The Silent Helper: How Implicit Regularization Enhances Group Robustness

Nahal Mirzaie[†] NAHAL.MIRZAIE@SHARIF.EDU
Sharif University of Technology, Tehran, Iran

Mahdi Ghaznavi[†] MAHDI.GHAZNAVI@CE.SHARIF.EDU
Sharif University of Technology, Tehran, Iran

Hosna Oyarhoseini[†] HOSNA.OYARHOSEINI79@SHARIF.EDU
Sharif University of Technology, Tehran, Iran

Alireza Alipanah ALIREZA.ALIPANAH46@SHARIF.EDU
Sharif University of Technology, Tehran, Iran

Erfan Sobhaei ERFAN.SOBHAEI11@SHARIF.EDU
Sharif University of Technology, Tehran, Iran

Ali Abbasi A.ABBASI@SHARIF.EDU
Sharif University of Technology, Tehran, Iran

Amirmahdi Farzaneh AMIRMFARZANE@UT.AC.IR
University of Tehran, Tehran, Iran

Hossein Jafarinia JAFARINIA@SHARIF.EDU
Sharif University of Technology, Tehran, Iran

Parsa Sharifi Sadeh PAR.SHAR21@SHARIF.EDU
Sharif University of Technology, Tehran, Iran

Arefeh Boushehrian AREFE.BOUSHEHRIAN82@SHARIF.EDU
Sharif University of Technology, Tehran, Iran

Mahdieh Soleymani Baghshah SOLEYMANI@SHARIF.EDU
Sharif University of Technology, Tehran, Iran

Mohammad Hossein Rohban ROHBAN@SHARIF.EDU
Sharif University of Technology, Tehran, Iran

[†]*These authors contributed equally to this work.*

Abstract

The implicit regularization effect of Stochastic Gradient Descent (SGD) is known to enhance the generalization of deep neural networks and becomes stronger with higher learning rates and smaller batch sizes. However, its role in improving group robustness, defined as a model's ability to perform well on underrepresented subpopulations, remains underexplored. In this work, we study the impact of SGD's implicit regularization under group imbalance characterized by spurious correlations. Through extensive experiments on various datasets, we show that increasing the strength of implicit regularization improves worst-group accuracy (WGA). Crucially, this improvement is not merely a byproduct of better overall generalization, but a targeted enhancement in robustness to spurious features. Moreover, our analysis reveals that this phenomenon also con-

tributes to improved feature learning in deep networks. These findings offer a new perspective on the role of SGD's implicit regularization, showing that it not only supports generalization but also plays a central role in achieving robustness to spurious correlations.

1. Introduction

Distribution shifts pose significant challenges to model robustness, particularly when minority groups that were underrepresented during training become more prevalent at test time [29, 30]. A critical factor that undermines generalization in these scenarios is the presence of *spurious correlations*, i.e., correlations between labels and easily learnable but non-causal patterns that models exploit during training but that cause them to fail to generalize at test time [1, 5]. Reliance on such spurious attributes disproportionately harms the performance of minority groups, leading to reduced *worst-group accuracy (WGA)* despite high overall accuracy [13, 24, 29].

Previous works have demonstrated that certain optimization hyperparameters, which govern the dynamics of the training process, can influence robustness to group shifts. Idrissi et al. [10] showed that careful tuning of hyperparameters, including learning rate and batch size, enables standard Empirical Risk Minimization (ERM) to achieve worst-group accuracy comparable to more sophisticated two-stage methods such as Just Train Twice (JTT) [18]. Building on this insight, Puli et al. [22] found that simply increasing the learning rate during vanilla training can reduce susceptibility to shortcut learning.

In this work, we are, to the best of our knowledge, the first to establish a connection between the implicit regularization behavior of Stochastic Gradient Descent (SGD) and robustness to correlation shifts. While prior work has shown that increasing the learning rate or decreasing the batch size amplifies implicit regularization and improves generalization in standard settings [26], its role in robustness under subpopulation shifts—particularly those involving spurious correlations—remains underexplored.

We investigate this connection by studying how SGD's implicit regularization behaves on datasets exhibiting spurious features. Our findings suggest that the implicit regularization of SGD does indeed improve group robustness, and this improvement goes beyond general accuracy gains. Specifically, it enhances the model's ability to learn and rely on core, invariant features, leading to better performance on minority groups and, as a result, improving WGA.

