IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 November 2025, accepted 5 December 2025, date of publication 11 December 2025,
date of current version 24 December 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3643350

RESEARCH ARTICLE

Improving Worst-Group Accuracy With a Filtering-Based Method

MINGI KIM[1], JONGWON RYU[1], AND JUNYEONG KIM[2], (Member, IEEE)
[1]AI Graduate School, Chung-Ang University, Seoul 06974, Repulic of Korea
[2]Department of AI, Chung-Ang University, Seoul 06974, Repulic of Korea

Corresponding author: Junyeong Kim (junyeongkim@cau.ac.kr)

This work was supported by the Chung-Ang University Graduate Research Scholarship in 2023 and supported by the Institute of Information and Communication Technology Planning and Evaluation (IITP) grant funded by the Korea Government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)].

ABSTRACT Spurious correlations in training data often cause models to underperform on minority groups. Existing methods that improve worst-group accuracy (WGA) typically rely on group annotations, complex optimization objectives, or multi-stage pipelines, limiting their practical application. We make a key observation: robust classifiers utilize a sparse set of features, suggesting they selectively focus on core, task-essential features. Based on this insight, we propose a single-stage, annotation-free method that co-trains a learnable binary mask to filter input features. End-to-end training is enabled via a Straight-Through Estimator (STE) and a Gradient Adjustment technique, with the mask learned solely from the main task loss. While many de-biasing methods improve WGA, they often overlook the critical trade-off with overall accuracy. We therefore propose a novel Trade-off Ratio (ToR) metric that quantifies the efficiency of WGA gains relative to the loss in overall accuracy. On four benchmarks (Waterbirds, CelebA, CivilComments, MultiNLI), our method achieves state-of-the-art WGA on CivilComments (81.9%) and the highest ToR score on Waterbirds (12.8).

INDEX TERMS Classification, de-biasing, group robustness, spurious correlation.

I. INTRODUCTION

Classification tasks are widely used in real-world domains, such as sentiment analysis [1] and facial [2] recognition. These tasks often suffer from generalization issues, leading to subpar performance on specific groups [3], [4]. This problem usually arises from spurious correlations, which refer to the frequent co-occurrence of features that are not causally related to the target label.

Neural networks trained with Empirical Risk Minimization (ERM) are vulnerable to these spurious correlations. This vulnerability leads the model to overfit to the majority groups, where these spurious features are prevalent, thereby causing poor generalization performance on minority groups [5], [6], [7]. For example, in classifying water birds from land birds, a model may incorrectly associate the background (water or land) with the class label, since most training images of water birds are taken near water.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhan-Li Sun.

This over-reliance on non-core background cues, as illustrated in Figure 1, inevitably leads to performance degradation for minority groups where the spurious correlation does not hold, such as in classifying a water bird on land or vice versa [8]. Mitigating the models' reliance on such shortcuts is crucial for building fair and robust systems. To this end, a body of research aims to improve group robustness, which is commonly measured by Worst Group Accuracy (WGA), the performance on the most disadvantaged subgroup.

One of the most effective strategies is to leverage group attribute labels directly during training to improve performance for minority groups [10]. However, obtaining such group information is often challenging and raises ethical concerns, limiting its real-world application.

More recent studies have shifted focus by offering a critical insight that the feature representations learned by an ERM-trained model are often sufficient for robust classification, and the primary failure point lies within the final classification layer [11], [12], [13], [14], [15]. This has led to a family of methods that retain only the classifier

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

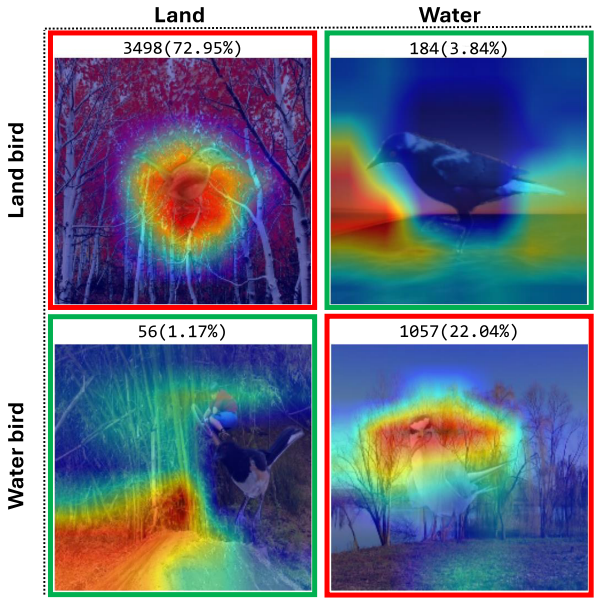


FIGURE 1. We visualized the grad-CAM [9] for the training data of the WaterBirds dataset [10] and annotated the proportions of each group. The red bounding box indicates the major group, which is aligned with the spurious features(background), while the green indicates the minor group, which is not. The model captures core features well for the major group, but for the minor group, it tends to focus on spurious or other irrelevant features.

to correct for spurious correlations. A prominent example, *Deep Feature Reweighting* (DFR) [13], demonstrated the efficacy of this approach by fine-tuning the last layer on a small, group-balanced dataset, achieving impressive group robustness without modifying the base feature extractor.

Despite their effectiveness, these last-layer retraining methods still present several challenges for real-world application. They fundamentally depend on:(1) a held-out dataset with group annotations, which is often unavailable; and (2) a complex two-stage process involving initial training followed by fine-tuning. Furthermore, many de-biasing methods still introduce an undesirable trade-off between improving worst-group accuracy and maintaining overall accuracy [16]. This motivates our central research question: can we mitigate spurious correlations in a single, end-to-end process without requiring any group annotations, while effectively balancing robustness and overall accuracy?

Our method is motivated by a key insight from analyzing the weights of robust classifiers: they consistently utilize a sparse set of features [5], [13]. We interpret this sparsity as an indication that they achieve robustness by selectively prioritizing core, task-essential features while suppressing spurious ones. Based on this insight, we propose a novel dimension filtering method that employs a learnable mask to mitigate spurious correlations within a single end-to-end training process. This mask forces the classifier to focus on core features, thereby mitigating spurious correlations without requiring complex architectural changes. Moreover, our approach achieves competitive worst-group accuracy (WGA) without requiring group annotations and, crucially, minimizes the trade-off between overall accuracy and WGA. To better analyze this balance, we also introduce a simple metric to quantify the efficiency of WGA gains relative to the loss in the overall accuracy. Our main contributions are:

- A simple filtering method for improving group robustness without group labels.
- An analysis demonstrating that the learned filter effectively removes confounding feature dimensions.
- A novel metric to quantify the trade-off between worst-group and overall accuracy.

