mathematics

MDPI

Article

Achieving Distributional Robustness with Group-Wise Flat Minima

Seowon Ji [1], Seunghyun Moon [2], Jiyoon Shin [3] and Sangwoo Hong [1,*]

[1] Department of Computer Science and Engineering, Konkuk University, Seoul 05029, Republic of Korea; seowonji@konkuk.ac.kr
[2] Department of Electrical and Electronics Engineering, Konkuk University, Seoul 05029, Republic of Korea; smoon@konkuk.ac.kr
[3] Department of Artificial Intelligence, Korea Aerospace University, Goyang 10540, Republic of Korea; jiyoonshin@kau.ac.kr
* Correspondence: swhong06@konkuk.ac.kr

Abstract

Improving robustness under distributional shifts remains a central challenge in machine learning. Although Sharpness-Aware Minimization (SAM) has proven effective in finding flatter minima for better generalization, it overlooks the heterogeneity in sharpness across different subpopulations, which can exacerbate performance gaps for minority or vulnerable groups. To address this challenge, we propose Group-gap Guided SAM (G^2-SAM), a new optimization framework that promotes distributional robustness by steering flatness-seeking directions according to intergroup loss disparities. Our method estimates group-wise sharpness and adaptively refines perturbation strategies to minimize the worst-group loss while preserving model consistency. Through comprehensive experiments across various datasets, we show that G^2-SAM achieves superior Worst-Group Accuracy and robustness, outperforming previous baselines. These findings highlight the importance of addressing group-specific geometry in the loss landscape to build more reliable and equitable neural networks.

Keywords: distribution shift; spurious correlation; robust optimization; flat minima

MSC: 68T07; 65D15

check for updates

Academic Editor: Simeon Reich

Received: 2 September 2025
Revised: 14 October 2025
Accepted: 17 October 2025
Published: 20 October 2025

Citation: Ji , S.; Moon, S.; Shin, J.; Hong, S. Achieving Distributional Robustness with Group-Wise Flat Minima. *Mathematics* **2025**, *13*, 3343. https://doi.org/10.3390/math13203343

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Deep neural networks trained using Empirical Risk Minimization (ERM) have achieved remarkable success in various machine learning tasks, such as classification [1], segmentation [2], speech recognition [3,4], 3D vision [5,6], and natural language processing [7,8]. However, many of these advancements have faced challenges of trustworthiness, which include issues such as explainability [9,10], fairness [11,12], and data imbalance [13]. Most importantly, many of these models frequently suffer from poor generalization performance under distributional shift [14,15] due to their tendency to exploit spurious features. Spurious features refer to features that correlate strongly with labels in the training data but do not represent meaningful relationships in the true underlying distribution. Such spurious correlations disproportionately degrade performance for minority or vulnerable subgroups, exacerbating fairness [16] and reliability issues in critical real-world applications like medical imaging and facial recognition [17,18]. In this work, we focus on validating

our approach on vision classification tasks that contain clear subgroup structures, laying the groundwork for potential future exploration in other domains.

Several approaches have been proposed to mitigate the adverse effects of spurious correlations. Early research primarily relied on explicitly annotated bias attributes to identify and rectify biases during training. Methods such as Group Distributionally Robust Optimization [15] (GDRO) explicitly optimize for worst-group performance. Recently, there has been significant interest in developing techniques that avoid explicit bias labels. Methods like Disagreement Probability-based Resampling [19] (DPR) identify bias-conflicting samples by measuring the disagreement between predictions from biased models and the true labels, subsequently upsampling these samples to counteract bias. On the other hand, some studies [20] have tried to utilize augmentation strategies to mitigate bias and enhance fairness by balancing dataset distribution, effectively reducing the performance gap between groups. However, all these methods exhibit significantly diminished performance when applied to scenarios where the spurious feature is more complex (non-binary).

Parallel to these debiasing efforts, optimization strategies such as Sharpness-Aware Minimization [21] (SAM) have emerged as powerful methods for improving generalization. By steering neural network training towards flat minima, where the parameter space is characterized by low curvature, SAM achieves remarkably better robustness against distributional shifts. Nevertheless, despite its effectiveness, SAM inherently neglects the heterogeneity of sharpness across different data subgroups, potentially amplifying performance disparities by disproportionately benefiting majority groups.

To bridge this gap, we propose Group-gap Guided Sharpness-Aware Minimization (G^2-SAM), a novel optimization framework that incorporates intergroup loss disparities into sharpness-aware training. G^2-SAM explicitly estimates group-wise sharpness and adaptively adjusts optimization strategies to achieve group-wise flat minima. By doing so, G^2-SAM ensures that flat minima are achieved specifically in regions beneficial to the minority, thus directly addressing the pitfalls of conventional SAM in biased scenarios.

Our contributions can be summarized as follows:

- We introduce a principled approach to incorporate group-wise sharpness considerations into the SAM framework, promoting equitable optimization across diverse subpopulations.
- We empirically demonstrate that our proposed G^2-SAM significantly enhances Worst-Group Accuracy and robustness against distributional shifts, consistently outperforming existing state-of-the-art methods across various benchmark datasets.
- Importantly, G^2-SAM achieves these improvements without requiring group labels during inference, thereby enhancing its practical applicability and scalability for vision tasks with known subgroup structures.

Our results underscore the necessity and effectiveness of addressing subgroup-specific geometric properties in the loss landscape, laying the foundation for building fairer, more reliable, and more robust machine learning models.

2. Related Work

2.1. Mitigating Spurious Correlation

In the field of machine learning, a significant amount of effort is being dedicated to mitigating spurious correlations. The goal is to make models more adaptable to distributional shifts. Group Distributionally Robust Optimization (GDRO) [15] enhances the performance of the worst-performing group by explicitly minimizing the maximum group risk. However, this approach can cause overfitting to the minority data. Moreover, since it only optimizes for the worst-group loss, it tends to show significantly decreased performance when there exists a large number of groups, making it challenging to apply

in real-world scenarios. More recently, methods without explicit bias labels [19,22] have been proposed. They usually adopt a bias-identifying strategy using biased models' predictions, and upweight bias-conflicting samples. Other strategies include using contrastive learning [23] to debias neural networks or fine-tuning only the last layer [24] of a network with a balanced validation dataset, which has been shown to be effective in mitigating spurious correlations.

2.2. Optimization Techniques for Generalization

Beyond traditional approaches for mitigating spurious correlation, optimization techniques aiming at improving generalization have also attracted significant attention. Sharpness-Aware Minimization [21] (SAM) is a prominent method that encourages convergence towards flat minima, which correlates with improved generalization capabilities. SAM enhances generalization by minimizing the worst-case loss within a neighborhood around the model parameters, thus favoring broader minima. However, SAM overlooks group-specific sharpness variations, which may unintentionally intensify disparities among subgroups. More recently, variations of SAM have been proposed. The authors of [25] proposed adaptive-SAM methods, and in [26], additional steps were incorporated to minimize the surrogate gap between original and perturbed losses explicitly, thereby enhancing the sharpness-aware optimization process further. Nonetheless, these methods still lack direct consideration of intergroup sharpness disparities, leaving a crucial gap in achieving true distributional robustness across diverse populations.

In this paper, we address this limitation by proposing G^2-SAM. Our method explicitly guides sharpness-aware training to consider subgroup-specific geometry, which leads to robust training against spurious correlations and a reduction in performance disparities.

