On Multi-Domain Long-Tailed Recognition, Generalization and Beyond

Yuzhe Yang[1] Hao Wang[2] Dina Katabi[1]
[1]MIT CSAIL [2]Rutgers University

Abstract

Real-world data often exhibit imbalanced label distributions. Existing studies on data imbalance focus on single-domain settings, i.e., samples are from the same data distribution. However, natural data can originate from distinct domains, where a minority class in one domain could have abundant instances from other domains. We define Multi-Domain Long-Tailed Recognition (MDLT), which learns from multi-domain imbalanced data, addresses label imbalance*,* domain shift*, and* divergent label distributions across domains*, and generalizes to all domain-class pairs. We first develop the* domain-class transferability graph*, and show that such transferability governs the success of learning in MDLT. We then propose* BoDA*, a theoretically grounded learning strategy that tracks the upper bound of transferability statistics, and ensures* balanced *alignment and calibration across imbalanced domain-class distributions. We curate five MDLT benchmarks, and compare* BoDA *to twenty algorithms that span different learning strategies. Extensive and rigorous experiments verify the superior performance of* BoDA*. Further, as a byproduct,* BoDA *establishes new state-of-the-art on Domain Generalization benchmarks, improving generalization to unseen domains.*

1. Introduction

Real-world data often exhibit label imbalance – i.e., instead of a uniform label distribution over classes, in reality, data are by their nature imbalanced [5,6]. This phenomenon poses a challenge for deep recognition models, and has motivated several prior solutions [6, 10, 33, 39, 52, 53]. Such prior solutions focus on *single domain* scenarios, i.e., samples are from the same data distribution.

In contrast, this paper formulates the problem of *Multi-Domain Long-Tailed Recognition* (MDLT) as learning from multi-domain imbalanced data, with each domain having its own imbalanced label distribution, and generalizing to a test set that is balanced over all domain-class pairs. MDLT is a natural extension of the single domain case. It arises in real-world scenarios, where data targeted for one task can originate from different domains. For example, in visual recognition problems, minority classes from "photo" images could be complemented with potentially abundant samples from "sketch" images. Similarly, in autonomous driving, the minority accident class in "real" life could be enriched with accidents generated in "simulation". In the above examples, different data types act as distinct *domains*, and such multi-domain data could be leveraged to tackle the inherent data imbalance within each domain.

We note that MDLT has key differences from its single-domain counterpart:

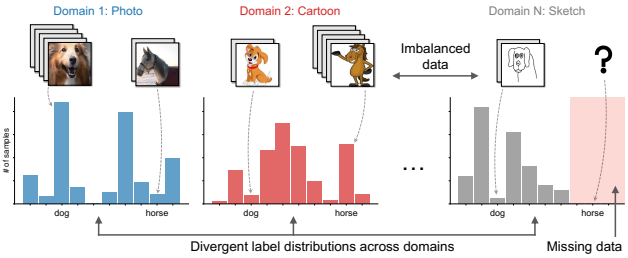

Figure 1. Multi-Domain Long-Tailed Recognition (MDLT) aims to learn from imbalanced data from multiple distinct domains, tackle label imbalance, domain shift, and divergent label distributions across domains, and generalize to all domain-class pairs.

1. The label distribution for each domain is likely different from other domains. For example, in Fig. 1, both "Photo" and "Cartoon" domains exhibit imbalanced label distributions; Yet, the "horse" class in "Cartoon" has many more samples than in "Photo". This creates challenges with *divergent label distributions across domains*.
2. Multi-domain data inherently involves *domain shift*. Simply treating different domains as a whole and applying traditional imbalanced methods is unlikely to yield the best results, as the domain gap can be arbitrarily large.
3. MDLT naturally motivates *zero-shot generalization within and across domains* – i.e., to generalize to both in-domain missing classes (Fig. 1 right part), as well as new domains with no training data, where the latter case is typically denoted as Domain Generalization (DG).

To deal with these challenges, we develop *domain-class transferability graph*, which quantifies the transferability between different domain-class pairs under data imbalance. We show that the transferability graph dictates the performance of imbalanced learning across domains. Inspired by this, we design BoDA, a new loss function that connects disparate and imbalanced domain-class distributions in a balanced manner. Analytically, we prove that minimizing the BoDA loss optimizes a tight upper bound of the balanced transferability statistics, which corroborates the effectiveness of BoDA for learning multi-domain imbalanced data.

