Calibrating Multi-modal Representations: A Pursuit of Group Robustness without Annotations

Chenyu You†, Yifei Min†, Weicheng Dai†,
Jasjeet S. Sekhon, Lawrence Staib, James S. Duncan
Yale University

Abstract

Fine-tuning pre-trained vision-language models, like CLIP, has yielded success on diverse downstream tasks. However, several pain points persist for this paradigm: (i) directly tuning entire pre-trained models becomes both time-intensive and computationally costly. Additionally, these tuned models tend to become highly specialized, limiting their practicality for real-world deployment; (ii) recent studies indicate that pre-trained vision-language classifiers may overly depend on spurious features – patterns that correlate with the target in training data, but are not related to the true labeling function; and (iii) existing studies on mitigating the reliance on spurious features, largely based on the assumption that we can identify such features, does not provide definitive assurance for real-world applications. As a piloting study, this work focuses on exploring mitigating the reliance on spurious features for CLIP without using any group annotation. To this end, we systematically study the existence of spurious correlation on CLIP and CILP+ERM. We first, following recent work on Deep Feature Reweighting (DFR), verify that last-layer retraining can greatly improve group robustness on pretrained CLIP. In view of them, we advocate a lightweight representation calibration method for fine-tuning CLIP, by first generating a calibration set using the pretrained CLIP, and then calibrating representations of samples within this set through contrastive learning, all without the need for group labels. Extensive experiments and in-depth visualizations on several benchmarks validate the effectiveness of our proposals, largely reducing reliance and significantly boosting the model generalization. Our codes will be available in here

1. Introduction

In recent years, large-scale pre-trained vision-language models (VLMs) [3, 18, 29, 48, 62, 70] have showcased impressive capabilities across various downstream tasks, including visual understanding [11, 27, 28, 49], image-text generation [39, 44] and more [13, 26, 74]. Leveraging these well-learned and rich representations, fine-tuning pre-trained VLMs has been the dominant methodology, starting from pre-training from extensive web-crawled data and then tuning it towards specific downstream tasks. However, when relying on standard Empirical Risk Minimization (ERM) for training, it raises a risk of inadvertently amplifying spurious correlations, which may compromise robustness, especially for underrepresented groups [9]. Consequently, even these advanced VLMs are not exempt from the challenges posed by *spurious correlation*, where patterns may correlate with the target class without truly pertaining to the classification function. This can inadvertently sideline certain minority groups within the training data, which poses practical challenges and limits the efficacy of these models in safety-critical applications. For example, on Waterbirds dataset [51], tasked with classifying "landbird" and "waterbird", there exists a bias where an "water (land)" background is spuriously correlated with the "waterbird (landbird)" class, leading to a minority groups of "waterbird on land" and "landbird on water".

As a result, a considerable body of work has aimed at improving **group robustness** of *vision* models [6, 17, 23, 33, 40, 41, 51]. Specifically, to ensure learning models do not depend on spurious correlations, which can lead to high error rates in certain data groups, we align with mainstream practices and focus on enhancing the *worst-case accuracy* (WGA) across different groups. Yet, this remains underexplored in the *vision-language* context. We therefore attempt to explore the central question motivating this work:

> ***(Q)*** *Does there exist an efficient way to mitigate multi-modal spurious correlations without retraining the entire model, thereby enhancing its group robustness without relying on any group annotations?*

To address **(Q)**, we aims to achieve two key objectives: (i) ***parameter-efficient fine-tuning*** – traditional *fine-tuning* typically involves updating a large proportion of or even all the parameters of the pre-trained model. Yet, in many practi-

† equal contribution.

cal scenarios, this paradigm becomes challenging owing to the significant memory and computational demands; and (ii) ***group label efficiency*** – since many real-world problems inherently contain spurious correlations, the existing methods often require prior group information to adapt large-scale pre-trained models for specific downstream datasets, which poses impediments to the deployment in real-world resource-constrained settings. Additionally, even when such spurious features are identifiable, the task of annotating vast datasets with group labels becomes prohibitively demanding.

In this paper, our research trajectory unfolds as follows: the **first** part of our work provides comprehensive analysis to ascertain the presence of spurious correlations within CLIP:

- ***Identifying spurious correlation issues in CLIP:*** Our investigation uncovers spurious correlation issues within the large pre-trained multi-modal models. In plain words (detailed analysis in Sec. 4), we use *t*-SNE [59] and UMAP [36] to inspect group-wise embeddings from benchmark datasets such as Waterbirds and CelebA. Our findings clearly indicate the presence of unintended spurious correlations in both the pre-trained CLIP and CLIP+ERM representations. This limitation arises from an over-reliance on spurious features, indicating a need for more robust feature calibrator tailored to specific downstream tasks.
- ***Verifying the efficacy of feature extractor in CLIP:*** To tackle the challenge of spurious correlations without incurring substantial computational overhead, we draw inspiration from [23] to calibrate feature representation quality by re-training of CLIP's final layer. This approach allows seamless adaptation of large pre-trained multi-modal models to specific downstream benchmarks. Empirical results suggest the feasibility of recalibrating the pre-trained feature representation to neutralize spurious correlations. These findings further motivates us to explore more parameter-efficient methods devoid of group information reliance.

The **second** part of our work is to put forth a streamlined approach to address the aforementioned concerns. To this end, we present a robust representation calibration approach that functions *without the need for any group annotations*. This novel framework, **C**ontrastive **F**eature **R**ecalibration (**CFR**), integrates a contrastive learning paradigm into representation calibration, as shown in Figure 5 (Appendix).

- Using the pre-trained CLIP, we construct a calibration set curated from the training data. The samples within this set act as pivotal anchor points. By calibrating the representation of these anchors, we aim to enhance robustness across the entire dataset. In curating this set, CFR employs an intuitive strategy, selecting samples misclassified by the pre-trained CLIP.
- With the calibration set established, CFR refines the sample representations, aligning them more closely with the centroid of their designated class in the feature space and distancing them from opposing class centroids. This calibration is adeptly achieved using a contrastive loss.