3. Preliminary Information

We consider a supervised learning scenario where we are given a dataset consisting of input–label pairs drawn independently and identically distributed (i.i.d.) from an unknown joint training distribution $\mathcal{D}_{\text{train}}$.

Here, $\mathbf{x} \in \mathcal{X}$ denotes the input feature vector and $y \in \mathcal{Y}$ is the corresponding label. The goal of Empirical Risk Minimization (ERM) is to learn a parameterized model f_θ, which minimizes the expected loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{train}}}[\ell(f_\theta(\mathbf{x}), y)],$$

where $\ell(\cdot,\cdot)$ denotes a suitable loss function, and θ represents the parameters of the model.

However, ERM-based models often suffer from poor generalization under distributional shifts, primarily because they tend to exploit spurious correlations. Spurious correlations refer to non-causal features that correlate strongly with the labels in the training data but do not generalize.

To express this formally, let us define the test data distribution as $\mathcal{D}_{\text{test}}$. The core assumption in an environment with a distributional shift can be expressed as

$$\mathcal{D}_{\text{train}}(x, y) \neq \mathcal{D}_{\text{test}}(x, y).$$

This difference in distributions often arises because the conditional probability between a spurious attribute a and the label y changes from training to testing, i.e., $\mathcal{D}_{\text{train}}(y|a) \neq \mathcal{D}_{\text{test}}(y|a)$.

Standard ERM only minimizes the expected loss over the training distribution $\mathcal{D}_{\text{train}}$:

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}}[\ell(f_\theta(x), y)]$$

If the model learns to use the spurious attribute a as a shortcut, this shortcut is no longer valid in the $\mathcal{D}_{\text{test}}$ environment, leading to a sharp degradation in performance. This phenomenon leads to significant performance degradation for minority or vulnerable subgroups within the data.

Formally, we can express the dataset as a collection of subgroups $\{\mathcal{D}_g\}_{g=1}^{G}$, where each subgroup represents different combinations of target and spurious attributes. The training and test distributions can be seen as mixtures of these same underlying groups but with different mixing proportions (π_g):

$$\mathcal{D}_{\text{train}} = \sum_{g=1}^{G} \pi_g^{\text{train}} P(x, y|g) \quad \text{and} \quad \mathcal{D}_{\text{test}} = \sum_{g=1}^{G} \pi_g^{\text{test}} P(x, y|g), \quad \text{where} \quad \pi^{\text{train}} \neq \pi^{\text{test}}$$

The robustness of the model is then evaluated by its performance on the worst-performing subgroup:

$$\mathcal{L}_{\text{worst}}(\theta) = \max_{g \in \{1,\ldots,G\}} \mathbb{E}_{(x,y)\sim\mathcal{D}_g}[\ell(f_\theta(x), y)]$$

A method like GDRO tries to achieve models with enhanced generalization performance by directly optimizing $\mathcal{L}_{\text{worst}}^{\text{train}}(\theta)$. This strategy intentionally improves performance on minority groups during training to reduce the model's reliance on spurious correlations.

On the other hand, another body of work has also tried to address the generalization issue, which is called Sharpness-Aware Minimization (SAM). SAM aims to find parameters θ located in flatter regions of the loss landscape, which are empirically linked to better generalization. To achieve this, SAM seeks to minimize the loss in the worst-case scenario within a neighborhood of the current parameters.

The core idea is encapsulated in the following min-max optimization problem:

$$\min_{\theta} \mathcal{L}_{\text{SAM}}(\theta) \quad \text{where} \quad \mathcal{L}_{\text{SAM}}(\theta) = \max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}(\theta + \epsilon).$$

Here, $\mathcal{L}(\theta)$ is the standard training loss (e.g., cross-entropy), and ρ defines the radius of the ℓ_2-norm ball around the current parameters θ. Solving the inner maximization problem exactly is computationally expensive. Therefore, SAM approximates the solution by taking a first-order Taylor expansion of the loss function with respect to ϵ:

$$\mathcal{L}(\theta + \epsilon) \approx \mathcal{L}(\theta) + \epsilon^T \nabla_\theta \mathcal{L}(\theta).$$

To maximize this approximated loss, the perturbation ϵ should be aligned with the gradient. The solution that maximizes the inner objective is thus found by scaling the gradient to the boundary of the neighborhood:

$$\epsilon(\theta) = \rho \frac{\nabla_\theta \mathcal{L}(\theta)}{\|\nabla_\theta \mathcal{L}(\theta)\|_2}.$$

By substituting this 'worst-case' perturbation back into the objective, the final SAM loss becomes:

$$\mathcal{L}_{\text{SAM}}(\theta) = \mathcal{L}(\theta + \epsilon(\theta)).$$

Minimizing this objective encourages convergence towards flatter minima where the loss remains low even after this gradient-based perturbation, thereby enhancing generalization.

4. Method: Group-Gap Guided Sharpness-Aware Minimization (G^2-SAM)

We now propose a novel method, Group-gap Guided Sharpness-Aware Minimization (G^2-SAM), that extends the SAM framework by incorporating subgroup-specific considerations into the optimization process using group-wise SAM loss gaps. By utilizing group labels, G^2-SAM explicitly addresses subgroup-specific sharpness disparities to improve distributional robustness and fairness.

4.1. Step 1: Identification of Worst-Performing Subgroups

First, we calculate the standard Empirical Risk Minimization (ERM) loss for each subgroup g:

$$\mathcal{L}_g(\theta) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_g}[\ell(f_\theta(\mathbf{x}), y)]. \tag{1}$$

Based on these loss values, we identify the set of worst-performing subgroups, denoted as G_{worst}, which contains the groups whose losses are in the top κth percentile.

$$G_{\text{worst}} = \{g \mid \mathcal{L}_g(\theta) \text{ is in the top } \kappa\text{th percentile}\}. \tag{2}$$

This initial step focuses our sharpness-aware optimization on the most vulnerable subgroups. By selecting a set of the worst-performing groups (the top κth percentile) rather than just the single worst group, our approach gains stability against outliers and can address systemic biases that may affect multiple minority subgroups simultaneously.

4.2. Step 2: Focused Sharpness Calculation and Regularization

Next, for the worst-performing subgroups identified in the previous step ($g \in G_{\text{worst}}$), we calculate the group-wise SAM loss gap, Δ_g. This is defined as the difference between the SAM-perturbed loss and the original loss for that subgroup:

$$\Delta_g = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_g}\Big[\ell(f_{\theta+\epsilon_g}(\mathbf{x}), y) - \ell(f_\theta(\mathbf{x}), y)\Big], \tag{3}$$

where the perturbation ϵ_g is computed only for $g \in G_{\text{worst}}$ as

$$\epsilon_g = \rho \cdot \frac{\nabla_\theta \mathcal{L}_g(\theta)}{\|\nabla_\theta \mathcal{L}_g(\theta)\|_2}. \tag{4}$$

We then formulate the intergroup gap regularization term, $\mathcal{R}_{gap}$, by averaging the sharpness gaps of these selected subgroups:

$$\mathcal{R}_{gap} = \frac{1}{|G_{\text{worst}}|} \sum_{g\in G_{\text{worst}}} \Delta_g. \tag{5}$$

This approach directly targets the geometric instability of the groups that are empirically struggling the most. The use of a percentile (κ) makes our method adaptive to datasets with varying numbers of subgroups. Instead of relying on a single, potentially noisy worst group, we focus on the tail of the performance distribution, ensuring the model addresses systematic disparities. For all experiments in this paper, we set $\kappa = 25$, a value that empirically provided a good balance between focusing on the most critical groups and maintaining training stability across our benchmarks.