2. Related Work

2.1. Group Robustness and the Role of Hyperparameters

In the context of robustness to correlation shifts [30], numerous approaches have been proposed [13, 15, 18, 23, 24]. A key limitation of many such methods is their reliance on group annotations—i.e., knowing which data points exhibit spurious (non-causal) patterns. This reliance restricts their applicability in real-world scenarios, where such annotations are often unavailable. To address this, recent works have explored methods that improve robustness to correlation shifts without requiring group annotations by leveraging implicit group identification [4, 6, 28].

Improving empirical risk minimization (ERM) to maximize WGA is thus a critical objective. Gulrajani and Lopez-Paz [7] observe that ERM could outperform other domain generalization methods with proper tuning. Notably, careful tuning of training hyperparameters such as batch size and learning rate can significantly affect worst-group performance [10]. For instance, Li et al. [17]

showed that low learning rates lead models to memorize spurious shortcuts, harming generalization. In contrast, higher learning rates delay shortcut reliance, encouraging more robust, generalizable learning. While separate from spurious correlation, recent work has shown that batch size matters under class imbalance, smaller batches often perform better [25]. However, its specific impact in group-imbalanced settings remains underexplored, highlighting a key gap in the literature.

2.2. Implicit Regularization of SGD and Its Impact on Group Robustness

Stochastic Gradient Descent (SGD) often demonstrates superior generalization that traditional convergence rate analyses do not fully capture [20, 31]. For example, models trained using moderately high learning rates or small batch sizes frequently achieve improved accuracy in test datasets [12, 16]. A well-established explanation attributes this behavior to the discrete update steps of SGD, which deviate from the continuous gradient flow that directly minimizes the loss function [3]. Instead, SGD implicitly follows a modified trajectory, approximating gradient descent on an altered cost function that includes an *implicit regularization* term proportional to the ratio of the learning rate to the batch size [26]. Formally, let the empirical risk be defined over a data distribution $\mathcal{D}$:

$$C(\omega) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(f_\omega(x), y)\right],$$

where ω denotes the model parameters, ℓ is the per-example loss function, and $f_\omega(x)$ represents the model's prediction for input x. While full-batch gradient descent directly minimizes this objective via deterministic updates, SGD, with its inherent stochasticity from mini-batch, implicitly optimizes a modified cost function. As shown by Smith et al. [27], the average iterate of SGD after one epoch with a small learning rate closely follows the dynamics of gradient flow with respect to a perturbed objective:

$$\widetilde{C}_{\text{SGD}}(\omega) = C(\omega) + \frac{\epsilon}{4m}\sum_{k=0}^{m-1}\|\nabla\hat{C}_k(\omega)\|^2 = C(\omega) + \frac{\epsilon}{4}\|\nabla C(\omega)\|^2 + \frac{N-B}{N-1}\frac{\epsilon}{4B}\Gamma(\omega), \quad (1)$$

where $\hat{C}_k(\omega)$ is the k-th minibatch cost, $\Gamma(\omega) = \frac{1}{N}\sum_{i=1}^{N}\|\nabla C_i(\omega) - \nabla C(\omega)\|^2$, ϵ is the learning rate, m is the number of batches, B is the batch size, N is the total number of samples, and $\nabla C_i(\omega)$ denotes the gradient computed on the i-th datapoint. Focusing on the last expression, the first additional term, $\frac{\epsilon}{4}\|\nabla C(\omega)\|^2$, penalizes regions with large gradient magnitudes, discouraging convergence to sharp minima. The second term, $\frac{N-B}{N-1}\frac{\epsilon}{4B}\Gamma(\omega)$, suppresses parameter directions exhibiting high variance across individual datapoints.