Our extensive experiments on CLIP (underexplored in semi-supervised spurious correlation literature till date) across multiple datasets illustrate that CFR not only significantly improves group robustness compared to semi-supervised methods but also rivals the performance of supervised approaches. Furthermore, utilizing *t*-SNE and UMAP, we observe that our proposed method exhibits substantially better class separation patterns compared to the pre-trained CLIP and CLIP fine-tuned with ERM. In addition, referring to the training-validation curves of different methods across four benchmark datasets (Figure 4), it becomes evident that CFR maintains its superiority in the ability to converge towards an optimal solution when compared to other methods. Collectively, these experiments provide strong support for the efficacy of CFR in addressing spurious correlations, all without reliance on group-specific information.

2. Related Work

On spurious correlations. There has been recently a burgeoning interest in examining the role of spurious correlations in the context of deep learning, particularly as it pertains to a wide range of real-world challenges. Existing work [9] shows that neural networks typically exhibit an inherent inclination to emphasize the intended features over the shortcut – shallow features that are spuriously correlated with the classification targets, which may be particularly problematic in high-stakes and safety-critical scenarios. In vision, models often hinge on semantically irrelevant attributes such as an image's background [26, 38, 51, 63], texture [8], and secondary objects [50, 54, 55], and other semantically irrelevant features [4, 14, 30]. Particularly concerning is their use in high-stakes areas like medical imaging, where networks might erroneously focus on hospital-specific tokens [69] or incidental cues [43] rather than actual disease symptoms. Similarly, in natural language processing (NLP), pre-trained models often exhibit a reliance on superficial features. This reliance allows them to perform well on benchmarks, even when not genuinely comprehending the tasks. For instance, models might leverage basic syntactic patterns, like lexical similarities between sentences, to deduce their interrelationship [12, 19, 35, 42]. Broader introductions are provided in [9, 67]. In this paper, we take a step further to meet more practical requirements – *without relying on any group annotations*, but with the help of language attributes on a pre-trained VLM with fine-tuning contrastive feature calibration. To our best knowledge, this training paradigm is underexplored in the spurious correlation literature, offering a new view to improve group robustness of large-scale pre-trained multi-modal models.

Robustness and group annotations. Group robustness has recently garnered significant attention owing to the preva-

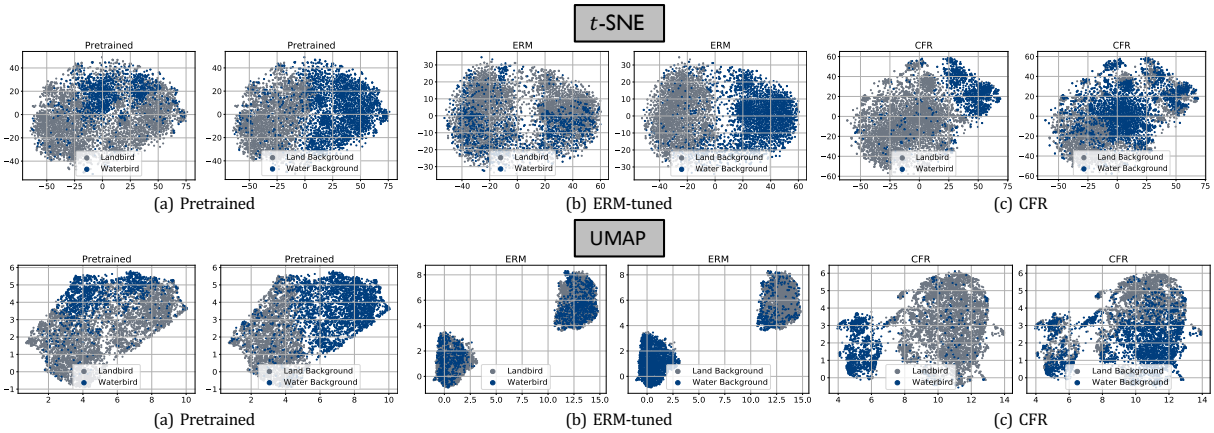

Figure 1. **t-SNE and UMAP visualizations for pre-trained CLIP, ERM-tuned CLIP, and CFR (ours) on Waterbirds.** We observe that both the pre-trained and ERM-tuned CLIP exhibit noticeable spurious correlations, with feature separations inappropriately aligned with spurious attributes, specifically the background, rather than the target class. In contrast, our method, as visualized through t-SNE and UMAP, demonstrates a significantly improved class separations, underscoring the robustness of our method in reducing spurious correlations.

lence of spurious correlations, and hereby we focus our discussion on closely related works, both those that rely on group annotations and those that do not. (i) *Annotation-dependent methods:* A body of research has explored leveraging the group annotation. Much of this has aimed to improve robustness based on some heuristics like minimizing the worst-group loss [51], learn invariant or diverse features [1, 10, 32, 64, 71], class or group balancing or weighting [7, 15, 17, 23, 28, 37], or contrastive learning [58, 66]. It usually leads to competitive performance but is hard to deploy in the real-world applications due to annotation costs. (ii) *Annotation-free methods:* Another line of work, which discards the use of group annotations, can be roughly categorized into various groups. For instance, existing studies leverage auxiliary models to pseudo-label the minority group or spurious features [2, 6, 21, 22, 25, 27, 40, 56, 57, 65, 73]. Others have emphasized upweighting samples misclassified by an early-stopped model [33], reweighting or subsampling classes [15, 47], or employing robust losses and regularizations [46, 68, 72]. These methods seek to bolster group robustness, albeit often necessitating held-out group annotations, dual-phase training, or the deployment of additional auxiliary models. To the best of our understanding, the integration of feature recalibration with the absence of group labels has been underexplored for the efficient fine-tuning of pre-trained CLIP models.

Spurious correlation in vision-language models. The advent of transformer-based architecture has led to the development of advanced large-scale pre-trained multi-modal models, aiming to enhance the effectiveness of these models. Some prior work has introduced language features aimed at making vision classifiers more robust, including attention maps [45], changes to feature attributes [75]. Several pioneering efforts [66, 72] have been made to obtain pre-trained multi-modal models that are robust to spurious correlations. For example, [72] designs a new contrastive adapter, integrating it with transfer learning to enhance group robustness. Nevertheless, this approach does not always guarantee improved performance, particularly for specialized downstream tasks. [66] for the first time introduces a fine-tuning approach specifically tailored to mitigate spurious correlations in pre-trained multi-modal frameworks. In contrast to this work, our goal is to devise a pragmatic training paradigm that functions without relying on any group annotations.

