Multi-Task Learning with Self-Supervised Objectives can Improve Worst-Group Outcomes

Atharva Kulkarni*, **Lucio M. Dery***, **Amrith Setlur, Aditi Raghunathan,**
Ameet Talwalkar, and Graham Neubig
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{atharvak, ldery, asetlur, raditi, atalwalk, gneubig}@cs.cmu.edu

Abstract

In order to create machine learning systems that serve a variety of users well, it is important to not only achieve high performance on average but also ensure equitable outcomes across diverse groups. In this paper, we explore the potential of multi-task learning (MTL) with self-supervised objectives as a tool to address the challenge of group-wise fairness. We show that by regularizing the joint representation space during multi-tasking, we are able to obtain improvements on worst-group error. Through comprehensive experiments across NLP and CV datasets, we demonstrate that regularized multi-tasking with self-supervised learning competes favorably with state-of-the-art distributionally robust optimization methods. Our approach – without introducing data external to the end-task – improves worst-case group accuracy over empirical risk minimization by as much as $\sim 4\%$ on average in settings where group annotations are completely unavailable.

1 Introduction

Multi-tasking [1, 2, 3] with self-supervised learning-based auxiliary objectives (MTL-SSL) has proven to be a powerful tool for improving a model's aggregate performance on a desired end-task [4, 5, 6]. Congruently, as learned systems exert ever-increasing influence on the real world, it is paramount that they not only perform well on aggregate but also exhibit equitable outcomes across diverse subgroups characterized by attributes such as race [7, 8], gender [7, 9], and geographies [10, 11]. However, despite the strides made from leveraging SSL via multi-task learning and the widely acknowledged importance of equitable machine learning, previous work has neither examined the effect of MTL-SSL on worst-group outcomes nor sought to design effective strategies that leverage said approach to achieve more equitable outcomes. In this paper, we attempt to fill this gap.

Traditionally, the problem of poor group-based outcomes has been tackled *explicitly* via methods such as distributionally robust optimization (DRO) [12, 13, 14, 15]. In DRO, instead of finding parameters θ that minimize the expected risk ℓ over the training distribution (i.e., empirical risk minimization - ERM), the worst-case risk over a pre-defined set of distributions (the *uncertainty set*) is instead minimized. Since designing an appropriate uncertainty set generally requires *explicit* access to group annotations at training time (which are unavailable in many practical settings), several works have been proposed to tackle the challenges of worst-group generalization with little to no access to group annotations. These efforts often fall into two categories. The first category consists of methods that employ a two-stage training framework where examples with high loss are identified in the first stage and a second re-training stage appropriately refocuses on these examples [16, 17, 18, 19, 20, 21]. By

*Equal contribution.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

virtually requiring two passes over the data, these methods tend to increase training overhead. On the other hand, the second category involves the unsupervised discovery of latent groups [22, 23], which can increase complexity at training time and thus deter their wide usage. In contrast, this paper empirically demonstrates that MTL-SSL can *implicitly* improve the worst-group performance without incurring the overhead of two-stage training, costly acquisition of group labels, or the instantiation of complex models.

In general, poor group outcomes occur when a model relies on spurious features (the ones irrelevant to the task's group-based test accuracy) instead of core features (those necessary to robustly solve the task) during training. A carefully chosen auxiliary task would be one that, when multi-tasked with the end task, encourages a model to use core features as opposed to spurious ones. In this paper, we explore self-supervised learning tasks as candidates for multi-tasking. Our choice of MTL-SSL as an intervention for worst-case group performance is inspired by previous work like [24] that shows that transfer learning can improve model robustness. Moreover, given the widespread adoption of SSL objectives in both pre-training and multi-tasking to enhance average-case performance, it prompts the question of how this approach affects worst-case group performance. Taken together, it is reasonable to anticipate that a thoughtfully designed variant of MTL-SSL will similarly enhance worst-case accuracy. Note that the MTL-SSL as a counter-measure alone may not be effective if the model has enough capacity to use the spurious features for the end task and the core features for the self-supervised task. We, therefore, augment the MTL-SSL framework with ℓ_1 regularization on the model's final layer activation order to restrict capacity right before end-task prediction. Regularization causes the SSL task to act as a constraint, forcing the model (due to limited capacity) to ignore spurious features whilst focusing on core features in order to do well on both tasks.

We empirically validate our MTL-SSL intervention against worst-case group performance across experiments in natural language processing (NLP) and computer vision (CV). Through a battery of experiments, we demonstrate that MTL using the self-supervision-based pre-training objective with ℓ_1 regularization on a pre-trained model's final layer activations is competitive when pitted against state-of-the-art DRO approaches like JTT (Just-Train-Twice) [17] and BR-DRO (Bitrate-Constrained DRO) [15]. In settings where validation group annotations are available across 3 widely used datasets, we outperform JTT and BR-DRO on 3 and 2 tasks, respectively, using our regularized MTL approach with end-task data only. Our approach improves worst-case group accuracy over ERM (by as much as $\sim 4\%$) and JTT (by $\sim 1\%$) in settings where group annotations are entirely unavailable. We ablate our method across a wide range of settings: varying the choice of SSL auxiliary task, using only end-task data versus external data, and training with a pre-trained model versus training a model from scratch. Our results demonstrate that multi-tasking the end-task with a self-supervised learning objective can be a versatile and robust tool against poor group-based outcomes.