II. RELATED WORK

A. DEBIASING METHODS

Significant research efforts have been dedicated to achieving fair performance across all groups in classification tasks. While methods such as GroupDRO [10] use group labels during training, they face practical limitations due to the scarcity of such annotations. Consequently, many studies have focused on minimizing the use of group labels during training. Several approaches in this category employ intentionally overfitted models to capture bias features, thereby improving the worst group accuracy. [17], [18], [19]. Other approaches either infer group information for the entire dataset from a small subset or generate proxy group labels [20], [21], [22], [23]. Alternatively, methods such as [24] and [25] assign higher weights to features or logits for worst-performing samples, encouraging the model to focus on these hard examples. However, these methods still pose practical challenges, as they typically require additional models to identify hard samples or rely on proxy group information for robustness.

Recent studies have demonstrated that when spurious correlations exist, the primary cause of poor minority group performance in ERM-trained models resides in the final classification layers rather than the feature extractor [11], [12], [26]. Specifically, instead of fine-tuning the entire model, retraining only the latter layers has been shown to improve robustness [13], [14]. However, these approaches typically rely on a small number of group labels or auxiliary models to construct a balanced dataset for fine-tuning. In contrast, we propose a method that eliminates the need for such multi-stage pipelines, such as separate fine-tuning or re-training steps.

B. FEATURE SELECTION

Feature selection is a fundamental step in machine learning, designed to identify the most relevant features to enhance model performance [27], [28], [29]. While traditional approaches, such as filter [30], [31] and wrapper [32] methods, have been widely adopted to optimize accuracy and interpretability, they primarily prioritize predictive performance. Crucially, these methods often overlook fairness considerations or the influence of protected attributes [33], [34].

To address this limitation, recent studies have introduced various feature selection techniques to improve model fairness and robustness. For instance, [35] proposed a robust unsupervised method utilizing graph representations, while [36] introduced a strategy to exclude protected and redundant features. Additionally, [37] proposed a framework

to evaluate the fairness of input features using human moral judgments.

However, applying these existing methods directly to high-dimensional unstructured data, such as images and text, presents significant challenges. Furthermore, they often necessitate complex additional constraints or auxiliary models. In contrast, our approach effectively mitigates bias against specific groups by jointly training a learnable filtering vector within the model in a single, end-to-end process.

III. METHOD

A. PROBLEM FORMULATION

We assume that the training dataset contains spurious correlations in the classification task. Specifically, given inputs $x \in \mathcal{X}$ and class labels $y \in \mathcal{Y}$, sensitive attributes $a \in \mathcal{A}$ frequently co-occur with specific class labels y. Based on the class labels y and sensitive attributes a, we can separate the training dataset into specific groups $g \in \mathcal{G}(:= \mathcal{Y} \times \mathcal{A})$. These groups are unequally represented in the training dataset, with the majority and minority groups depending on the class labels and sensitive attributes. This paper aims to find the parameters θ of a model $f(x;\theta)$ that maximizes the worst group accuracy (WGA) [10]. Here, $\mathrm{Acc}(g;\theta)$ denotes the accuracy of the model for the group g.

$$\theta^* = \arg\max_{\theta} \min_{g \in \mathcal{G}} \mathrm{Acc}(g;\theta) \tag{1}$$

Challenges: However, achieving this objective in practice presents two significant challenges. First, many state-of-the-art methods require group annotations to guide training, but such labels are scarce or unavailable, making it difficult to estimate the distribution of groups. Second, previous approaches, such as last-layer retraining [13], [26], rely on complex, multi-stage pipelines that are inefficient and cumbersome. These practical hurdles motivate the development of a new approach that addresses:

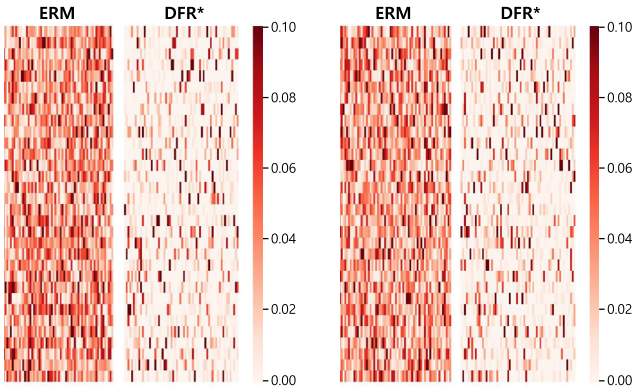
- **Scarcity of Group Annotations**: A dependency on group labels that are not available in most practical scenarios.
- **Complex Multi-Stage Pipelines**: The inefficiency of multi-stage training processes that hinder straightforward deployment.

TABLE 1. ERM and group balanced last layer fine-tuning result.

	Waterbirds		CivilComments	
	ACC(%)	WGA(%)	ACC(%)	WGA(%)
ERM	$98.1_{\pm 0.1}$	$75.3_{\pm 0.01}$	$91.9_{\pm 0.00}$	$66.7_{\pm 0.46}$
DFR(*our impl.*)	94.59	92.37	88.60	79.74

B. FROM WEIGHT SPARSITY TO FEATURE FILTERING

Recent studies show that ERM models tend to learn spurious correlations from data, leading to poor performance for specific groups, whereas robust models mitigate this issue [5]. This phenomenon is clearly illustrated in Table 1, which shows that Deep Feature Reweighting (DFR) [13], known as a state-of-the-art method, achieves a significantly higher Worst-Group Accuracy(WGA) than ERM on the Waterbird dataset [10].



(a) CelebA Dataset (b) Waterbirds Dataset

FIGURE 2. Weight heatmap of the last layer for each dataset. DFR* denotes our implementation of DFR.

Furthermore, these robust classifiers are characterized by highly sparse weights, as visualized in Figure 2. We interpret this sparsity as strong evidence that a robust classifier achieves its goal not by learning a complex decision boundary across all features, but by identifying and focusing only on a concise set of core, task-essential features.

Therefore, we aim to implement this selective utilization more directly. Unlike existing methods that modify the loss function [10], [38] or remove weights post-hoc (as in pruning) [39], [40], we propose a simple filtering method that explicitly blocks features during training. We implement this using a learnable binary vector that serves as a feature mask, trained end-to-end with the model, as detailed in the following section.