For MDLT evaluation, we curate five MDLT benchmarks based on datasets widely used for domain generalization (DG). These datasets naturally exhibit heavy class imbalance within each domain and data shift across domains, highlighting that the MDLT problem is widely present in current benchmarks. We compare BoDA against twenty algorithms. Extensive experiments verify that BoDA consistently outperforms all these baselines on all datasets.

Additionally, we examine how BoDA performs in DG. We show that combining BoDA with the DG state-of-the-

art (SOTA) yielding a new SOTA for DG. These results shed light on how label imbalance can affect out-of-distribution generalization and highlight the importance of integrating label imbalance into practical DG algorithm design.

Our contributions are as follows:

- We define the MDLT task as learning from multi-domain imbalanced data & generalizing to all domain-class pairs.
- We introduce the domain-class transferability graph, a unified model for investigating MDLT. We further show that the transferability statistics induced from such graph are crucial and govern the success of MDLT algorithms.
- We design BoDA, a simple, effective, and interpretable loss function for MDLT. We prove theoretically that minimizing the BoDA loss is equivalent to optimizing an upper bound of balanced transferability statistics.
- Through extensive experiments on benchmark datasets, we verify the superior and consistent performance of BoDA. Further, we combine BoDA with DG algorithms establishing a new SOTA for domain generalization.

2. Domain-Class Transferability Graph

When learning from MDLT, a natural question arises: How do we model MDLT in the presence of both *domain shift* and *class imbalance within and across domains*? We argue that in contrast to single-domain imbalanced learning where the basic unit one cares about is a *class* (i.e., minority *vs.* majority classes), in MDLT, the basic unit naturally translates to a **domain-class pair**.

Problem Setup. Given a multi-domain classification task with a discrete label space $\mathcal{C} = \{1, \dots, C\}$ and a domain space $\mathcal{D} = \{1, \dots, D\}$, let $\mathcal{S} = \{(\mathbf{x}_i, c_i, d_i)\}_{i=1}^N$ be the training set, where $\mathbf{x}_i \in \mathbb{R}^l$ denotes the input, $c_i \in \mathcal{C}$ is the class label, and $d_i \in \mathcal{D}$ is the domain label. We denote as $\mathbf{z} = f(\mathbf{x}; \theta)$ the representation of $\mathbf{x}$. The final prediction $\widehat{c} = g(\mathbf{z})$ is given by function $g : \mathcal{Z} \to \mathcal{C}$. We denote the set of samples belonging to domain d and class c (i.e., the domain-class pair (d, c)) as $\mathcal{S}_{d,c} \subseteq \mathcal{S}$, with $N_{d,c} \triangleq |\mathcal{S}_{d,c}|$ as the number of samples. We use $\mathcal{M} = \mathcal{D} \times \mathcal{C} := \{(d, c) : d \in \mathcal{D}, c \in \mathcal{C}\}$ to denote the set of all domain-class pairs.

Definition 1 (Transferability)**.** *Given a learned model and a distance function* $\mathsf{d} : \mathbb{R}^h \times \mathbb{R}^h \to \mathbb{R}$ *in feature space, the transferability from domain-class pair* (d, c) *to* (d', c') *is:*

$$\text{trans}\big((d, c), (d', c')\big) \triangleq \mathbb{E}_{\mathbf{z} \in \mathcal{Z}_{d,c}} \big[\mathsf{d}\left(\mathbf{z}, \boldsymbol{\mu}_{d',c'}\right)\big],$$

where $\boldsymbol{\mu}_{d',c'} \triangleq \mathbb{E}_{\mathbf{z}' \in \mathcal{Z}_{d',c'}}[\mathbf{z}']$ *is the first order statistics (i.e., mean) of* (d', c').