2 Experimental Details

Datasets. We conduct experiments across three primary datasets: Waterbirds [13], MultiNLI [25] and CivilComments [26, 27]. To relieve the burden of compute, we introduce a fourth dataset, which is a smaller, sub-sampled version of one of the original CivilComments dataset, which we call CivilComments-Small. Please see Appendix A.1 for more details about the datasets we explore.

Multitask Model and Training Details. Let $\mathbf{T}_{\text{prim}}$ denote the end-task and $\mathbf{T}_{\text{ssl}}$ denote the self-supervised auxiliary task. We follow the parameter sharing paradigm [28, 29] where both $\mathbf{T}_{\text{prim}}$ and $\mathbf{T}_{\text{ssl}}$ share the same model body, parameterized by θ_{base}. We instantiate task-specific heads, parameterized by θ_{prim} and θ_{ssl}, respectively. We introduce ℓ_1 regularization to the final layer activations, positioned just before the per-task prediction heads. This regularization constraint encourages sparsity in the shared representation, ultimately enhancing the alignment of core features between the two tasks and reducing the prominence of spurious features. To facilitate such constrained multi-task training, we adopt a task-heterogeneous batching scheme [30], where each parameter update is performed using aggregated gradients across tasks. Let h^{prim}, $h^{\text{ssl}} \in \mathbb{R}^d$ be the output representations generated by the base model, which are fed into their respective task-specific heads. Then, our final multi-task learning objective can be expressed in terms of equation 1. Note that whilst we optimize $\mathcal{L}_{\text{final}}$ we care only about improving worst-group error on $\mathbf{T}_{\text{prim}}$.

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{prim}} + w_{\text{ssl}} \cdot \mathcal{L}_{\text{trans}} + w_{\text{reg}} \left(\|h^{\text{prim}}\|_1 + \|h^{\text{ssl}}\|_1 \right) \tag{1}$$

We use the pre-trained $\text{BERT}_{\text{base}}$ [31] and ViT_{base} [32] as the shared base models for NLP and CV tasks, respectively. We leverage the base models' self-supervised pretraining objectives, namely, masked language modeling (MLM) and masked image modeling (MIM), for our auxiliary transfer task $\mathbf{T}_{\text{ssl}}$. Please see Appendix A.3 for more details about training and hyper-parameters.

Baseline Methods. We explore four baselines: Empirical Risk Minimization **(ERM)**, Just Train Twice **(JTT)** [17], Bit-rate Constrained DRO **(BR-DRO)**[15] and **Group-DRO** [13]. Appendix section A.4 contains a more detailed description of these baseline methods.

Evaluation Details. We evaluate all the methods and datasets based on worst-group accuracy. We adopt two model selection strategies: 1) ***Val-GP:*** Assuming access to group annotations in the validation data, we choose the model checkpoint that maximizes the worst-group accuracy on the validation set. 2) ***No-GP:*** We checkpoint models based on average validation accuracy, assuming no access to any group annotations.

3 Results And Discussion

3.1 SSL based Multi-Tasking is Competitive with Bespoke DRO Methods

Table 1: Regularized multitask learning consistently reduces the gap between ERM and groupDRO.

Method	Group Labels	Civilcomments	MNLI	Waterbirds
ERM	Val Only	$61.3_{2.0}$	$67.6_{1.2}$	$85.4_{1.4}$
JTT	Val Only	$67.8_{1.6}$	$67.5_{1.9}$	$85.9_{2.5}$
BR-DRO	Val Only	$\mathbf{68.9}_{0.7}$	$68.5_{0.8}$	$86.7_{1.3}$
ERM + MT + L1	Val Only	$68.2_{3.2}$	$\mathbf{69.7}_{1.5}$	$\mathbf{87.5}_{2.7}$
groupDRO (Upper Bound)	Train and Val	$69.9_{1.2}$	$78.0_{0.7}$	$93.9_{0.7}$

Table 1 details the performance comparison with previously proposed methods for worst-group generalization [*]. We consider GroupDRO as an upper bound on performance since it uses group annotations at training time. Our MTL-SSL approach outperforms JTT and BR-DRO on two datasets (MNLI and Waterbirds) while performing comparatively with BR-DRO on the CivilComments dataset. Given the competitive results in Table 1, we argue that our regularized MTL-SSL formulation is an attractive option over JTT and BR-DRO based on its simplicity. Unlike JTT, multi-task learning does not require training the model twice. Also, our approach is simple compared to BR-DRO, which requires jointly training the task model and an adversary model in a min-max fashion.