C. FEATURE FILTERING WITH LEARNABLE MASK

As illustrated in Figure 3, the core of our method is a learnable filtering vector m, which matches the dimensionality of the feature vector z and is positioned between the feature extractor h and the classifier c. Our framework comprises three key mechanisms: (1) In the forward pass, m is binarized into a hard mask $\tilde{m}$ applied to z to explicitly filter feature dimensions. (2) To resolve the non-differentiability of this step, we employ the Straight-Through Estimator (STE) [41] during the backward pass to approximate the gradients of m. (3) Finally, a gradient adjustment technique is implemented to ensure training stability by preventing weight freezing when the mask is zero. We detail these components in the following sections.

1) FORWARD PROCESS: HARD FILTERING

The forward process describes how the learnable filtering vector m generates a binary mask that structurally filters the feature representation z. This procedure consists of three sequential steps: vector initialization, mask generation, and feature filtering.

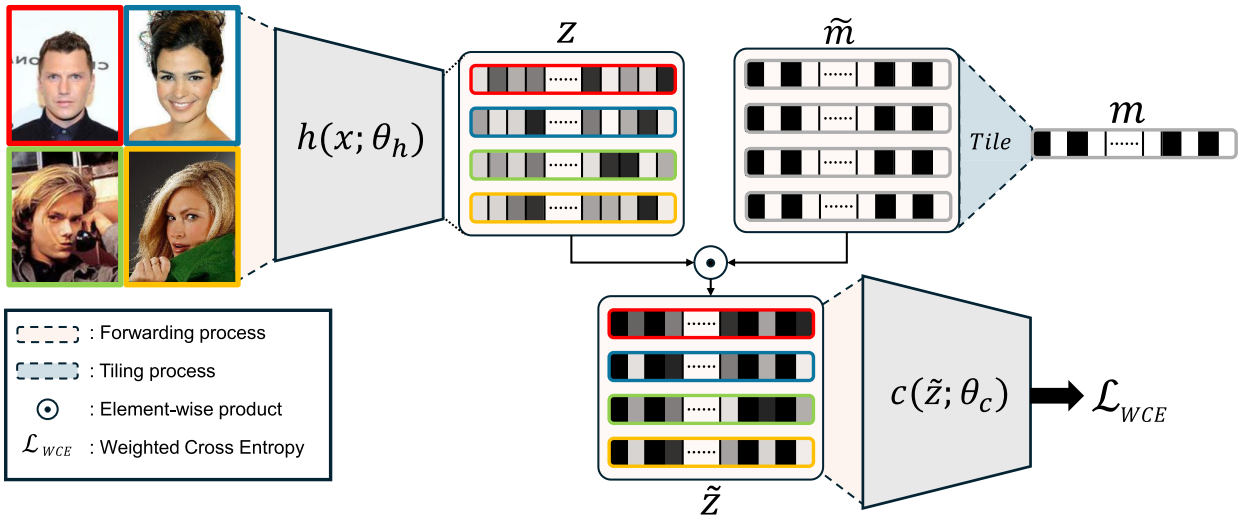


FIGURE 3. Method overview. The entire network f consists of h and g, where h denotes the feature extractor and c denotes the classifier. m is a learnable filtering vector with the same dimensions as embedding vectors z. *Tile* represents the tiling process, which replicates the filtering vector m across the batch dimension n times.

a: FILTER VECTOR INITIALIZATION

The training process can become unstable if the mask aggressively filters out dimensions during the initial stages of training. To mitigate this, we initialize the learnable vector $m \in \mathbb{R}^k$ with values near zero. Specifically, each element m_i is sampled from a normal distribution $\mathcal{N}(0, \sigma^2)$ using a small standard deviation (e.g., $\sigma = 0.1$). Consequently, passing these values through the subsequent sigmoid function yields probabilities close to 0.5. This strategy ensures that no dimension is decisively discarded in the early stages, thereby facilitating stable learning.

b: BINARY MASK GENERATION

The initialized vector m is transformed into a binary mask $\tilde{m}$ through a two-step process. First, we apply the element-wise sigmoid function, $\delta(\cdot)$, to map the values of m to the interval $(0, 1)$. These mapped values serve as probabilities representing the importance of each feature dimension. Subsequently, we obtain the deterministic binary mask $\tilde{m}$ by applying a fixed threshold of 0.5:

$$\tilde{m} = \mathbb{I}[\delta(m) \geq 0.5] \tag{2}$$

where $\mathbb{I}$ denotes the indicator function. This binarization compels the model to make a hard decision for each dimension: either to retain it or discard it entirely.

c: FEATURE FILTERING AND LOSS

The generated binary mask $\tilde{m}$ is applied to the feature vector $z \in \mathbb{R}^{n \times k}$ via an element-wise product. Since m is a vector and z is a batch of n vectors, where n represents the batch size, the mask is broadcast (or tiled) [42] across the batch dimension:

$$\tilde{z} = z \odot \text{Tile}(\tilde{m}, n) \tag{3}$$

Finally, the filtered representation $\tilde{z}$ is passed to the classifier to compute the final loss using a Weighted Cross-Entropy (WCE) as shown in Figure 3.

We note that some studies have raised concerns that traditional class-balancing techniques can inadvertently exacerbate group unfairness [43]. However, in our framework, WCE is not the primary mechanism for fairness. Instead, its purpose is to ensure training stability by mitigating the class imbalance in the data, thereby allowing the model to learn meaningful representations in the first place. We hypothesize that our proposed learnable filtering vector acts as the crucial de-biasing component. It is specifically designed to suppress spurious correlations actively, thereby counteracting the very fairness issues that class balancing might otherwise amplify. This synergistic relationship enables our model to leverage the stability of WCE while relying on the filter to ensure group fairness.

2) BACKWARD PROCESS: LEARNING THE MASK

The binarization step, while crucial for the forward pass, poses a challenge for training because the indicator function is non-differentiable, which blocks the gradient flow needed to update the filtering vector m. To this end, we employ the *Straight Through Estimator* (STE) [41]. The core principle of STE is to define a custom gradient for the non-differentiable operation.

In the backward pass, STE approximates the gradient by treating the binarization function as an identity function. Essentially, the gradient computed with respect to the output of the sigmoid function, $\frac{\partial L}{\partial \delta(m)}$, is used as a direct estimator for the gradient with respect to the binarized output $\tilde{m}$. The chain rule is then applied to compute the final gradient for the underlying vector m:

$$\frac{\partial L}{\partial m} = \underbrace{\frac{\partial L}{\partial \tilde{m}} \cdot \frac{\partial \tilde{m}}{\partial m}}_{\text{True Gradient (Undefined)}} \approx \overbrace{\frac{\partial L}{\partial \tilde{m}} \cdot \delta'(m)}^{\text{STE Approximation}} \tag{4}$$

where $\delta'(\cdot)$ is the derivative of the sigmoid function. This approximation provides a meaningful gradient, allowing the filtering vector m to be updated using standard gradient descent:

$$m_{\text{new}} \longleftarrow m_{\text{old}} - \eta \frac{\partial L}{\partial m}, \tag{5}$$

where η is the learning rate.