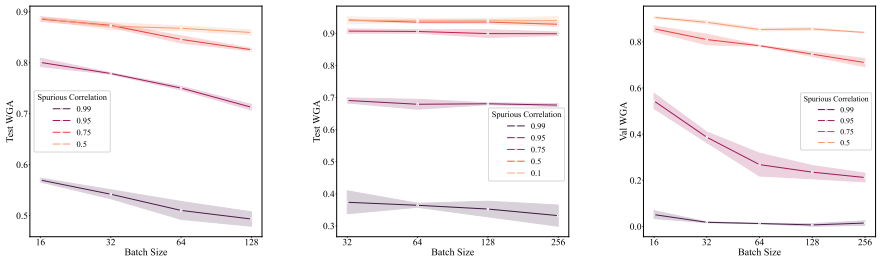
Reducing gradient variance promotes more uniform learning across training samples. This discourages solutions that overfit majority samples while underfitting minority ones, even if the overall loss remains low. Instead, it favors solutions that fit all groups more evenly, leading to better generalization under group imbalance settings. As a result, the added regularization term promotes the discovery of feature representations that capture structure shared across groups, rather than those specific to the majority. This helps prevent the optimizer from converging to solutions that overfit the majority group while ignoring minority groups, even when the overall training loss appears low. As a result, stochastic gradient descent (SGD) is guided toward solutions that perform more uniformly across groups, thereby enhancing generalization under group imbalance.

3. Results

3.1. Experimental Setups

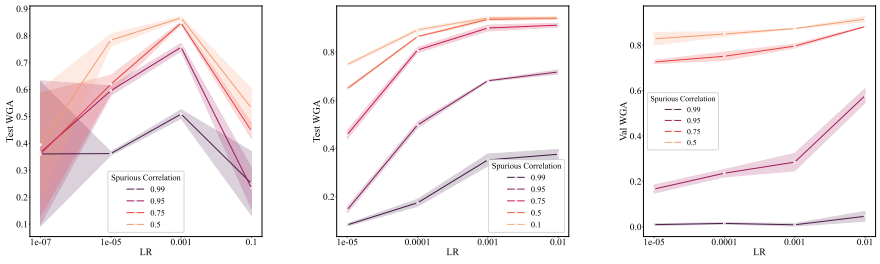
We evaluate our methods on a diverse set of synthetic datasets, Waterbirds [24], Colored MNIST [2], CIFAR-10 [19], and Domino [21]. These datasets introduce controlled spurious features (e.g., background, color, or patch) that correlate with labels during training but not at test time (Supp 5.1).

3.2. Robustness to Spurious Correlation



(a) Waterbirds **(b)** CMNIST **(c)** Domino

Figure 1: **Effect of Batch Size on Worst-Group Accuracy (WGA)**. Decreasing the batch size consistently improves WGA across Waterbirds, CMNIST, and Domino datasets, suggesting enhanced group robustness.



(a) Waterbirds **(b)** CMNIST **(c)** Domino

Figure 2: **Effect of Learning Rate on Worst-Group Accuracy**. Increasing the learning rate also leads to higher WGA, up to a point. Large learning rates (e.g., 0.1) result in a decline in WGA in Waterbirds dataset, reflecting findings in implicit regularization literature that overly high learning rates can harm generalization.

We observe that the implicit regularization of SGD can significantly improve worst-group accuracy under spurious correlations. As shown in Figure 1, reducing the batch size consistently enhances worst-group accuracy across datasets. A similar trend is observed in Figure 2 for increasing the learning rate, except in cases where very large learning rates prevent the model from converging on some datasets.

This effect is substantially stronger than what has been previously reported in standard training scenarios [27]. In Tables 1 and 2, we compare improvements in overall accuracy and worst-group accuracy on the test set across two settings: small versus large batch sizes, and high versus low

learning rates. As shown, worst-group accuracy consistently improves and, notably, its improvement exceeds that of overall accuracy in all cases.

Table 1: Test WGA and Accuracy under Small vs. Large Batch Sizes.

Dataset	WGA			Accuracy		
	Small	Large	Δ	Small	Large	Δ
Waterbirds	$80.1_{\pm 0.8}$	$70.8_{\pm 0.9}$	**+9.3**	$90.7_{\pm 0.8}$	$87.4_{\pm 0.2}$	+3.3
CMNIST	$69.1_{\pm 0.8}$	$68.1_{\pm 0.3}$	**+1.0**	$79.6_{\pm 0.1}$	$79.0_{\pm 0.1}$	+0.6
Domino	$54.4_{\pm 3.3}$	$23.5_{\pm 2.9}$	**+30.9**	$77.2_{\pm 0.4}$	$62.7_{\pm 0.7}$	+14.5

Table 2: Test WGA and Accuracy under High vs. Low Learning Rates.

