Are all classes created equal? Domain Generalization for Domain-Linked Classes

Kimathi Kaai[1,2] **Saad Hossain**[1] **Sirisha Rambhatla**[1]
[1]Critical ML Lab, University of Waterloo
[2]Vision and Image Processing Lab, University of Waterloo
{kkaai, s42hossain, sirisha.rambhatla}@uwaterloo.ca

Abstract

Domain generalization (DG) focuses on transferring domain-invariant knowledge from multiple source domains (available at train time) to an *a priori* unseen target domain(s). This task implicitly assumes that a class of interest is expressed in multiple source domains (*domain-shared*), which helps break the spurious correlations between domain and class and enables domain-invariant learning. However, we observe that this results in extremely poor generalization performance for classes only expressed in a specific domain (*domain-linked*). To this end, we develop a contrastive and fairness based algorithm – FOND – to learn generalizable representations for these domain-linked classes by transferring useful representations from domain-shared classes. We perform rigorous experiments against popular baselines across benchmark datasets to demonstrate that given a sufficient number of domain-shared classes FOND achieves SOTA results for domain-linked DG.

1 Introduction

Domain generalization (DG) aims to learn discriminative representations that can generalize to data distributions (domains) different from those observed during training, i.e. *out-of-distribution*. Specifically, here the target domain is assumed to be *unseen* during training. Given this goal, the guiding principle in modern DG algorithms is to learn representations that are invariant to source domains, and hence generalizable to unseen targets [Ye et al., 2021].

As a result, recent works aim to explicitly reduce the representation discrepancy between multiple source-domains [Zhou et al., 2023]. However, these methods rely on classes being observed in multiple source-domains and/or focus only on the overall accuracy. In the real-world however, classes of interest may often be observed in a specific domain (*domain-linked*, $\mathcal{Y}_L$), setting it apart from those observed in multiple domains (*domain-shared*, $\mathcal{Y}_S$).These lead to generalization challenges in applications including, healthcare [Chen et al., 2021], autonomous driving [Piva et al., 2023], and fraud detection [Ataabadi et al., 2022], where classes/anomalies of interest may *only* have been observed in particular demographics, regions etc.

Models which aim to utilize domain-linked data often encounter spurious correlations between the domain and the class [Lynch et al., 2023, Zhang et al., 2022a]. Unsurprisingly, existing DG approaches yield large performance discrepancies between $\mathcal{Y}_L$ and $\mathcal{Y}_S$ classes; see Fig. 1. Therefore, we seek to specifically improve the generalizability of domain-linked classes. This task, to the best of our knowledge has not been extensively studied in the literature.

Notwithstanding these challenges, recent advances in real-world machine learning draw from the success of pretraining with classes/objectives different from downstream tasks [He et al., 2022]. This begs the question – *can we transfer useful representations from domain-shared classes to domain-linked classes?* We answer this in the affirmative and propose FOND, (***F**air and c**ON**trastive **D**omain-linked learning*). Specifically, we draw insights from recent works on *fairness* [Pham et al., 2023, Makhlouf et al., 2021, Wang et al., 2020] to learn generalizable representations from $\mathcal{Y}_S$ for

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

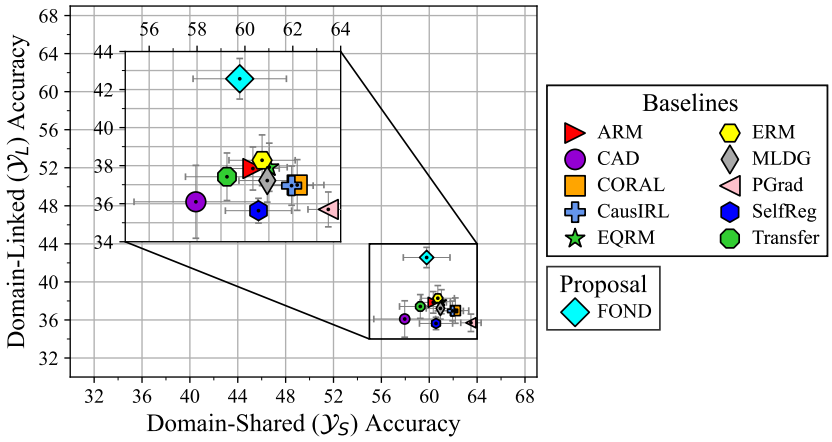


Figure 1: **Performance discrepancies between domain-linked and domain-shared classes.** Values are averaged across datasets (PACS, OfficeHome, VLCS) and shared-class settings (*High*, *Low*).

