Multi-Domain Long-Tailed Learning by Augmenting Disentangled Representations

Anonymous Author(s)
Affiliation
Address
email

1 Introduction

Deep classification models can struggle when the number of examples per class varies dramatically [5, 45]. This *long-tailed* setting arises frequently in practice, such as wildlife recognition [5]. Classifiers tend to be biased towards majority classes and perform poorly on class-balanced test distributions, i.e. when there is a shift in the label distribution between training and test. Existing approaches focus on single-domain long-tailed learning, while we study *multi-domain long-tailed learning*, where each domain has its own long-tailed distribution and the classifiers need to handle distribution shift amidst class imbalance. Here, we focus on two types of distribution shift: subpopulation shift and domain shift. In subpopulation shift, we train a model on data from multiple domains and evaluate the model on a test set with balanced domain-class pairs. A machine learning model trained on the entire population may fail on the test set when this correlation does not hold anymore. In domain shift, we expect the trained model to generalize well to completely new test domains.

Prior long-tailed classification methods work well in single-domain settings, but may perform poorly when the test data is from underrepresented domains or novel domains. Meanwhile, invariant learning approaches alleviate cross-domain performance gaps by learning representations or predictors that are invariant across different domains [3, 23]. Yet, these approaches are mostly evaluated in class-balanced settings, where models must be trained on plenty of examples from each class even if augmentation strategies are applied [40]. With multi-domain long-tailed data, learning a class-unbiased domain-invariant model is not trivial since the imbalance can exist within a domain or across domains. We aim to address these challenges in this work, leading to a novel method named **TALLY**.

TALLY empower augmentation to balance examples over domains and classes by decomposing and reassembling example pairs, combining the class-relevant semantic information of one example with the domain-associated nuisances of another. Specifically, TALLY first decouples the representation of each example into semantic information and nuisances with instance normalization. To further mitigate the effects of nuisances, we first average out domain information over examples of the same class and construct class prototype representations. Each semantic representation is then linearly interpolated with a corresponding class prototype, leading to the prototype-enhanced semantic representation. The domain-associated factors are similarly interpolated with class-agnostic domain factors to improve training stability and remove noise. Finally, TALLY produces augmented representations to benefit the training process by reassembling the prototype-enhanced semantic representation and domain-associated nuisances among examples. To further achieve balanced augmentation, we additionally propose a selective balanced sampling strategy to draw example pairs for augmentation.

In summary, our major contributions are: we investigate and formalize an important yet less explored problem – multi-domain long-tailed learning, and propose an effective augmentation algorithm called TALLY to simultaneously address the class-imbalance issue and learn domain-invariant predictors. We empirically demonstrate the effectiveness of TALLY under subpopulation shift and domain shift. We observe that TALLY outperforms both prior single-domain long-tailed learning and domain-invariant learning approaches, with a 5.18% error decrease over all datasets. Furthermore, TALLY is capable of capturing stronger invariant predictors compared with prior invariant learning approaches.

Submitted to the Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

2 Preliminaries and Method

2.1 Multi-Domain Long-Tailed Learning.

In this paper, we investigate the setting where one predicts the class label $y \in \mathcal{C}$ based on the input feature $x \in \mathcal{X}$, where $\mathcal{C} = \{1, \ldots, C\}$. Given a machine learning model f parameterized by parameter θ and a loss function ℓ, empirical risk minimization (ERM) trains such a model by minimizing average loss over all training examples as

$$\min_{\theta} \mathbb{E}_{(x,y)\sim P^{tr}}[\ell(f_\theta(x), y)], \tag{1}$$

which works well when the label distribution is approximately uniform. In multi-domain long-tailed learning, the overall data distribution is drawn from a set of domains $\mathcal{D} = \{1, \ldots, D\}$ and each domain d is associated with a class-imbalanced dataset $\{(x_i, y_i, d)\}_{i=1}^{N_d}$ drawn from domain-specific distribution p_d. Following [2, 19], both training and test distribution can be formulated as a mixture distribution over domain space $\mathcal{D}$, i.e., $P^{tr} = \sum_{d=1}^{D} \eta_d^{tr} P_{tr}^d$ and $P^{ts} = \sum_{d=1}^{D} \eta_d^{ts} P_{ts}^d$. The corresponding training and test domains are $\mathcal{D}^{tr} = \{d \in \mathcal{D} | \eta_d^{tr} > 0\}$ and $\mathcal{D}^{ts} = \{d \in \mathcal{D} | \eta_d^{ts} > 0\}$, respectively, where η_d^{tr} and η_d^{ts} represent the mixture probability. For each domain d, we define the number of training examples in each class as $\{n_{1,d}^{tr}, \ldots, n_{C,d}^{tr}\}$, sorted by cardinality. The imbalance ratio ρ^{tr} is extended to domain-level ratio as $\rho_d^{tr} = n_{C,d}^{tr}/n_{1,d}^{tr}$. During test time, we consider two kinds of test distributions, corresponding to two categories of distribution shifts – subpopulation shift and domain shift. In subpopulation shift, the test domains have been observed during training time, but the test distribution is class-balanced and domain-balanced, i.e., $\mathcal{D}^{ts} \subseteq \mathcal{D}^{tr}$ and $\{\eta_d^{ts} = 1/|\mathcal{D}^{ts}| | \forall d \in \mathcal{D}^{ts}\}$. In domain shift, the test domains are disjoint from the training domains, i.e., $\mathcal{D}^{tr} \cap \mathcal{D}^{ts} = \emptyset$.

2.2 Detailed Descriptions of TALLY

To improve robustness in multi-domain long-tailed learning, we would like method that can learn class-unbiased domain-invariant representations. To accomplish this, we introduce TALLY to do balanced augmentation over classes and domains.

Representation Disentanglement and Reassembly As described above, TALLY reassembles augmented examples from pairs of examples by combining the semantic representation of one with the domain-related nuisance factors of the other. Motivated by style transfer [15], we use instance normalization (InstanceNorm) to perform the required disentanglement of semantic and nuisance information. Concretely, given an example (x, y, d) we denote the hidden representation at layer r as $s = f^r(x) \in \mathbb{R}^{C\times H\times W}$, where C, H, and W denote channel, height, and width dimensions, respectively. Ignoring affine parameters, InstanceNorm normalizes the example as:

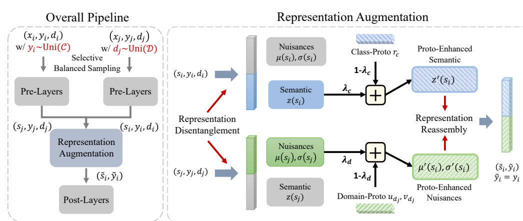


Figure 1: An illustration of TALLY.

$$z(s) = \mathrm{InstanceNorm}(s) = \frac{s - \mu(s)}{\sigma(s)}, \quad \text{where } z(s), \mu(s), \sigma(s) \in \mathbb{R}^C \tag{2}$$

Following Huang and Belongie [15], we treat the normalized example $z(s)$ as the semantic representation, and regard $\mu(s)$ and $\sigma(s)$ as the domain-associated nuisances.

After decoupling representations, we produce an augmented representation from a pair of examples (x_i, y_i, d_i) and (x_j, y_j, d_j) by swapping semantic representations and domain-associated nuisances:

$$\tilde{s} = \sigma(s_j)\left(\frac{s_i - \mu(s_i)}{\sigma(s_i)}\right) + \mu(s_j),\ \tilde{y} = y_i. \tag{3}$$

Since the semantic content of the augmented representation $\tilde{s}$ is from example (x_i, y_i, d_i), we label our augmented example with $\tilde{y} = y_i$. By reassembling disentangled representations, we can augment representations for minority domains or minority classes.