Our approach stands distinct from prior studies, anchored in **two key insights**: (i) *the pivotal role of feature recalibration in bolstering robustness without the need for group annotations*, and (ii) *the advantageous influence of language attributes on vision classifiers' group robustness*. Our approach directly refines the sample representations using a contrastive loss with specially sampled positive and negative batches. Consequently, we achieve enhanced robustness against spurious correlations, optimizing group accuracy, efficiency, and practicality.

3. Preliminaries

Setting. In this work, we consider classification tasks within the group robustness setting [51], wherein the input is denoted as $\mathbf{x} \in \mathcal{X}$ and target classes as $\mathbf{y} \in \mathcal{Y}$. Specifically, we assume the data distribution consists of multiple *groups* $g \in \mathcal{G}$. Typically, these groups are defined by a combination of the class label $\mathbf{y} \in \mathcal{Y}$ and a spurious attribute $s \in \mathcal{S}$. Consider the Waterbirds dataset [51]. Here, the classification task involves classifying $\mathbf{y} \in \{\text{landbird}, \text{waterbird}\}$, and the background depicted in the image serves as the spurious attributes $s \in \{\text{land}, \text{water}\}$. Consequently, the groups are formulated from the combinations of the class label and the spurious attribute, denoted as $\mathcal{G} = \mathcal{Y} \times \mathcal{S}$.

The attribute s is considered spurious when it correlates

with $\mathbf{y}$ but lacks a causal relationship. For example, within the Waterbirds dataset, approximately 95% of data points labeled as $\mathbf{y} =$ waterbird possess the spurious attribute $s =$ water. Consequently, models trained on this dataset may heavily depend on the background (water) to predict the class (waterbird), leading to reduced performance on the minority group $g = (\text{water}, \text{landbird})$.

To safeguard against models relying on spurious correlations, we align with mainstream practices [6, 17, 23, 33, 40, 41, 51], and utilize *worst group accuracy* (WGA) as our evaluation metric, which denotes the minimum predictive accuracy of our model across all groups.

Access to spurious attributes. Many current approaches addressing robustness against spurious correlations presuppose the availability of spurious attributes s within the training data [51, 66] or at least within a subset of data designated for model training [23, 41, 57]. In contrast, we delve into a more challenging scenario where the group information remains inaccessible for fine-tuning.

CLIP. Contrastive Language-Image Pre-training (CLIP) [48][1] learns from over 400M image-caption pairs collected from the web[2] by maximizing the similarity between the image and text. Specifically, CLIP consists of (i) a visual encoder, (ii) a text encoder, and (iii) the dot product of their outputs serves as "alignment score" between the input image and text. Formally, given a batch of N images and their associated captions, each image representation $\mathbf{v}$ should align with its corresponding text representation $\mathbf{u}$. The likelihood of image i aligning with caption j is expressed as $\exp(\beta \mathbf{v}_i^T \mathbf{u}_j) / \sum_{k=1}^{N} \exp(\beta \mathbf{v}_i^T \mathbf{u}_k)$, with β being a hyperparameter[3].

4. Representation of Pretrained Models

In this section, we investigate the inherent spurious correlations present in the CLIP model. Utilizing widely-used feature visualization techniques, including t-SNE, UMAP [36], and GradCAM [52], our objective is to substantiate the presence of spurious correlations within the CLIP framework. We chart our investigation as below:

(i) ***Revealing spurious correlations in pre-trained CLIP:*** Our initial step involves illustrating spurious correlations in pre-trained CLIP. We utilize dimensionality reduction techniques such as t-SNE and UMAP to unveil the separation between class features and spurious attributes in both ERM and pre-trained CLIP. Our t-SNE visualization, as shown in Figure 1, indicates that the pre-trained CLIP model inadequately separates between classes, yet efficiently identifies spurious attributes, thus corroborating the existence of spurious correlations. Remarkably, a similar pattern emerges when examining ERM-tuned CLIP. The resemblance between pre-trained CLIP and ERM-tuned CLIP in their t-SNE visualizations underscores the severe issue of spurious correlation in the pre-trained models, as ERM training is well-known to be prone to such correlations. The UMAP visualization further confirms our findings with a similar discernment (See Figure 1).

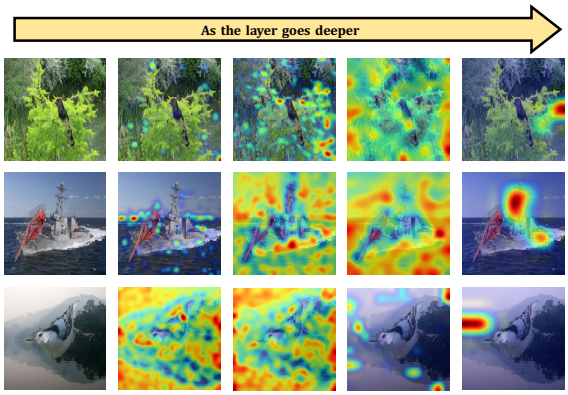

Figure 2. **Layer-by-layer GradCAM analysis of the CLIP-ResNet50.** Each row starts with the original image on the left, followed by four GradCAM visualizations corresponding to the four successive layers of the ResNet-50, with the depth of the layers increasing from left to right.

(ii) ***Investigating the underlying cause of spurious correlations:*** To further substantiate the presence of spurious correlation, we employ GradCAM to analyze various layers of the pre-trained CLIP model. This analysis reveals that pre-trained CLIP tends to focus on spurious attributes in the data, rather than the desired complex features (*e.g.*, the bird) for classification. In Figure 2, we present the GradCAM results from each of the four layers of the ResNet-50 backbone in CLIP. A noticeable trend emerges: in shallower layers, the model's attention spans a broader region of the image. As we delve deeper into the layers, the model progressively narrows its focus to smaller regions of the image. Notably, in the final layer, the model concentrates its attention on an extremely limited portion of the image, which is often the background rather than the object of interest (*e.g.*, the bird). This observation underscores the model's tendency to concentrate on spurious attributes, providing a plausible explanation for the prevalence of spurious correlation in pre-trained CLIP.

