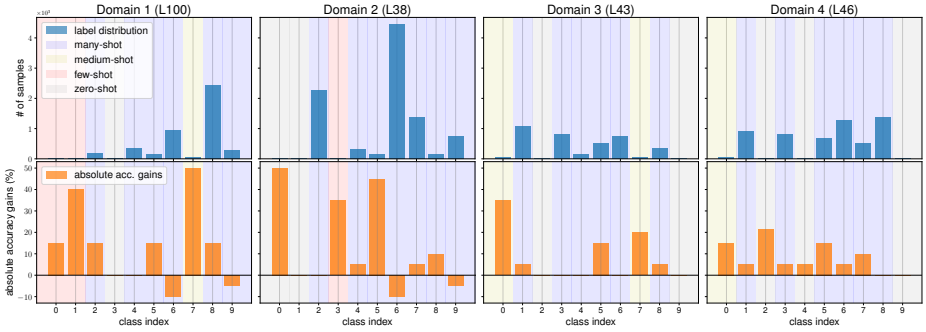

Fig. 12. The absolute accuracy gains of `BoDA` *vs.* ERM over all domain-class pairs on `TerraInc-MLT`.

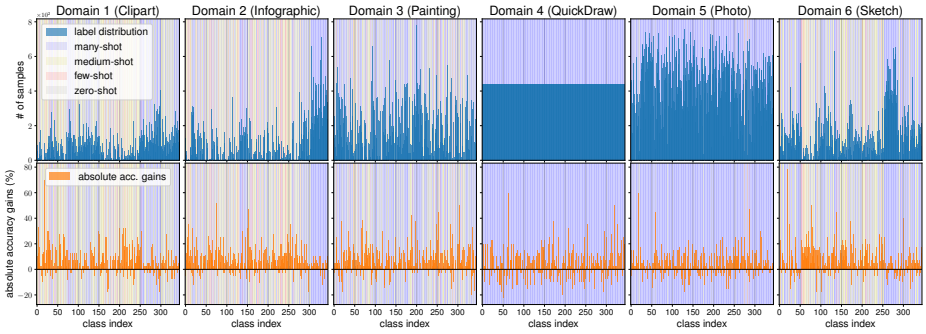

Fig. 13. The absolute accuracy gains of `BoDA` *vs.* ERM over all domain-class pairs on `DomainNet-MLT`.

H.3 Robustness to Diverse Skewed Label Distributions

We investigate how `BoDA` performs under arbitrary label imbalance across domains, especially when the cross-domain label distributions are both *imbalance* and *divergent*. We again employ the `Digits-MLT` dataset, and manually vary the label proportions for each domain.

As Fig. 14 demonstrates, when the label distributions for two domains are balanced and identical, both ERM and `BoDA` maintains discriminative representations. If the label distributions become imbalanced but still identical across domains, ERM is still able to align similar classes in the two domains, but with majority classes being closer in terms of transferability than minority classes. In contrast, `BoDA` maintains consistent transferability regardless of number of samples within each class. Finally, as the label distributions become further mismatched across domains, ERM is not able to align the domains and produces a clear gap; by contrast, `BoDA` maintains consistent and transferable representations even under severe data imbalance. As a result, `BoDA` substantially boosts the performance upon ERM, with an average gains of 6.4% across all label configurations.