Selective Balanced Sampling. In the process of representation disentanglement and reassembly, finding a suitable strategy of sampling examples from the training distribution is crucial to solving the class-domain imbalance problem. In multi-domain long-tailed learning, the most straightforward way is up-sampling examples from minority domain-class groups, which is named *balanced sampling* here. In practice, for each example (x_i, y_i, d_i), the label y_i and domain d_j are uniformly sampled a joint uniform distribution over all domain-class combinations, i.e., $(y_i, d_i) \sim \text{Uniform}(\mathcal{C}, \mathcal{D})$.

However, to transfer the knowledge between different domain-class groups in TALLY, using such a sampling strategy may overemphasize the importance of minority domain-class groups. To augment minority groups, balanced sampling tends to repeatedly draw examples from the same minority group. We do not expect this because of two reasons: first, it limits the sample diversity in knowledge transfer; second, minority groups typically perform worse than majority groups, which may make the knowledge transfer less reliable. Hence, we propose a selective balanced sampling strategy in TALLY. Concretely, for a pair of examples (x_i, y_i, d_i) and (x_j, y_j, d_j), the label y_i of example i is uniformly sampled from all classes ($y_i \sim \text{Uniform}(\mathcal{C})$) and the domain d_j of example j is uniformly sampled from all domains ($d_j \sim \text{Uniform}(\mathcal{D})$).

Prototype-guided Invariant Learning. Since the semantic representation $z(s)$ (Eqn. 2) should contain only class-relevant information, it should ideally be *domain-invariant*. However, per-instance statistics can be noisy and instance normalization may not perfectly disentangle the semantic information from the domain-related nuisances. To improve robustness, we can "average out" domain information over many examples of the same class from different domains. However, merely averaging over examples would remove the diversity that distinguishes different examples of the same class. We balance diversity and domain-invariance by interpolating $z(s)$ with the corresponding *class prototype representation*. We define the class prototype representation r_c as the average *semantic* representation over examples belonging to class c regardless of domain: $r_c = \frac{1}{n_c^{tr}} \sum_{i=1}^{n_c^{tr}} z(s_i) = \frac{1}{n_c^{tr}} \sum_{i=1}^{n_c^{tr}} \frac{s_i - \mu(s_i)}{\sigma(s_i)}$.

For each example (x_i, y_i, d_i) with $y_i = c$, we obtain the prototype-enhanced semantic representation by linearly interpolating $z(s_i)$ with the corresponding class prototype r_c:

$$z'(s_i) = \lambda_c z(s_i) + (1 - \lambda_c) r_c, \tag{4}$$

where $\lambda_c \sim \mathrm{Beta}(\alpha_c, \alpha_c)$ is the interpolation coefficient. By applying this class prototype-based interpolation strategy, we are capable of capturing invariant knowledge and keeping the diversity of instance-level semantic representation when swapping information.

We also desire that the disentangled $\mu(s)$ and $\sigma(s)$ (Eqn. 2) contain only domain-related nuisance information. However, for similar reasons as with $z(s)$, they may still contain some class-related semantic information which we would like to remove by "averaging out." In this case, we remove semantic information by averaging over examples from *different classes* within the *same domain*: $u_d = \frac{1}{n_d^{tr}} \sum_{i=1}^{n_d^{tr}} \mu(s_i),\ v_d = \frac{1}{n_d^{tr}} \sum_{i=1}^{n_d^{tr}} \sigma(s_i)$, where n_d^{tr} represents the number of training examples in domain d. Then, for each example, we linearly interpolate its domain-associated nuisances with the above class-agnostic nuisances as:

$$\mu'(x_i) = \lambda_d \mu(x) + (1 - \lambda_d) u_d,\ \sigma'(x_i) = \lambda_d \sigma(x) + (1 - \lambda_d) v_d, \tag{5}$$

where the interpolation ratio is $\lambda_d \sim \mathrm{Beta}(\alpha_d, \alpha_d)$.

By replacing the original semantic representation and domain-associated nuisances in Eqn. 3 with the prototype-guided ones, we obtain the enhanced augmented representation as follows:

$$\tilde{s}' = \sigma'(s_j) z'(s_i) + \mu'(s_j),\ \tilde{y}' = y_i. \tag{6}$$

Finally, we replace the original training data with the augmented ones. We summarize the overall framework of TALLY in Algorithm 1 in Appendix.

3 Experiments

In this section, we conduct extensive experiments to evaluate how TALLY performs. To answer Q1, we compare TALLY to two categories of algorithms. The first category includes single-domain long-tailed learning methods such as Focal [24], LDAM [6], CRT [17], MiSLAS [47], and Remix [8]. Focal, LDAM, and CRT are up-weighting or up-sampling approaches, while MiSLAS and Remix are data augmentation strategies. The second category includes approaches for improving robustness to distribution shift: IRM [3], GroupDRO [29], LISA [40], MixStyle [50], DDG [43], and BODA [39]. Follow Yang et al. [39], we use a ResNet-50 architecture for all algorithms, and detail the baselines in Appendix B. All hyperparameters are selected via cross-validation. Due to space limitation, we only show ERM and top-5 baselines here and put full results in Appendix. We also provide comprehensive analysis to understand the results in Appendix E.

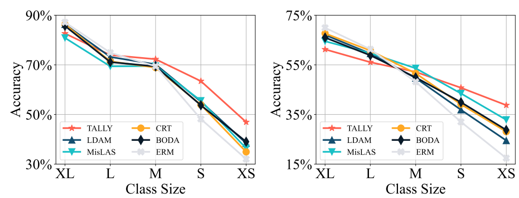
Table 1: Results of subpopulation shifts and domain shifts on synthetic data (Full table Appx. C.3).

	Subpopulation Shift				Domain Shift			
	VLCS-LT	PACS-LT	OH-LT	DN-LT	VLCS-LT	PACS-LT	OH-LT	DN-LT
ERM	73.33%	90.40%	61.07%	44.33%	67.62%	76.27%	51.95%	33.21%
Focal	74.83%	90.44%	62.57%	47.35%	69.38%	75.29%	54.03%	35.23%
MiSLAS	71.83%	90.99%	61.38%	49.15%	68.64%	77.94%	52.86%	36.18%
CORAL	71.67%	88.22%	59.10%	43.92%	66.54%	75.62%	50.74%	33.44%
MixStyle	74.30%	91.55%	62.26%	43.59%	67.75%	79.78%	52.47%	33.71%
BODA	74.83%	91.03%	62.79%	47.61%	69.63%	78.81%	53.32%	35.85%
TALLY (ours)	**76.83%**	**92.38%**	**67.00%**	**50.15%**	**70.60%**	**81.55%**	**55.69%**	**36.45%**

3.1 Evaluation on Long-Tailed Variants of Domain Generalization Benchmarks

Datasets. We curate four *multi-domain long-tailed* datasets by modifying four existing domain-generalization benchmarks: VLCS [11], PACS [22], OfficeHome [33], and DomainNet [27]. We modify the prior datasets by removing training examples so that each domain has a long-tailed label distribution (overall imbalance ratio: 50) and call the resulting datasets **VLCS-LT**, **PACS-LT**, **OfficeHome-LT**, and **DomainNet-LT**. See Appendix C for more details and evaluation protocol.

Results. The overall performance of TALLY and prior methods for tackling subpopulation shift and domain shift is reported in Table 1. For subpopulation shift, we report the average performance over all domains. We observe that TALLY consistently outperforms all methods, verifying its effectiveness in improving the robustness to subpopulation shifts. In addition, Figure 2 shows performance broken down by class size for OfficeHome-LT and DomainNet-LT under subpopulation shift, where we split all classes into five levels according to their cardinality. We compare TALLY with ERM, and four strongest baselines. The results show that TALLY's performance improvements arise from larger improvements on smaller classes rather than performance improvements across the board, hence indicating that it is particularly well-suited for class-imbalanced problems.



(a) : OfficeHome-LT (b) : DomainNet-LT

Figure 2: Performance w.r.t. Class Size. XL and XS represent the largest and smallest classes.