3.2 MTL-SSL improves Worst Group Performance even without Group Annotated Validation Data

While prior studies often assume access to group annotations in the validation set [17, 33], our focus is on scenarios where no such annotations are available. This is particularly relevant to practical tasks where obtaining group annotations for the validation set can be prohibitively expensive in terms of cost and human effort. Figure 1 summarizes our findings, illustrating performance across five random seeds. Without group annotations, the MTL-SSL approach outperforms JTT and achieves a $\sim 2\% - 3\%$ improvement over ERM. This improvement persists even when validation group annotations are introduced, but a more significant boost of $\sim 5\% - 15\%$ is observed when using group-labeled data. This boost is valuable for practitioners with resources to obtain at least some group annotations.

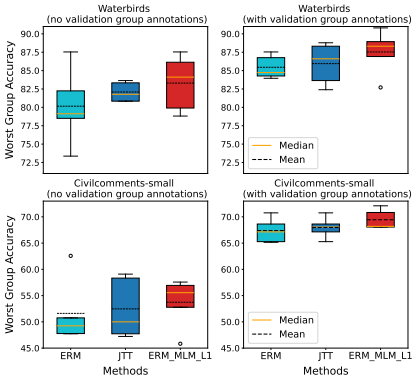

Figure 1: Regularized MTL-SSL boosts worst-group accuracy, regardless of group annotations.

[*]Values in tables are means from 5 random seeds with subscripts indicating standard deviations.

3.3 Are both Regularization and Multi-Tasking Jointly Necessary?

Table 2: Waterbirds Ablation

Method	Group Labels	
	None	Validation
ERM	$80.1_{4.6}$	$85.4_{1.4}$
+ L1	$82.0_{5.4}$	$86.4_{1.4}$
+ MIM	$80.1_{4.6}$	$85.3_{2.4}$
+ MIM + L1	$\mathbf{83.3}_{3.4}$	$\mathbf{87.5}_{2.7}$

Table 3: Civilcomments-small Ablation

Method	Group Labels	
	None	Validation
ERM	$51.6_{5.6}$	$67.4_{2.1}$
+ L1	$51.6_{4.0}$	$66.3_{1.6}$
+ MLM	$\mathbf{58.3}_{6.6}$	$68.5_{0.4}$
+ MLM + L1	$53.7_{4.3}$	$\mathbf{69.4}_{1.7}$

We conduct an ablation to verify if **both** ingredients are indeed necessary for improved worst-group accuracy. Our results are captured in Tables 2 and 3. Regularizing the final embedding space during ERM can result in worse performance compared to training via ERM solely ($66.3_{1.6}$ vs $67.4_{2.1}$ on CivilComments-small with validation group labels). On the other hand, multi-tasking without regularization can fail to improve over ERM, as evidenced by the experiments on Waterbirds. The regularized MTL-SSL approach stands out as the only configuration consistently enhancing performance on both datasets, whether with or without validation group annotations. We discuss more about the effect of w_{ssl} and w_{reg} in Appendix A.7.

3.3.1 Going Beyond the Pre-Training Objective

Table 4: Waterbirds: MIM $\rightarrow$ SimCLR

Method	Group Labels	
	None	Validation
ERM	$80.1_{4.6}$	$85.4_{1.4}$
+ MIM + L1	$83.3_{3.4}$	$\mathbf{87.5}_{2.7}$
+ SimCLR + L1	$\mathbf{84.0}_{3.4}$	$87.2_{1.6}$

Table 5: Civilcomments-small: MLM $\rightarrow$ CLM

Method	Group Labels	
	None	Validation
ERM	$51.6_{5.6}$	$67.4_{2.1}$
+ MLM + L1	$\mathbf{53.7}_{4.3}$	$\mathbf{69.4}_{1.7}$
+ CLM + L1	$50.9_{4.9}$	$67.3_{1.4}$

In this section, we explore other auxiliary objectives outside the model's original pre-training objective during multitasking. For the Waterbirds dataset, we experiment with SimCLR – a constrastive prediction task based on determining whether two distinct augmented images originate from the same base image [34]. For Civilcomments-small, we substitute the standard masked language modeling (MLM) task with causal language modeling (CLM) as the auxiliary task. Results are mixed when we deviate from the pre-training objective as the auxiliary task. From the results in Tables 4 and 5, we observe that SimCLR's performance closely resembles that of the MIM pre-training objective, whereas CLM shows relatively inferior results compared to MLM. We hypothesize that BERT's intrinsic bidirectional attention mechanism and non-autoregressive nature are not ideally suited for causal language modeling [35], resulting in the model underperforming on both the end-task and CLM. Given the sensitivity of model performance to the choice of the replacement objective, we proffer a practical recommendation to practitioners: use the pre-training objective as the auxiliary task. This is in line with recent work on best practices for fine-tuning pre-trained models [36]. We provide additional insights, such as the effect on average accuracy and introduction of external data, in Appendix A.5 and A.8.