4.3. Step 3: Final Optimization Objective

The final optimization objective of our revised method integrates the average loss of the worst-performing groups with the proposed focused regularization term. This results in the following combined optimization problem:

$$\min_{\theta} \left[\frac{1}{|G_{\text{worst}}|} \sum_{g \in G_{\text{worst}}} \mathcal{L}_g(\theta) + \lambda \cdot \mathcal{R}_{gap} \right], \tag{6}$$

where λ is a hyperparameter controlling the strength of the regularization.

Through this method, our approach effectively reduces the performance disparity between subgroups by ensuring the model finds flat minima beneficial to the most vulnerable groups, ultimately achieving improved robustness and fairness under distributional shifts. The overall process is summarized in Algorithm 1.

Algorithm 1 G^2-SAM training algorithm

Require: Training data $\mathcal{D} = \bigcup_{g=1}^{G} \mathcal{D}_g$, model with parameters θ, learning rate η, neighborhood size ρ, regularization strength λ.

for each training step **do**
// 1. Calculate ERM loss for all groups in the minibatch to identify the worst-performing ones
  for each group $g \in \{1, \ldots, G\}$ **do**
    $\mathcal{L}_g(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_g}[l(f_\theta(x), y)]$
  end for
// 2. Select top κ% groups based on their high ERM loss
  Identify the set of groups with the highest losses:
  $G_{\text{worst}} \leftarrow \{g \mid \mathcal{L}_g(\theta)$ is in the top κth percentile$\}$
// 3. Calculate sharpness gaps only for the selected worst-performing groups
  for each selected group $g \in G_{\text{worst}}$ **do**
    Calculate group-specific perturbation: $\epsilon_g \leftarrow \rho \cdot \frac{\nabla_\theta \mathcal{L}_g(\theta)}{\|\nabla_\theta \mathcal{L}_g(\theta)\|_2}$
    Calculate group-wise SAM loss gap: $\Delta_g \leftarrow \mathcal{L}_g(\theta + \epsilon_g) - \mathcal{L}_g(\theta)$
  end for
// 4. Calculate the regularization term from the gaps of the selected groups
  Calculate the intergroup gap regularization term:
  $\mathcal{R}_{gap} \leftarrow \frac{1}{|G_{\text{worst}}|} \sum_{g \in G_{\text{worst}}} \Delta_g$
// 5. Define the final objective using the losses and gaps of the selected groups
  Define the final optimization objective:
  $\mathcal{L}_{\text{final}}(\theta) \leftarrow \frac{1}{|G_{\text{worst}}|} \sum_{g \in G_{\text{worst}}} \mathcal{L}_g(\theta) + \lambda \cdot \mathcal{R}_{gap}$
// 6. Update the model parameters
  Update the model parameters:
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{final}}(\theta)$
end for

5. Experimental Results

We now validate the performance of G^2-SAM. By experimenting on the various benchmark datasets with spurious correlations, we show that promoting geometric flatness across all subgroups via G^2-SAM leads to superior distributional robustness. We structure our validation across four key analyses. In Section 5.2, a controlled study on synthetic data (CMNIST) is conducted to examine the effect of our approach across varying bias strengths, including a qualitative visualization of the induced loss landscapes. We provide an evaluation on standard binary spurious correlation benchmarks (Waterbirds and CelebA) in Section 5.3. We also provide a more challenging evaluation using non-binary, multigroup benchmarks like UTKFace and FairFace to test the scalability of our approach for real-world applications Section 5.4. Finally, we conduct an ablation study on the regularization strength to analyze the effect of G^2-SAM's regularization term Section 5.5.

5.1. Experimental Setup

5.1.1. Implementation Details

All models were implemented using the PyTorch 3.10 framework and trained on a single NVIDIA RTX 4090 GPU. For a fair and robust comparison, we used a ResNet-50 backbone pretrained on the ImageNet dataset for all experiments.

Training configurations: We used the SGD optimizer with a constant learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}. All models were trained for 50 epochs with a batch size of 64. For all sharpness-aware methods, the neighborhood size ρ was set to 0.05. The regularization hyperparameter λ in the G^2-SAM objective was set to 0.1 for all datasets.

Data augmentation: During training, we applied standard data augmentation, including random resized cropping to 224×224 with random horizontal flipping. For validation and testing, we only applied a center crop followed by normalization.

Evaluation protocol: All reported results are the mean and standard deviation computed over three independent runs. Our primary evaluation metrics are Worst-Group Accuracy, which directly measures robustness to distributional shifts, and Average Accuracy to assess overall performance.

5.1.2. Baselines

We compare G^2-SAM against the following baselines to demonstrate its performance.

- ERM (Empirical Risk Minimization): The standard training paradigm that minimizes the average loss over the training data. It is known to be highly susceptible to learning spurious correlations.
- GDRO [15] (Group Distributionally Robust Optimization): A strong robust optimization baseline that explicitly minimizes the loss of the worst-performing group.
- SUBG [27] (Subgroup Balancing): A resampling method that increases the loss contribution of samples from minority groups to create a balanced training objective.
- DFR [24] (Deep Feature Reweighting): A two-stage method that first trains a model with ERM and then fine-tunes the final linear classifier on a group-balanced dataset.
- SAM [21] (Sharpness-Aware Minimization): A key baseline that seeks flat minima in the loss landscape to improve generalization.
- ASAM [25] (Adaptive Sharpness-Aware Minimization): A variant of SAM that introduces an adaptive perturbation size, making the optimization process invariant to the scaling of model weights and improving training stability.
- GSAM [26] (Surrogate-Gap Minimization): An extension of SAM that incorporates an additional objective to explicitly minimize the surrogate gap, defined as the difference between the original loss and the perturbed loss.

5.1.3. Datasets

We perform experiments with five datasets. Three of them are datasets with binary spurious attributes, and two of them are more complex datasets with non-binary spurious attributes.

- CMNIST: We construct a variant of the MNIST dataset to create a controlled environment for analyzing the impact of spurious correlations. The task is binary digit classification (digits 0–4 vs. 5–9). We introduce a spurious color cue by correlating digit color with the label. The bias ratio $p \in \{0.9, 0.95, 0.99\}$ defines the strength of this correlation; for a fraction p of the training data, digits < 5 are red and digits ≥ 5 are green. For the remaining $1 - p$ fraction, this correlation is reversed. The test set is synthetically balanced, with colors distributed equally across both label classes, thereby removing the spurious cue and evaluating the model's reliance on invariant features (the digit's shape).