D. GRADIENT ADJUSTMENT: STABILIZING THE CLASSIFIER

A significant challenge arises from the use of a discrete binary mask. In the forward pass, if a dimension is masked out (i.e., $\tilde{z}_i = 0$), the standard chain rule dictates that the gradient with respect to the corresponding classifier weight θ_c must also be zero. This results in dead weights that cease to update, potentially destabilizing the training process.

To address this limitation, we introduce a gradient-adjustment mechanism based on a proxy gradient approach. The core idea is to decouple the forward and backward passes. While we use the hard binary mask $\tilde{m}$ for explicit filtering in the forward pass, we substitute it with the continuous soft mask $\delta(m)$ during the backward pass for the classifier weights. Specifically, we compute the gradients for θ_c as if the input features were scaled by the soft mask probabilities rather than zeroed out. The adjusted update rule can be formulated as:

$$\theta_c \leftarrow \theta_c - \eta \cdot \hat{\nabla}_{\theta_c} L(\delta(m)) \tag{6}$$

where $\hat{\nabla}_{\theta_c} L(\delta(m))$ denotes the proxy gradients computed using the soft-masked features. By maintaining a continuous learning signal via the soft mask, our substitution strategy prevents premature weight freezing and ensures stable convergence even for heavily filtered dimensions. As we will demonstrate in the following section, this conceptually simple algorithm enhances group robustness with minimal impact on overall accuracy.

E. TRADE-OFF METRIC

Numerous studies have proposed algorithms for group robustness that improve the accuracy of specific groups. However, they often overlook the trade-off between group robustness and overall accuracy. To quantitatively evaluate the trade-off between improving worst group accuracy (WGA) and maintaining overall accuracy (ACC), we introduce a revised Trade-off Ratio (ToR) metric. This metric measures the efficiency of WGA gains relative to concurrent losses in overall accuracy. Let $\Delta\text{WGA} = \text{WGA}_{\text{new}} - \text{WGA}_{\text{base}}$ and $\Delta\text{ACC}_{\text{loss}} = \text{ACC}_{\text{base}} - \text{ACC}_{\text{new}}$. The new ToR is defined as:

$$\text{ToR} = \begin{cases} \infty & \text{if } \Delta\text{WGA} > 0 \text{ and } \Delta\text{ACC}_{\text{loss}} \leq 0 \\ \dfrac{\Delta\text{WGA}}{\Delta\text{ACC}_{\text{loss}}} & \text{if } \Delta\text{WGA} > 0 \text{ and } \Delta\text{ACC}_{\text{loss}} > 0 \\ \Delta\text{WGA} & \text{if } \Delta\text{WGA} \leq 0 \end{cases} \tag{7}$$

This formulation provides an intuitive score for all possible outcomes, enabling transparent and fair comparisons across methods. It assigns an infinite score to the ideal case in which WGA improves without loss of overall accuracy, thereby identifying this as the best possible result. For the more common scenario in which a trade-off occurs, the ToR yields a positive value that quantifies the trade-off's efficiency, specifically, how much WGA was gained per unit of ACC sacrificed. Therefore. A higher value indicates a more effective de-biasing method. Finally, the metric assigns a direct penalty to any method that fails to improve WGA. In this case, the ToR is simply the (negative or zero) change in WGA, ensuring that any degradation in robustness is reflected as a poor score, regardless of the change in overall accuracy. This consistent framework enables a more nuanced and logically sound comparison of different de-biasing approaches.

Algorithm 1 Classification With Filtering Vector

Input: Input data $\mathcal{X}$, Labels $\mathcal{Y}$, Network $f = h \circ c$, Max epoch T
Initialize: Parameters θ_h, θ_c. Filter vector $m_i \sim \mathcal{N}(0, \sigma^2)$
Output: Trained parameters θ_h, θ_c, Trained filter vector m
for $t \leftarrow 1$ *to* T **do**
  for each batch (x, y) in dataset **do**
    $z \leftarrow h(x; \theta_h)$ ⋯ Extract features
    Forward Pass
    $\tilde{m} \leftarrow \mathbb{I}(\delta(m) \geq 0.5)$ ⋯ Binarize the filter vector
    $\tilde{z} \leftarrow z \odot \text{Tile}(\tilde{m}, n)$ ⋯ Apply filter
    $L \leftarrow \mathcal{L}_{WCE}(c(\tilde{z}; \theta_c), y)$ ⋯ Calculate loss
    Backward Pass
    Compute gradients $\nabla_{\theta_h} L, \nabla_{\theta_c} L, \nabla_m L$
    Apply STE for m
    $m \leftarrow m - \eta_m \nabla_m L$
    Apply Gradient Adjustment
    $\theta_c \leftarrow \theta_c - \eta \cdot (\nabla_{\theta_c} L \odot \delta(m))$
    $\theta_h \leftarrow \theta_h - \eta \cdot (\nabla_{\theta_h} L$ adjusted similarly)
  end for
end for

TABLE 2. Summary of spurious correlation benchmark datasets.

Dataset	Domain	Target Label	Spurious Attribute
Waterbirds	Image	Water/Land Bird	Background (Water/Land)
CelebA	Image	Hair Color (Blond)	Gender (Male/Female)
CivilComments	Text	Toxicity	Demographic Identity
MultiNLI	Text	NLI Relation	Negation Words

IV. EXPERIMENTS

A. EXPERIMENT SETTING

1) DATASETS

We evaluate our method on four spurious correlation benchmarks. First, **Waterbirds** is a binary image classification task that distinguishes between land birds and water birds.

TABLE 3. Distribution of class and spurious feature in training and held-out datasets for various classification datasets. The percentages next to the values represent the proportion of each group relative to the total, with the smallest group highlighted in bold.