4 Conclusion

This works presents an empirical investigation on how multi-tasking with self-supervised objective can improve worst-group performance. Specifically, we show that constraining the shared representations in the multi-tasking framework with ℓ_1 regularization limits the usage of spurious features, thereby improving generalization. Consequently, As a result, our MTL-SSL approach consistently outperforms the conventional ERM training paradigm and either improves upon or competes effectively with previously proposed methods for worst-group generalization, regardless of the availability of group annotations. These findings emphasize the versatility of our approach, providing a robust solution for enhancing model performance in challenging worst-case scenarios across various practical applications.

References

[1] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

[2] Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[3] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, pages 15–18, 2019.

[4] Lucio M Dery, Paul Michel, Ameet Talwalkar, and Graham Neubig. Should we be pre-training? an argument for end-task aware training as an alternative. *arXiv preprint arXiv:2109.07437*, 2021.

[5] Lucio M Dery, Yann Dauphin, and David Grangier. Auxiliary task update decomposition: The good, the bad and the neutral. *arXiv preprint arXiv:2108.11346*, 2021.

[6] Charlotte Loh, Thomas Christensen, Rumen Dangovski, Samuel Kim, and Marin Soljačić. Surrogate-and invariance-boosted contrastive learning for data-scarce applications in science. *Nature Communications*, 13(1):4223, 2022.

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[8] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR, 2021.

[9] Tejas Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 77–85, Seattle, Washington, July 2022. Association for Computational Linguistics.

[10] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 52–59, 2019.

[11] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021.

[12] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[13] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[14] Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. Examining and combating spurious features under distribution shift. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12857–12867. PMLR, 18–24 Jul 2021.

[15] Amrith Setlur, Don Dennis, Benjamin Eysenbach, Aditi Raghunathan, Chelsea Finn, Virginia Smith, and Sergey Levine. Bitrate-constrained DRO: Beyond worst case robustness to unknown group shifts. In *The Eleventh International Conference on Learning Representations*, 2023.

[16] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20673–20684. Curran Associates, Inc., 2020.

[17] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6781–6792. PMLR, 18–24 Jul 2021.

[18] 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 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 26484–26516. PMLR, 17–23 Jul 2022.

[19] Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *International Conference on Learning Representations*, 2022.

[20] Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28448–28467. PMLR, 23–29 Jul 2023.

[21] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, 2022.

[22] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19339–19352. Curran Associates, Inc., 2020.

[23] Bhargavi Paranjape, Pradeep Dasigi, Vivek Srikumar, Luke Zettlemoyer, and Hannaneh Hajishirzi. AGRO: Adversarial discovery of error-prone groups for robust optimization. In *The Eleventh International Conference on Learning Representations*, 2023.

[24] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International conference on machine learning*, pages 2712–2721. PMLR, 2019.

[25] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[26] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA, 2019. Association for Computing Machinery.

[27] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021.

[28] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[29] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[30] Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[32] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[33] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

[34] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[35] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: Masked sequence to sequence pre-training for language generation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR, 09–15 Jun 2019.

[36] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19338–19347, 2023.

[37] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[38] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[40] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022.

[41] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[42] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 336–351. PMLR, 11–13 Apr 2022.

[43] 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 (CVPR)*, June 2016.

[44] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[45] Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

[46] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori Hashimoto, Sergey Levine, Chelsea Finn, and Percy Liang. Extending the WILDS benchmark for unsupervised adaptation. In *International Conference on Learning Representations*, 2022.

[47] Karan Goel, Albert Gu, Yixuan Li, and Christopher Re. Model patching: Closing the subgroup performance gap with data augmentation. In *International Conference on Learning Representations*, 2021.

[48] 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.

[49] Kaidi Cao, Yining Chen, Junwei Lu, Nikos Arechiga, Adrien Gaidon, and Tengyu Ma. Heteroskedastic and imbalanced deep learning with adaptive regularization. In *International Conference on Learning Representations*, 2021.

[50] Yujia Bao, Shiyu Chang, and Regina Barzilay. Predict then interpolate: A simple algorithm to learn stable classifiers. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 640–650. PMLR, 18–24 Jul 2021.

[51] Yuge Shi, Jeffrey Seely, Philip Torr, Siddharth N, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *International Conference on Learning Representations*, 2022.

[52] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

[53] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 872–881. PMLR, 09–15 Jun 2019.

[54] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8346–8356. PMLR, 13–18 Jul 2020.

A Appendix

A.1 Datasets

Waterbirds: This image classification dataset was introduced by [13]. The task is to distinguish between species of land and water birds. It consists of bird images sourced from the CUB dataset [37] and superimposed on land or water backgrounds from the Places dataset [38]. The label (type of bird) is spuriously correlated with the background, resulting in 4 groups. Since this is a small dataset (4,795 train examples), we also use it for ablations.

MultiNLI: This is a natural language inference dataset. The task is to classify whether the second sentence is entailed by, contradicts, or is neutral with respect to the first sentence [25]. Following [13], we utilize the presence of negation words as a spurious attribute, leading to the creation of a total of 6 groups.