- Waterbirds: Waterbirds is a widely used benchmark designed specifically to test robustness against spurious correlations. It was created by combining images of birds from the CUB dataset with backgrounds from the Places dataset. The task is to classify waterbirds vs. landbirds. The spurious correlation is the background: most waterbirds are shown on water backgrounds, and most landbirds are on land backgrounds. The minority groups, such as landbirds on water, are the primary challenge for standard ERM models.
- CelebA: CelebA is a large-scale dataset of celebrity faces with various attribute annotations. Following standard experimental protocols in the literature [15,19,22,24], we select 'Blond Hair' as the target attribute and 'Gender' as the spurious (bias) attribute. The dataset exhibits a strong natural correlation where the majority of individuals with blond hair are female and the majority of individuals with non-blond hair are male. The model must learn to classify hair color without relying on gender cues.
- UTKFace: UTKFace is a large-scale face dataset with annotations for age, gender, and race. To evaluate performance in a non-binary bias setting, we use 'Gender' (binary) as the target attribute and 'Age' as the bias attribute. Age is categorized into 8 distinct groups (0–20, 11–20, 21–30, . . . , 71–80, 81+), resulting in $2 \times 8 = 16$ distinct subgroups. This setup tests the model's ability to maintain fair gender classification performance across a wide age distribution.
- FairFace: FairFace is a face attribute dataset specifically designed to be balanced across seven racial categories (White, Black, Latino and Hispanic, East Asian, Southeast Asian, Indian, Middle Eastern). This inherent balance makes it an ideal benchmark for evaluating the effectiveness of bias mitigation methods. Unlike datasets that require simple data balancing, FairFace allows us to isolate and measure whether an advanced learning algorithm can successfully reduce a model's reliance on spurious correlations. We use Gender (binary) as the target attribute and Race as the bias attribute, creating $2 \times 7 = 14$ subgroups. This benchmark directly tests a model's ability to achieve robust performance across all subgroups, demonstrating the efficacy of methods that go beyond simple distribution balancing techniques.

5.2. Analysis in Controlled Bias Environments on CMNIST

In this section, we first perform experiments on the CMNIST dataset with varying bias strength. This experiment serves as a controlled analysis to directly probe and visualize the effect of G^2-SAM's core mechanism. By systematically increasing the bias ratio (the proportion of the data that follows spurious correlation between color and digit), we observe how different optimizers behave under increasing pressure to learn the spurious color correlation. A visualized example for the CMNIST dataset is given in Figure 1.

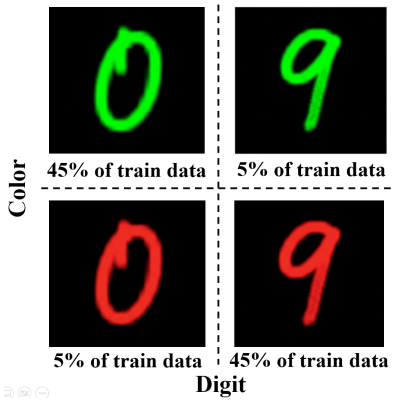


Figure 1. Example for the CMNIST dataset. The digit '0' is spuriously correlated with the color red, and '9' is spuriously correlated with the color green.

5.2.1. Quantitative Analysis

Table 1 presents the performance of G^2-SAM against key baselines on CMNIST. The results clearly show that as the bias ratio increases, the performance of group-agnostic sharpness-aware methods deteriorates significantly. While SAM maintains a high Average Accuracy by successfully classifying the majority groups, its Worst-Group Accuracy deteriorates significantly, indicating a complete reliance on the spurious feature color. For instance, at a bias ratio of 0.99, SAM's Worst-Group Accuracy drops to 38.0%, showing lower accuracy than the random guess. In contrast, both GDRO and G^2-SAM demonstrate remarkable resilience to distributional shift. G^2-SAM consistently achieves the highest Worst-Group Accuracy across all bias levels, even surpassing the GDRO. This result provides strong initial evidence that regulating group-wise sharpness is a highly effective strategy for mitigating the impact of strong spurious correlations. Moreover, the widening performance gap between SAM and G^2-SAM on Worst-Group Accuracy as bias intensifies directly supports the central claim that group-agnostic flatness is insufficient for achieving true distributional robustness.

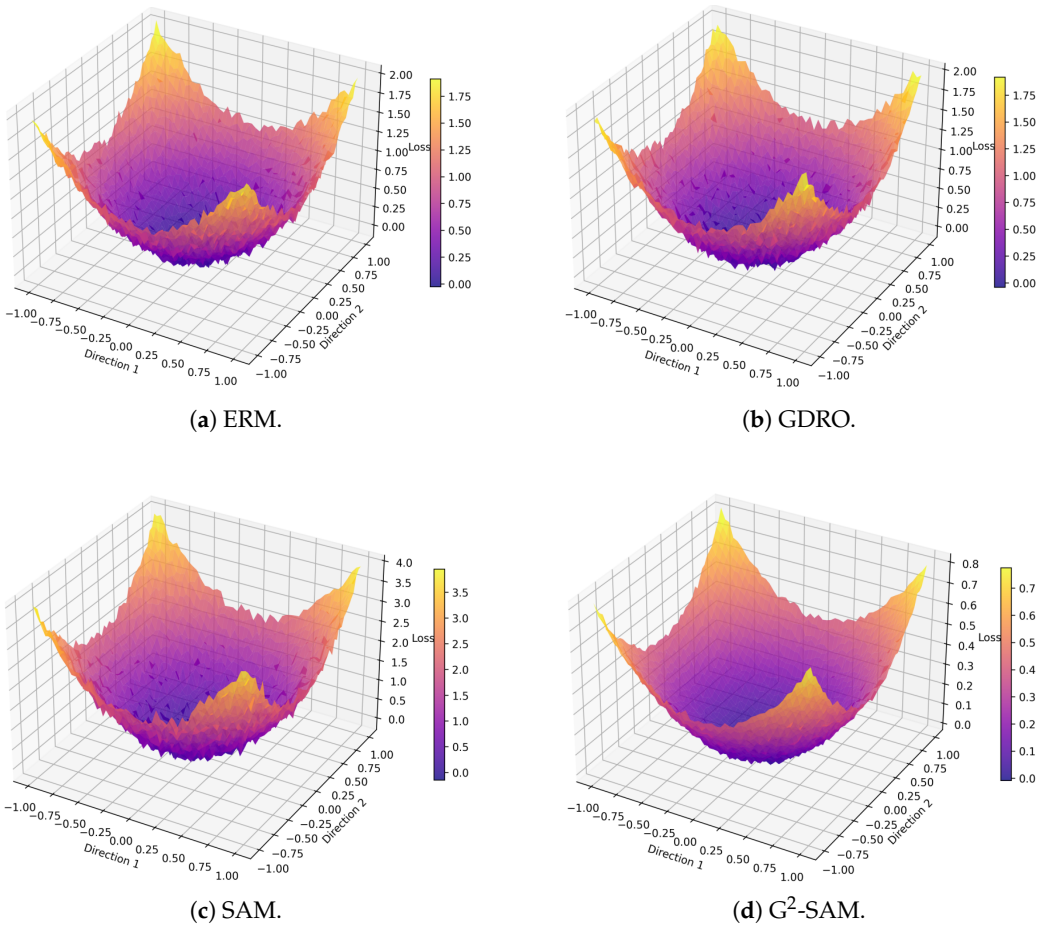
Table 1. Performance comparison on the CMNIST dataset under varying bias ratios. We report Average Accuracy (%) and Worst-Group Accuracy (%). G^2-SAM consistently achieves the highest Worst-Group Accuracy, demonstrating its superior robustness as the spurious correlation becomes stronger. We use bold text to indicate the best results.