Intuitively, the transferability between two domain-class pairs is the average distance between their learned representations, characterizing how close they are in the feature space. By default, d is chosen as the Euclidean distance, but it can also represent the higher order statistics of (d, c). For example, the Mahalanobis distance [11] uses the covariance $\boldsymbol{\Sigma}_{d,c} \triangleq \mathbb{E}_{\mathbf{z} \in \mathcal{Z}_{d,c}} \left[(\mathbf{z} - \boldsymbol{\mu}_{d,c})(\mathbf{z} - \boldsymbol{\mu}_{d,c})^\top\right]$. In the remaining paper, with a slight abuse of the notation, we allow $\boldsymbol{\mu}_{d,c}$ to represent both the first and higher order statistics for (d, c).

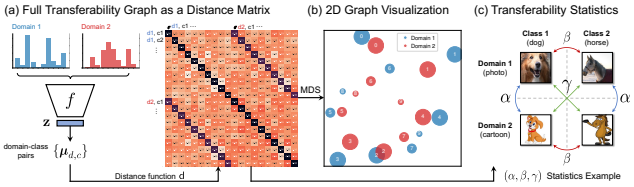

Figure 2. Overall framework of transferability graph. **(a)** Distribution statistics $\{\boldsymbol{\mu}_{d,c}\}$ is computed for all domain-class pairs, by which we generate a full transferability matrix. **(b)** MDS is used to project the graph into a 2D space for visualization. **(c)** We define (α, β, γ) statistics to further describe the whole graph.

Definition 2 (Transferability Graph)**.** *The transferability graph for a learned model is defined as* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$*, where the vertices,* $\mathcal{V} \subseteq \{\boldsymbol{\mu}_{d,c}\}$*, represents the domain-class pairs, and the edges,* $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$*, are assigned weights equal to* $\text{trans}\left((d, c), (d', c')\right)$.

Transferability Graph Visualization. It is convenient to visualize the transferability graph of a learned model in a 2D Cartesian space. To do so, we use the average of $\text{trans}\left((d, c), (d', c')\right)$ and $\text{trans}\left((d', c'), (d, c)\right)$ as a similarity measure, and visualize this similarity and the underlying transferability graph using multidimensional scaling (MDS) [8]. Figs. 2a and 2b show this process, where for each (d, c) pair, we estimate its distribution statistics $\{\boldsymbol{\mu}_{d,c}\}$ from the learned model, then compute the model transferability graph as a distance matrix. We then use MDS to project it into a 2D space, where each dot refers to one (d, c), and the distance represents transferability.

Definition 3 ((α, β, γ) Transferability Statistics)**.** *The transferability graph can be summarized by the following transferability statistics:*

$$\textit{Diff-domain-same-class: } \alpha = \mathbb{E}_c \mathbb{E}_d \mathbb{E}_{d' \neq d} \left[\text{trans}\big((d, c), (d', c)\big)\right].$$
$$\textit{Same-domain-diff-class: } \beta = \mathbb{E}_d \mathbb{E}_c \mathbb{E}_{c' \neq c} \left[\text{trans}\big((d, c), (d, c')\big)\right].$$
$$\textit{Diff-domain-diff-class: } \gamma = \mathbb{E}_d \mathbb{E}_{d' \neq d} \mathbb{E}_c \mathbb{E}_{c' \neq c} \left[\text{trans}\big((d, c), (d', c')\big)\right].$$

As illustrated in Fig. 2c, (α, β, γ) captures the similarity between features of the same class across domains and different classes within and across domains.

3. What Yields Good MDLT Representations?

3.1. Label Divergence Hampers Transferability

Motivating Example. We construct Digits-MLT, a two-domains toy MDLT dataset that combines two digit datasets: MNIST-M [15] and SVHN [36]. Details of the datasets are in Appendix F. We manually vary the number of samples for each domain-class pair to simulate different label distributions, and train a plain ResNet-18 [21] using

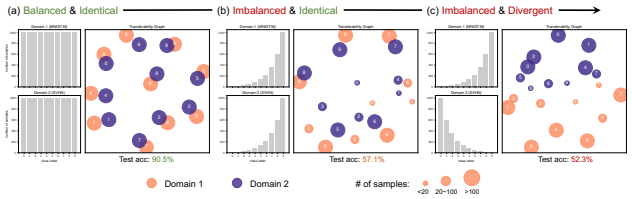

Figure 3. The evolving pattern of transferability graph when varying label proportions of `Digits-MLT`. Label distribution for two domains are **(a)** balanced and identical; **(b)** imbalanced but identical; **(c)** imbalanced and *divergent*.

empirical risk minimization (ERM) for each case. We keep all test sets balanced and identical.