Civilcomments: The Civilcomments dataset is a toxicity classification dataset that contains comments from online forums [26, 27]. Along with the toxicity label, each text is annotated with additional overlapping sub-group labels of 8 demographic identities: male, female, LGBTQ, Christian, Muslim, other religions, Black, and White. As per [27] and [13], we define 16 overlapping groups by taking the Cartesian product of the binary toxicity label and each of the above eight demographic identities.

Civilcomments-small: As Civilcomments is a large dataset of about 448,000 data points, we create a sub-group stratified subset of 5% for conducting ablations and other detailed experiments. Our subset contains 13770, 2039, and 4866 data points in our train, validation, and test split, respectively.

A.2 Model Details

For $\mathbf{T}_{\text{prim}}$, we employ a classification head of a single-layer multi-layer perceptron (MLP). For $\mathbf{T}_{\text{aux}}$, we leverage the pre-trained MLM and MIM heads from BERT and ViT, respectively. We utilize the embedding of the [CLS] token from the base model through this MLP for classification. To ensure proper scaling, the L1 loss is normalized by the number of parameters in the shared representation.

A.3 Training Details

For training, we vary the fine-tuning learning rate within the ranges of $\{10^{-3}, 10^{-4}\}$ for Waterbirds, and $\{10^{-4}, 10^{-5}\}$ for the text datasets. We experiment with batch sizes in the set $\{4, 8, 16, 32\}$ for all the datasets. We use the same batch sizes for $\mathbf{T}_{\text{prim}}$ and $\mathbf{T}_{\text{trans}}$. We train for 50 epoch for the NLP datasets and 200 epochs for Waterbirds, with an early stopping patience of 10, as per the check-pointing scheme explained in section 2. We use the Adam optimizer for NLP datasets with decoupled weight decay regularization of 10^{-2} [39]. Consistent with the recent studies on ViT [32, 40], we use SGD with momentum of 0.9 [41] to fine-tune Waterbirds. We cross-validate w_{ssl} and w_{reg} from $\{e^{-1}, 1, e^{1}\}$. We run each hyperparameter configuration across 5 seeds and report the averaged results. We report the ERM, JTT, and groupDRO results for Civilcomments and MultiNLI from [42] as the authors conducted extensive hyperparameter tuning across all these methods. However, since [42] report results on Waterbirds using a ResNet-50 model [43] and our experiments employ ViT, we re-run all baselines using ViT with a consistent set of hyperparameters, as mentioned above.

A.4 Baseline Methods

Empirical Risk Minimization (ERM) This is the standard approach of minimizing the average loss over all the training data. No group information is used during training except when the **Val-GP** strategy is used for model selection.

Just Train Twice (JTT): JTT presents a two step approach for worst group generalization [17]. JTT first trains a standard ERM model for T epochs to identify misclassified data points. Then, a second model is trained on a reweighted dataset constructed by upweighting the misclassified examples by λ_{up}. It does not use group information during training except when the **Val-GP** strategy is used for model selection.

Bit-rate Constrained DRO (BR-DRO): Traditionally, in the two-player formulation of DRO (e.g., CVaR-DRO), the adversary can propose arbitrarily complex reweighting functions, which leads to overly pessimistic solutions for the game equilibrium. On the contrary, BR-DRO [15] restricts the complexity class of the adversary (max player), where the complexity is defined with an information-theoretic constraint under a data-independent prior. This leads to a weaker form of robustness since the game equilibrium does not come with performance guarantees for arbitrarily reweighted training points. On the other hand, the BR-DRO solution is also less pessimistic and is robust to "simpler" distribution shifts, where the shift is characterized by a reweighting function contained in a simpler complexity class. BR-DRO does not use group information during training except for the **Val-GP** setting.

Group-DRO: Group distributionally robust optimization minimizes the maximum loss across all the sub-groups [13]. This optimization method incorporates group annotations during training. Consequently, we regard its results as an upper bound for methods that do not leverage group labels during the training process.

A.5 Impact on Average Accuracy under Different Settings

A.5.1 Multi-tasking Improves Average Performance

Figure 2 illustrates the performance of ERM, JTT, and our approach concerning average accuracy. Constrained multi-tasking with the pre-training objective consistently improves performance across both datasets when validation group annotations are available. Therefore, when considering both Figure 1 and 2, it becomes evident that our method not only yields superior average performance but also excels in worst-case scenarios. In cases where group annotations are absent, our method achieves slightly lower results, though still exceeding 97%. However, juxtaposing the results in Figure 1 and 2, we see that our approach achieves a better trade-off between average and worst-case accuracy across both datasets. Thus, our empirical findings align with our theoretical investigation, demonstrating that our method effectively reduces spurious features in shared representations and enhances generalization across all sub-groups.