3.2 Evaluation on Naturally Imbalanced Multi-Domain Data

Datasets. To further evaluate TALLY and prior methods, we study two multi-domain datasets that are naturally imbalanced: Terra Incognita (TerraInc) [4] and iWildCam [5], both of which aim to classify wildlife across different camera traps. More details of these datasets and class distribution are described in Appendix D. To better capture performance on rare species, we use macro F1 score as the primary evaluation metric following Koh et al. [19], but we also report average accuracy. We list all hyperparameters in Appendix D.2.

Table 2: Results of Domain Shifts on Real-world Data (full results: Appx. D.3)

	TerraInc		iWildCam	
	Macro F1	Acc	Macro F1	Acc
ERM	42.35%	54.81%	32.0%	69.0%
Focal	43.54%	56.62%	33.2%	74.7%
MiSLAS	40.68%	52.96%	30.5%	59.8%
CORAL	45.43%	58.10%	32.8%	73.3%
MixStyle	44.73%	57.55%	32.4%	74.9%
BODA	44.47%	57.52%	32.9%	70.5%
TALLY (ours)	**46.23%**	**59.89%**	**34.4%**	73.4%

Results. We report the results over all test domains in Table 2. The conclusions are largely consistent with the results from Sec. 3.1, where TALLY consistently improves the performance over all baselines and enhances the robustness of multi-domain long-tailed learning. The superiority of TALLY over prior augmentation techniques is further evidence of the effectiveness of balanced augmentation.

4 Conclusion

In this paper we investigate multi-domain imbalanced learning, a natural extension of classical single-domain imbalanced learning. We propose a novel balanced augmentation algorithm called TALLY to achieve robust imbalanced learning that can overcome distribution shifts. TALLY introduces a prototype enhanced disentanglement procedure for separating semantic and nuisance information, and then mixes the enhanced semantic and domain-associated nuisance information among examples. The results demonstrate its effectiveness of TALLY.

References

[1] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. 2021.

[2] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2019.

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

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[5] Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020.

[6] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

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

[8] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: rebalanced mixup. In *European Conference on Computer Vision*, pages 95–110. Springer, 2020.

[9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[10] Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20(1):18–36, 2004.

[11] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013.

[12] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2021.

[13] Ruocheng Guo, Pengchuan Zhang, Hao Liu, and Emre Kiciman. Out-of-distribution prediction with invariant risk minimization: The limitation and an effective fix. *arXiv preprint arXiv:2101.07732*, 2021.

[14] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6626–6636, 2021.

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[16] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7610–7619, 2020.

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

[18] Kia Khezeli, Arno Blaas, Frank Soboczenski, Nicholas Chia, and John Kalantari. On invariance penalties for risk minimization. *arXiv preprint arXiv:2106.09777*, 2021.

[19] Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[20] Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. *arXiv preprint arXiv:2008.01883*, 2020.

[21] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[22] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[23] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018.

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[25] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2970–2979, 2020.

[26] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2): 539–550, 2008.

[27] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[28] Alexander Robey, George J Pappas, and Hamed Hassani. Model-based domain generalization. *Advances in Neural Information Processing Systems*, 34:20210–20229, 2021.

[29] 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.

[30] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9624–9633, 2021.

[31] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[32] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[33] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[34] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. *Advances in Neural Information Processing Systems*, 30, 2017.

[35] Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3622–3626. IEEE, 2020.

[36] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, pages 247–263. Springer, 2020.

[37] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6502–6509, 2020.

[38] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

[39] Yuzhe Yang, Hao Wang, and Dina Katabi. On multi-domain long-tailed recognition, generalization and beyond. *arXiv preprint arXiv:2203.09513*, 2022.

[40] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. *arXiv preprint arXiv:2201.00299*, 2022.

[41] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5704–5713, 2019.

[42] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2100–2110, 2019.

[43] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. Towards principled disentanglement for domain generalization. In *CVPR*, 2022.

[44] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. 2018.

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

[46] Zizhao Zhang and Tomas Pfister. Learning fast sample re-weighting without reward data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 725–734, 2021.

[47] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16489–16498, 2021.

[48] Allan Zhou, Fahim Tajwar, Alexander Robey, Tom Knowles, George J Pappas, Hamed Hassani, and Chelsea Finn. Do deep networks transfer invariances across classes? 2022.

[49] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020.

[50] Fan Zhou, Zhuqing Jiang, Changjian Shui, Boyu Wang, and Brahim Chaib-draa. Domain generalization with optimal transport and metric learning. *arXiv preprint arXiv:2007.10573*, 2020.

[51] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13025–13032, 2020.

Algorithm 1 TALLY Training Process

Require: learning rates η; warm start epochs T_0; prototype momentum γ; model $f_\theta(\cdot)$ with hidden representation $f_\theta^r(\cdot)$ at layer r; Dataset $\mathcal{D}^{tr} = \{(x, y, d)\}$

Initialize domain-agnostic prototypes $\{r_c^{(0)}\}_{c=1}^C$ and class-agnostic statistics $\{(u_d^{(0)}, v_d^{(0)})\}_{d=1}^D$

Train f_θ with ERM for $t < T_0$

for $t = T_0$ to T **do**

  $d_i, d_j \sim \mathrm{Uniform}(\mathcal{D})$, $y_i, y_j \sim \mathrm{Uniform}(\mathcal{C})$ ▷ Randomly sample domains and classes

  $(x_i, y_i, d_i) \sim \{\mathcal{D}^{tr} | y = y_i, d = d_i\}$, $(x_j, y_j, d_j) \sim \{\mathcal{D}^{tr} | y = y_j, d = d_j\}$

  $(s_i, s_j) \leftarrow (f^r(x_i), f^r(x_j))$ ▷ Compute hidden representations

  $z(s_i) \leftarrow \mathrm{InstanceNorm}(s_i)$ ▷ Disentangle semantic factor (Eqn. 2)

  $z'(s_i) \leftarrow \lambda_{y_i} z(s_i) + (1 - \lambda_{y_i}) r_{y_i}^t$ ▷ Enhance semantic factor (Eqn. 4)

  $(\mu'(s_j), \sigma'(s_j)) \leftarrow \lambda_{y_j}(\mu(s_j), \sigma(s_j)) + (1 - \lambda_{y_j})(u_{y_j}^t, v_{y_j}^t)$ ▷ Enhance nuisances (Eqn. 5)

  $(\tilde{s}', \tilde{y}') \leftarrow (\sigma'(s_j) z'(s_i) + \mu'(s_j),\ y_i)$ ▷ Generate augmented example (Eqn. 6)

  $\mathrm{Optimize}[\ell(f_\theta^{L-r}(\tilde{s}'), \tilde{y}')]$ ▷ Train on augmented example

  Estimate the current prototypes and feature statistics $\{r_c\}_{c=1}^C$, $\{(u_d, v_d)\}_{d=1}^D$

  for $c = 1$ to C **do**

    $r_c^{(t+1)} \leftarrow \gamma r_c^t + (1 - \gamma) r_c$ ▷ Update domain-agnostic prototypes

  for $d = 1$ to D **do**

    $(u_d^{(t+1)}, v_d^{(t+1)}) \leftarrow \gamma(u_d^t, v_d^t) + (1 - \gamma)(u_d, v_d)$ ▷ Update class-agnostic statistics

A Related Work

A.1 Long-Tailed Learning

Training a well-performed machine learning model on class-imbalanced data has been widely studied. A typical setting of imbalanced learning is the long-tailed class distribution, where the model can be easily biased towards majority classes [45]. A lot of approaches have been proposed under this setting, including over-sampling minority classes or under-sampling majority classes [7, 10, 17, 26, 46], adjusting loss functions or logits for different classes during training [6, 9, 14, 16, 24], transferring knowledge from head classes to tail classes [34, 25, 41, 48], directly augmenting tail classes [8, 17, 47], and ensembling models with different sampling or loss weighting strategies [36, 49]. Unlike single-domain imbalanced learning, Yang et al. [39] targets on the multi-domain imbalanced learning scenario by encouraging invariant representation learning with a domain-class calibrated regularizer. However, BODA focuses on subpopulation shift with the imbalanced distribution for each domain, while the overall distribution among all classes are relatively balanced. TALLY instead studies more kinds of distribution shifts with conceptually different direction to alleviate domain-associated nuisances via balanced augmentation. It relaxes the explicit constraint on internal representations and leads to stronger empirical performance.