Fig. 3 reveals interesting findings. When the per-domain label distributions are balanced and *identical*, although a domain gap exists, it does not prohibit the model from learning discriminative features of high accuracy (Fig. 3a, 90.5%). If label distributions are imbalanced but *identical* (Fig. 3b), ERM is still able to align similar classes in the two domains, where majority classes (class 9) are closer in terms of transferability than minority classes (class 0). In contrast, when labels are both imbalanced and *mismatched* across domains (Fig. 3c), the learned features are no longer transferable, resulting in a clear gap across domains and worst accuracy. This is because *divergent label distributions* across domains produce an undesirable shortcut; the model can minimize classification loss simply by separating the two domains.

Transferable Features are Desirable. As the results indicate, *transferable* features across (d, c) pairs are needed, especially when imbalance occurs. In particular, the transferability link between the same class across domains should be greater than that between different classes within or across domains. This can be captured via the (α, β, γ) transferability statistics, as we show next.

3.2. Transferability Characterizes Generalization

Motivating Example. Again, we use `Digits-MLT` with varying label distributions. We use three imbalance types to compose different label configurations: (1) **Uniform** (i.e., balanced labels), (2) **Forward-LT**, where the labels exhibit a long tail over class ids, and (3) **Backward-LT**, where labels are inversely long-tailed with respect to the class ids. For each configuration, we train 20 ERM models with varying hyperparameters. We then calculate the (α, β, γ) statistics for each model, and plot its accuracy against $(\beta+\gamma)-\alpha$.

Fig. 4 reveals the following findings: (1) *The (α, β, γ) statistics characterize model performance in MDLT.* In particular, the $(\beta + \gamma) - \alpha$ quantity displays a very strong correlation with test performance across every label configuration. (2) *Data imbalance increases the risk of learning less transferable features.* When the label distributions are similar across domains (Fig. 4a), the models are robust to varying parameters, clustering in the upper-right region. However, as labels become imbalanced (Figs. 4b, 4c) and

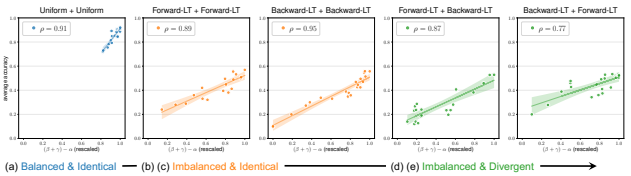

Figure 4. Correspondence between $(\beta + \gamma) - \alpha$ and test accuracy across different label configurations of `Digits-MLT`. Each plot refers to a specific label configuration. Each point corresponds to a model trained with ERM using different hyperparameters.

further divergent (Figs. 4d, 4e), chances that model learns non-transferable features (lower $(\beta+\gamma)-\alpha$) increase, leading to a large performance drop.

3.3. A Loss that Bounds Transferability Statistics

We use the above findings to design a new loss function particularly suitable for MDLT. We will first introduce the loss function then prove that it minimizes an upper bound of the (α, β, γ) statistics. We start from a simple loss inspired by the metric learning objective [17, 44]. We call this loss $\mathcal{L}_{\texttt{DA}}$ since it aims for Domain Alignment, i.e., aligning the features of the same class across domains. Let $(\mathbf{x}_i, c_i, d_i)$ denote a sample with feature $\mathbf{z}_i$. Given a set of training samples with feature set $\mathcal{Z}$, we have

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

Intuitively, $\mathcal{L}_{\texttt{DA}}$ tackles label *divergence*, as (d, c) pairs that share same class would be pulled closer, and vice versa. It is also related to (α, β, γ) because the numerator represents *positive* cross-domain pairs (α), and the denominator represents *negative* cross-class pairs (β, γ). A probabilistic interpretation of $\mathcal{L}_{\texttt{DA}}$ is provided in Appendix D.2.

But, $\mathcal{L}_{\texttt{DA}}$ does not address label *imbalance*. Note that (α, β, γ) is defined in a *balanced* way, independent of the number of samples of each (d, c). However, given an imbalanced dataset, most samples will come from majority domain-class pairs, which would dominate $\mathcal{L}_{\texttt{DA}}$, and cause minority pairs to be overlooked.