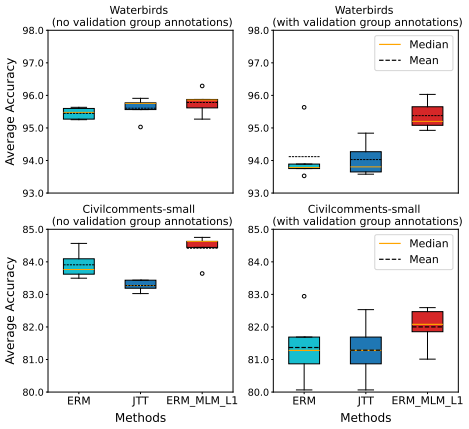

Figure 2: Regularized MTL improves average group accuracy when group annotations are available

A.5.2 Are both Regularization and Multi-Tasking Jointly Necessary?

Table 6: Waterbirds Ablation

Method	**Group Labels**	
	None	Validation
ERM	$95.5_{0.2}$	$94.1_{0.7}$
+ L1	$95.6_{0.3}$	$94.7_{0.9}$
+ MIM	$95.3_{0.4}$	$95.0_{0.6}$
+ MIM + L1	$\mathbf{95.8}_{0.3}$	$\mathbf{95.4}_{0.4}$

Table 7: Civilcomments-small Ablation

Method	**Group Labels**	
	None	Validation
ERM	$83.9_{0.4}$	$81.4_{1.0}$
+ L1	$83.7_{0.4}$	$80.3_{0.7}$
+ MLM	$83.9_{1.2}$	$81.1_{1.1}$
+ MLM + L1	$\mathbf{84.4}_{0.4}$	$\mathbf{82.0}_{0.5}$

Tables 6 and 7 present ablation studies focusing on average accuracy. Consistent with the findings in Section A.5.2, we observe that incorporating L1 regularization into multi-task learning with the pre-trained objective produces the highest average accuracy when validation annotations are accessible. In line with the theoretical findings, the synergy between L1 regularization and MLM/MIM results in stable trade-offs between average and worst-group accuracy.

Table 8: Waterbirds: MIM → SimCLR

Method	Group Labels	
	None	Validation
ERM	$95.5_{0.2}$	$94.1_{0.7}$
+ MIM + L1	$95.8_{0.3}$	$95.4_{0.4}$
+ SimCLR + L1	$\mathbf{96.1}_{0.3}$	$\mathbf{95.5}_{0.7}$

Table 9: Civilcomments-small: MLM → CLM

Method	Group Labels	
	None	Validation
ERM	$83.9_{0.4}$	$81.3_{0.9}$
+ MLM + L1	$\mathbf{84.4}_{0.4}$	$\mathbf{82.0}_{0.5}$
+ CLM + L1	$83.3_{0.7}$	$81.1_{0.9}$

A.5.3 Going Beyond the Pre-Training Objective

The results presented in Tables 8 and 9 elucidate the impact of multi-tasking using various transfer learning objectives on average accuracy. In line with the observations in Section 3.3.1, SimCLR consistently delivers results that are either superior or comparable to those achieved with the masked image modeling objective. CLM, on the other hand, gives lower performance than ERM and ERM+MLM+L1 on both checkpointing schemes.

A.6 Impact of Pre-training

Table 10: Waterbirds: Finetuning vs training from scratch.

Pretrianed	Method	No Group Annotations		Val Group Annotations	
		Avg Acc	WG Acc	Avg Acc	WG Acc
No	ERM	$65.1_{0.5}$	$4.5_{1.6}$	$53.3_{0.7}$	$10.1_{2.9}$
	JTT	$67.0_{5.3}$	$\mathbf{10.8}_{12.2}$	$\mathbf{56.2}_{2.1}$	$\mathbf{49.9}_{4.0}$
	ERM + MIM + L1	$\mathbf{67.0}_{2.3}$	$1.65_{0.7}$	$53.5_{2.7}$	$12.0_{3.2}$
yes	ERM	$95.5_{0.2}$	$80.1_{4.6}$	$94.1_{0.7}$	$85.4_{1.4}$
	JTT	$95.6_{0.3}$	$82.1_{1.2}$	$94.0_{0.5}$	$85.9_{2.5}$
	ERM + MIM + L1	$\mathbf{95.8}_{0.3}$	$\mathbf{83.3}_{3.4}$	$\mathbf{95.4}_{0.4}$	$\mathbf{87.5}_{2.7}$

Table 11: Civilcomments-small : Fine-tuning versus training from scratch.