$\mathcal{Y}_L$ classes. Note that this is different from the goal of classic DG fairness – achieve similar outcomes across protected attributes (e.g., gender). On the contrary, we use fairness to ensure that the model can be moved away from learning domain-specific features that can lead to spurious correlations for $\mathcal{Y}_L$. To complement this objective, we develop a contrastive learning objective that regularizes the relationships between same-class-inter-domain and different-class-intra-domain training samples.

We rigorously evaluate `FOND` on three standard DG benchmark datasets – PACS [Li et al., 2017], VLCS [Fang et al., 2013], and, OfficeHome [Venkateswara et al., 2017] – across ten DG baselines, include the SOTA [Wang et al., 2023, Eastwood et al., 2022]. We find that indeed domain-linked class performance improves with the presence of a high enough number of domain-shared classes, thus accomplishing domain-invariant representation learning. We observe that that ERM is still a strong baseline, but `FOND` achieve a remarkable overall performance improvement of $+9.3$ over ERM on an average (26.9% improvement), with a gain of 39.2% on VLCS. This demonstrates that even observing other classes in diverse domains can immensely help domain-linked classes!

2 Related works

We briefly introduce DG works related to this paper and identify the addressable research gap.

Data manipulation techniques primarily focus on data augmentation and generation techniques. Typical augmentations include affine transformations in conjunction with additive noise, cropping and so on [Shorten and Khoshgoftaar, 2019]. Other methods include simulations [Yue et al., 2019, Tremblay et al., 2018], gradient-based perturbations like CrossGrad [Shankar et al., 2018], and image mixing [Mancini et al., 2020, Zhang et al., 2018, Shu et al., 2021]. Furthermore, generative models are also popular techniques for diverse data generation [Zhou et al., 2020, Somavarapu et al., 2020]. Since a model's generalizability is a function of training data diversity [Vapnik, 2000], `FOND` and other approaches should be used in conjunction.

Multi-domain feature alignment techniques primarily align features across source-domains through explicit feature/distribution alignment [Wang et al., 2023, Rame et al., 2022, Eastwood et al., 2022, Chevalley et al., 2022, Sun and Saenko, 2016, Peng et al., 2019, Hu et al., 2019, Zhou et al., 2021, Wang et al., 2021]. Other approaches perform domain-discriminative adversarial training [Zhu et al., 2022, Yang et al., 2021, Shao et al., 2019, Li et al., 2018b, Gong et al., 2018]. Additionally, meta-learning approaches improve find success imitating the generalization tasks through meta-train and meta-test objectives [Qin et al., 2023, Zhang et al., 2021a,b, Zhong et al., 2022].

Contrastive learning methods aim to learn representations, such that similar samples are embedded close to each other while distancing dissimilar samples [Huang et al., 2020, Ruan et al., 2022, Kim et al., 2021, Khosla et al., 2020, Chen et al., 2020b]. These methods focus on multi-domain comparisons to identify domain-invariant representations [Zhou et al., 2023].

Fairness notions in DG [Makhlouf et al., 2021] involve reducing the performance discrepancy between protected attributes (e.g. gender) [Pham et al., 2023, Wang et al., 2020]. While the goals may be different, fairness offers an avenue to learn useful representations from domain-shared classes.

Research Gap. Existing DG approaches presuppose all classes are expressed in multiple domains and/or seek to maximize average generalization; thus ignoring the large performance discrepancies between domain-linked and domain-shared classes. Since these domain-linked classes may be of interest in real-world settings, there is a need to understand the factors that impact their performance, and to build models which can generalize to unseen domains.

3 Problem formulation

We formalize some definitions for the purposes of the paper. First, we define a domain as follows.