Overall findings: Our comprehensive examinations confirm the presence of spurious correlations in the features learned by the pre-trained CLIP. These insights prompt us to recalibrating CLIP model's features. Encouraged by these findings, we introduce our proposed approach, *Calibrated Feature Refinement* (CFR), in the subsequent section, aimed to mitigate the spurious correlation issue within the model.

[1] https://github.com/openai/CLIP
[2] This dataset is not public.
[3] $\mathbf{v}_i$ and $\mathbf{u}_j$ are normalized prior to the dot product calculation.

5. Feature Recalibration

In this section, we introduce CFR, a novel representation calibration method, aimed at enhancing group robustness in pre-trained CLIP models, without the need for group annotations. As shown in Figure 5 (Appendix), CFR unfolds in two pivotal steps: the initial step involves the assembly of a calibration set, thoughtfully curated from the training data. In the second step, we augment the robustness of the CLIP feature by calibrating the representation using the samples from the previously curated set.

5.1. Calibration Set Formulation

Sample selection strategies have been widely explored in the context of addressing spurious correlations [23, 25, 33, 72, 73]. Such issues often arise when using the full training data. To mitigate this issue, researchers have focused on identifying more group-balanced subsets for refining models. In real-world scenarios where group information is commonly inaccessible, a widely accepted solution is to generate pseudo-group-labels, commonly achieved by using an auxiliary model trained with Empirical Risk Minimization (ERM). Existing works in this direction have taken three angles: (1) disagreement-based methods (*e.g.*, selecting samples where ERM predictions contradict group-truth labels [25]; (2) uncertainty-based method (*e.g.*, opting for samples with high uncertainty in ERM predictions [47]), and (3) clustering-based method (*e.g.*, interpreting clusters formed by ERM-learned features as pseudo-groups [56, 73]), *etc.*

Inspired by these, we propose a straightforward yet highly effective approach tailored for vision-language models. Rather than training an entire model from the ground up using ERM, we opt to utilize the feature representation inherent in the pre-trained CLIP. Our strategy involves training only the projection head of CLIP using cross-entropy loss, subsequently selecting samples that are misclassified by this ERM-tuned CLIP for further rectification. Our empirical results underscore the efficacy of this approach, showcasing its impressive performance in recalibrating CLIP. Furthermore, we recognize the potential of exploring alternative selection methods as a promising avenue for future research.

5.2. Contrastive Feature Recalibration

The subsequent phase is centered around recitifying the features of samples within this calibration set. By recalibrating these specific samples, our objective is to improve the quality of feature representation across the entire training dataset. Inspired by insights gained from the feature visualization of the pre-trained CLIP and ERM-tuned CLIP in Sec. 4, our proposed CFR method is capable of recalibrating the feature of a given sample. This is achieved by pulling the sample's feature closer to the centroid of its designated class while simultaneously pushing it away from centroids associated with opposing classes.

Take an arbitrary sample $(\mathbf{x}, \mathbf{y})$ from the calibration set as the *anchor*. We define $\mathbf{v}$ as the visual feature of $\mathbf{x}$, encoded by the pre-trained CLIP, situated prior to the final-layer projection head. Note that $\mathbf{v}$ remains constant throughout the recalibration process as updates are exclusively applied to the model parameters within the projection head. For ease of reference, we encapsulate a single sample with the tuple $(\mathbf{x}, \mathbf{y}, \mathbf{v})$. Subsequently, $\mathbf{v}$ is processed through the projection head, denoted as $f_\theta(\cdot)$, with θ representing the parameter. This yields the final output feature $f_\theta(\mathbf{v})$. CFR fine-tunes f_θ, resulting in the recalibration of the output visual feature $f_\theta(\mathbf{v})$. It is crucial to note that $\mathbf{v}$ itself remains unaltered during training, as all layers preceding the projection head are frozen. Let $c_\mathbf{y}$ represent the 'optimal' centroid of class $\mathbf{y}$ within the feature space, which remains *unknown* and is an idealized concept. The centroid of a class can be conceptualized as the geometric center of all samples pertaining to that class within the feature space. CFR is designed to enhance the similarity between $f_\theta(\mathbf{v})$ and $c_\mathbf{y}$, while ensuring $f_\theta(\mathbf{v})$ remains distant from samples belonging to opposing classes.

We would like to emphasize that these class centroids are unknown a priori, necessitating their estimation or on-the-fly learning during the training process. In the subsequent section, we present an in-depth discussion on the implementation of centroid estimation, employing various sample selection strategies. At a conceptual level, our approach draws inspiration from the state-of-the-art contrastive learning framework [5], which guides our feature recalibration process. Specifically, we conduct feature recalibration by strategically selecting positive and negative samples within each batch, which will serve as the anchor points for the recalibration process. This method allows us to refine and enhance the features in a manner consistent with [5].

Estimation via Sample Selection. In our pursuit of refining the sample features within the calibration set, a naïve approach is to estimate the class centroid $c_\mathbf{y}$ as the mean representation of all samples within class $\mathbf{y}$, that is, $c_\mathbf{y} = \sum_{\mathbf{y}'=\mathbf{y}} f_\theta(\mathbf{v}')/M_\mathbf{y}$, where $M_\mathbf{y}$ represents the total sample count in class $\mathbf{y}' = \mathbf{y}$ within the training dataset. However, this method has two significant limitations: (1) It does not ensure the accuracy of $c_\mathbf{y}$ as an optimal centroid, given that the averaging includes samples that may be incorrectly predicted, leading to potential misalignments in the feature space. (2) The iterative updates to $f_\theta(\cdot)$ render the exact computation resource-intensive. To address the first concern, we refine our strategy to consider solely those samples in class y accurately identified by the pre-trained model, ensuring $\widehat{\mathbf{y}}' = \mathbf{y}$. From this curated subset, we randomly select a subset of data, denoted as $P(\mathbf{x})$ (with P standing for 'positive'), to mitigate the influence of misaligned samples. To tackle the second (computational) challenge, we employ an Exponential Moving Average (EMA) update, formulating

the class centroid updating scheme as follows:

$$c_{\mathbf{y}} \leftarrow (1-\gamma)c_{\mathbf{y}} + \gamma \sum_{\mathbf{x}' \in P(\mathbf{x})} f_\theta(\mathbf{v}')/|P(\mathbf{x})|, \quad (5.1)$$

where $|P(\mathbf{x})|$ denotes the cardinality of $P(\mathbf{x})$.