A.2 Domain Generalization and Out-of-Distribution Robustness

To improve out-of-distribution robustness, one line of works aims to learn domain-invariant representations by 1) minimizing the discrepancy of feature representations across all training domains [23, 31, 32, 50]; 2) leveraging domain augmentation methods to generate more training domains and improve the consistency of feature representations between the original and augmented domains [30, 35, 37, 38, 42, 51]; 3) disentangling feature representations to semantic and domain-varying ones and minimizing the semantic differences across training domains [28, 43]. Another line of works focuses on strengthening the correlations between representations and labels, leading to stronger invariant predictors. These works introduce various regularizers in learning invariant predictors, including minimizing the variances of risks across domains [21], encouraging a predictor that performs well over all domains [1, 3, 13, 18], and matching the gradient across different domains [20]. Besides explicitly involving regularizers, data interpolation is also a promising approach for learning invariant predictors [40, 50]. Unlike previous augmentation methods that require sufficient training examples for each class to learn invariance, TALLY tackles the class-imbalanced issue in domain generalization and employs a domain-balanced augmentation strategy to learn class-unbiased invariant representation.

B Detailed Description of Baselines

In this paper, we compare TALLY with two types of approaches: long-tailed classification methods and invariant learning approaches. We detail these methods here:

B.1 Long-tailed Classification Methods

We compare TALLY with Focal [24], LDAM [6], CRT [17], MiSLAS [47], and Remix [8]. Here, Focal and LDAM up-weight the loss for minority classes. CRT uses up-sampling strategy to fine-tune the classifier. MiSLAS and Remix modify the vanilla mixup [44] and make it suitable to long-tailed distribution.

B.2 Invariant Learning

We further compare TALLY with invariant learning approaches, i.e., IRM [3], GroupDRO [29], LISA [40], MixStyle [50], DDG [43], and BODA [39]. IRM learns invariant predictors that perform well across different domains. GroupDRO optimizes the worst-domain loss. LISA cancels out domain-associated information by mixing examples with the same label but different domains. MixStyle decomposes the feature representation into content information and style information. It then mixes the style information and generates new examples. Unlike MixStyle, TALLY generates examples of minority classes or domains, and uses prototypes to improve the model robustness, which is more suitable for long-tailed multi-domain learning. DDG uses an extra network to disentangle original examples and generate more. Finally, BODA is a concurrent work for long-tailed multi-domain learning with an explicit regularizer. Unlike BODA, TALLY studies a conceptually different direction to cancel out domain-associated nuisances by domain-class balanced augmentation, leading to stronger empirical performance.

C Additional Results of Synthetic Data

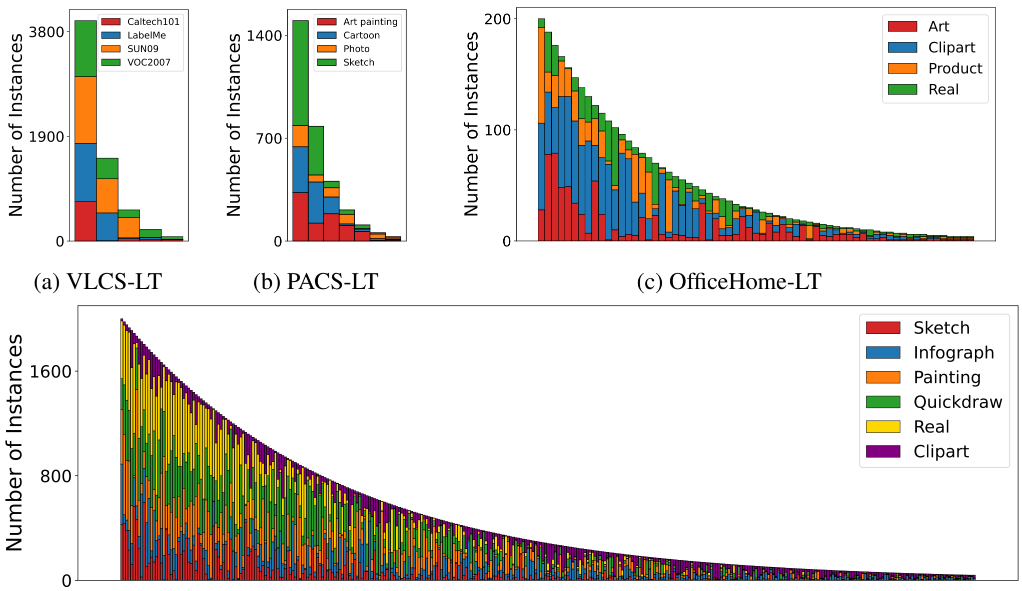
C.1 Detailed Dataset Description

VLCS-LT contains examples from 4 different domains, including Caltech101, LabelMe, SUN09, VOC2007. To create the long-tailed class distribution, we modify the original dataset by removing training examples. The dataset contains 5 classes with 6,361 images of dimension (224,224,3). The long-tailed training distribution is visualized in Figure 3a. In subpopulation shift, the number of examples of each class per domain for validation and testing is 5, 10, respectively.

PACS-LT includes 3,097 images collected from 4 domains (Art painting, Cartoon, Photo, Sketch) and 7 classes. Similar to VLCS-LT, we construct PACS-LT with long-tailed training distribution illustrated in Figure 3b. The validation set size and test set size of each class per domain in subpopulation shift are 15 and 30 respectively.

OfficeHome-LT is built upon the original OfficeHome dataset, including 3280 images of 65 classes collected from four domains – Art, Clipart, Product, Real. The long-tailed training distribution is shown in Figure 3c and the number of examples for each class per domain in validation and test sets are 4, 8, respectively.

DomainNet-LT. Similar to the other three datasets, DomainNet-LT covers 173,200 examples from Sketch, Infograph, Painting, Quickdraw, Real, Clipart. There are 345 classes in DomainNet-LT. In subpopulation shift, the number of examples of each class per domain is 3, 6, respectively. We illustrate the long-tailed training distribution in Figure 3d.



(a) VLCS-LT (b) PACS-LT (c) OfficeHome-LT

(d) DomainNet-LT

Figure 3: Long-tailed training distributions for all synthetic datasets. Here, the x-axis represents sorted class indices.

C.2 Detailed Hyperparameters

We evaluate performance under both subpopulation shift and domain shift. In subpopulation shift, the test set is balanced across both domains and classes, which means that each domain-class pair contains the same number of test examples. In domain shift, we use the classical domain generalization setting [43]. More specifically, we alternately use one domain as the test domain, and the rest as the training domains. Results are averaged over all combinations. We list the hyperparameters in Table 3 for the above four synthetic datasets.

Table 3: Hyperparameters for experiments on synthetic data.

Hyperparameters	VLCS-LT	PACS-LT	OfficeHome-LT	DomainNet-LT
Learning Rate	1e-5	1e-5	3e-5	3e-5
Weight Decay	1e-6	1e-6	1e-6	1e-6
Batch Size	18	18	18	18
Epochs	15	15	15	15
Steps	200	500	500	1000
Warm Start Epochs	7	7	7	7
γ in feat. estimation	0.8	0.8	0.8	0.8
class prototype mixup parameter α_c	0.2	0.5	0.5	0.5
domain prototype mixup parameter α_d	0.2	0.5	0.5	0.5

C.3 Full Results

The full results of subpopulation shift are reported in Table 4. In domain shift, we report the results of each domain for VLCS-LT, PACS-LT, OfficeHome-LT and DomainNet-LT in Table 5, 6, 7, 8, respectively. In the domain shift scenario of VLCS-LT, though TALLY only performs best in VOC2007 (VLCS), the results of TALLY is relatively more stable compared to other approaches, leading to the best averaged performance.