Dataset		Train(Org.)		Class Counts		Held-out(Org.)		Class Counts
		Male	Female			Male	Female	
CelebA	Not Blond	66874(*41.08%*)	71629(*44.01%*)	138503	Not Blond	8276(*41.66%*)	8535(*42.96%*)	16811
	Blond	1387(***0.85%***)	22880(*14.06%*)	24267	Blond	182(***0.92%***)	2874(*14.47%*)	3056
		Land	Water			Land	Water	
Waterbirds	Land bird	3489(*79.5%*)	184(*3.84%*)	3673	Land bird	467(*38.95%*)	466(*38.87%*)	933
	Water bird	56(***1.17%***)	1057(*22.04%*)	1113	Water bird	133(***11.09%***)	133(***11.09%***)	266
		No Identity	Identity			No Identity	Identity	
CivilComments	Non Toxic	148186(*57.21%*)	90337(*34.87%*)	238523	Non Toxic	25159(*55.69%*)	14966(*33.13%*)	40125
	Toxic	12731(***4.91%***)	17784(*6.87%*)	30515	Toxic	2111(***4.67%***)	2944(*6.52%*)	5055
		No negation	Negation			No negation	Negation	
MultiNLI	Contradiction	57498(*27.89%*)	11158(*5.41%*)	68656	Contradiction	22814(*27.67%*)	4634(*5.62%*)	27448
	Entailment	67376(*32.68%*)	1521(***0.74%***)	68897	Entailment	26949(*32.68%*)	613(***0.74%***)	27562
	Neutral	66630(*32.32%*)	1992(*0.97%*)	68622	Neutral	26655(*32.32%*)	797(*0.97%*)	27452

The primary spurious correlation in this task is the background—either land or water—where the bird is located [10], [44]. Similarly, **CelebA** involves binary image classification to differentiate between Blond and Non-Blond individuals, with gender (Male, Female) serving as a spurious correlation [10], [45]. **Civilcomments** is focused on binary text classification to determine the toxicity of comments. The spurious attribute in this dataset is the presence of one of eight demographic identities: male, female, LGBTQ, Christian, Muslim, other religions, Black, and Whitečiteborkan2019nuanced. Lastly, **MultiNLI** is a multi-class text classification task that predicts the relationship between a premise and a hypothesis, with the presence of negation noted as a spurious attribute [46]. The key characteristics of each benchmark are summarized in Table 2.

As shown in Table 3, the training set for each benchmark is intentionally skewed. For instance, in the MultiNLI dataset, the presence of negation words is highly correlated with the contradiction label. This skew, which creates challenging minority groups, is the primary problem our method aims to solve. Accordingly, our primary evaluation metric is Worst-Group Accuracy (WGA), the performance on the most challenging subgroup, supplemented by overall accuracy.

TABLE 4. ERM and our proposed filtering method hyperparameters.

	learning rate	scheduler	batch size	weight decay
Waterbirds	0.003	cosine	32	0.00001
Celeba	0.003	cosine	100	0.00001
CivilComment	0.00001	step	16	0.0001
MultiNLI	0.00001	step	16	0.0001

2) BASELINES

We compare our method against several baselines that represent different de-biasing strategies. For instance, JTT [24] is a two-stage method that upweights training examples misclassified by an initial ERM model. Other approaches, such as DFR [13] and SELF [14], focus on retraining only the final layer. DFR freezes the feature extractor and retrains the classifier on a group-balanced dataset, while SELF offers a lightweight alternative by automatically constructing a fine-tuning set of hard examples. More complex two-stage frameworks include SSA [23] and CnC [8]. SSA trains a model to predict pseudo-labels for the spurious attribute, enabling fully-supervised training. Meanwhile, CnC uses an ERM model's predictions to guide a contrastive learning process that aligns representations. In contrast to these baselines, which all rely on a multi-stage process, our proposed method is a single-stage, end-to-end framework that learns to filter feature representations directly during training.

3) MODELS

We follow similar experimental settings as in previous studies [12], [13], [14], [15]. For image classification, we use pre-trained ResNet-50 [47] as the backbone network, and for text classification, we use pre-trained BERT [48]. In all experiments, we report the performance of the best accuracy checkpoint on the validation data without using early stopping.

4) IMPLEMENTATION DETAILS

To ensure a fair comparison, the hyperparameters for our method were set to be identical to those of the ERM baseline, as detailed in Table 4.

We used the SGD optimizer [49] for image classification tasks (Waterbirds, CelebA) and the AdamW optimizer [50] for text classification tasks (CivilComments, MultiNLI). Furthermore, the final linear classifier was initialized with 98% sparsity, with non-zero entries sampled from a normal distribution with mean 0 and variance 0.01. We found this

TABLE 5. Overall results. All baselines were from their original papers. We run ERM and our methods in 3 random seeds and report the mean and the standard deviation of *ACC* and *WGA* respectively. To calculate the trade-off ratio(TOR) metric, we take the average of *ACC* and *WGA*. We highlighted the best scores in *WGA* and TOR in bold.

Method	Waterbirds			CelebA			CivilComments			MultiNLI		
	ACC(%)	*WGA(%)*	TOR	*ACC(%)*	*WGA(%)*	TOR	*ACC(%)*	*WGA(%)*	TOR	*ACC(%)*	*WGA(%)*	TOR
JTT	93.3	86.7	2.38	88.0	81.1	4.69	78.6	72.6	0.44	91.1	69.3	4
CnC	$90.9_{\pm 0.1}$	$88.5_{\pm 0.3}$	1.8	$89.9_{\pm 0.5}$	$88.5_{\pm 0.9}$	7.6	-	-	-	$81.7_{\pm 0.5}$	$68.9_{\pm 2.1}$	1.3
SSA	$92.2_{\pm 0.87}$	$89.0_{\pm 0.55}$	2.3	$92.8_{\pm 0.11}$	$\mathbf{89.8_{\pm 1.28}}$	**16.3**	$79.9_{\pm 0.87}$	$76.6_{\pm 0.66}$	0.8	$88.2_{\pm 1.95}$	$69.9_{\pm 2.02}$	0.3
DFR	$94.2_{\pm 0.4}$	$92.9_{\pm 0.2}$	4.5	$91.3_{\pm 0.3}$	$88.3_{\pm 1.1}$	8.6	$82.1_{\pm 0.2}$	$74.7_{\pm 0.7}$	0.82	$87.2_{\pm 0.3}$	$70.1_{\pm 0.8}$	0.3
SELF	$94.0_{\pm 1.7}$	$\mathbf{93.0_{\pm 0.3}}$	4.3	$91.7_{\pm 0.4}$	$83.9_{\pm 0.9}$	10.0	$87.7_{\pm 0.6}$	$79.1_{\pm 2.1}$	1.00	$81.2_{\pm 0.7}$	$\mathbf{70.7_{\pm 2.5}}$	**2.6**
ERM	$98.1_{\pm 0.1}$	$75.3_{\pm 0.01}$	-	$95.5_{\pm 0.19}$	$45.9_{\pm 2.92}$	-	$91.9_{\pm 0.05}$	$66.7_{\pm 0.46}$	-	$82.1_{\pm 0.04}$	$68.4_{\pm 0.23}$	-
Ours	$97.3_{\pm 0.32}$	$85.5_{\pm 0.00}$	**12.8**	$92.3_{\pm 0.12}$	$80.4_{\pm 1.9}$	10.8	$88.7_{\pm 0.01}$	$\mathbf{81.9_{\pm 0.01}}$	**4.8**	$81.2_{\pm 0.04}$	$67.8_{\pm 10.7}$	-

initialization helped stabilize the training of the filtering vector.