Definition 3.1 (Domain)**.** *Let $\mathcal{X}$ denote an nonempty input space (e.g. images, text, etc) and $\mathcal{Y}$ an output label space. We denote as specific domain as $S = \{(\mathbf{x}_j, y_j)\}_{j=1}^{n} \sim \mathcal{D}^S : \mathcal{X}^S \times \mathcal{Y}^S$, where $\mathbf{x} \in \mathcal{X}^S \subseteq \mathbb{R}^d$ and $y \in \mathcal{Y}^S \subset \mathbb{Z}$.*

Given this definition of a domain, the DG task – which entails learning representations from multiple source-domains to generalize to unseen target-domain(s) – can be formalized as shown below.

Definition 3.2 (Domain generalization)**.** *Given K training (source) domains $\mathcal{S} = \{S^i \mid i = 1, ..., K\}$ where $S^i = \{(\mathbf{x}_j^i, y_j^i)\}_{j=1}^{n_i}$ denotes the i-th source domain with n_i samples, and the joint distributions between each pair of domains are different: $\mathcal{D}^{S^i} \neq \mathcal{D}^{S^j} : 1 \leq i \neq j \leq K$. Then the goal is to learn a predictive function from $\mathcal{S}$ for reliable performance on an unseen, out-of-distribution target-domain $T \sim \mathcal{D}^T : \mathcal{X}^T \times \mathcal{Y}^T$ (i.e. $\mathcal{D}^T \neq \mathcal{D}^{S^i}$ for $i \in \{1, ..., K\}$).*

We evaluated methods for *closed-set* domain generalization (i.e. $\mathcal{Y}^T = \bigcup_{i=1}^{K} \mathcal{Y}^{S^i}$) where no source-domain expresses all target classes. During training there exists a set of classes expressed in only one source-domain (i.e. *domain-linked* classes $\mathcal{Y}_L$) and in multiple domains (i.e. *domain-shared* classes $\mathcal{Y}_L$). Note, $\mathcal{Y}^T = \mathcal{Y}_L \cup \mathcal{Y}_S$ and $\mathcal{Y}_L \cap \mathcal{Y}_S = \emptyset$.

4 Methodology

We now introduce the FOND objective, which learns domain-invariant representations to improve domain-linked $\mathcal{Y}_L$ class generalization. We achieve this by minimizing the following objective:

$$\mathcal{L}_{\texttt{FOND}} = \mathcal{L}_{task} + \lambda_{xdom} \cdot \mathcal{L}_{xdom} + \lambda_{fair} \cdot \mathcal{L}_{fair}. \tag{1}$$

Here, $\mathcal{L}_{xdom}$ encourages domain-invariant representation learning by focusing on negative (different-class) intra-domain comparisons (β), and positive inter-domain pairwise comparisons (α):

$$\mathcal{L}_{xdom} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\alpha \cdot \exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{j \in I \setminus \{i\}} \beta \cdot \exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}, \tag{2}$$

$$\alpha = \begin{cases} a, & S(\mathbf{z}_i) \neq S(\mathbf{z}_p), \text{ where } a \geq 1 \\ 1, & \text{otherwise} \end{cases}, \beta = \begin{cases} b, & S(\mathbf{z}_i) = S(\mathbf{z}_j), y_i \neq y_j, \text{ where } b \geq 1 \\ 1, & \text{otherwise} \end{cases}$$

FOND introduces α which is motivated from works which aim to maximize mutual information between positive (same-class) inter-domain samples guides domain-invariant learning [Chen et al., 2020a, Ren et al., 2023]. While β encodes "hard negative mining" which has found success in representation learning [Robinson et al., 2021, Zhang et al., 2022b, Liu et al., 2023]. Here, $S(\mathbf{z}_i)$ denotes the domain $S \in \mathcal{S}$ that $\mathbf{z}_i$ belongs to, and a, b are hyper-parameters.

We propose to leverage $\mathcal{L}_{fair}$ to encourage the model to learn $\mathcal{Y}_S$ features from classes that are *also* generalizable for $\mathcal{Y}_L$ classes. Specifically, we aim to make the prediction *error rate* between domain-linked and domain-shared classes similar. We accomplish this by penalizing the violation of the the absolute difference between their classification losses as follows, where $\mathcal{L}_{task}$ is cross-entropy,

$$\mathcal{L}_{fair} = |\mathcal{L}_{task}^L - \mathcal{L}_{task}^S|, \text{ where, } \mathcal{L}_{task} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{S}} \left[-\mathbf{y} \cdot \log\left(M(\mathbf{x})\right)\right]. \tag{3}$$

5 Experiments

We specifically focus on analyzing domain-linked class performance under varying 1) shared-class distribution settings (Sec. 5), 2) inter-domain variation types, and 3) number of classes. For rigorous, fair and reproducible evaluation, we mirrored the DomainBed [Gulrajani and Lopez-Paz, 2021] test-bed to be consistent with the experimentation in DG literature.