Method	**Bias Ratio = 0.90**		**Bias Ratio = 0.95**		**Bias Ratio = 0.99**	
	Worst Acc.	**Avg. Acc.**	**Worst Acc.**	**Avg. Acc.**	**Worst Acc.**	**Avg. Acc.**
ERM	92.4 ± 1.5	96.5 ± 0.9	86.4 ± 2.1	94.0 ± 1.1	41.8 ± 1.3	71.9 ± 2.4
GDRO [15]	93.8 ± 1.2	96.1 ± 0.7	88.4 ± 2.3	93.5 ± 1.0	78.4 ± 1.6	82.5 ± 1.9
SAM [21]	91.0 ± 1.4	95.5 ± 0.8	86.0 ± 2.2	93.3 ± 0.6	38.0 ± 0.9	72.0 ± 1.8
G^2-SAM (proposed)	$\mathbf{95.6 \pm 0.6}$	$\mathbf{97.6 \pm 0.5}$	$\mathbf{93.4 \pm 1.1}$	$\mathbf{96.6 \pm 0.7}$	$\mathbf{86.4 \pm 1.0}$	$\mathbf{88.6 \pm 1.3}$

5.2.2. Qualitative Analysis: Loss Landscape Visualization

To understand the geometric rationale behind our quantitative results, we visualize the loss landscapes for ERM, SAM, GDRO, and G^2-SAM under a high bias ratio (0.99) in Figure 2. These visualizations provide a clear, intuitive explanation for the performance disparities. For SAM, the training algorithm fails to find a flat basin because it does not account for the geometric disparity between groups. Similarly, both ERM and GDRO converge to solutions located in sharp, irregular minima. Although GDRO focuses on improving worst-group performance, it does so without considering the solution's geometric stability, thus finding a precarious optimum sensitive to minor data perturbations.

Conversely, the landscape induced by G^2-SAM demonstrates the effectiveness of its regularization. The $\mathcal{R}_{gap}$ term penalizes the divergence in group-wise sharpness, forcing the optimizer to find a solution that resides in a region that is simultaneously flat for both the majority and minority groups. This shared flat minimum may not be the absolute flattest point for the majority group alone, but it represents a far more robust and equitable solution. This visual evidence validates the core premise of G^2-SAM: enforcing group-wise geometric similarity is the key to robust generalization. This analysis reframes G^2-SAM not merely as a debiasing algorithm that mitigates spurious correlation but as a more fundamental geometric regularizer. Unlike reweighting schemes that only manipulate the magnitude of group losses, G^2-SAM directly shapes the curvature of the loss surface, yielding solutions that are inherently more stable for all subgroups.



(**a**) ERM. (**b**) GDRO.

(**c**) SAM. (**d**) G^2-SAM.

Figure 2. Loss landscape of trained neural networks.

5.3. Robustness on Binary Spurious Correlation Benchmarks

Having established the mechanism of G^2-SAM on synthetic data, we now evaluate its efficacy on standard, real-world benchmarks. Success on CelebA and Waterbirds is critical for demonstrating practical relevance in the field of fairness and robustness. For these binary datasets, we adopt an approach of fine-tuning the last layer after the main training with G^2-SAM. Table 2 presents a comprehensive comparison of G^2-SAM against the full suite of baselines. On both datasets, G^2-SAM achieves state-of-the-art Worst-Group Accuracy, significantly outperforming all other methods. On CelebA, G^2-SAM achieves a Worst-Group Accuracy of 90.2%, surpassing the strong GDRO baseline by over 3 percentage points. On the challenging Waterbirds dataset, it achieves a Worst-Group Accuracy of 92.9%, again establishing a new benchmark. These results powerfully illustrate the limitations of group-agnostic flatness. As hypothesized, SAM improves Average Accuracy over ERM but provides only marginal gains in Worst-Group Accuracy. This demonstrates that seeking 'general' flatness without accounting for subgroup structure is not a viable strategy for distributional robustness. G^2-SAM successfully bridges this gap, inheriting the generalization benefits of sharpness-aware optimization while incorporating the group-specific focus required for robust performance. Its ability to outperform GDRO, which directly optimizes the worst-group loss, suggests that optimizing for group-wise geometric properties can be a more effective path to robustness than simply optimizing the worst-group loss magnitude, potentially leading to better solutions with more favorable optimization trade-offs.

Table 2. Performance comparison on binary spurious correlation benchmarks. We report mean ± std over 3 runs for Average Accuracy (%) and Worst-Group Accuracy (%). G^2-SAM achieves the highest Worst-Group Accuracy on both datasets. We use bold text to indicate the best results.

Dataset Method	Waterbirds		CelebA	
	Worst Acc.	Average Acc.	Worst Acc.	Average Acc.
ERM	77.1 ± 2.4	89.2 ± 0.8	46.5 ± 3.1	**95.8 ± 0.1**
SubG [27]	83.0 ± 1.0	91.6 ± 0.6	85.7 ± 0.8	93.0 ± 0.4
GDRO [15]	91.4	93.5	88.9	92.4
DFR [24]	89.7 ± 2.6	92.2 ± 1.1	88.8 ± 0.5	91.5 ± 0.3
SAM [21]	78.5 ± 1.0	89.6 ± 0.4	50.2 ± 5.8	**95.8 ± 0.2**
ASAM [25]	63.2 ± 1.1	93.6 ± 0.7	48.8 ± 5.6	95.6 ± 0.2
GSAM [26]	75.4 ± 1.4	91.6 ± 0.6	45.0 ± 7.3	95.6 ± 0.3
G^2-SAM (proposed)	**91.9 ± 0.9**	**94.1 ± 0.5**	**89.7 ± 0.7**	91.8 ± 0.4

5.4. Robustness on Non-Binary Spurious Correlation Benchmarks

Real-world biases are often complex and non-binary. Spurious correlations can exist across numerous attributes like age and race, which are non-binary attributes. We omit the last-layer fine-tuning step for multigroup benchmarks, as ensuring sufficient samples from all 14–16 subgroups per minibatch would require a prohibitively large batch size, making the training process computationally infeasible. This set of experiments tests the scalability and generalizability of G^2-SAM's core mechanism, minimizing the gap between the maximum and minimum group sharpness in more complex, multigroup settings.

The results on UTKFace (16 groups) and FairFace (14 groups) are presented in Table 3. Once again, G^2-SAM demonstrates superior performance, achieving the highest Worst-Group Accuracy on both datasets. This is particularly significant because methods that perform well in a simple majority–minority binary case may struggle when balancing performance across many groups. The formulation of G^2-SAM's regularizer, which focuses on the minor groups in terms of sharpness, proves to be highly scalable. It does not depend on the total number of groups, making it computationally efficient and stable.

Table 3. Performance comparison on non-binary spurious correlation benchmarks. We report mean ± std over 3 runs for Average Accuracy (%) and Worst-Group Accuracy (%). G^2-SAM demonstrates superior scalability and robustness in multigroup scenarios. We use bold text to indicate the best results.