Dataset	WGA			Accuracy		
	High	Low	Δ	High	Low	Δ
Waterbirds	$75.7_{\pm 1.4}$	$59.6_{\pm 1.8}$	**+16.1**	$88.3_{\pm 0.2}$	$83.3_{\pm 0.8}$	+5.0
CMNIST	$68.1_{\pm 0.3}$	$14.5_{\pm 1.3}$	**+53.6**	$79.0_{\pm 0.1}$	$38.9_{\pm 1.1}$	+40.1
Domino	$28.5_{\pm 3.9}$	$16.6_{\pm 1.8}$	**+11.9**	$64.1_{\pm 1.9}$	$59.3_{\pm 0.9}$	+4.8

3.2.1. EXPLICITLY PROMOTING THE IMPLICIT REGULARIZATION OF SGD

Following Smith et al. [27], to isolate the effect of SGD's implicit regularization, we augment the loss with a controlled explicit term: the ℓ_2 norm of the mini-batch gradient, scaled by a positive coefficient. This modification amplifies the regularization effect during training. We evaluate its impact on CMNIST and CIFAR-10 (with 95% spurious correlation), using a fixed learning rate of 0.001 and a batch size of 128. As shown in Table 3, explicitly encouraging the implicit bias of SGD leads to improvements in WGA.

Table 3: Effect of Explicit Regularization on Test WGA

Dataset	WGA (w/o Reg)	WGA (w/ Reg)	Δ WGA
CMNIST	$66.5_{\pm 0.01}$	$70.9_{\pm 0.03}$	**+4.4**
CIFAR-10	$47.8_{\pm 0.01}$	$56.7_{\pm 0.04}$	**+8.9**

3.3. Improving Feature Learning

We have investigated how does implicit regularization of SGD affect the quality of learned features? While neural networks learn core and spurious features during training, empirical evidence shows they preferentially rely on spurious features for prediction when such features are present in the input [13]. We employ decoded WGA [9] as our primary metric for evaluating feature quality. This measure amount of information about the core (non-spurious) features that can be decoded from the representations learned by standard ERM [11] (see details of experiments in Supp 5.4).

Figure 3: **Worst-Group Accuracy (WGA) and Decoded WGA** on the CMNIST testset across batch sizes. We observe that increasing the batch size generally leads to improved decoded WGA.

In Figure 3, we plot both the worst-group accuracy (WGA) of standard ERM and the decoded WGA on the CMNIST test set across different batch sizes. Overall, decoded test WGA tends to improve as the batch size decreases, with this improvement being much more pronounced in settings with high spurious correlation. While decoded WGA decreases slightly as the spurious correlation increases, standard ERM's WGA drops significantly when the spurious correlation reaches 0.95 or higher. These patterns remain consistent across all training batch sizes, with smaller batch sizes (i.e., stronger implicit regularization) producing higher WGA in both ERM and decoded settings.

Moreover, the variations observed in test WGA are more pronounced than those in decoded test WGA, especially under stronger spurious correlations. This suggests that reducing batch size shifts the model's reliance from spurious features toward core features.

4. Conclusion

We investigate how the implicit regularization of SGD influences group robustness and find that its variance reducing dynamics systematically enhance worst group accuracy under group imbalance, more than overall accuracy. Our analysis further reveals that implicit regularization encourages reliance on core, invariant features. Finally, we show that explicitly promoting this regularization term can reproduce similar robustness gains even under standard batch size and learning rate.

References

[1] Herbert A. Simon and. Spurious correlation: A causal interpretation*. *Journal of the American Statistical Association*, 49(267):467–479, 1954. doi: 10.1080/01621459.1954.10483515. URL https://doi.org/10.1080/01621459.1954.10483515.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL https://arxiv.org/abs/1907.02893.

[3] David G. T. Barrett and Benoit Dherin. Implicit gradient regularization. *CoRR*, abs/2009.11162, 2020. URL https://arxiv.org/abs/2009.11162.

[4] Reza Bayat, Mohammad Pezeshki, Elvis Dohmatob, David Lopez-Paz, and Pascal Vincent. The pitfalls of memorization: When memorization hurts generalization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vVhZh9ZpIM.