Balanced Domain-class Distribution Alignment (BoDA). To tackle data imbalance across (d, c) pairs, we modify the `DA` loss to the `BoDA` loss:

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

`BoDA` scales the original d by a factor of $1/N_{d_i,c_i}$, i.e., it counters the effect of imbalanced domain-class pairs by introducing a *balanced* distance measure $\widetilde{\mathsf{d}}$.

Theorem 1 ($\mathcal{L}_{\texttt{BoDA}}$ 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, and (α, β, γ) be the transferability statistics for $\mathcal{S}$ defined in Definition 3. It holds that*

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

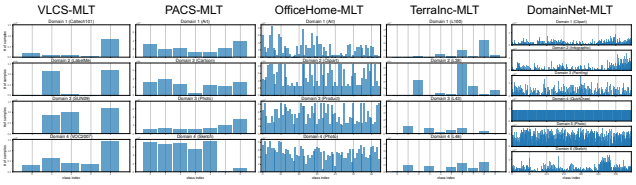

Figure 5. Overview of training set label distribution for five MDLT datasets. We set up MDLT benchmarks from datasets traditionally used for DG, and make validation/test sets balanced across all domain-class pairs. More details are provided in Appendix F.

Theorem 1 has the following interesting implications: (1) $\mathcal{L}_{\texttt{BoDA}}$ *upper-bounds* (α, β, γ) *statistics in a desired form that naturally translates to better performance.* By minimizing $\mathcal{L}_{\texttt{BoDA}}$, we ensure a low α (attract same classes) and high β, γ (separate different classes), which are essential conditions for generalization in MDLT. (2) *The constant factors correspond to how much each component contributes to the transferability graph.* Zooming on the arguments of $\exp(\cdot)$, we observe that the objective is proportional to $\alpha - (\frac{1}{|\mathcal{D}|}\beta + \frac{|\mathcal{D}|-1}{|\mathcal{D}|}\gamma)$. Note that α summarizes data similarity for the same class, while $(\frac{1}{|\mathcal{D}|}\beta + \frac{|\mathcal{D}|-1}{|\mathcal{D}|}\gamma)$ summarizes data similarity across different classes, using weighted average of β and γ, with weights proportional to the number of associated domains (1 for β, $(|\mathcal{D}|-1)$ for γ).

BoDA Variants: Matching Higher Order Statistics. Distance d can be Euclidean distance $\sqrt{(\mathbf{z}-\boldsymbol{\mu}_{d,c})^\top(\mathbf{z}-\boldsymbol{\mu}_{d,c})}$, which captures first order statistics. To match higher order statistics such as covariance, $\sqrt{(\mathbf{z}-\boldsymbol{\mu}_{d,c})^\top\boldsymbol{\Sigma}_{d,c}^{-1}(\mathbf{z}-\boldsymbol{\mu}_{d,c})}$ is used for d, resembling the Mahalanobis distance [11]. We refer to these variants as BoDA and BoDA-M.

4. Benchmarking MDLT

Datasets. We curate five multi-domain datasets typically used in DG and adapt them for MDLT evaluation. For each dataset, we create two balanced datasets one for validation and the other for testing, and leave the rest for training. The size of the validation and test data sets is 5% and 10% of original data, respectively. Table 4 in Appendix F provides the statistics of each MDLT dataset. Fig. 5 shows the label distributions across domains in the five datasets.

Competing Algorithms. We compare BoDA to a large number of algorithms that span different learning strategies and categories, including (1) *vanilla:* **ERM** [46], (2) *distributionally robust optimization:* **GroupDRO** [40], (3) *data augmentation:* **Mixup** [50], **SagNet** [35], (4) *meta-learning:* **MLDG** [27], (5) *domain-invariant feature learning:* **IRM** [1], **DANN** [15], **CDANN** [31], **CORAL** [45], **MMD** [29], (6) *transfer learning:* **MTL** [4], (7) *multi-task learning:* **Fish** [42], and (8) *imbalanced learning:* **Focal** [32], **CBLoss** [10], **LDAM** [6], **BSoftmax** [39], **SSP** [52], **CRT** [23]. We provide detailed descriptions in Appendix G.