To intensify the recalibration effect, we pair the positive subset $P(\mathbf{x})$ with the centroid $\{c_{\mathbf{y}}\}$ to form a positive mini-batch for a data point $(\mathbf{x}, \mathbf{y}, \mathbf{v})$. That is, the positive mini-batch is given as $P(\mathbf{x}) \cup \{c_{\mathbf{y}}\}$. We term this sample selection strategy for the positive mini-batch as ***D**ynamic **P**ositive Centroid **S**ampling* (**DPS**).

For distancing the feature representation $f_\theta(\mathbf{v})$ from different classes, CFR selects a negative mini-batch $N(\mathbf{x})$ through two strategies: (1) ***R**andom **N**egative **S**ampling* (**RNS**): randomly choosing negative samples outside the anchor's class, and (2) *Nearest-neighbour **N**egative **S**ampling* (**NNS**): selecting negative samples from the top-k instances closest to the anchor within the feature space.[4] While the latter strategy, initially validated by [72], is effective, it also incurs significant computational load on large datasets. To address this, we implement a batch sampling method followed by a top-k selection within the batch.

Combining the positive and negative sample selection, we now have two options: {DPS+RNS} and {DPS+NNS}, which differ only in how the negative batch is selected.

Calibration loss. With the positive and negative batches in place, CFR applies a contrastive loss for recalibration. For an individual instance $(\mathbf{x}, \mathbf{y}, \mathbf{v})$ as the anchor, its loss is:

$$\mathcal{L}_{\text{cal}}(\mathbf{x}) = -\frac{1}{|P(\mathbf{x})|+1} \sum_{\mathbf{v}^+ \in P(\mathbf{x}) \cup \{c_{\mathbf{y}}\}} \log \frac{e^{z_+}}{e^{z_+} + \sum_{\mathbf{v}^- \in N(\mathbf{x})} e^{z_-}}, \quad (5.2)$$

where each z_+ replies on $\mathbf{v}^+$ and is given by $z_+ = \langle f_\theta(\mathbf{v}), f_\theta(\mathbf{v}^+) \rangle / \tau$. Similarly $z_- = \langle f_\theta(\mathbf{v}), f_\theta(\mathbf{v}^-) \rangle / \tau$. In other words, $\mathcal{L}_{\text{Cal}}$ is a contrastive loss applied to the anchor $(\mathbf{x}, \mathbf{y}, \mathbf{v})$ with the positive and negative batch selected by either of the two strategies (*i.e.*, DPS+RNS and DPS+NNS).

Holistic Data Integration. Determining the size and distribution of the calibration set beforehand is a challenge, since the composition of this set can significantly differ across various downstream tasks. Particularly when dealing with a limited-sized calibration set, there is a considerable risk of model overfitting if training relies solely on this small subset, jeopardizing the model's ability to generalize.

Recognizing this potential pitfall, we incorporate the entire training dataset into the loss function for adjustment. Specifically, we start by selecting a mini-batch from the entire training dataset. For each sample $(\mathbf{x}, \mathbf{y}, \mathbf{v})$ within this mini-batch, we identify positive examples (those belonging to the same class) and negative examples (those from different classes). To optimize the model, we adopt cosine similarity loss, aiming to reduce the distance between positive pairs while increasing the distance between negative pairs. The loss is formulated as follows:

$$\mathcal{L}_{\text{CS}}(\mathbf{x}) = -\sum_{p=1}^{P} \frac{\mathbf{u}^\top \mathbf{u}_p}{\|\mathbf{u}\| \cdot \|\mathbf{u}_p\|} + \sum_{j=1}^{J} \frac{\mathbf{u}^\top \mathbf{u}_j}{\|\mathbf{u}\| \cdot \|\mathbf{u}_j\|}, \quad (5.3)$$

where $\mathbf{u}$, $\{\mathbf{u}_p\}_{p=1}^{P}$ and $\{\mathbf{u}_j\}_{j=1}^{J}$ denote the final visual representation of the image $\mathbf{x}$, images from the same group, and images from other groups, respectively. That is, $\mathbf{u} = f_\theta(\mathbf{v})$, $\mathbf{u}_p = f_\theta(\mathbf{v}_p)$, and $\mathbf{u}_j = f_\theta(\mathbf{v}_j)$.

Final loss function. The final loss combines the cosine similarity loss $\mathcal{L}_{\text{CS}}$ (Eq. 5.3) and the calibration loss $\mathcal{L}_{\text{cal}}$ (Eq. 5.2):

$$\mathcal{L}_{\text{total}} = \lambda \sum_{\mathbf{x} \in \mathcal{D}_{\text{cal}}} \mathcal{L}_{\text{cal}}(\mathbf{x}) + \sum_{\mathbf{x} \in \mathcal{D}_{\text{all}}} \mathcal{L}_{\text{CS}}(\mathbf{x}), \quad (5.4)$$

where λ is a tunable parameter and $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}_{\text{all}}$ are the calibration set and the entire training set, respectively.

6. Experiments

In this section, we evaluate CFR across various benchmarks with spurious correlations, accompanied by comprehensive ablations on design choices and hyperparameter settings. Due to limited space, additional dataset and implementation details are discussed in Appendix A and B.

Evaluated methods. As baseline methods, we compare against other semi-supervised methods, including JTT [33], CnC [73] and AFR [47]. We also compare against methods that require group annotations, including GroupDRO [51], DFR [23], S-CS/S-CL [66].