B. MAIN RESULTS

The main results in Table 5 demonstrate that our simple, single-stage filtering method achieves highly competitive performance against more complex, multi-stage baselines.

Our method demonstrates robust performance on the CivilComments benchmark. It not only achieves the highest WGA of 81.9% among all compared methods but also secures the best ToR of 4.8, establishing it as both the most robust and efficient approach for this task. The trend of high efficiency continues on the Waterbirds dataset, where our method obtains a ToR of 12.8, significantly higher than all other baselines. On CelebA, our method remains highly efficient, achieving a competitive ToR of 10.8, second only to SSA.

However, the results also highlight the limitations of our approach. While demonstrating a substantial trade-off, our method's absolute WGA on CelebA (80.4%) is surpassed by several multi-stage baselines. Furthermore, on the MultiNLI dataset, our method did not improve WGA over the ERM baseline and exhibited a high standard deviation (± 10). This suggests that our single-stage approach, while efficient, may be less effective on datasets like MultiNLI. A potential reason for this instability is the extreme imbalance in the MultiNLI training data [14]. As detailed in Table 3, the worst-minority groups constitute less than 1% of the entire data. Our filtering mask m is trained by the main task loss $\mathcal{L}_{WCE}$. However, despite using $\mathcal{L}_{WCE}$, the extreme scarcity (0.74%) of the worst-group in MultiNLI provided a high-variance learning signal for the mask m. This signal instability might hinder stable convergence, as empirically validated by the high standard deviation (± 10.7) in our results.

In summary, our method consistently delivers competitive-to-state-of-the-art robustness across most benchmarks. This indicates that a simple, end-to-end filtering approach can be a highly effective and efficient alternative to more complex de-biasing pipelines.

TABLE 6. WGA performance across varying pruning ratios. All experiments were conducted using the same seed across all datasets.

Ratio (%)	Waterbirds	CelebA	CivilComments	MultiNLI
ERM(no pruning)	91.0	65.6	63.0	27.7
10	92.2	67.8	62.7	28.6
20	92.4	68.3	62.8	24.7
30	92.2	67.2	59.8	22.4
40	92.8	66.8	56.6	19.2
50	92.5	67.2	55.7	14.3
60	**93.1**	67.8	51.1	12.1
70	91.5	67.8	41.2	10.9
80	90.1	75.6	24.1	0.84
90	85.2	73.3	37.8	0.72
Ours	84.9	**81.7**	**81.8**	**67.4**

C. ANALYSIS AND ABLATION STUDIES

To validate our key design choices and analyze the behavior of our method, we conduct a series of analyses and ablation studies. We first compare our method against simpler regularization techniques and then conduct ablations to isolate the importance of the key components of our approach.

1) COMPARISON WITH UNSTRUCTURED PRUNING

We begin by distinguishing our method from unstructured magnitude pruning [39], [40]. The key difference is that pruning statically removes low-magnitude weights from a pre-trained classifier. In contrast, our method dynamically learns a filter on feature representations in an end-to-end process. To create a strong baseline, we follow the procedure of prior work: We first train an ERM model, prune its classifier at varying ratios, and then retrain it on a group-balanced held-out set. We then compare our single-stage method, which doesn't use any held-out data for training, against this strengthened baseline. As shown in Table 6,

our single-stage method—which does not use the held-out set for training—is a more effective and reliable de-biasing tool. While aggressive pruning performs anomalously well on the unique Waterbirds benchmark, its failure on the other three datasets underscores the need for a more adaptive approach, such as our learnable filter.

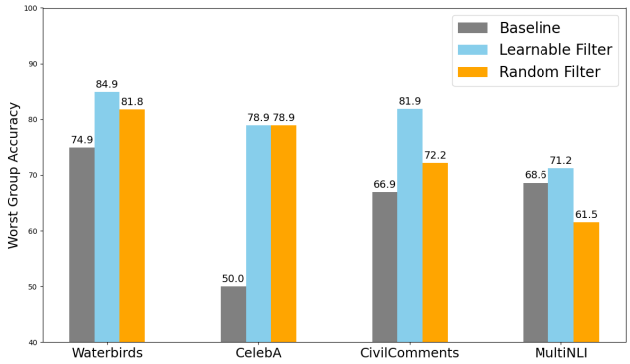


FIGURE 4. The effectiveness of a learnable filter versus a random one. For a direct comparison, the random filter vector was initialized with the same seed as the learnable filter.

2) THE IMPORTANCE OF A LEARNABLE FILTER

To demonstrate that the success of our approach critically depends on its ability to learn a filter, we compare it against two random filtering baselines.

First, we test against a fixed random filter. The results in Figure 4 demonstrate the superiority of our learnable approach. Across all benchmarks, our Learnable Filter outperforms the Random Filter and the ERM Baseline. Interestingly, on CelebA, the Random Filter yields meaningful improvements, likely serving as a regularizer that reduces the model's reliance on spurious features. Notably, on MultiNLI, the Random Filter is detrimental, underperforming the baseline, underscoring the necessity of a data-driven learning process.

Second, we test against a dynamic random filter by applying dropout at inference time. As shown in Figure 5, dropout yields only marginal and inconsistent improvements, and its performance is volatile and significantly inferior to our method (dashed lines). These experiments confirm that the effectiveness of our method is not due to random feature removal, but to the principled, learned filter it produces. While aggressive pruning achieves a high WGA on Waterbirds (up to 93.1%), this result appears anomalous and may be specific to the benchmark's unique properties. As noted in prior work [5], [13], the held-out set has a distribution that differs substantially from the training set, so nearly any strong regularization that forces the model to relearn on this balanced data can yield high performance. The failure of this simple heuristic on the other three benchmarks, however, underscores the need for a more adaptive and principled approach, such as our learnable filter.