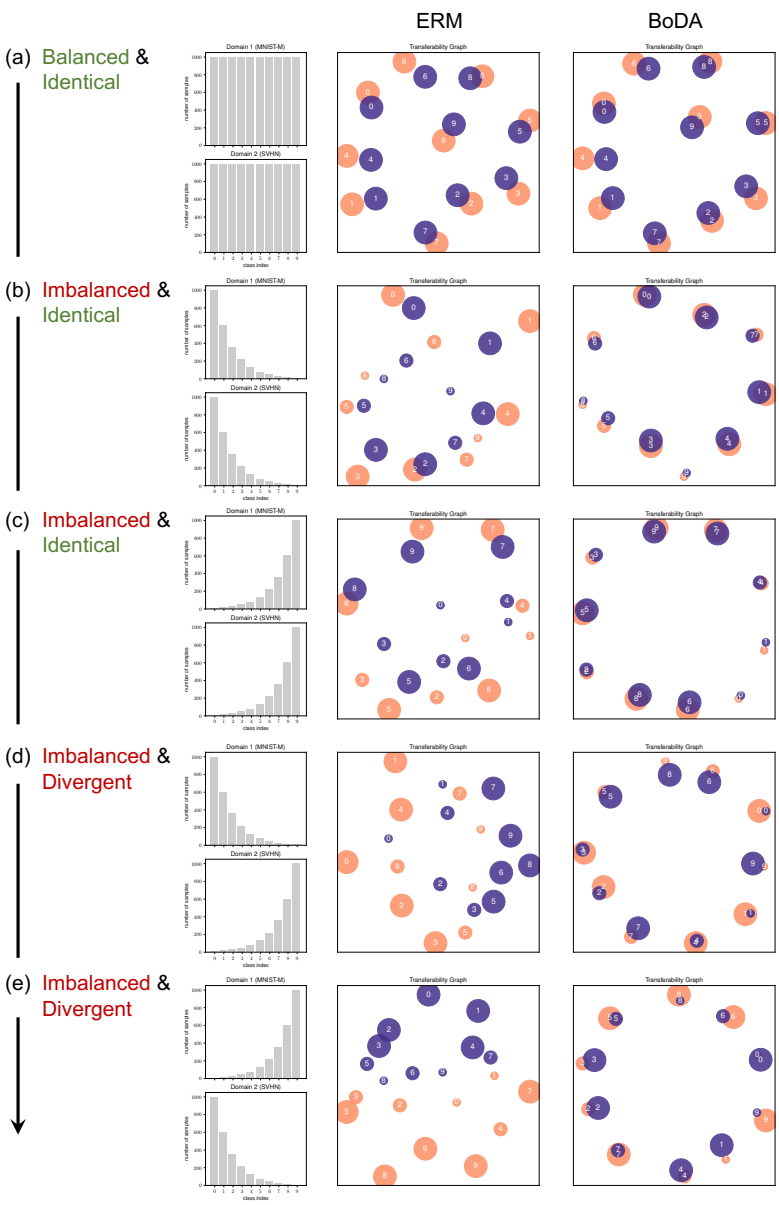

Fig. 14. The evolving patterns of the transferability graph of `BoDA` *vs.* ERM across different label configurations on `Digits-MLT`. Label distributions for two domains are **(a)** balanced and identical; **(b)(c)** imbalanced and identical; **(d)(e)** imbalanced and divergent. `BoDA` maintains consistent and transferable representations across all label configurations, and leads to much better test accuracy.

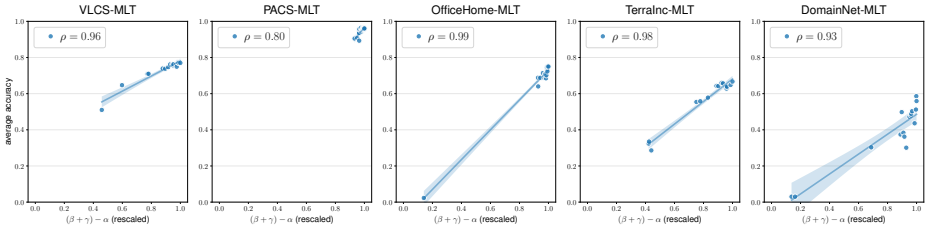

Fig. 15. Correspondence between $(\beta+\gamma)-\alpha$ quantity and test accuracy across different MDLT datasets. Each point within each plot corresponds to a model trained with ERM using different hyperparameters.

H.4 Transferability *vs.* Generalization on More Datasets

We provide further results on transferability statistics *vs.* generalization on real MDLT datasets, in addition to results on `Digits-MLT` as we showed in the main text.

Specifically, on all five MDLT datasets, we train 20 ERM models with varying hyperparameters, calculate the (α, β, γ) statistics for each model, and plot its classification accuracy against $(\beta+\gamma)-\alpha$. Fig. 15 reveals similar and consistent findings, that the (α, β, γ) statistics characterize model performance in MDLT. Across all datasets, the $(\beta+\gamma)-\alpha$ quantity displays a very strong correlation with test performance across the entire range, suggesting that the (α, β, γ) statistics govern the success of learning in MDLT.

H.5 Additional Visualization of Feature Discrepancy

We provide additional results for understanding `BoDA`, i.e., how `BoDA` calibrates the feature statistics. Fig. 16 shows the feature discrepancy of `BoDA` *vs.* ERM across different label configurations on `Digits-MLT`. In addition to the mean distance we showed in the main text, we show also the feature covariance distance between training and test data, and plot them for both domains. Similarly, solid lines plot the distance between training and test data from the same domain-class pairs. Dashed lines plot the distance between test data from a particular domain-class pair and the training data with which it shares the same class but differs in the domain. The figure also shows regions with different data densities using colors blue, yellow, red.