Dataset Method	UTKFace		FairFace	
	Worst Acc.	Average Acc.	Worst Acc.	Average Acc.
ERM	75.5 ± 0.6	91.1 ± 0.2	93.0 ± 0.3	94.6 ± 0.2
SubG [27]	77.1 ± 0.5	91.3 ± 0.4	93.2 ± 0.3	95.5 ± 0.1
GDRO [15]	66.1 ± 4.1	82.4 ± 3.9	91.0 ± 1.4	94.1 ± 0.5
DFR [24]	63.8 ± 4.3	90.6 ± 0.5	92.8 ± 0.3	95.7 ± 0.1
SAM [21]	76.2 ± 0.4	90.9 ± 0.3	93.2 ± 0.4	95.6 ± 0.2
ASAM [25]	76.7 ± 0.9	91.6 ± 0.7	92.4 ± 0.4	95.4 ± 0.6
GSAM [26]	78.0 ± 0.7	91.3 ± 0.4	93.5 ± 0.8	95.7 ± 0.4
G^2-SAM (proposed)	**78.1 ± 0.9**	**91.8 ± 0.3**	**93.6 ± 0.5**	**95.8 ± 0.1**

The strong performance in these multigroup scenarios suggests that G^2-SAM is not a specialized solution for binary bias but rather a general principle for robust optimization. While GDRO must actively balance the losses of all 10 or 7 groups, G^2-SAM's focus on the sharpness spectrum may provide a more stable optimization target, preventing the optimizer from being pulled in too many conflicting directions. This success hints at the potential for G^2-SAM to address even more complex intersectional fairness challenges,

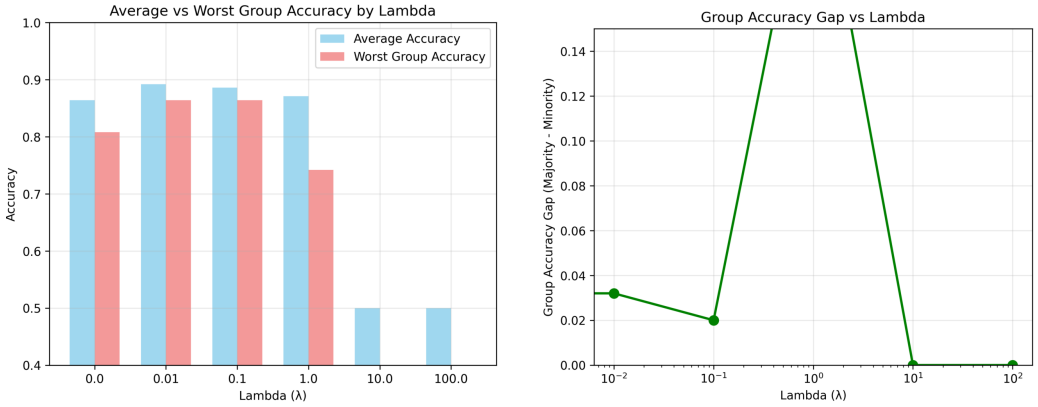
where groups are defined by the intersection of multiple sensitive attributes and the total number of subgroups can become very large.

5.5. Ablation Study: Impact of the Regularization Hyperparameter λ

To isolate and understand the contribution of our proposed intergroup gap regularization, we conduct an ablation study on the hyperparameter λ. This coefficient, as defined in our final objective function (6), controls the strength of the regularization term $\mathcal{R}_{gap}$, which is designed to minimize the disparity in sharpness across subgroups.

5.5.1. Quantitative Analysis

We evaluate the performance of G^2-SAM on the CMNIST datasets while varying λ across the set $\{0, 0.1, 1.0, 10.0, 100.0\}$. The case where $\lambda = 0$ removes our proposed intergroup gap regularizer, $\mathcal{R}_{gap}$. Consequently, the objective is reduced to minimizing the average loss of the groups belonging to the top κth percentile with the highest ERM loss ($\mathcal{L}_g(\theta)$). This approach is fundamentally different from GDRO, which targets the single group with the highest loss value, whereas our $\lambda = 0$ baseline targets a set of groups with the highest loss. The results are presented in Figure 3.



(**a**) Major and minor group accuracy. (**b**) Group accuracy gap.

Figure 3. Performance of CMNIST dataset with varying λ values.

As shown, setting $\lambda = 0$ results in a 5.2% drop in Worst-Group Accuracy compared to the optimized G^2-SAM (when $\lambda = 0.1$), confirming that simply minimizing the worst-group perturbed loss is insufficient. As we increase λ from 0 to 0.1, we observe a steady decrease in the performance gap between major and minor groups, which is indicated in Figure 3b. This demonstrates the direct positive impact of the $\mathcal{R}_{gap}$ term. However, when λ is increased further to 100.0, the performance begins to degrade. This suggests that an excessively large regularization term can over-constrain the optimization, lowering the performance for both minor and major groups. This study validates that our intergroup gap regularization is a critical component of G^2-SAM's success and that its influence can be effectively tuned.

5.5.2. Impact on the Loss Landscape

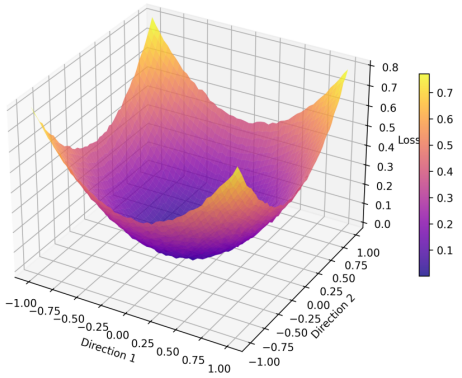
We now provide the quantitative results to analyze the effect of λ on the geometry of the loss landscape and indicate in Figure 4.

- Low λ (e.g., 0): With little to no regularization, the optimizer is not strongly incentivized to close the sharpness gap between groups. The resulting landscape resembles that of standard GDRO, where the solution may be flat for the majority group but remains sharp and precarious for the minority group. The optimizer finds a region

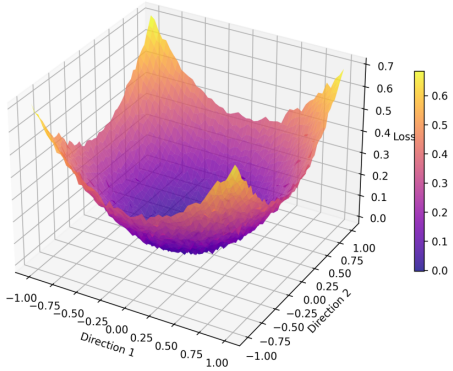
of low loss for the worst group, but this region lacks the geometric stability that ensures robustness.

- Optimal λ (e.g., 0.1): In this regime, the regularization term is strong enough to compel the optimizer to find a solution that is simultaneously flat for all subgroups. It actively penalizes solutions where sharpness is high for any single group, even if the loss magnitude is low. This leads to the desired outcome: a shared, wide, and flat basin in the loss landscape that provides robust generalization for both majority and minority groups.
- High λ (e.g., 10.0, 100.0): When the regularization is too dominant, the optimizer prioritizes making the group-wise sharpness values identical above all else. This can force the model into a suboptimal region of the parameter space. The search for perfect geometric homogeneity overrides the primary goal of accurate classification, leading to a big spike in the loss landscape.

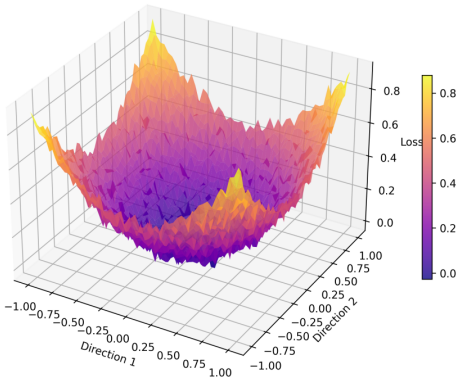
This analysis confirms that the $\mathcal{R}_{gap}$ regularizer directly and effectively manipulates the loss landscape's geometry to achieve distributional robustness, and its strength must be appropriately tuned to achieve the best results.