Benchmark Results on MDLT Datasets. We report the main results in this section for all MDLT datasets. The complete results and all additional experiments are provided in Appendix H and J. The performance of all methods on the curated VLCS-MLT, PACS-MLT, OfficeHome-MLT, TerraInc-MLT and DomainNet-MLT benchmarks are in Table 1. First, BoDA consistently achieves the best average accuracy across all datasets. Moreover, on certain datasets (e.g., OfficeHome-MLT), MDL methods perform better, while on others (e.g., TerraInc-MLT), imbalanced methods are better; Yet, regardless of the dataset, BoDA outperforms all methods, highlighting its effectiveness for the MDLT task. Finally, BoDA substantially boosts the performance across all benchmarks compared to ERM.

Table 1. Results over all MDLT benchmarks.

Algorithm	VLCS-MLT	PACS-MLT	OfficeHome-MLT	TerraInc-MLT	DomainNet-MLT	Avg
ERM [46]	76.3 ±0.4	97.1 ±0.1	80.7 ±0.0	75.3 ±0.3	58.6 ±0.2	77.6
IRM [1]	76.5 ±0.2	96.7 ±0.2	80.6 ±0.4	73.3 ±0.7	57.1 ±0.1	76.8
GroupDRO [40]	76.7 ±0.4	97.0 ±0.1	80.1 ±0.3	72.0 ±0.4	53.6 ±0.1	75.9
Mixup [50]	75.9 ±0.1	96.7 ±0.2	81.2 ±0.2	71.1 ±0.7	57.6 ±0.1	76.5
MLDG [27]	76.9 ±0.2	96.6 ±0.1	80.4 ±0.2	76.6 ±0.2	58.5 ±0.0	77.8
CORAL [45]	75.9 ±0.5	96.6 ±0.5	81.9 ±0.1	76.4 ±0.5	59.4 ±0.1	78.0
MMD [29]	76.3 ±0.6	96.9 ±0.1	78.4 ±0.4	73.3 ±0.4	56.7 ±0.0	76.3
DANN [15]	77.5 ±0.1	96.5 ±0.0	79.2 ±0.2	68.7 ±0.9	55.8 ±0.1	75.5
CDANN [31]	76.6 ±0.4	96.1 ±0.1	79.0 ±0.2	70.3 ±0.5	56.0 ±0.1	75.6
MTL [4]	76.3 ±0.3	96.7 ±0.2	79.5 ±0.2	75.0 ±0.7	58.6 ±0.1	77.2
SagNet [35]	76.3 ±0.2	**97.2** ±0.1	80.9 ±0.1	75.1 ±1.6	58.9 ±0.0	77.7
Fish [42]	77.5 ±0.3	96.9 ±0.2	81.3 ±0.3	75.3 ±0.5	59.6 ±0.1	78.1
Focal [32]	75.6 ±0.4	96.5 ±0.2	77.9 ±0.0	75.7 ±0.4	57.8 ±0.2	76.7
CBLoss [10]	76.8 ±0.3	96.9 ±0.1	79.8 ±0.2	78.0 ±0.4	58.9 ±0.1	78.1
LDAM [6]	77.5 ±0.1	96.5 ±0.2	80.3 ±0.2	74.7 ±0.9	59.2 ±0.0	77.7
BSoftmax [39]	76.7 ±0.5	96.9 ±0.3	80.4 ±0.2	76.7 ±1.0	58.9 ±0.1	77.9
SSP [52]	76.1 ±0.3	96.9 ±0.2	81.1 ±0.3	78.5 ±0.7	59.7 ±0.0	78.5
CRT [23]	76.3 ±0.2	96.3 ±0.1	81.2 ±0.0	81.6 ±0.1	60.4 ±0.2	79.2
BoDA	77.3 ±0.2	**97.2** ±0.1	82.3 ±0.1	82.3 ±0.3	**61.7** ±0.1	80.2
BoDA-M	**78.2** ±0.4	97.1 ±0.2	**82.4** ±0.2	**83.0** ±0.4	**61.7** ±0.2	**80.5**
BoDA *vs.* ERM	+1.9	+0.1	+1.7	+7.7	+3.1	+2.9

Table 2. BoDA strengthens performance on Domain Generalization (DG) benchmarks. The full tables including detailed results for each dataset are in Appendix I.