Pretrianed	Method	No Group Annotations		Val Group Annotations	
		Avg Acc	WG Acc	Avg Acc	WG Acc
No	ERM	$80.7_{0.8}$	$31.1_{7.2}$	$\mathbf{74.4}_{0.9}$	$54.0_{3.7}$
	JTT	$79.6_{0.6}$	$\mathbf{34.9}_{8.9}$	$74.3_{1.2}$	$\mathbf{58.7}_{1.3}$
	ERM+MLM+L1	$\mathbf{80.7}_{0.6}$	$31.3_{7.6}$	$74.2_{0.3}$	$56.2_{0.9}$
Yes	ERM	$83.9_{0.4}$	$51.6_{5.6}$	$81.4_{1.0}$	$67.4_{2.1}$
	JTT	$83.3_{0.2}$	$52.5_{5.2}$	$81.3_{0.8}$	$68.0_{1.8}$
	ERM+MLM+L1	$\mathbf{84.4}_{0.4}$	$\mathbf{53.7}_{4.3}$	$\mathbf{82.0}_{0.6}$	$\mathbf{69.4}_{1.7}$

Finetuning pre-trained models is arguably the de-facto paradigm in machine learning [31, 32, 4]. Consequently, our experiments so far have exclusively focused on pre-trained models. In this section, we wish to understand the effect of deviating from this paradigm on our MTL approach. We thus compare against JTT and ERM when the model is trained from scratch instead of starting with a pre-trained model.

Tables 10 and 11 depict our results on Waterbirds and Civilcomments-small, respectively. Our results show that pre-training is critical for setting up regularized MTL as a viable remedy against poor worst-group outcomes. We posit the following explanation for this outcome. Solving the MLM and MIM tasks effectively from scratch with only task data is difficult. This poor performance on the auxiliary task translates to an inability to constrain the use of spurious features on the end task. Our recommendation to practitioners is that our approach be used **during finetuning of pre-trained models** in order to be maximally effective.

A.7 Impact of w_{ssl} and w_{reg}

Figures 3 and 4 depict the influence of w_{ssl} and w_{reg} on the trade-off between average and worst group accuracy. In both the Waterbirds and Civilcomments-small datasets, we consistently observe

that setting w_{reg} to e^{-1} results in the best performance. This configuration consistently outperforms $w_{\text{reg}} = 1$ and $w_{\text{reg}} = 0$, highlighting the importance of L1 regularization in enhancing both average and worst-group performance, albeit with a smaller weight coefficient. Notably, when w_{reg} exceeds 1, there is a significant drop in performance, indicating a substantial reduction in the representation capacity of the shared representation layer. The influence of w_{ssl} varies between the datasets, and it is challenging to precisely quantify the effect of different w_{ssl} values.

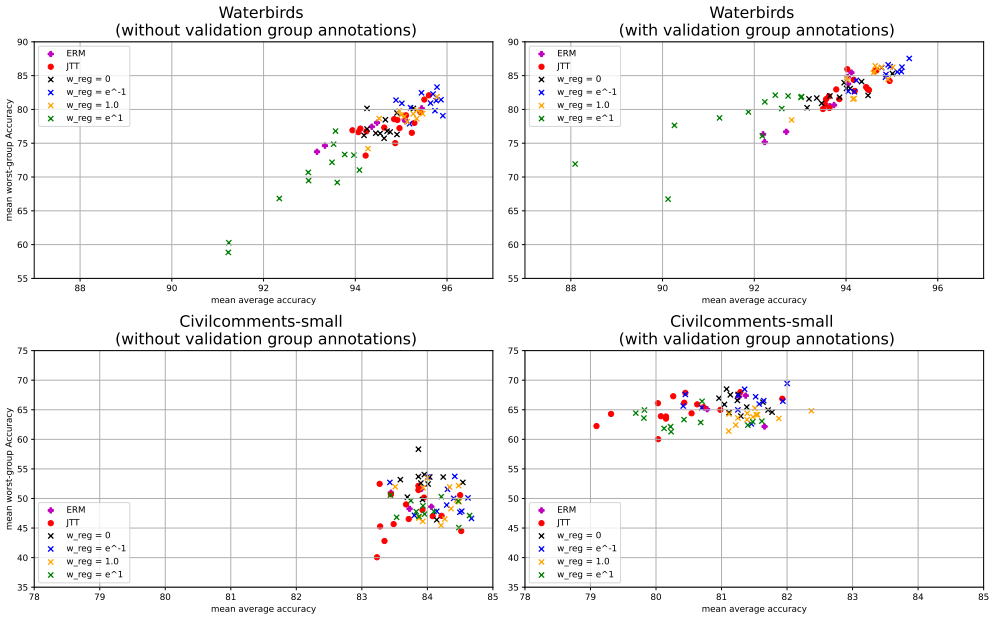

Figure 3: Effect of L1 regularization weight on the trade-off between average and worst-group performance.