Table 1: *High* setting out-of-distribution test accuracy (%) for $\mathcal{Y}_L$ classes.

Algorithm	VLCS	PACS	OfficeHome	**Average**	**% Improvement**
ERM	51.8 ± 3.3	14.7 ± 2.2	37.5 ± 0.6	34.6	*(0.0%)*
CORAL [Sun and Saenko, 2016]	49.8 ± 4.2	13.7 ± 1.0	38.9 ± 0.2	34.1	*(-1.4%)*
MLDG [Li et al., 2018a]	45.2 ± 3.4	13.8 ± 0.5	37.4 ± 0.7	32.1	*(-7.2%)*
ARM [Zhang et al., 2021b]	49.0 ± 1.4	16.2 ± 2.9	38.4 ± 0.2	34.5	*(-0.3%)*
SelfReg [Kim et al., 2021]	41.9 ± 0.2	13.4 ± 1.2	39.5 ± 0.6	31.6	*(-8.7%)*
CAD [Ruan et al., 2022]	51.7 ± 5.8	13.1 ± 0.7	36.4 ± 1.4	33.7	*(-2.6%)*
Transfer [Zhang et al., 2021a]	48.9 ± 3.0	16.0 ± 1.6	36.8 ± 0.2	33.9	*(-2.0%)*
CausIRL [Chevalley et al., 2022]	48.9 ± 2.5	13.3 ± 1.5	39.2 ± 0.2	33.8	*(-2.3%)*
EQRM [Eastwood et al., 2022]	45.4 ± 3.5	17.9 ± 2.0	37.8 ± 0.1	33.7	*(-2.6%)*
PGrad [Wang et al., 2023]	40.2 ± 1.8	12.6 ± 1.4	39.0 ± 0.8	30.6	*(-11.6%)*
FOND	**72.1 ± 3.5**	**19.1 ± 0.6**	**40.6 ± 0.4**	**43.9**	*(+26.9%)*

Table 2: *Low* setting out-of-distribution test accuracy (%) for $\mathcal{Y}_L$ classes.

Algorithm	**VLCS**	**PACS**	**OfficeHome**	**Average**	**% Improvement**
ERM	50.7 ± 1.0	36.5 ± 0.5	38.5 ± 0.4	41.9	*(0.0)*
CORAL [Sun and Saenko, 2016]	45.5 ± 1.6	33.3 ± 0.8	40.7 ± 0.2	39.8	*(-5.0)*
MLDG [Li et al., 2018a]	**50.8 ± 2.0**	**38.0 ± 0.1**	38.1 ± 0.1	**42.3**	*(+1.0)*
ARM [Zhang et al., 2021b]	47.7 ± 0.9	36.8 ± 1.3	39.0 ± 0.1	41.2	*(-1.7)*
SelfReg Kim et al. [2021]	46.6 ± 1.2	32.4 ± 0.4	40.0 ± 0.3	39.6	*(-5.5)*
CAD [Ruan et al., 2022]	45.5 ± 1.5	33.0 ± 0.9	36.9 ± 1.2	38.5	*(-8.1)*
Transfer [Zhang et al., 2021a]	48.3 ± 0.6	36.4 ± 1.8	38.1 ± 0.3	40.9	*(-2.4)*
CausIRL [Chevalley et al., 2022]	45.8 ± 0.9	33.6 ± 0.9	**40.9 ± 0.2**	40.1 7	*(-4.3)*
EQRM [Eastwood et al., 2022]	49.1 ± 1.1	37.4 ± 0.7	39.9 ± 0.2	42.1	*(+0.5)*
PGrad [Wang et al., 2023]	49.0 ± 0.7	34.4 ± 0.5	39.0 ± 0.3	40.8	*(-2.6)*
FOND	48.0 ± 0.4	35.3 ± 1.2	40.3 ± 0.3	41.2	*(-1.7)*

Datasets. To evaluated $\mathcal{Y}_L$ performance with respect to inter-domain variations and the number of classes, we required keeping consistent a) dataset sizes and b) number of domains. Therefore we chose DG literature gold standard datasets [Zhou et al., 2023]: PACS, VLCS, and OfficeHome.