Table 4: Full results of subpopulation shifts on long-tailed variants of domain generalization benchmarks. The standard deviation is computed across three seeds.

		VLCS-LT	PACS-LT	OfficeHome-LT	DomainNet-LT
Avg.	ERM	73.33 ± 0.76%	90.40 ± 0.88%	61.07 ± 0.73%	44.33 ± 0.14%
	Focal	74.83 ± 0.29%	90.44 ± 0.06%	62.57 ± 0.50%	47.35 ± 0.09%
	LDAM	73.83 ± 1.04%	90.91 ± 0.15%	63.57 ± 0.08%	46.71 ± 0.33%
	CRT	73.83 ± 1.89%	89.17 ± 1.47%	61.92 ± 0.54%	47.37 ± 0.83%
	MiSLAS	71.83 ± 1.25%	90.99 ± 0.90%	61.38 ± 0.19%	49.15 ± 0.69%
	Remix	74.16 ± 0.76%	90.83 ± 0.77%	61.59 ± 0.44%	47.56 ± 0.25%
	IRM	50.50 ± 8.18%	65.24 ± 7.57%	45.48 ± 4.30%	35.57 ± 5.76%
	GroupDRO	72.50 ± 0.50%	89.80 ± 0.70%	59.79 ± 0.43%	43.86 ± 0.33%
	CORAL	71.67 ± 0.28%	88.22 ± 0.67%	59.10 ± 0.20%	43.92 ± 0.36%
	LISA	74.67 ± 0.76%	90.08 ± 0.45%	57.39 ± 0.59%	43.17 ± 0.53%
	MixStyle	74.30 ± 1.04%	91.55 ± 0.25%	62.26 ± 0.22%	43.59 ± 0.57%
	DDG	73.00 ± 1.63%	89.60 ± 0.40%	58.80 ± 0.57%	44.46 ± 0.06%
	BODA	74.83 ± 1.84%	91.03 ± 0.31%	62.79 ± 0.45%	47.61 ± 0.04%
	TALLY (ours)	**76.83 ± 1.04%**	**92.38 ± 0.26%**	**67.00 ± 0.47%**	**50.15 ± 0.46%**
Worst	ERM	52.67 ± 2.31%	83.81 ± 2.43%	54.48 ± 0.89%	25.36 ± 0.63%
	Focal	52.67 ± 1.15%	84.44 ± 0.81%	56.41 ± 1.34%	27.68 ± 0.13%
	LDAM	51.33 ± 2.31%	85.24 ± 1.03%	58.07 ± 0.82%	27.23 ± 0.26%
	CRT	52.00 ± 0.00%	83.02 ± 1.12%	55.51 ± 0.79%	27.55 ± 0.49%
	MiSLAS	52.00 ± 3.46%	86.03 ± 0.90%	52.82 ± 0.39%	29.42 ± 0.15%
	Remix	51.33 ± 3.05%	86.98 ± 0.59%	53.85 ± 0.54%	28.13 ± 0.99%
	IRM	32.63 ± 7.03%	59.38 ± 5.93%	40.58 ± 4.42%	20.48 ± 3.71%
	GroupDRO	51.33 ± 1.15%	83.02 ± 0.59%	54.04 ± 0.30%	25.02 ± 0.73%
	CORAL	49.33 ± 1.15%	81.59 ± 0.81%	53.53 ± 0.60%	24.50 ± 0.68%
	LISA	53.33 ± 1.15%	83.01 ± 0.81%	49.04 ± 0.40%	24.05 ± 0.48%
	MixStyle	54.00 ± 2.00%	86.98 ± 0.98%	55.19 ± 1.10%	22.65 ± 0.22%
	DDG	51.33 ± 0.94%	82.70 ± 2.38%	51.99 ± 0.55%	24.35 ± 0.20%
	BODA	54.00 ± 2.83%	85.08 ± 1.37%	55.70 ± 0.50%	26.94 ± 0.44%
	TALLY (ours)	**56.00 ± 2.00%**	**89.21 ± 0.22%**	**60.45 ± 0.09%**	**29.55 ± 0.19%**

Table 5: Domain shift results on VLCS-LT.

	Caltech101	LabelMe	SUN09	VOC2007	Avg
ERM	92.39 ± 0.35%	47.74 ± 1.13%	59.79 ± 2.70%	70.55 ± 1.51%	67.62%
Focal	**97.12 ± 1.06%**	48.83 ± 0.38%	58.66 ± 2.31%	72.91 ± 1.51%	69.38%
LDAM	95.55 ± 1.65%	47.61 ± 1.12%	61.34 ± 3.02%	73.17 ± 1.33%	69.41%
CRT	92.39 ± 1.82%	47.74 ± 1.52%	55.10 ± 2.61%	67.45 ± 1.54%	65.67%
MiSLAS	95.24 ± 1.51%	47.00 ± 0.99%	56.03 ± 1.29%	76.28 ± 1.66%	68.64%
Remix	92.66 ± 1.42%	48.77 ± 1.31%	57.98 ± 2.62%	71.45 ± 0.87%	67.71%
IRM	74.10 ± 2.67%	37.07 ± 1.87%	34.33 ± 1.77%	47.78 ± 3.59%	48.32%
GroupDRO	93.79 ± 1.01%	49.63 ± 1.09%	**62.25 ± 1.89%**	71.06 ± 0.55%	69.18%
CORAL	93.94 ± 1.53%	48.29 ± 1.08%	56.12 ± 1.84%	67.82 ± 1.43%	66.54%
LISA	90.28 ± 0.68%	48.51 ± 1.58%	58.82 ± 2.41%	68.09 ± 1.53%	66.42%
MixStyle	96.58 ± 0.84%	48.15 ± 1.20%	58.82 ± 1.94%	68.09 ± 1.98%	67.75%
DDG	95.46 ± 1.19%	50.42 ± 1.45%	57.44 ± 2.07%	70.21 ± 1.33%	68.38%
BODA	95.60 ± 1.37%	**51.42 ± 1.31%**	59.93 ± 1.97%	71.57 ± 1.18%	69.63%
TALLY (ours)	95.22 ± 0.92%	50.07 ± 1.17%	60.13 ± 2.17%	**76.98 ± 0.57%**	**70.60%**

Table 6: Domain shift results on PACS-LT.

	Art painting	Cartoon	Photo	Sketch	Avg
ERM	80.41 ± 1.21%	70.21 ± 1.14%	94.46 ± 0.19%	60.00 ± 5.04%	76.27%
Focal	80.92 ± 0.51%	69.58 ± 0.64%	93.81 ± 0.80%	56.83 ± 2.04%	75.29%
LDAM	81.82 ± 1.14%	71.64 ± 0.66%	95.34 ± 0.32%	61.30 ± 4.83%	77.53%
CRT	78.14 ± 0.99%	67.17 ± 0.73%	94.33 ± 0.78%	55.62 ± 6.57%	73.82%
MiSLAS	81.31 ± 0.49%	71.15 ± 0.28%	93.51 ± 1.40%	65.78 ± 2.13%	77.94%
Remix	82.79 ± 1.21%	69.10 ± 1.13%	92.09 ± 1.01%	57.00 ± 4.13%	75.25%
IRM	51.87 ± 4.93%	50.27 ± 5.77%	69.11 ± 4.43%	39.13 ± 9.65%	52.60%
GroupDRO	80.20 ± 0.57%	70.61 ± 1.41%	94.58 ± 0.90%	61.61 ± 1.48%	76.75%
CORAL	77.60 ± 0.79%	68.19 ± 0.73%	93.88 ± 0.40%	62.82 ± 2.67%	75.62%
LISA	81.09 ± 0.61%	65.68 ± 0.87%	94.40 ± 0.33%	56.69 ± 1.99%	74.47%
MixStyle	83.45 ± 0.90%	72.84 ± 0.59%	95.20 ± 0.49%	67.61 ± 0.83%	79.78%
DDG	79.67 ± 0.77%	68.30 ± 0.34%	94.72 ± 0.50%	61.20 ± 0.82%	75.97%
BODA	81.13 ± 0.59%	72.03 ± 0.65%	95.73 ± 0.56%	66.34 ± 1.55%	78.81%
TALLY (ours)	**85.86 ± 0.40%**	**74.20 ± 0.30%**	**96.56 ± 0.20%**	**69.58 ± 0.62%**	**81.55%**

Table 7: Domain shift results on OfficeHome-LT.