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
Current SOTA [45]	**78.8** ±0.6	86.2 ±0.3	68.7 ±0.3	47.6 ±1.0	41.5 ±0.1	64.5
BoDA	78.5 ±0.3	**86.9** ±0.4	**69.3** ±0.1	**50.2** ±0.4	**42.7** ±0.1	**65.5**
BoDA + Current SOTA [45]	79.1 ±0.1	87.9 ±0.5	69.9 ±0.2	50.7 ±0.6	43.5 ±0.3	66.2
BoDA *vs.* ERM	+1.6	+2.4	+3.4	+4.6	+2.6	+2.9

5. Beyond MDLT: Domain Generalization

Domain generalization (DG) refers to learning from multiple domains and generalizing to unseen domains. Since naturally the learning domains differ in their label distributions and even have class imbalance within each domain, we study whether BoDA can improve performance for DG. Note that all datasets we adapted for MDLT are standard benchmarks for DG, which confirms that data imbalance is an intrinsic problem in DG, but has yet been overlooked.

We follow the DG evaluation protocol in [19]. Table 2 reveals the following findings: First, BoDA alone can improve upon the current SOTA on four out of the five datasets, and achieves notable performance gains. Moreover, combined with the current SOTA, BoDA further boosts the result by a notable margin across all datasets, suggesting that label imbalance is orthogonal to existing DG-specific algorithms. The intriguing results shed light on the importance of integrating label imbalance for practical DG algorithm design.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV*, 2018.

[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[4] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of Machine Learning Research*, 22(2):1–55, 2021.

[5] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[6] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019.

[7] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019.

[8] J Douglas Carroll and Phipps Arabie. Multidimensional scaling. *Measurement, judgment and decision making*, pages 179–250, 1998.

[9] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[10] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019.

[11] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000.

[12] Qi Dong, Shaogang Gong, and Xiatian Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1367–1381, 2019.

[13] Mark Dredze, Alex Kulesza, and Koby Crammer. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning*, 79(1):123–149, 2010.

[14] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *ICCV*, 2013.

[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(1):2096–2030, 2016.

[16] Amir Globerson, Gal Chechik, Fernando Pereira, and Naftali Tishby. Euclidean embedding of co-occurrence data. In *NeurIPS*, 2004.

[17] Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. Neighbourhood components analysis. In *NeurIPS*, 2004.

[18] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[19] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

[20] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *IEEE international joint conference on neural networks*, pages 1322–1328, 2008.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[22] Chen Huang, Yining Li, Change Loy Chen, and Xiaoou Tang. Deep imbalanced learning for face recognition and attribute prediction. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[23] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *ICLR*, 2020.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[25] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

[26] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[27] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018.

[28] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017.

[29] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018.

[30] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. *arXiv preprint arXiv:2111.13998*, 2021.

[31] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *AAAI*, 2018.

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[33] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019.

[34] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013.

[35] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *CVPR*, 2021.

[36] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[37] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[38] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019.

[39] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. In *NeurIPS*, 2020.

[40] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

[41] Alice Schoenauer-Sebag, Louise Heinrich, Marc Schoenauer, Michele Sebag, Lani F Wu, and Steve J Altschuler. Multi-domain adversarial learning. In *ICLR*, 2019.

[42] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

[43] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *arXiv preprint arXiv:1902.07379*, 2019.

[44] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*, 2016.

[45] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016.

[46] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[47] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017.

[48] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2021.

[49] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016.

[50] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI*, 2020.

[51] Yongxin Yang and Timothy M Hospedales. A unified perspective on multi-domain and multi-task learning. In *ICLR*, 2015.

[52] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In *NeurIPS*, 2020.

[53] Yuzhe Yang, Kaiwen Zha, Ying-Cong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *ICML*, 2021.

[54] Marvin Zhang, Henrik Marklund, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group shift. *arXiv preprint arXiv:2007.02931*, 2020.

[55] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *ICCV*, 2017.

[56] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. *arXiv preprint arXiv:2107.09249*, 2021.

[57] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021.

[58] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. *CVPR*, 2020.

[59] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization in vision: A survey. *arXiv preprint arXiv:2103.02503*, 2021.

[60] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2021.