[5] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *CoRR*, abs/2004.07780, 2020. URL https://arxiv.org/abs/2004.07780.

[6] Mahdi Ghaznavi, Hesam Asadollahzadeh, Fahimeh Hosseini Noohdani, Soroush Vafaie Tabar, Hosein Hasani, Taha Akbari Alvanagh, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Exploiting what trained models learn for making them robust to spurious correlations without group annotations. In *Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions*, 2025. URL https://openreview.net/forum?id=8vo1YSAt6g.

[7] 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.

[8] 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)*, pages 770–778, 2016.

[9] Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9995–10006. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/71e9c6620d381d60196ebe694840aaaa-Paper.pdf.

[10] 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. URL https://proceedings.mlr.press/v177/idrissi22a.html.

[11] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations, 2022. URL https://arxiv.org/abs/2210.11369.

[12] Nitish Shirish Keskar, Jorge Nocedal, Ping Tak Peter Tang, Dheevatsa Mudigere, and Mikhail Smelyanskiy. On large-batch training for deep learning: Generalization gap and sharp minima. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*, 2017. 5th International Conference on Learning Representations, ICLR 2017 ; Conference date: 24-04-2017 Through 26-04-2017.

[13] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

[14] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 2009. URL http://www.cs.toronto.edu/~kriz/cifar.html.

[15] Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations, 2023. URL https://arxiv.org/abs/2309.08534.

[16] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. *Efficient BackProp*, pages 9–48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_3. URL https://doi.org/10.1007/978-3-642-35289-8_3.

[17] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks, 2020. URL https://arxiv.org/abs/1907.04595.

[18] 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. URL https://proceedings.mlr.press/v139/liu21f.html.

[19] Ekdeep Singh Lubana, Eric J Bigelow, Robert P Dick, David Krueger, and Hidenori Tanaka. Mechanistic mode connectivity. In *International Conference on Machine Learning*, pages 22965–23004. PMLR, 2023.

[20] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. *arXiv preprint arXiv:1712.06559*, 12 2017. doi: 10.48550/arXiv.1712.06559.

[21] Nihal Murali, Aahlad Puli, Ke Yu, Rajesh Ranganath, and Kayhan Batmanghelich. Beyond distribution shift: Spurious features through the lens of training dynamics. *Transactions on machine learning research*, 2023:https–openreview, 2023.

[22] Aahlad Puli, Lily Zhang, Yoav Wald, and Rajesh Ranganath. Don't blame dataset shift! shortcut learning due to gradients and cross entropy. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 71874–71910. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e35460304fdf6df523f068a59aaf8829-Paper-Conference.pdf.

[23] Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and Fast Group Robustness by Automatic Feature Reweighting. *International Conference on Machine Learning (ICML)*, 2023.

[24] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

[25] Ravid Shwartz-Ziv, Micah Goldblum, Yucen Lily Li, C. Bayan Bruss, and Andrew Gordon Wilson. Simplifying neural network training under class imbalance. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=iGmDQn4CRj.

[26] Samuel Smith, Benoit Dherin, David Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent, 01 2021.

[27] Samuel L. Smith, Benoit Dherin, David G. T. Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. *CoRR*, abs/2101.12176, 2021. URL https://arxiv.org/abs/2101.12176.

[28] Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying spurious biases early in training through the lens of simplicity bias. In *International Conference on Artificial Intelligence and Statistics*, pages 2953–2961. PMLR, 2024.

[29] Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. 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 39584–39622. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/yang23s.html.

[30] Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zhenguo Li, and Jun Zhu. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7947–7958, June 2022.

[31] Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George Dahl, Chris Shallue, and Roger B Grosse. Which algorithmic choices matter at

which batch sizes? insights from a noisy quadratic model. 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. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/e0eacd983971634327ae1819ea8b6214-Paper.pdf.