	Art	Clipart	Product	Real	Avg
ERM	45.20 ± 0.73%	41.94 ± 0.17%	59.21 ± 0.44%	61.44 ± 0.27%	51.95%
Focal	47.06 ± 0.24%	43.29 ± 0.71%	62.34 ± 0.16%	63.45 ± 0.19%	54.03%
LDAM	47.08 ± 0.37%	42.89 ± 0.18%	61.48 ± 0.55%	62.93 ± 0.24%	53.60%
CRT	47.17 ± 0.26%	42.62 ± 0.55%	61.37 ± 0.12%	63.31 ± 0.25%	53.62%
MiSLAS	45.22 ± 0.52%	41.36 ± 0.09%	62.28 ± 0.49%	62.56 ± 0.25%	52.86%
Remix	44.26 ± 0.49%	39.18 ± 0.34%	60.70 ± 0.28%	61.58 ± 0.42%	51.43%
IRM	33.55 ± 4.21%	34.34 ± 3.74%	49.54 ± 5.30%	51.95 ± 4.64%	42.34%
GroupDRO	44.62 ± 0.51%	41.84 ± 0.68%	58.40 ± 0.43%	59.63 ± 0.53%	51.12%
CORAL	43.93 ± 0.56%	42.71 ± 0.59%	56.91 ± 0.45%	59.40 ± 0.94%	50.74%
LISA	41.80 ± 0.36%	36.96 ± 0.45%	56.51 ± 0.16%	57.62 ± 0.39%	48.22%
MixStyle	45.11 ± 0.18%	**45.52 ± 0.20%**	58.32 ± 0.64%	60.92 ± 0.22%	52.47%
DDG	43.89 ± 0.39%	42.79 ± 0.91%	57.92 ± 0.15%	59.69 ± 0.30%	51.07%
BODA	47.08 ± 0.25%	44.38 ± 0.77%	59.58 ± 0.26%	62.25 ± 0.10%	53.32%
TALLY (ours)	**49.79 ± 0.76%**	44.22 ± 0.45%	**63.02 ± 0.52%**	**65.71 ± 0.26%**	**55.69%**

Table 8: Domain shift results on DomainNet-LT.

	Sketch	Infograph	Painting	Quickdraw	Real	Clipart	Avg
ERM	39.22 ± 0.24%	18.96 ± 0.37%	34.71 ± 0.49%	10.70 ± 0.06%	50.87 ± 0.43%	44.83 ± 0.25%	33.21%
Focal	41.01 ± 0.61%	19.99 ± 0.14%	36.42 ± 0.45%	10.29 ± 0.23%	**55.63 ± 0.56%**	48.09 ± 0.72%	35.23%
LDAM	40.44 ± 0.26%	19.06 ± 0.20%	36.32 ± 0.52%	11.38 ± 0.29%	52.91 ± 0.52%	46.38 ± 0.44%	34.42%
CRT	40.78 ± 0.35%	20.41 ± 0.42%	39.01 ± 0.35%	**11.41 ± 0.17%**	55.26 ± 0.69%	50.00 ± 0.83%	36.14%
MiSLAS	41.34 ± 0.39%	19.89 ± 0.25%	39.85 ± 0.56%	11.00 ± 0.13%	55.50 ± 0.36%	49.49 ± 0.29%	36.18%
Remix	40.01 ± 0.51%	19.17 ± 0.46%	38.93 ± 0.32%	11.39 ± 0.20%	53.43 ± 0.60%	47.94 ± 0.71%	35.14%
IRM	34.65 ± 0.75%	15.41 ± 0.83%	28.18 ± 1.26%	7.69 ± 0.40%	40.83 ± 0.72%	42.36 ± 1.25%	28.19%
GroupDRO	38.47 ± 0.27%	18.63 ± 0.07%	34.23 ± 0.15%	10.26 ± 0.35%	50.80 ± 0.47%	42.85 ± 0.63%	32.54%
CORAL	39.42 ± 0.42%	19.30 ± 0.33%	35.15 ± 0.70%	10.61 ± 0.22%	51.05 ± 0.28%	45.15 ± 0.38%	33.44%
LISA	40.75 ± 0.46%	18.47 ± 0.14%	37.99 ± 0.19%	9.98 ± 0.09%	54.33 ± 0.49%	48.42 ± 0.42%	34.99%
MixStyle	40.99 ± 0.60%	18.64 ± 0.32%	35.86 ± 0.39%	11.03 ± 0.15%	50.26 ± 0.53%	45.49 ± 0.84%	33.71%
DDG	40.66 ± 0.58%	19.08 ± 0.34%	35.61 ± 0.63%	11.39 ± 0.29%	50.93 ± 0.41%	45.95 ± 0.47%	33.94%
BODA	41.95 ± 0.45%	**20.65 ± 0.58%**	37.98 ± 0.27%	11.02 ± 0.23%	55.22 ± 0.65%	48.26 ± 0.33%	35.85%
TALLY (ours)	**42.66 ± 0.32%**	19.26 ± 0.09%	**40.49 ± 0.34%**	11.15 ± 0.21%	54.79 ± 0.62%	**50.36 ± 0.41%**	**36.45%**

D Additional Results of Real-world Data

D.1 Detailed Dataset Description

TerraInc. Building upon the original Terra Incognita [4], we select images from 10 classes and split the entire dataset to training, validation and test domains, which includes images from 38,042, 6,783, 7,303 camera traps, respectively.

iWildCam is a wildlife recognition datasets. It is a multi-class species classification, where the training data are collected from 243 domains and the test data includes images from 164 domains. We follow Koh et al. [19] to split the data and construct training, validation and test sets.

D.2 Detailed Hyperparameters

We list the hyperparameters in Table 9 for both TerraInc and iWildCam datasets.

Table 9: Hyperparameters for experiments on real-world data.

Hyperparameters	TerraInc	iWildCam
Learning Rate	3e-5	3e-5
Weight Decay	1e-6	0
Batch Size	18	16
Epochs	15	15
Steps	1000	1000
Warm Start Epochs	7	7
γ in feat. estimation	0.8	0.8
class prototype mixup parameter α_c	0.5	0.5
domain prototype mixup parameter α_d	0.5	0.5

D.3 Full Results

The full results on Real-world Data are reported in Table 10.

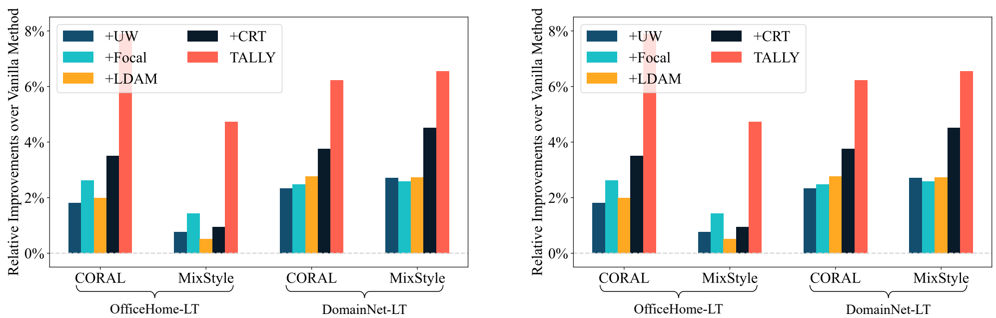
Table 10: Full Results of Domain Shifts on Real-world Data.