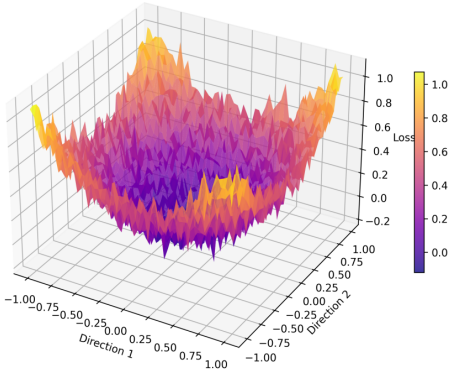
(**a**) $\lambda = 0.1$.



(**b**) $\lambda = 1.0$.



(**c**) $\lambda = 10$.



(**d**) $\lambda = 100$.

Figure 4. Loss landscape of trained neural networks with G^2-SAM with varying λ values.

5.6. Ablation Study: Impact of the Group Selection Strategy

5.6.1. Ablation on the Worst-Group Selection Percentile (κ)

To analyze the impact of our group selection strategy, we conducted an ablation study on the hyperparameter κ, which defines the percentile of worst-performing groups to be included in the regularization term $\mathcal{R}_{gap}$. The choice of κ is critical: a value that is too small may lead to noisy and unstable gradient estimates from focusing on too few samples, while a value that is too large could dilute the optimizer's focus on the most vulnerable

subgroups. We evaluated the model's performance on the FairFace dataset while varying κ from 7% to 50%. The results are presented in Table 4.

Table 4. Performance on the CMNIST (bias = 0.99) dataset as a function of the Worst-Group Selection percentile, κ. Both Worst-Group and Mean Accuracy improve as a more stable set of groups is selected. We use bold text to indicate the best results.

Worst-Group Percentile (κ)	Worst Accuracy	Mean Accuracy
7.1%	76.6	85.9
14.2%	89.60	94.5
28.5%	**93.6**	**95.8**
50%	91.4	94.9

The results clearly demonstrate that performance improves as κ increases within this range. A very small κ (7%) results in significantly lower accuracy, which confirms our hypothesis that relying on too few groups can lead to unstable training. By increasing the percentile, we average the sharpness gaps over a larger and more stable set of groups, which provides a more reliable optimization signal and substantially boosts both Worst-Group and Mean Accuracy. The performance peaked at 28% in this specific setting, so we chose 25% for our main experiments as it provided a robust and effective balance across all tested datasets.

5.6.2. Ablation on Group Selection Strategy

We conducted an additional ablation study to verify that the core mechanism of our method, explicitly targeting the worst-performing subgroups, is essential for its success. In this experiment, we compared our proposed strategy against a baseline where the group-wise sharpness regularization is applied not to the worst-performing groups but to a randomly selected subset of groups at each training step. This helps isolate the contribution of our 'guided' approach. We performed this comparison on the FairFace dataset, and the results are summarized in Table 5.

Table 5. Comparison of our proposed Worst-Group Selection strategy against a Random Group Selection baseline on the FairFace dataset. We use bold text to indicate the best results.

Selection Strategy	Worst Accuracy	Mean Accuracy
Random Selection	92.5	95.1
Worst-Group Selection (ours)	**93.6**	**95.8**

The results clearly indicate that our targeted Worst-Group Selection strategy outperforms the Random Selection baseline on both metrics. Specifically, by focusing the sharpness-aware optimization on the most vulnerable groups, our method achieves a higher worst accuracy (93.6 vs. 92.5) and a better Mean Accuracy (95.8 vs. 95.1). This finding provides strong evidence that the 'group-gap guided' component of G^2-SAM is the primary driver of its improved performance. It confirms that merely applying group-wise regularization is insufficient. The key is to direct the optimization toward the subgroups that need it most, thereby effectively enhancing both robustness and fairness.

5.7. Further Analysis on Fairness Metrics

To provide a broader perspective on our model's performance beyond robustness, we evaluated it against several key fairness metrics. In addition to Worst-Group Accuracy, we measured Equalized Odds and Equal Opportunity. These metrics are crucial for assessing whether a model's performance is equitable across different demographic groups.

For these fairness metrics, a lower score indicates better performance, with a value of zero representing perfect fairness. The experiments were conducted on the Waterbirds dataset. The results, presented in Table 6, demonstrate that our proposed method not only excels in robustness but also significantly improves fairness. Our model, G^2-SAM, achieves the lowest (best) scores for both Equalized Odds (0.0523) and Equal Opportunity (0.0155). This indicates that our model provides more equitable predictions for both positive and negative classes across the majority and minority groups. In contrast, standard baselines like ERM and SAM exhibit very high values, highlighting their tendency to produce biased outcomes. This analysis reinforces our central claim that by addressing group-specific geometry in the loss landscape, our method achieves a solution that is not only robust but also substantially fairer.

Table 6. Performance comparison on the Waterbirds dataset, focusing on fairness metrics. For Equalized Odds and Equal Opportunity, lower values indicate better fairness. We use bold text to indicate the best results.

Method	Equalized Odds	Equal Opportunity
ERM	0.6091	0.3084
GDRO	0.1373	0.0561
SAM	1.0214	0.4143
DFR	0.0881	0.0327
G^2-SAM (ours)	**0.0523**	**0.0155**

5.8. Ablation on the Last-Layer Retraining Method

To better isolate the contribution of our optimization strategy, we conducted an ablation study to analyze the impact of the last-layer retraining technique. This study helps to distinguish the performance gains from the main training phase versus the fine-tuning phase. The results on the Waterbirds dataset are presented in Table 7.

The results show that applying DFR provides a significant performance boost to all methods, underscoring its effectiveness as a fine-tuning strategy. Notably, when combined with DFR, standard SAM achieves a very strong Worst-Group Accuracy, demonstrating that a flat minima-seeking optimizer provides a good foundation for subsequent debiasing. While the Worst-Group Accuracy of SAM + DFR is highly competitive and marginally higher than our method, our G^2-SAM achieves the best overall performance, evidenced by its superior performance. This suggests that our G^2-SAM optimizer learns a more robust and generalizable feature representation during the initial training phase.

Table 7. Ablation study on the effect of last-layer retraining (DFR) combined with different optimization methods on the Waterbirds dataset. We use bold text to indicate the best results.

Method	Worst Accuracy	Mean Accuracy
ERM + DFR	89.5	92.6
GDRO + DFR	89.9	90.2
SAM + DFR	92.8	93.5
G^2-SAM (ours)	**91.9**	**94.1**

5.9. Ablation on the Gap Regularizer ($\mathcal{R}_{gap}$)

To quantify the contribution of our core technical idea, we conducted a diagnostic ablation study on the gap regularizer, $\mathcal{R}_{gap}$. For this experiment, we created a variant of our method that removes the $\mathcal{R}_{gap}$ term from the final objective function. This ablated model only minimizes their standard loss without explicitly regularizing the sharpness disparity between them.

We compared the performance of our full G^2-SAM method against this variant on FairFace datasets. The results are presented in Table 8.

Table 8. Ablation study results on the effect of the $\mathcal{R}_{gap}$ regularizer. Removing the regularizer leads to a substantial drop in Worst-Group Accuracy on both datasets. We use bold text to indicate the best results.