5. Supplementary Materials

5.1. Datasets

In our experiments, we evaluate models on a diverse set of datasets that are specifically designed to test robustness against spurious correlations. Each dataset introduces a known, controllable spurious feature that can confound standard training methods. Below, we briefly describe the datasets used in our experiments:

- **Waterbirds** [24] is a synthetic dataset constructed by overlaying bird images from the CUB dataset onto backgrounds from the Places dataset. It defines a binary classification task: waterbird versus landbird, where the background (water or land) is spuriously correlated with the bird type. This results in a strong dataset bias: the majority of waterbirds appear over water backgrounds, and most landbirds over land. As a consequence, models trained with standard techniques often learn to rely on the background rather than bird-specific features, leading to poor generalization to minority groups where this correlation is broken.

- **Colored-MNIST** [2] is a synthetic variant of the MNIST dataset, designed to study the reliance of the model on spurious correlations in simple visual settings. Each grayscale digit from the original MNIST dataset is colorized based on its label: with high probability, a fixed color is assigned to each digit class (e.g., all 0s are red, all 1s are green, etc.). To introduce some variation, a small proportion of samples are assigned colors uniformly at random from the remaining color set. This creates a controlled correlation between digit class and color, where color serves as a spurious attribute. Although the shape of the digits is the true predictive feature, models trained with standard techniques often exploit the more easily learnable color signal, leading to degraded performance when this correlation is broken at the test time. Colored MNIST thus provides a clean and interpretable setting for evaluating methods aimed at mitigating spurious correlations.

- **CIFAR-10 (Car vs. Truck)** [19] is a binary classification data set derived from the CIFAR-10 dataset [14], where only two classes are selected, car and truck. To introduce a spurious correlation, each image is augmented with a small colored square (cue) in the top left corner. The color of this square (e.g., red or blue) is highly correlated with the label: for example, most cars may have a red square and most trucks a blue one. This synthetic cue serves as a spurious attribute that a model might learn to exploit, rather than relying on shape-based features of the main object. A small subset of training samples breaks this correlation by assigning the opposite color or a random color, encouraging robustness evaluation under distribution shift. This setup enables systematic analysis of model reliance on spurious features and robustness to such correlations.

- **Domino** [21] is a synthetic dataset created to study model behavior in the presence of multiple potentially spurious features. It is constructed by concatenating CIFAR-10 images with Fashion-MNIST images such that each pair shares the same class label (e.g., a CIFAR-10 ”cat” image is paired with a Fashion-MNIST ”pullover”). This results in composite images where the CIFAR-10 object represents the primary visual cue, while the Fashion-MNIST segment introduces a structured but potentially spurious feature. Since both segments are class-consistent, a model can use either or both to make predictions. However, during training, the model may learn to rely disproportionately on the simpler-to-learn Fashion-MNIST

portion, which serves as a shortcut or spurious signal. Domino allows for controlled interventions such as removing, randomizing, or altering the Fashion-MNIST side, making it well-suited for studying the dynamics of spurious feature learning and evaluating robustness to such correlations.

Figure 4: Illustration of three datasets—Domino CIFAR10-MNIST, and Waterbirds, used in our experiments. Each dataset exhibits spurious correlations, where the majority groups are highlighted in red and blue, while the minority groups are shown in green and orange.

5.2. Architectures

We select architectures based on common practical choices aligned with dataset complexity and input resolution. For **Waterbirds**, we adopt a **ResNet-50** backbone [8]. For **CIFAR-10** and **Domino**, we use **ResNet-18**. For **Colored MNIST**, we use a lightweight **two-layer MLP**, which suffices due to the dataset's simplicity and lower resolution.

5.3. Hyperparameters

All models are trained using SGD with no learning rate scheduler, and a standard weight decay of 1×10^{-5}. These settings are kept constant across experiments to isolate the effect of batch size, learning rate, and spurious correlation on generalization and robustness.

5.4. Decoded Accuracy

In line with Izmailov et al. [11], we first train a neural network and then freeze all layers except the final one. Next, we retrain only this last layer with L1 regularization to its weights using a group-balanced dataset. This ensures that the learned representations remain fixed, and only the decision boundary is adapted to rely more on core-related features. A higher decoded WGA implies that the learned representations contain more useful core features, as the retrained classifier can generalize better across groups.

All models were trained for 500 epochs using various batch sizes, and the final checkpoint from each run was selected for a second-stage training on group-balanced data. During this second stage, the L1 regularization strength λ was tuned over the range 10^{-5} to 0.1, with the batch size fixed at 128 for all models.