3) EFFECT OF HARD VS. SOFT FILTERING

Our method's design includes a binarization step to create a hard binary mask. To validate this critical component,

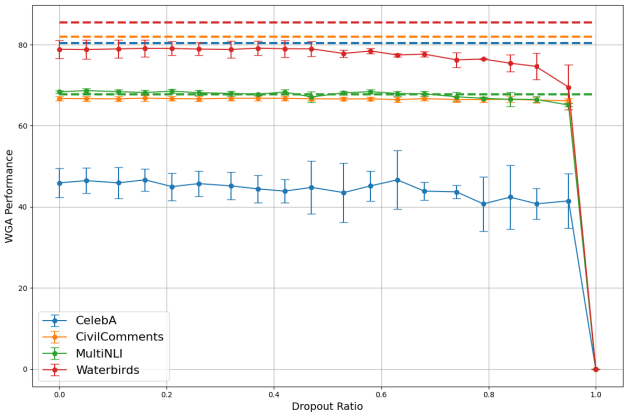


FIGURE 5. WGA performance versus inference-time dropout ratio. Applying dropout yields only marginal, unstable improvements over the baseline. Our method (dashed lines) consistently outperforms the random filtering baseline across all datasets.

TABLE 7. Performance of hard filtering versus soft filtering, evaluated on WGA (%). To ensure a fair comparison, a consistent seed was maintained across all experimental runs.

Dataset	Hard Filter (Ours)	Soft Filter
ERM Waterbirds	85.5	85.4
CelebA	81.6	61.1
CivilComments	81.8	67.8
MultiNLI	65.0	30.1

we conduct an ablation study comparing it with a soft-filtering variant. The soft filtering baseline omits the binarization step, instead multiplying the feature vector z directly by the continuous sigmoid outputs $\delta(m)$. All other experimental conditions remain identical.

As shown in Table 7, our hard filtering approach consistently and decisively outperforms the soft filtering baseline across all datasets. The performance gap is particularly pronounced on CelebA, where hard filtering yields an absolute WGA improvement of 20.5%.

This substantial difference suggests that forcing the model to make a hard, binary decision for each feature dimension is more effective than applying a continuous, soft filter. These results confirm that the binarization step is a crucial component of our method.

4) LEARNING RATE SENSITIVITY

To validate the stability of our approach, we conduct an ablation study on the learning rate (LR) using the Waterbirds dataset. As shown in Table 8, our method demonstrates high robustness across a wide range of learning rates. While our main experiments used an LR of 0.003 (to match the ERM baseline), reducing the LR to 0.001, 0.0005, and 0.0001 still yields strong and stable WGA scores of 85.4%, 86.1%, and 83.4%, respectively. This confirms that our filtering mechanism is not overly sensitive to this hyperparameter and provides the practicality of our method.

TABLE 8. Learning rate (LR) sensitivity analysis for our method on the Waterbirds dataset. All results are reported using seed 1001.

	ACC	WGA
ERM	$98.1_{\pm 0.1}$	$75.3_{\pm 0.01}$
Ours (LR=0.001)	97.8	85.4
Ours (LR=0.0005)	97.6	86.1
Ours (LR=0.0001)	96.9	83.4

V. QUALITATIVE ANALYSIS

To provide an intuitive understanding of how our method operates, we conduct a qualitative analysis of the learned filter and its downstream effects on the classifier.

A. FILTER'S FOCUS

First, we visualize what our filter learns to focus on within an image. Figure 6 (a) displays saliency maps for representative samples from the Waterbirds dataset's minority and majority groups. The maps clearly show that our filter directs the model's focus to the core feature (the bird) while consistently ignoring the spurious feature (the background). This provides direct visual evidence that our filter successfully identifies and removes spurious information as intended.

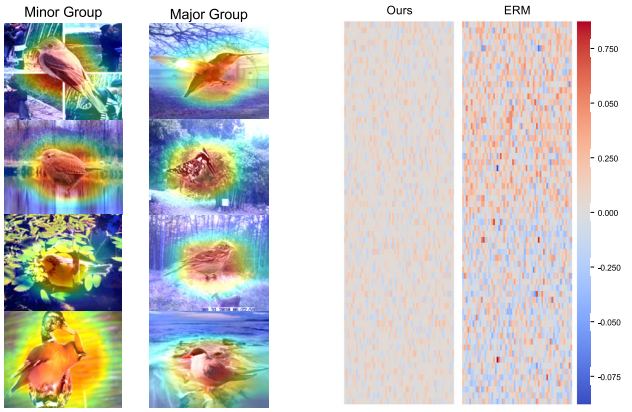


FIGURE 6. Qualitative analysis of our filtering method. (a) Saliency maps on samples from the Waterbirds dataset. Our model consistently learns to focus on the bird (core feature) while ignoring the spurious background across both minority and majority groups. (b) This effective filtering yields a final classifier with significantly calmer, more regularized weights than those of the ERM baseline.

B. RESULTING CLASSIFIER WEIGHT

Next, we analyze the downstream impact of this filtering on the final classifier's learned weights, shown in Figure 6 (b). The ERM baseline learns a noisy, complex weight structure characterized by a scattered mix of high-magnitude positive (red) and negative (blue) values. This suggests it must learn a volatile function to counteract the unfiltered spurious features. In contrast, the classifier trained with our method is significantly sparser, with most weights clustered near zero. This demonstrates that our filter allows the classifier to learn a more robust function, validating our initial motivation that a robust classifier can operate on a concise set of core features.

VI. CONCLUSION

In this paper, we introduce a simple and effective end-to-end method to improve group robustness to spurious correlations. Our approach trains a lightweight, learnable filtering vector that dynamically masks dimensions in the feature representation, guiding the model to focus on core features. Our experiments demonstrated that this single-stage framework is a powerful and highly efficient alternative to complex multi-stage pipelines, achieving competitive-to-state-of-the-art worst-group accuracy on several benchmarks. Furthermore, we proposed a revised Trade-off Ratio (ToR) metric that offers the community a more robust tool for evaluating the critical balance between group robustness and overall accuracy. We hope our work encourages further research into simple yet effective methods for building more reliable and fair models.

REFERENCES

[1] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022.

[2] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Comput. Vis. Image Understand.*, vol. 189, Dec. 2019, Art. no. 102805.

[3] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do ImageNet classifiers generalize to ImageNet," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5389–5400.

[4] P. W. Koh et al., "WILDS: A benchmark of in-the-wild distribution shifts," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5637–5664.

[5] P. Izmailov, P. Kirichenko, N. Gruver, and A. G. Wilson, "On feature learning in the presence of spurious correlations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 38516–38532.

[6] S. Beery, G. V. Horn, and P. Perona, "Recognition in Terra incognita," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 472–489.

[7] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002683.