Sampling Setup. In Section 5.2, we detailed our DPS strategy for creating a positive batch for an anchor. DPS combines $P(\mathbf{x})$, a subset of correctly predicted instances by pre-trained CLIP within the anchor's class $\mathbf{y}$, with $c_{\mathbf{y}}$, the Exponential Moving Average (EMA) estimated centroid of class $\mathbf{y}$ (Eq. 5.1). To assess the impact of incorporating the positive centroid $c_{\mathbf{y}}$ into the contrastive loss term $\mathcal{L}_{\text{cal}}$, we propose to try CFR with $c_{\mathbf{y}}$ removed from the loss term. Therefore, we explore an alternative to DPS, termed ***R**andom **P**ositive **S**ampling* (**RPS**). RPS involves simply removing $c_{\mathbf{y}}$ from the positive batch of the contrastive loss $\mathcal{L}_{\text{cal}}$, resulting in a modified calibration loss term[5]:

$$\mathcal{L}_{\text{cal}}(\mathbf{x}) = -\frac{1}{|P(\mathbf{x})|} \sum_{\mathbf{v}^+ \in P(\mathbf{x})} \log \frac{e^{z_+}}{e^{z_+} + \sum_{\mathbf{v}^- \in N(\mathbf{x})} e^{z_-}}.$$

[4] The feature space here refers to the feature before the final projection layer of CLIP's visual branch, represented as $\mathbf{v}$ (introduced in Sec. 5.2).

[5] This variant, compared to the DPS loss term (Eq. 5.2), simply excludes $\{c_{\mathbf{y}}\}$ from the summation.

Table 1. **Comparison results across various supervised methods, semi-supervised methods and our proposed four methods across the Waterbirds, CelebA, CheXpert and MetaShift benchmarks.** Best results within the *semi-supervised* group are in **bold**. Please refer to the text for discussion.

		ResNet-50								ViT							
		Waterbirds		CelebA		CheXpert		MetaShift		Waterbirds		CelebA		CheXpert		MetaShift	
	Method	WGA	Avg	WGA	Avg	WGA	Avg	WGA	Avg	WGA	Avg	WGA	Avg	WGA	Avg	WGA	Avg
supervised	ERM [60]	45.64	94.08	52.78	93.88	18.46	90.01	73.85	90.05	57.91	97.60	23.33	94.30	14.07	90.48	89.23	97.37
	GroupDRO [51]	75.08	83.84	84.09	89.54	68.29	75.04	83.19	87.30	90.82	96.37	88.33	91.24	67.02	73.53	93.85	97.37
	S-CS [66]	77.51	83.16	75.24	80.38	67.34	74.74	81.15	89.82	89.09	95.69	86.11	89.29	65.26	74.48	92.31	97.14
	S-CL [66]	75.23	85.96	75.56	80.56	64.49	75.89	81.54	88.79	89.93	96.04	87.78	90.51	66.26	74.19	93.14	96.89
	DFR [23]	73.22	83.82	82.22	91.57	60.64	74.96	83.08	88.33	89.69	97.80	85.56	90.80	68.09	76.59	92.31	97.03
semi-sup	AFR [47]	48.38	89.31	53.44	94.25	45.21	59.41	76.92	86.84	73.42	88.17	70.00	85.17	48.72	74.99	90.31	97.14
	JTT [33]	61.68	90.63	60.16	79.93	45.89	59.01	78.46	89.36	83.64	97.29	75.56	93.25	50.95	73.96	91.21	94.16
	CnC [73]	61.21	87.14	63.89	90.34	45.10	57.52	78.31	87.07	84.49	97.51	79.22	89.33	58.89	74.46	92.15	94.74
	Con-Adapter [72]	69.89	70.51	63.98	90.19	42.78	59.12	77.92	85.47	86.14	95.54	76.11	93.06	49.59	71.98	91.29	93.36
	○ DPS+RNS	**76.93**	77.61	**73.66**	81.07	**54.44**	62.76	**81.54**	89.52	**88.23**	96.79	**84.77**	87.81	**64.11**	73.48	**93.72**	95.54
	○ RPS+RNS	73.08	76.66	72.78	80.52	49.50	61.02	80.13	84.55	85.67	94.74	83.78	88.17	62.03	74.22	91.15	94.05
	• DPS+NNS	76.63	78.93	73.21	77.52	52.67	62.67	**81.54**	89.59	87.58	96.40	84.11	86.67	63.69	71.62	93.41	95.31
	• RPS+NNS	72.43	77.26	68.44	70.99	46.25	60.96	81.15	89.24	84.89	96.23	82.72	88.05	62.64	73.97	92.23	94.98

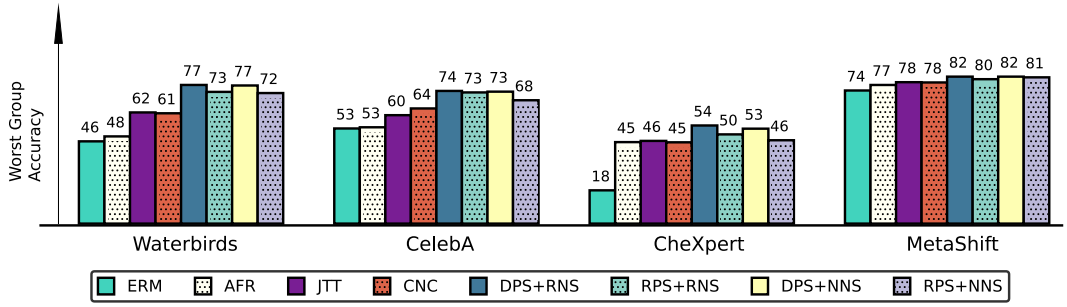

Figure 3. **Comparison of methods using the CLIP-ResNet50 architecture on four benchmark datasets.** We use Worst Group Accuracy to evaluate the performance for various methods, including ERM, semi-supervised baselines (*i.e.*, AFR [47], CnC [73], JTT [33]), and our proposed methods. We observe that CFR combined with the sample selection strategies (*i.e.*, {DPS, RPS}×{RNS,NNS}) outperforms all semi-supervised baselines across all benchmarks.

Considering the setting of {DPS, RPS} × {RNS, NNS}, we evaluate in total 4 sample selection strategies for CFR: {DPS+RNS}, {RPS+RNS}, {DPS+NNS}, {RPS+NNS}.

6.1. Results

We now present the results of the methods across multiple benchmarks, using WGA as the performance metric.