Method	Worst Accuracy	Mean Accuracy
G^2-SAM	**93.6**	**95.8**
G^2-SAM without $\mathcal{R}_{gap}$	90.6	94.9

The results demonstrate the critical importance of the $\mathcal{R}_{gap}$ term. Its removal leads to a dramatic degradation in Worst-Group Accuracy. This confirms that simply targeting the worst-performing groups with a standard loss objective is insufficient for achieving distributional robustness. The explicit minimization of the intergroup sharpness gap is the key mechanism driving the performance gains, validating the core hypothesis of our work.

6. Computational Cost Analysis

To provide a clear picture of the practical implications of our proposed method, this section presents an analysis of its computational overhead. While G^2-SAM requires additional computations to assess group-wise sharpness, we empirically demonstrate that this overhead is a reasonable trade-off for the substantial gains in distributional robustness.

Theoretically, standard training paradigms like ERM and GDRO require a single forward and backward pass per optimization step. Sharpness-Aware Minimization (SAM) performs an additional ascent step to find the worst-case perturbation, effectively doubling the cost to two forward and two backward passes. Our method, G^2-SAM, computes gradients specifically for the set of worst-performing subgroups, G_{worst}, to calculate the group-specific perturbations ϵ_g and the subsequent sharpness gaps Δ_g. Consequently, the computational cost scales with the number of selected worst groups, $|G_{\text{worst}}|$, rather than the total number of groups, G. This design ensures that G^2-SAM remains scalable even in scenarios with a large number of subgroups. To quantify the actual cost, we measured the single-epoch runtime and peak VRAM usage for each method on the Waterbirds dataset using a single NVIDIA RTX 4090 GPU. The results are summarized in Table 9.

Table 9. Comparison of VRAM usage and training time per epoch. All experiments were conducted on a single NVIDIA RTX 3090 GPU. Time is reported in seconds.

Method	VRAM Usage (GB)	Runtime (s)
ERM	5.62	87.52
GDRO	5.58	91.78
SAM	5.82	93.08
G^2-SAM (ours)	8.48	111.96

As shown in Table 9, G^2-SAM incurs a 51% increase in VRAM usage and a 26% increase in runtime compared to the ERM training. This is an expected consequence of the additional gradient computations required for the worst-performing groups. However, we argue that this increase in computational cost is a worthwhile and justifiable trade-off for the significant improvements in Worst-Group Accuracy demonstrated throughout our experiments.

7. Limitation and Future Work

Despite its strong performance, a practical limitation of our framework is its sensitivity to the regularization hyperparameter, λ, which requires careful manual tuning to balance the optimization objectives effectively. This highlights a valuable direction for future research. Developing an adaptive method to automatically tune this hyperparameter would enhance the framework's usability and robustness. Further research could also explore the application of G^2-SAM to other domains such as natural language processing or medical imaging, as well as extending the framework to automatically tune for hyperparameters such as κ and λ and developing the theoretical justification of our proposed regularizer.

8. Conclusions

In this work, we introduced Group-gap Guided Sharpness-Aware Minimization (G^2-SAM), a novel optimization framework designed to achieve distributional robustness by addressing the heterogeneity of sharpness across different data subgroups. We identified that group-agnostic methods often fail to find flat minima that are flat for both majority and minority groups, potentially exacerbating performance disparities. G^2-SAM remedies this by explicitly minimizing the disparity in group-wise sharpness, guiding the optimizer towards solutions that are uniformly flat for all subgroups. Our comprehensive experiments validated the superiority of this approach, with G^2-SAM consistently establishing a new state of the art in worst-group performance across synthetic, binary, and challenging non-binary benchmarks. The success of G^2-SAM underscores the critical importance of considering group-specific geometry in the loss landscape for building fair and reliable models.

Author Contributions: Methodology, S.J.; Software, S.M.; Validation, J.S.; Writing—Original Draft, S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: This paper was supported by Konkuk University in 2025.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
2. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment Anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 4–6 October 2023; pp. 4015–4026.
3. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; Volume 48, pp. 173–182.
4. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 12449–12460.
5. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. NeRF: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **2021**, *65*, 99–106. [CrossRef]
6. Shin, J.; Kim, Y.; Hong, S.; Lee, J. Learning Dual Hierarchical Representation for 3D Surface Reconstruction. In Proceedings of the Asian Conference on Computer Vision (ACCV), Hanoi, Vietnam, 8–12 December 2024; pp. 4422–4438.

7. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
8. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
9. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
10. Koh, P.W.; Nguyen, T.; Tang, Y.S.; Mussmann, S.; Pierson, E.; Kim, B.; Liang, P. Concept Bottleneck Models. In Proceedings of the 37th International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 5338–5348.
11. Jung, S.; Chun, S.; Moon, T. Learning fair classifiers with partially annotated group labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10348–10357.
12. Hashimoto, T.; Srivastava, M.; Namkoong, H.; Liang, P. Fairness without demographics in repeated loss minimization. In Proceedings of the 2018 International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1929–1938.
13. Xu, Z.; Chai, Z.; Yuan, C. Towards Calibrated Model for Long-Tailed Visual Recognition from Prior Perspective. *Adv. Neural Inf. Process. Syst.* **2021**, *9*, 7139–7152.
14. Duchi, J.C.; Hashimoto, T.; Namkoong, H. Distributionally robust losses against mixture covariate shifts. *arXiv* **2019**, arXiv:2007.13982.
15. Sagawa, S.; Koh, P.W.; Hashimoto, T.B.; Liang, P. Distributionally Robust Neural Networks. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
16. Joo, H.; Han, H.; Kim, S.; Hong, S.; Lee, J. Constructing Fair Latent Space for Intersection of Fairness and Explainability. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 4156–4165. [CrossRef]
17. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
18. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–10 June 2015.
19. Han, H.; Kim, S.; Joo, H.; Hong, S.; Lee, J. Mitigating Spurious Correlations via Disagreement Probability. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 74363–74382.
20. Chuang, C.Y.; Mroueh, Y. Fair Mixup: Fairness via Interpolation. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 3–7 May 2021.
21. Foret, P.; Kleiner, A.; Mobahi, H.; Neyshabur, B. Sharpness-aware Minimization for Efficiently Improving Generalization. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 3–7 May 2021.
22. Liu, E.Z.; Haghgoo, B.; Chen, A.S.; Raghunathan, A.; Koh, P.W.; Sagawa, S.; Liang, P.; Finn, C. Just train twice: Improving group robustness without training group information. In Proceedings of the 2021 International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 6781–6792.
23. Zhang, M.; Sohoni, N.S.; Zhang, H.R.; Finn, C.; Re, C. Correct-N-Contrast: A Contrastive Approach for Improving Robustness to Spurious Correlations. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; Volume 162, pp. 26484–26516.
24. Kirichenko, P.; Izmailov, P.; Wilson, A.G. Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
25. Kwon, J.; Kim, J.; Park, H.; Choi, I.K. ASAM: Adaptive Sharpness-Aware Minimization for Scale-Invariant Learning of Deep Neural Networks. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; Volume 139, pp. 5905–5914.
26. Zhuang, J.; Gong, B.; Yuan, L.; Cui, Y.; Adam, H.; Dvornek, N.C.; Tatikonda, S.; Duncan, J.; Liu, T. Surrogate Gap Minimization Improves Sharpness-Aware Training. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
27. Idrissi, B.Y.; Arjovsky, M.; Pezeshki, M.; Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. In Proceedings of the First Conference on Causal Learning and Reasoning, Eureka, CA, USA, 11–13 April 2022; Volume 177, pp. 336–351.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.