[8] M. Zhang, N. S. Sohoni, H. R. Zhang, C. Finn, and C. Ré, "Correct-N-contrast: A contrastive approach for improving robustness to spurious correlations," 2022, *arXiv:2203.01517*.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[10] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," 2019, *arXiv:1911.08731*.

[11] Y. Lee, A. S. Chen, F. Tajwar, A. Kumar, H. Yao, P. Liang, and C. Finn, "Surgical fine-tuning improves adaptation to distribution shifts," 2022, *arXiv:2210.11466*.

[12] Y. Yang, H. Zhang, D. Katabi, and M. Ghassemi, "Change is hard: A closer look at subpopulation shift," 2023, *arXiv:2302.12254*.

[13] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Last layer re-training is sufficient for robustness to spurious correlations," 2022, *arXiv:2204.02937*.

[14] T. LaBonte, V. Muthukumar, and A. Kumar, "Towards last-layer retraining for group robustness with fewer annotations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 11552–11579.

[15] S. Qiu, A. Potapczynski, P. Izmailov, and A. G. Wilson, "Simple and fast group robustness by automatic feature reweighting," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 28448–28467.

[16] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7472–7482.

[17] N. Kim, S. Hwang, S. S. Ahn, J. Park, and S. Kwak, "Learning debiased classifier with biased committee," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 18403–18415.
[18] J. Lee, J.-H. Park, D. Kim, J. Lee, E. Choi, and J. Choo, "Revisiting the importance of amplifying bias for debiasing," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 12, pp. 14974–14981.
[19] H. Bahng, S. Chun, S. Yun, J. Choo, and S. J. Oh, "Learning de-biased representations with biased representations," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 528–539.
[20] E. Kim, J. Lee, J. Y. Lee, J. Lee, and J. Choo, "Learning debiased representation via disentangled feature augmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 25123–25133.
[21] V. V. Ramaswamy, S. S. Y. Kim, and O. Russakovsky, "Fair attribute classification through latent space de-biasing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9297–9306.
[22] S. Han, S. Lee, F. Wu, S. Kim, C. Wu, X. Wang, X. Xie, and M. Cha, "DualFair: Fair representation learning at both group and individual levels via contrastive self-supervision," in *Proc. ACM Web Conf.*, Apr. 2023, pp. 3766–3774.
[23] J. Nam, J. Kim, J. Lee, and J. Shin, "Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation," 2022, *arXiv:2204.02070*.
[24] E. Liu, B. Haghgoo, A. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn, "Just train twice: Improving group robustness without training group information," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 6781–6792.
[25] J. H. Nam, H. Cha, S. Ahn, J. Lee, and J. Shin, "Learning from failure: De-biasing classifier from biased classifier," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 20673–20684.
[26] Y. Mao, Z. Deng, H. Yao, T. Ye, K. Kawaguchi, and J. Zou, "Last-layer fairness fine-tuning is simple and effective for neural networks," 2023, *arXiv:2304.03935*.
[27] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *Proc. Sci. Inf. Conf.*, Aug. 2014, pp. 372–378.
[28] P. Dhal and C. Azad, "A comprehensive survey on feature selection in the various fields of machine learning," *Appl. Intell.*, vol. 52, no. 4, pp. 4543–4581, Mar. 2022.
[29] D. Theng and K. K. Bhoyar, "Feature selection techniques for machine learning: A survey of more than two decades of research," *Knowl. Inf. Syst.*, vol. 66, no. 3, pp. 1575–1637, Mar. 2024.
[30] H. Liu and R. Setiono, "A probabilistic approach to feature selection—A filter solution," in *Proc. ICML*, vol. 96, 1996, pp. 319–327.
[31] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine Learning Proceedings 1992*. Amsterdam, The Netherlands: Elsevier, 1992, pp. 249–256.
[32] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, Dec. 1997.
[33] Z. Yang, Z. Wang, C. Huang, and X. Yao, "An explainable feature selection approach for fair machine learning," in *Proc. Int. Conf. Artif. Neural Netw.*, 2023, pp. 75–86.
[34] S. Galhotra, K. Shanmugam, P. Sattigeri, and K. R. Varshney, "Causal feature selection for algorithmic fairness," in *Proc. Int. Conf. Manage. Data*, Jun. 2022, pp. 276–285.
[35] M. You, A. Yuan, M. Zou, D. He, and X. Li, "Robust unsupervised feature selection via multi-group adaptive graph representation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 3030–3044, Mar. 2023.
[36] G. Dorleon, I. Megdiche, N. Bricon-Souf, and O. Teste, "Feature selection under fairness and performance constraints," in *Proc. Int. Conf. Big Data Anal. Knowl. Discovery*, 2022, pp. 125–130.
[37] N. Grgić-Hlača, M. B. Zafar, K. P. Gummadi, and A. Weller, "Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 51–60.
[38] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," 2019, *arXiv:1907.02893*.
[39] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," 2015, *arXiv:1510.00149*.
[40] H. Cheng, M. Zhang, and J. Q. Shi, "A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10558–10578, Dec. 2024.
[41] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," 2013, *arXiv:1308.3432*.
[42] H. Li, Y. Liu, H. Zhang, and B. Li, "Mitigating and evaluating static bias of action representations in the background and the foreground," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 19854–19866.
[43] S. Yan, H.-T. Kao, and E. Ferrara, "Fair class balancing: Enhancing model fairness without observing sensitive attributes," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 1715–1724, doi: 10.1145/3340531.3411980.
[44] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-UCSD birds-200-2011 dataset," California Inst. Technol. (Caltech), Pasadena, CA, USA, Tech. Rep. 1, 2011.
[45] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.
[46] A. Williams, N. Nangia, and S. R. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," 2017, *arXiv:1704.05426*.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[48] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[49] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*.
[50] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.



MINGI KIM received the B.S. degree in business administration from Chung-Ang University, Seoul, Republic of Korea, in 2023, where he is currently pursuing the M.S. degree. His research interests include fairness, natural-language processing, and various multi-modal learning.



JONGWON RYU received the B.S. degree in mathematics from Dankook University, Cheonan, Republic of Korea, in 2023. He is currently pursuing the M.S. degree with Chung-Ang University. His research interests include natural language processing and various multi-modal reasoning.



JUNYEONG KIM (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from KAIST, Republic of Korea, in 2015, 2017, and 2021, respectively. He was a Postdoctoral Research Associate with the Artificial Intelligence and Machine Learning Laboratory, School of Electrical Engineering, KAIST. He is currently an Assistant Professor with the Department of AI, Chung-Ang University, Republic of Korea. His research interests include visual-language reasoning, visual question answering, and various video-based problems.