Main Results. We adopt classical visual backbones in CLIP, *i.e.*, {ResNet, ViT}, and train them in the combination of *positive* and *negative* sampling on {DPS+RNS, DPS+NNS, RPS+RNS, RPS+NNS}. Main results are in Table 1, Figure 3, and Figure 6 (Appendix). Of note, we refer more discussion in Appendix C The following observations can be drawn: ❶ Our CFR demonstrates superior performance compared to all other semi-supervised training algorithms. Specifically, CFR-ResNet with DPS+RNS obtains $\{15.25 \sim 28.55,\ 9.77 \sim 20.22,\ 8.55 \sim 9.34,\ 3.08 \sim 4.62\}$ WGA improvements across the datasets over AFR, JTT, CnC, respectively. This validates the effective of our proposed method. ❷ Appropriate selection strategies show consistent performance benefits across all two network backbones. Moreover, our DPS+RNS strategy surpasses the completely random strategy (RPS+RNS) in CLIP, which aligns well with our expectations as our assignment process is implicitly "optimized" by leveraging the naturally evolved feature embeddings. ❸ When using ViT backbone, CFR-ViT with DPS+RNS has up to $\{3.74 \sim 14.81,\ 5.55 \sim 14.77,\ 5.22 \sim 15.39,\ 1.57 \sim 3.41\}$ compared to all three semi-supervised baselines, respectively. ❹ As is shown in Table 1, we observe that the improvements are more significant using ViT backbone than ResNet backbone. Our DPS+RNS approach the benchmarks set by fully supervised models. Besides, when using a ResNet backbone, a performance gap remains, suggesting room for further improvement. A possible reason is that, under the guidance of multi-modality information during fine-tuning, using ViT are less prone to capture spurious correlation features. ❺ By visualizing the training-validation curve[6] (See Figure 4) of our method and other semi-supervised baselines, we observe that CFR converges to a better optimal solution at a faster convergence rate, demonstrating the effectiveness of our approach.

Analysis of Sample Selection. We conduct an extensive

[6]WGA results on the validation dataset during the training process.

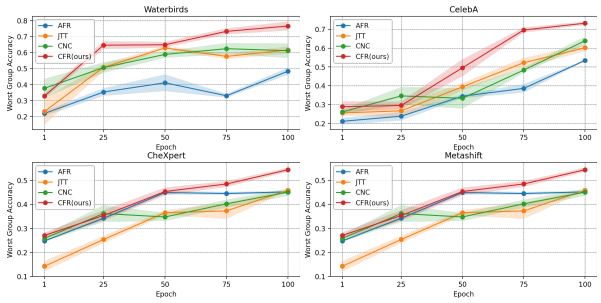

Figure 4. **Training-validation curves of various semi-supervised methods using CLIP-ResNet.** We plot WGA on a validation dataset at regular intervals of 25 epochs throughout the training process. Results are averaged across 3 random seeds.

study to understand the performance benefits of different selection strategies in terms of WGA by comparing them against multiple state-of-the-art methods. Through comparison upon the following perspectives: (1) *positive selection*: method w/o and w/ DPS strategy; (2) *negative selection*: method w/o and w/ NNS strategy, we have the following findings. **❶ DPS generally contributes to competitive performance gains.** With the assistance of an adaptive sample selection process, DPS can lead to more robust and accurate models by effectively addressing challenges such as data imbalance and improving the model's focus on more minority-group features. As shown in Figure 3 and Table 1, equipped with DPS+RNS, CLIP+ResNet achieves {3.85, 0.88, 4.94, 1.44} WGA accuracy gains compared to RPS+RNS on Waterbirds, CelebA, CheXpert, and Metashift, respectively. Furthermore, we alter the model backbone to ViT, and observe a similar phenomenon happens to DPS+RNS with CLIP-ViT. A comparison between DPS+RNS and RPS+RNS shows that DPS+RNS performs better than RPS+RNS when using the ViT backbone. Similarly, as in Figure 6 (Appendix) and Table 1, we observe that DPS+NNS outperforms RPS+NNS in terms of WGA across four benchmarks. This is clear evidence that DPS is able to better capture minority-group than RPS. **❷ In a principled way, RNS further boosts CLIP performance with DPS across all the evaluated benchmark datasets.** Given the advancements of NNS, it is naturally expected that it would lead to better performance. However, as depicted in Figure 3 and Table 1, the utilization of DPS+RNS with CLIP-ResNet, compared to the baseline using DPS+NNS, achieves better performance across Waterbirds, CelebA, CheXpert, and Metashift. Furthermore, our observations also reveal comparable improvements in WGA when utilizing CLIP-ViT (Appendix Figure 6). This underscores the effectiveness of DPS+RNS compared to using DPS+NNS. Our findings suggest that employing RNS during the fine-tuning of CLIP-ResNet50 facilitates the model's ability to seeking the most informative features, particularly those are more correlated with minority-group features. We hypothesize that the observed decrease in effectiveness with NNS may be attributed to its sensitivity to the hyperparameter k. Unlike RNS, top-k sampling is less random and, as indicated by [72], might require a larger batch size to be more effective. Nevertheless, it is important to note that a larger batch size for fine-tuning VLMs corresponds to increased computational resource requirements, which contradicts our objective of achieving lightweight fine-tuning. We hope that our finding inspires future work to further explore utilizing larger batch sizes for NNS in VLMs when computational budget allows.

Table 2. Ablation on the loss component $\mathcal{L}_{\text{CS}}$ from Holistic Data Integration in Sec. 5.2. Adding $\mathcal{L}_{\text{CS}}$ brings significant performance gain, especially with ResNet-50.

	ResNet-50			ViT		
	WGA			WGA		
Method	◦ w. $\mathcal{L}_{\text{CS}}$ (**ours**)	w/o $\mathcal{L}_{\text{CS}}$	Gain↑	◦ w. $\mathcal{L}_{\text{CS}}$ (**ours**)	w/o $\mathcal{L}_{\text{CS}}$	Gain↑
DPS+RNS	76.93	69.67	7.26	88.23	87.07	1.16
RPS+RNS	73.08	65.98	7.10	85.67	85.23	0.44
DPS+NNS	76.63	69.14	7.49	87.58	86.61	0.97
RPS+NNS	72.43	70.40	2.03	84.89	83.02	1.87

6.2. Ablation

In this subsection, we conduct comprehensive ablation studies to gain deeper insights into the rationale behind our design choices. Note that all these experiments are conducted using the Waterbirds dataset with both CLIP-ResNet50 and CLIP-ViT models.