Defining Shared-Class Distribution Settings. We define two shared-class distribution settings – *Low* and *High* – denoting the relative number of shared classes $|\mathcal{Y}_S|$ with respect to the total $|\mathcal{Y}_T|$. In the *Low* setting $\sim 1/3$ of the classes are domain-shared; $\sim 2/3$ in the *High* setting.

Baselines. Baselines were selected to cover a variety of well represented DG methodologies, that have been benchmarked against DomainBed [Gulrajani and Lopez-Paz, 2021].

Experimental results. We report the performance under the low-and high-shared setting in Table. 2 and Table. 1 respectively. The lack of domain-shared classes in the *Low* setting leads to poor performance across the board. `FOND` still demonstrates top-4 performance in the *Low* setting. In the *High* setting we observe that `FOND` consistently outperforms all baselines as shown in Table 1. Strikingly, `FOND` results in a 39% performance improvement over the best baseline (ERM) for VLCS. VLCS domains are real-world while for PACS and OfficeHome domain differences are more obvious.

6 Conclusion and Future Work

Domain generalization (DG) in real-world settings often suffers from data scarcity leading to classes which are only observed in specific domains, i.e. they are *domain-linked*. DG efforts primarily focus on improving the overall accuracy which can lead to critical failures in the real-world. Motivated from this challenge, we focused on robust out-of-distribution generalization for domain-linked classes and propose `FOND`; the first method for domain-linked domain generalization, which achieves state-of-the-art performance as compared to the current SOTA DG approaches. For future work we would like to theoretically analyze the factors impacting domain-linked class performances relative to domain-shared. Additionally, we seek to analyze the effects of different distribution shift scenarios and varying the absolute/relative number of domain-shared classes available. Developing methods that address *Low* setting can also be an independent line of research. Enabling transfer from different classes can help overcome the fundamental challenges of data scarcity and representation learning that DG faces today, opening an exciting avenue of research.

References

Parvin Esmaeili Ataabadi, Behzad Soleimani Neysiani, Mohammad Zahiri Nogorani, and Nazanin Mehraby. Semi-supervised medical insurance fraud detection by predicting indirect reductions rate using machine learning generalization capability. In *2022 8th International Conference on Web Research (ICWR)*, pages 176–182, 2022. doi: 10.1109/ICWR54782.2022.9786251.

Irene Y. Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4(1):123–144, 2021. doi: 10.1146/annurev-biodatasci-092820-114757. URL https://doi.org/10.1146/annurev-biodatasci-092820-114757.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020a. URL https://proceedings.mlr.press/v119/chen20j.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020b.

Mathieu Chevalley, Charlotte Bunne, Andreas Krause, and Stefan Bauer. Invariant causal mechanisms through distribution matching. *arXiv preprint arXiv:2206.11646*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

Cian Eastwood, Alexander Robey, Shashank Singh, Julius Von Kügelgen, Hamed Hassani, George J Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems*, 35:17340–17358, 2022.

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

Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2472–2481, 2018.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Shoubo Hu, Kun Zhang, Zhitang Chen, and Lai-Wan Chan. Domain generalization via multidomain discriminant analysis. *Uncertainty in artificial intelligence : proceedings of the ... conference. Conference on Uncertainty in Artificial Intelligence*, 35, 2019.

Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Daehee Kim, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9599–9608, 2021.

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.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018a. ISBN 978-1-57735-800-8.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex Chichung Kot. Domain generalization with adversarial feature learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018b.

Yue Liu, Xihong Yang, Sihang Zhou, Xinwang Liu, Zhen Wang, Ke Liang, Wenxuan Tu, Liang Li, Jingcan Duan, and Cancan Chen. Hard sample aware network for contrastive deep graph clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 8914–8922, 2023.

Aengus Lynch, Gbètondji JS Dovonon, Jean Kaddour, and Ricardo Silva. Spawrious: A benchmark for fine control of spurious correlation biases. *arXiv preprint arXiv:2303.05470*, 2023.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 58 (5):102642, 2021. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102642. URL `https://www.sciencedirect.com/science/article/pii/S0306457321001321`.