As the figure confirms, across different label distributions, `BoDA` consistently learns better representations especially for the tail data (i.e., the red regions), where the feature mean/covariance distance between training and test data becomes smaller and more aligned across domains. Comparing `BoDA` with ERM further demonstrates that `BoDA` maintains consistent and transferable representations with smaller feature discrepancy.

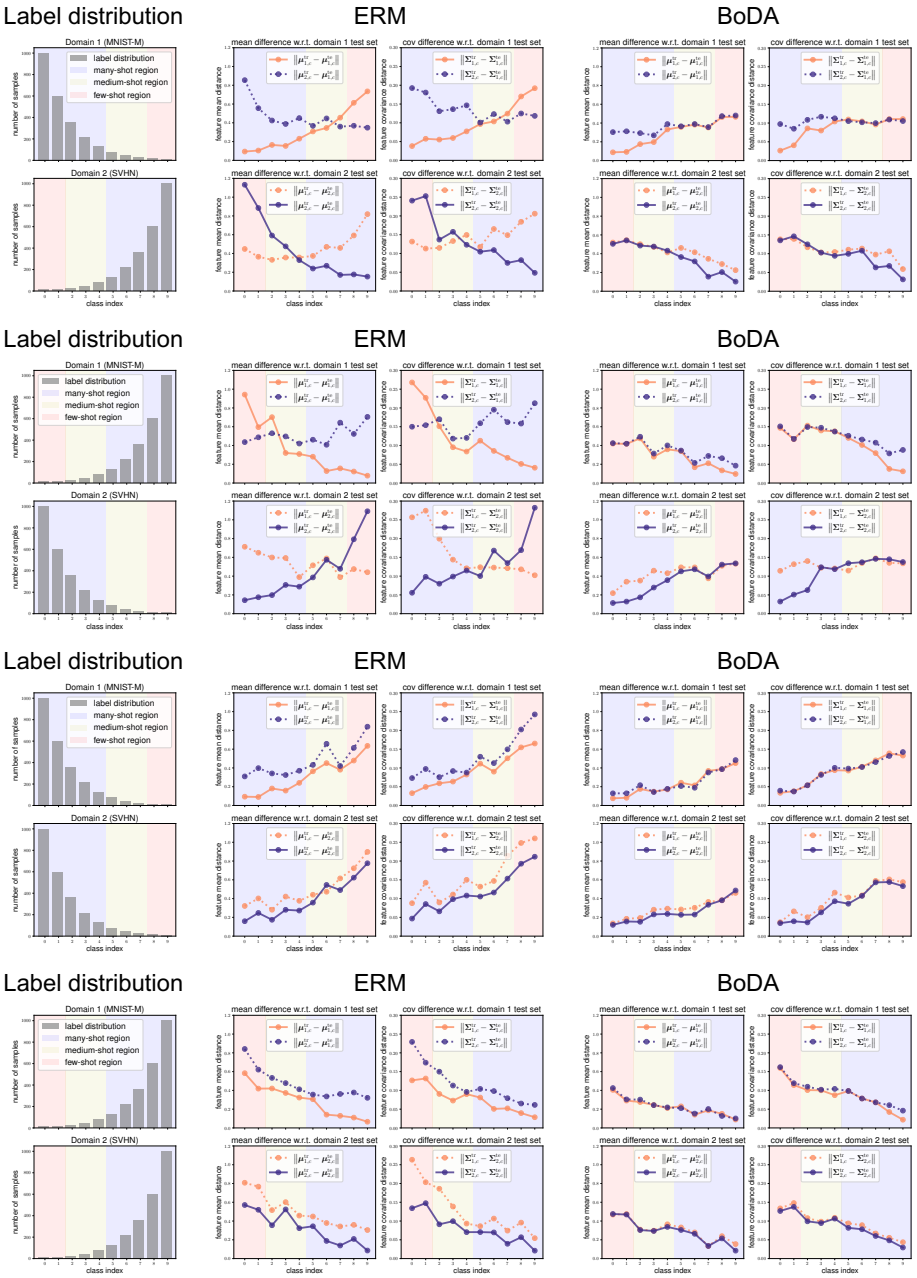

Fig. 16. Feature discrepancy of `BoDA` *vs.* ERM across different label configurations on `Digits-MLT`. Each row plots a per-domain label distribution, and the feature mean / covariance distance between training and test data on each domain for both ERM and `BoDA`. `BoDA` enables better learned tail (d, c) with smaller feature discrepancy.