Importance of Holistic Data Integration. We analyse the necessity of adding the loss component $\mathcal{L}_{\text{CS}}$ from the Holistic Data Integration introduced in Sec. 5.2. We evaluate the setting with and without $\mathcal{L}_{\text{CS}}$ using WGA as the metric, as shown in Table 2. Our findings reveal a notable performance gains when employing Holistic Data Integration, particularly when using CLIP-ResNet50. This underscores the significant role of $\mathcal{L}_{\text{CS}}$ in achieving performance improvements, especially considering that the calibration loss term $\mathcal{L}_{\text{cal}}$ is confined to a relatively small calibration set.

Extra Study. More studies on (1) weights of loss functions; (2) batch sizes in sample selection are in Appendix D.

7. Conclusion

In our work, we study into the group robustness of the CLIP model without using any group annotations. Our initial findings indicate that retraining the last layer can considerably improve the group robustness of a pre-trained CLIP. Building upon this, we introduce a novel and efficient representation calibration technique for fine-tuning CLIP. This method involves creating a calibration set with the pre-trained CLIP and subsequently refining the representations of the samples within this set via contrastive learning, all without the need for group labels. Through comprehensive experiments and detailed visualizations across multiple benchmarks, our method demonstrates its capability to achieve state-of-the-art results in robust classification.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore. In *NeurIPS*, 2022.

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.

[4] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *ICLR*, 2019.

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

[6] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *ICML*, 2021.

[7] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019.

[8] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

[9] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2020.

[10] Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. Model patching: Closing the subgroup performance gap with data augmentation. In *ICLR*, 2021.

[11] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022.

[12] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *NAACL*, 2018.

[13] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP*. IEEE, 2022.

[14] Yi He, Fudong Lin, Nian-Feng Tzeng, et al. Interpretable minority synthesis for imbalanced classification. In *International Joint Conferences on Artificial Intelligence*, 2021.

[15] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, 2022.

[16] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, 2019.

[17] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. In *NeurIPS*, 2022.

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

[19] Divyansh Kaushik and Zachary C Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *EMNLP*, 2018.

[20] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[21] Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *ICCV*, 2021.

[22] Nayeong Kim, Sehyun Hwang, Sungsoo Ahn, Jaesik Park, and Suha Kwak. Learning debiased classifier with biased committee. In *NeurIPS*, 2022.

[23] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *ICLR*, 2023.

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[25] Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. In *NeurIPS*, 2023.

[26] Zhengfeng Lai, Zhuoheng Li, Luca Cerny Oliveira, Joohi Chauhan, Brittany N Dugger, and Chen-Nee Chuah. Clipath: Fine-tune clip with visual feature fusion for pathology image analysis towards minimizing data collection efforts. In *ICCV*, 2023.

[27] Zhengfeng Lai, Noranart Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee Chuah. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In *ICCV*, 2023.

[28] Zhengfeng Lai, Haotian Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, et al. From scarcity to efficiency: Improving clip training via visual-enriched captions. *arXiv preprint arXiv:2310.07699*, 2023.

[29] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.

[30] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *ECCV*, 2018.

[31] Weixin Liang and James Zou. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. *arXiv preprint arXiv:2202.06523*, 2022.

[32] Fudong Lin, Xu Yuan, Lu Peng, and Nian-Feng Tzeng. Cascade variational auto-encoder for hierarchical disentangle-

ment. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1248–1257, 2022.

[33] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *ICML*, 2021.

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[35] R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *EMNLP*, 2019.

[36] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[37] Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. Overparameterisation and worst-case generalisation: friend or foe? In *ICLR*, 2020.

[38] Mazda Moayeri, Phillip Pope, Yogesh Balaji, and Soheil Feizi. A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. In *CVPR*, 2022.

[39] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[40] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *NeurIPS*, 2020.

[41] Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *ICLR*, 2022.

[42] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In *ACL*, 2019.

[43] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *CHIL*, 2020.

[44] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021.

[45] Suzanne Petryk, Lisa Dunlap, Keyan Nasseri, Joseph Gonzalez, Trevor Darrell, and Anna Rohrbach. On guiding visual attention with language specification. In *CVPR*, 2022.

[46] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *NeurIPS*, 2021.

[47] Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *ICML*, 2023.

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[49] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022.

[50] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

[51] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

[52] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[53] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.*, 2021.

[54] Rakshith Shetty, Bernt Schiele, and Mario Fritz. Not using the car to see the sidewalk–quantifying and controlling the effects of context in classification and segmentation. In *CVPR*, 2019.

[55] Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *ICLR*, 2022.

[56] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *NeurIPS*, 2020.

[57] Nimit S Sohoni, Maziar Sanjabi, Nicolas Ballas, Aditya Grover, Shaoliang Nie, Hamed Firooz, and Christopher Ré. Barack: Partially supervised group robustness with guarantees. *arXiv preprint arXiv:2201.00072*, 2021.

[58] Saeid A Taghanaki, Kristy Choi, Amir Hosein Khasahmadi, and Anirudh Goyal. Robust representation learning via perceptual similarity metrics. In *ICML*, 2021.

[59] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008.

[60] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

[61] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[62] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022.

[63] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *ICLR*, 2021.

[64] Yilun Xu, Hao He, Tianxiao Shen, and Tommi Jaakkola. Controlling directions orthogonal to a classifier. In *ICLR*, 2022.

[65] Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet, Timothy J Hazen, and Alessandro Sordoni. Increasing robustness to spurious correlations using forgettable examples. *arXiv preprint arXiv:1911.03861*, 2019.

[66] Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. Mitigating spurious correlations in multi-modal models during fine-tuning. In *ICML*, 2023.

[67] Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. In *ICML*, 2023.

[68] Yao-Yuan Yang, Chi-Ning Chou, and Kamalika Chaudhuri. Understanding rare spurious correlations in neural networks. *arXiv preprint arXiv:2202.05189*, 2022.

[69] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.*, 2018.

[70] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *NeurIPS*, 2021.

[71] Jianyu Zhang, David Lopez-Paz, and Léon Bottou. Rich feature construction for the optimization-generalization dilemma. In *ICML*, 2022.

[72] Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. In *NeurIPS*, 2022.

[73] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. In *ICML*, 2022.

[74] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR*, 2022.

[75] Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. *arXiv preprint arXiv:2302.04269*, 2023.

[76] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 2017.