	TerraInc		iWildCam	
	Macro F1	Acc	Macro F1	Acc
ERM	42.35 ± 1.25%	54.81 ±0.83%	32.0 ± 1.5%	69.0 ± 0.4%
Focal	43.54 ± 0.81%	56.62 ± 1.49%	<u>33.2 ± 1.2%</u>	74.7 ± 1.9%
LDAM	44.29 ± 1.41%	57.22 ± 0.92%	32.7 ± 0.9%	**75.2 ± 2.0%**
CRT	43.09 ± 0.79%	58.27 ± 1.35%	32.5 ± 1.8%	67.3 ± 1.3%
MiSLAS	40.68 ± 1.33%	52.96 ± 2.58%	30.5 ± 1.1%	59.8 ± 2.8%
Remix	43.72 ± 1.87%	<u>58.40 ± 2.57%</u>	28.4 ± 0.8%	65.8 ± 1.6%
IRM	31.17 ± 3.52%	49.27 ± 5.22%	15.1 ± 4.9%	59.8 ± 3.7%
GroupDRO	42.22 ± 0.87%	56.43 ± 1.63%	23.9 ± 2.1%	72.7 ± 2.0%
CORAL	<u>45.43 ± 0.92%</u>	58.10 ± 1.38%	32.8 ± 0.1%	73.3 ± 4.3%
LISA	39.27 ± 0.69%	54.92 ± 1.04%	27.6 ± 1.2%	64.9 ± 2.2%
MixStyle	44.73 ± 0.99%	57.55 ± 2.05%	32.4 ± 1.1%	<u>74.9 ± 2.7%</u>
DDG	40.47 ± 1.93%	53.61 ± 1.71%	29.8 ± 0.2%	69.7 ± 2.3%
BODA	44.47 ± 0.84%	57.52 ± 1.13%	32.9 ± 0.3%	70.5 ± 2.3%
TALLY (ours)	**46.23 ± 0.56%**	**59.89 ± 1.32%**	**34.4 ± 0.4%**	73.4 ± 1.8%

E Analysis of Performance

E.1 Can we simply combine invariant learning approaches with long-tailed learning techniques?

To further understand the performance gains of TALLY, we investigate whether combining existing invariant learning and long-tailed learning approaches can tackle multi-domain long-tailed distribution shifts. Specifically, we incorporate four up-weighting or up-sampling approaches (UW, Focal, LDAM, CRT) with two representative invariant learning methods (CORAL, MixStyle). We report the relative improvement of each combination over the vanilla methods in Figure 4. Here, we use Officehome-LT and DomainNet-LT to evaluate subpopulation shift and TerraInc and iWildCam to evaluate domain shift performance. We see that applying loss up-weighting or up-sampling approaches on performant invariant learning approaches does improve their performance, as evidenced by Figure 4. Nonetheless, the consistent improvements from TALLY indicates the importance of considering domain-class pair information to achieve balanced augmentation.

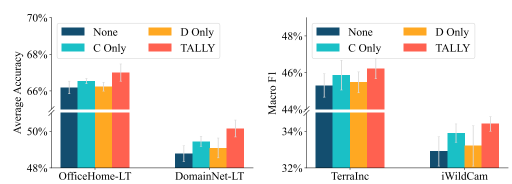


(a) Subpopulation Shift (Avg. Acc) (b) Domain Shift (Macro F1)

Figure 4: Comparison between TALLY and variants of two domain generalization approaches (CORAL, MixStyle), where we replace the losses of them with class re-weighting or re-sampling ones.

E.2 How do prototypes benefit invariant learning?

We analyze the effects of prototypes in alleviating domain-associated nuisances. Specifically, we compare TALLY with three variants: (1) without using any prototype information (None); (2) only applying class prototype (C Only); (3) only applying class-agnostic nuisances (D Only). We report the results in Figure 5. We observe that adding class prototype does improve the performance, especially the worst-domain accuracy in Officehome-LT. The class-agnostics domain factors also benefits the performance to some extent. In summary, TALLY outperforms its variants, demonstrating the effectiveness of prototype representation in mitigating domain-associated nuisances.



(a) Subpopulation Shift (b) Domain Shift

Figure 5: Analysis of prototype-guided invariant learning. C Only and D Only represent only using class prototype representation or class-agnostic domain factors, respectively.

E.3 Does TALLY lead to stronger domain invariance?

We analyze and compare the domain invariance of classifiers trained by ERM, TALLY, and other invariant learning approaches. Following [39, 40], we measure the lack of domain invariance as the accuracy of domain prediction (I_{acc}) and as the pairwise divergence of unscaled logits (I_{kl}). Specifically, for the accuracy of domain prediction, we perform logistic regression on top of the unscaled logits to predict the domain. For the pairwise divergence, we use kernel density estimation to estimate the probability density function $P(h^{c,d})$ of logits from domain-class pair (c, d) and calculate the KL divergence of the distribution of logits from different pairs. Formally, I_{kl} is defined as $\mathrm{I}_{kl} = \frac{1}{|\mathcal{C}||\mathcal{D}|^2} \sum_{c\in\mathcal{C}} \sum_{d',d\in\mathcal{D}} \mathrm{KL}(P(h^{c,d})|P(h^{c,d'}))$. We report the results of Officehome-LT and

Table 11: Invariance Analysis of TALLY. OH-LT and DN-LT represents Officehome-LT and DomainNet-LT, respectively.

Model	OH-LT		DN-LT	
	$\mathrm{I}_{acc} \downarrow$	$\mathrm{I}_{kl} \downarrow$	$\mathrm{I}_{acc} \downarrow$	$\mathrm{I}_{kl} \downarrow$
ERM	46.35%	2.030	70.00%	4.852
MixStyle	44.42%	2.169	67.11%	5.661
CORAL	42.21%	1.248	66.79%	4.593
BODA	40.10%	2.052	65.15%	6.810
TALLY	**39.52%**	**1.179**	**63.80%**	**3.956**

DomainNet-LT in Table 11. Smaller I_{acc} and I_{kl} values indicate more invariant representations with respect to the labels. The results show that TALLY does lead to greater domain-invariance compared to prior invariant learning approaches (e.g., BODA).

E.4 Analysis of Sampling Strategies

Finally, we compare the proposed selective balanced sampling in TALLY with domain-class balanced sampling. For an example pair $(x_i,\ y_i,\ d_i)$ and $(x_i,\ y_j,\ d_j)$, selective balanced sampling gets $y_i \sim \mathrm{Uniform}(\mathcal{C})$ and $d_j \sim \mathrm{Uniform}(\mathcal{D})$, while traditional balanced sampling get $(y_i, d_i), (y_j, d_j) \sim \mathrm{Uniform}(\mathcal{C}, \mathcal{D})$. The results of subpopulation shifts in OfficeHome-LT, DomainNet-LT and of domain shifts (Macro-F1) in TerraInc, iWildCam are reported in Table 12, indicating the effectiveness of selective balanced sampling in transferring knowledge over domains and classes.

Table 12: Comparison between sampling strategies.

	OfficeHome-LT		DomainNet-LT	
	Avg.	Worst	Avg.	Worst
Balanced Sampling	65.03 ± 0.91%	58.33 ± 0.24%	49.35 ± 0.21%	27.97 ± 0.25%
TALLY(Selective))	**67.00 ± 0.47%**	**60.45 ± 0.09%**	**50.15 ± 0.46%**	**29.55 ± 0.19%**
	TerraInc		iWildCam	
	Macro F1	Acc	Macro F1	Acc
Balanced Sampling	44.79 ± 0.62%	57.78 ± 0.36%	33.1 ± 0.4%	71.6 ± 1.8%
TALLY(Selective))	**46.23 ± 0.56%**	**59.89 ± 1.32%**	**34.4 ± 0.4%**	73.4 ± 1.8%

F Results on Standard Domain Generalization Benchmarks

In this section, we present the additional comparison on standard domain generalization benchmarks. Notice that the data distributions in these standard benchmarks are not long-tailed, which is thus not our focus in this paper. The goal is to compare our approach with other domain generalization methods. In Table 13-16, we present results on four standard benchmarks: VLCS, PACS, OfficeHome, DomainNet, respectively. Results for all algorithms except TALLY are directly copied from [12] and [39]. In Table 17, we summarize all results and show the comparison between different approaches. According to the results, TALLY can achieve comparable performance compared with state-of-the-art domain generalization approaches.