Massimiliano Mancini, Zeynep Akata, Elisa Ricci, and Barbara Caputo. Towards recognizing unseen categories in unseen domains. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII*, page 466–483, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58591-4. doi: 10.1007/978-3-030-58592-1_28. URL `https://doi.org/10.1007/978-3-030-58592-1_28`.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1406–1415, 2019.

Thai-Hoang Pham, Xueru Zhang, and Ping Zhang. Fairness and accuracy under domain generalization. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=jBEXnEMdNOL`.

Fabrizio J. Piva, Daan de Geus, and Gijs Dubbelman. Empirical generalization study: Unsupervised domain adaptation vs. domain generalization methods for semantic segmentation in the wild. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 499–508, January 2023.

Xiaorong Qin, Xinhang Song, and Shuqiang Jiang. Bi-level meta-learning for few-shot domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15900–15910, June 2023.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18347–18377. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/rame22a.html`.

Yuchen Ren, Zhendong Mao, Shancheng Fang, Yan Lu, Tong He, Hao Du, Yongdong Zhang, and Wanli Ouyang. Crossing the gap: Domain generalization for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2871–2880, June 2023.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations (ICLR)*, 2021.

Yangjun Ruan, Yann Dubois, and Chris J. Maddison. Optimal representations for covariate shift. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=Rf58LPCwJj0.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Dx7fbCW.

Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10031, 2019.

Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:1–48, 2019.

Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9619–9628, 2021.

Nathan Somavarapu, Chih-Yao Ma, and Zsolt Kira. Frustratingly simple domain generalization via image stylization. *ArXiv*, abs/2006.11207, 2020.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 443–450, Cham, 2016. Springer International Publishing. ISBN 978-3-319-49409-8.

Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, V. Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1082–10828, 2018.

Vladimir Naumovich Vapnik. The nature of statistical learning theory. In *Statistics for Engineering and Information Science*, 2000.

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.

Junchang Wang, Yang Li, Liyan Xie, and Yao Xie. Class-conditioned domain generalization via wasserstein distributional robust optimization. *ArXiv*, abs/2109.03676, 2021.

Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020.

Zhe Wang, Jake Grigsby, and Yanjun Qi. Pgrad: Learning principal gradients for domain generalization. In *The Eleventh International Conference on Learning Representations*, 2023.

Fu-En Yang, Yuan-Chia Cheng, Zu-Yun Shiau, and Yu-Chiang Frank Wang. Adversarial teacher-student representation learning for domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, pages 19448–19460. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a2137a2ae8e39b5002a3f8909ecb88fe-Paper.pdf.

Hao-Tong Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang. Towards a theoretical framework of out-of-distribution generalization. In *Neural Information Processing Systems*, 2021.

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

Guojun Zhang, Han Zhao, Yaoliang Yu, and Pascal Poupart. Quantifying and improving transferability in domain generalization. In *Advances in Neural Information Processing Systems*, 2021a. URL `https://openreview.net/forum?id=SQqKl8I6xD8`.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=r1Ddp1-Rb`.

Marvin Mengxin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021b. URL `https://openreview.net/forum?id=-zgb2v8vV_w`.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In *International Conference on Machine Learning*, pages 26484–26516. PMLR, 2022a.

Shaofeng Zhang, Meng Liu, Junchi Yan, Hengrui Zhang, Lingxiao Huang, Xiaokang Yang, and Pinyan Lu. M-mix: Generating hard negatives via multi-sample mixing for contrastive learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2461–2470, 2022b.

Tao Zhong, Zhixiang Chi, Li Gu, Yang Wang, Yuanhao Yu, and Jin Tang. Meta-dmoe: Adapting to domain shift by meta-distillation from mixture-of-experts. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22243–22257. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/8bd4f1dbc7a70c6b80ce81b8b4fdc0b2-Paper-Conference.pdf`.

Fan Zhou, Zhuqing Jiang, Changjian Shui, Boyu Wang, and Brahim Chaib-draa. Domain generalization via optimal transport with metric similarity learning. *Neurocomputing*, 456:469–480, 2021.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 561–578, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58517-4.