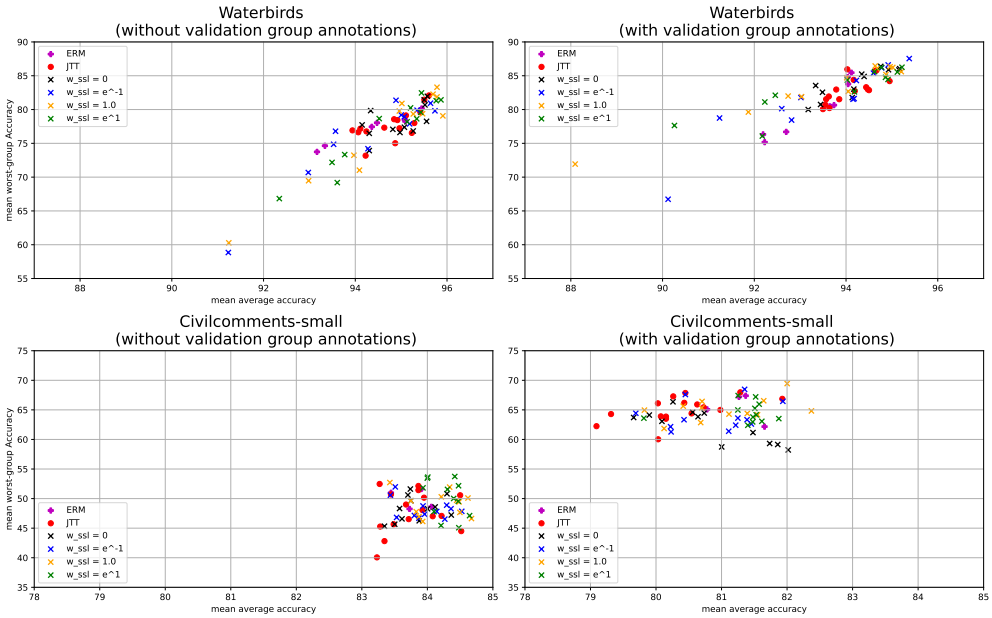

Figure 4: Effect of SSL objective weight on the trade-off between average and worst-group performance.

A.8 Introducing External Data

Table 12: Civilcomments-small : Unlabeled data also helps improve worst-group error.

Method	**Group Labels**	
	None	Validation
ERM	$51.6_{5.6}$	$67.4_{2.1}$
ERM + MLM + L1$_{task}$	$53.7_{4.3}$	$69.4_{1.7}$
ERM + MLM + L1$_{unlabeled}$	$\mathbf{57.2}_{2.5}$	$\mathbf{70.5}_{0.5}$

Leveraging unlabeled data has proven effective in improving adversarial robustness [44], generalization [45], and worst-group performance [46]. Consequently, we extend our constrained multitasking framework to incorporate unlabeled data. Among the three datasets considered, only Civilcomments has a readily available unlabeled corpus [46]. Therefore, we investigate the impact of using unlabeled data on the Civilcomments-small dataset. We sample unlabeled data of equivalent size to our training data. Instead of conducting experiments across five different seeds, we generate five distinct, non-overlapping subsets of unlabeled data and report results based on their average performance.

As evidenced from Table 12, unlabeled data leads to further improvements over the end-task data. Thus, transfer learning from unlabeled, which is supported by our theoretical investigation. We leave the experiments with a larger pool of unlabeled data for future work.

A.9 Related Work

Robustness using group demographics. Among the approaches that leverage group information, *Distributionally Robust Optimization* [13] is the most popular technique that tries to minimize the maximum loss over the sub-groups. [47] presented *Model Patching*, a data augmentation method designed to enhance the representation of minority groups. [48] and [49] developed training paradigms that impose heavy Lipschitz regularization around minority data points for worst-group performance improvement. The *Predict-then-Interpolate* approach established by [50], introduces an effective algorithm capable of learning correlations across stable environments, thereby enhancing worst-group generalization. *FISH*, proposed by [51], focuses on domain generalization via inter-domain gradient matching. [21] proffered *DFR*, a simple yet efficient method that involves retraining the final layer of the model using a held-out dataset where the spurious correlations are broken. Lastly, studies by [52], [53], and [54] have demonstrated that straightforward data reweighing and subsampling strategies can yield commendable worst-group performance.

Robustness without group demographics. Extensive research has been dedicated to addressing the challenges of worst-group generalization in a more realistic scenario, one where access to group annotations during training is not available. *GEORGE* [22] adopts a clustering-based methodology to unveil latent groups within the dataset and subsequently employs groupDRO for improved robustness. *Learning from Failure (LfF)* [16] introduces a two-stage strategy. In the first stage, an intentionally biased model aims to identify minority instances where spurious correlations do not apply. In the second stage, the identified examples are given increased weight during the training of a second model. *Just Train Twice (JTT)* method [17] follows a similar principle by training a model that minimizes loss over a reweighted dataset. This dataset is constructed by up-weighting training examples misclassified during the initial few epochs. *Correct-N-Contrast (CNC)* [18] builds upon JTT but employs a contrastive loss in the training of the final model. *Spread Spurious Attribute (SSA)* [19] follows a slightly different two-step approach. Initially, a pseudo-labeling phase trains a spurious attribute predictor per group. In the subsequent robust training phase, the spurious attribute predictor is used to pseudo-label (input, label) pairs with spurious attributes. *AGRO* [23] proposes a method for the adversarial discovery of error-prone groups to enhance robust optimization. Recently, [20] introduced *AFR*, a modification of *DFR* [21], which is both simple and efficient, requiring no group annotations.