Table 13: Comparison on the standard VLCS benchmark.

	Caltech101	LabelMe	SUN09	VOC2007	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	<u>66.1 ± 1.2</u>	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	<u>78.6</u>
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	<u>98.7 ± 0.2</u>	63.6 ± 0.7	71.3 ± 1.2	76.7 ± 0.6	77.6
VREx	98.4 ± 0.3	64.4 ± 1.4	<u>74.1 ± 0.4</u>	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**	<u>78.0 ± 0.6</u>	78.5
TALLY (ours)	97.5 ± 0.5	**67.2 ± 1.1**	73.8 ± 0.5	**79.2 ± 0.9**	**78.8**

Table 14: Comparison on the standard PACS benchmark.

	Art painting	Cartoon	Photo	Sketch	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	<u>97.6 ± 0.1</u>	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	<u>88.3 ± 0.2</u>	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	83.6
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	<u>97.6 ± 0.3</u>	78.2 ± 1.2	85.2
BODA	88.2 ± 0.2	**81.7 ± 0.3**	**97.8 ± 0.2**	<u>80.2 ± 0.3</u>	<u>86.9</u>
TALLY (ours)	**89.5 ± 0.8**	<u>81.2 ± 0.7</u>	97.0 ± 0.1	**81.7 ± 0.9**	**87.4**

Table 15: Comparison on the standard OfficeHome benchmark.

	Art	Clipart	Product	Real	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**	<u>77.1 ± 0.1</u>	**79.5 ± 0.3**	**69.3**
TALLY (ours)	<u>64.2 ± 0.5</u>	<u>55.1 ± 0.8</u>	**78.0 ± 1.1**	<u>79.2 ± 0.5</u>	<u>69.1</u>

Table 16: Comparison on the standard DomainNet benchmark.

	Sketch	Infograph	Painting	Quickdraw	Real	Clipart	Avg
ERM	49.8 ± 0.4	18.8 ± 0.3	46.7 ± 0.3	12.2 ± 0.4	59.6 ± 0.1	58.1 ± 0.3	40.9
IRM	42.3 ± 3.1	15.0 ± 1.5	38.3 ± 4.3	10.9 ± 0.5	48.2 ± 5.2	48.5 ± 2.8	33.9
GroupDRO	40.1 ± 0.6	17.5 ± 0.4	33.8 ± 0.5	9.3 ± 0.3	51.6 ± 0.4	47.2 ± 0.5	33.3
Mixup	48.2 ± 0.5	18.5 ± 0.5	44.3 ± 0.5	12.5 ± 0.4	55.8 ± 0.3	55.7 ± 0.3	39.2
MLDG	50.2 ± 0.4	19.1 ± 0.3	45.8 ± 0.7	13.4 ± 0.3	59.6 ± 0.2	59.1 ± 0.2	41.2
CORAL	50.1 ± 0.6	19.7 ± 0.2	46.6 ± 0.3	13.4 ± 0.4	59.8 ± 0.2	59.2 ± 0.1	41.5
MMD	28.9 ± 11.9	11.0 ± 4.6	26.8 ± 11.3	8.7 ± 2.1	32.7 ± 13.8	32.1 ± 13.3	23.4
DANN	46.8 ± 0.6	18.3 ± 0.1	44.2 ± 0.7	11.8 ± 0.1	55.5 ± 0.4	53.1 ± 0.2	38.3
CDANN	45.9 ± 0.5	17.3 ± 0.1	43.7 ± 0.9	12.1 ± 0.7	56.2 ± 0.4	54.6 ± 0.4	38.3
MTL	49.2 ± 0.1	18.5 ± 0.4	46.0 ± 0.1	12.5 ± 0.1	59.5 ± 0.3	57.9 ± 0.5	40.6
SagNet	48.8 ± 0.2	19.0 ± 0.2	45.3 ± 0.3	12.7 ± 0.5	58.1 ± 0.5	57.7 ± 0.3	40.3
ARM	43.5 ± 0.4	16.3 ± 0.5	40.9 ± 1.1	9.4 ± 0.1	53.4 ± 0.4	49.7 ± 0.3	35.5
VREx	42.0 ± 3.0	16.0 ± 1.5	35.8 ± 4.6	10.9 ± 0.3	49.6 ± 4.9	47.3 ± 3.5	33.6
RSC	47.8 ± 0.9	18.3 ± 0.5	44.4 ± 0.6	12.2 ± 0.2	55.7 ± 0.7	55.0 ± 1.2	38.9
BODA	**51.3 ± 0.3**	**20.5 ± 0.7**	**48.0 ± 0.1**	<u>13.8 ± 0.6</u>	**60.6 ± 0.4**	**62.1 ± 0.4**	**42.7**
TALLY (ours)	<u>50.5 ± 0.2</u>	<u>19.7 ± 0.1</u>	<u>47.7 ± 0.6</u>	**14.1 ± 0.3**	<u>60.0 ± 0.2</u>	<u>60.1 ± 0.5</u>	<u>42.0</u>

Table 17: Domain shift results over all four benchmarks.

	VLCS	PACS	OfficeHome	DomainNet	Avg
ERM	77.5 ± 0.4	85.5 ± 0.2	66.5 ± 0.3	40.9 ± 0.1	67.6
IRM	78.5 ± 0.5	83.5 ± 0.8	64.3 ± 2.2	33.9 ± 2.8	65.1
GroupDRO	76.7 ± 0.6	84.4 ± 0.8	66.0 ± 0.7	33.3 ± 0.2	65.1
Mixup	77.4 ± 0.6	84.6 ± 0.6	68.1 ± 0.3	39.2 ± 0.1	67.3
MLDG	77.2 ± 0.4	84.9 ± 1.0	66.8 ± 0.6	41.2 ± 0.1	67.5
CORAL	**78.8 ± 0.6**	86.2 ± 0.3	68.7 ± 0.3	41.5 ± 0.1	68.8
MMD	77.5 ± 0.9	84.6 ± 0.5	66.3 ± 0.1	23.4 ± 9.5	63.0
DANN	<u>78.6 ± 0.4</u>	83.6 ± 0.4	65.9 ± 0.6	38.3 ± 0.1	66.6
CDANN	77.5 ± 0.1	82.6 ± 0.9	65.8 ± 1.3	38.3 ± 0.3	66.3
MTL	77.2 ± 0.4	84.6 ± 0.5	66.4 ± 0.5	40.6 ± 0.1	67.2
SagNet	77.8 ± 0.5	86.3 ± 0.2	68.1 ± 0.1	40.3 ± 0.1	68.1
ARM	77.6 ± 0.3	85.1 ± 0.4	64.8 ± 0.3	35.5 ± 0.2	65.8
VREx	78.3 ± 0.2	84.9 ± 0.6	66.4 ± 0.6	33.6 ± 2.9	65.8
RSC	77.1 ± 0.5	85.2 ± 0.9	65.5 ± 0.9	38.9 ± 0.5	66.7
BODA	78.5 ± 0.3	<u>86.9 ± 0.4</u>	**69.3 ± 0.1**	**42.7 ± 0.1**	**69.4**
TALLY (ours)	**78.8 ± 0.4**	**87.4 ± 0.2**	<u>69.1 ± 0.4</u>	<u>42.0 ± 0.1</u>	<u>69.3</u>