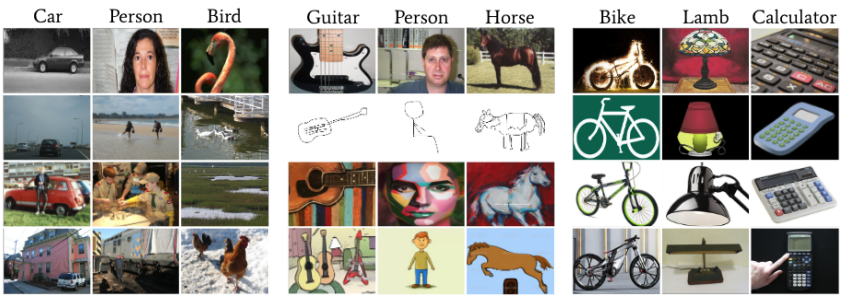
Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2023. doi: 10.1109/TPAMI.2022.3195549.

Wei Zhu, Le Lu, Jing Xiao, Mei Han, Jiebo Luo, and Adam P. Harrison. Localized adversarial domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7108–7118, June 2022.

A Experimental setup

A.1 Implementation Details

For consistency, all algorithms have a fine-tuned ResNet-18 backbone [He et al., 2016] pre-trained on ImageNet [Deng et al., 2009]. Additionally, the training data augmentations are: random size crops and aspect ratios, resizing to 224 × 224 pixels, random horizontal flips, random color jitter, and random gray-scaling. Our experiments ran on different GPUs: NVIDIA RTXA6000, NVIDIA GeForce RTX2080.



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

Figure 2: **Visualizing domains (rows) and classes (columns) for evaluation dataset.** Note how PACS and OfficeHome exhibit more obvious domain variations than VLCS.

A.2 Hyper-Parameter Search

For each algorithm we perform five random search attempts over a joint distributions of all their hyper-parameters. The performance of each hyper-parameter is evaluated using the strategy outlined in App. A.3. This is repeated for each of the five sets of hyper-parameters and the set maximizing the average domain-linked $\mathcal{Y}_L$ accuracy is selected. This search is performed for across three different seeds where all hyper-parameters are optimized anew for each algorithm, dataset and partial-overlap setting. The hyper-parameter search space for each algorithm is provided in the attached code.

A.3 Model Selection

Given K domains, we train K models, sharing the same hyper-parameters θ, but each model holds a different domain out. During the training of each model, 80% of the training data from each domain is used for training and the other 20% is used to determine the version that will be evaluated. We evaluate each model on its held-out domain data, and average the $\mathcal{Y}_L$ accuracy of these K models over their held-out domains. This provides us with an estimate of the quality of a given set of hyper-parameters. This strategy was chosen because it aligns with the goal of maximizing expected performance under out-of-distribution domain-variation without picking the model using the out-of-distribution data. The $\mathcal{Y}_L$ accuracy performance across held-out domains.

A.4 Datasets

PACS [Li et al., 2017] is a 9,991-image dataset consisting of four domains corresponding to four different image styles: photo (P), art-painting (A), cartoon (C) and sketch (S). Each of the four domains hold seven object categories; refer to Fig. 2b

VLCS [Fang et al., 2013] is a 10,729-image dataset consisting of four domains corresponding to four different datasets: VOC2007 (V), LabelMe (L), Caltech101 (C) and SUN09 (S). Each of the four domains hold five object categories: bird, car, chair, dog and person; refer to Fig. 2a

OfficeHome [Venkateswara et al., 2017] is a 15,588-image dataset consisting of images of everyday objects organized into four domains; art-painting, clip-art, images without backgrounds and real-world photos. Each of the domains holds 65 object categories; refer to Fig. 2c

Table 3: **Domain-shared ($\mathcal{Y}_S$) classes for each dataset and shared-class distribution setting.** The left table defines the number of $\mathcal{Y}_S$ classes and the right displays the corresponding $\mathcal{Y}_L$ classes.

	$\lvert\mathcal{Y}_S\rvert/\lvert\mathcal{Y}_T\rvert$			$\mathcal{Y}_L$		
Setting	**PACS**	**OfficeHome**	**VLCS**	**PACS**	**OfficeHome**	**VLCS**
Low	3/7	25/65	2/5	{0,1,3,5,6}	{0-13, 30-34, 44-64}	{0,1,4}
High	5/7	50/65	4/5	{1,6}	{0-4, 44-64}	{1}