Linear Calibration Approach to Knowledge-free Group Robust Classification

Ryota Ishizaki
4624501@ed.tus.ac.jp
Shunya Yamagami
4623533@ed.tus.ac.jp
Yuta Goto
4623511@ed.tus.ac.jp
Go Irie
goirie@ieee.org

Tokyo University of Science
Tokyo, Japan

Abstract

Large-scale pre-trained vision-language models such as CLIP have shown remarkable performance on various downstream tasks. However, such a model often learns not only the information that is truly useful for classification, but also *group attributes* that are spuriously correlated with classes, leading to misclassification of an image into a group with the same group attributes but with the wrong class. The goal of this paper is to develop a method for learning a classifier that is robust to the group attributes. Unlike existing methods, our method is (i) *knowledge-free*: does not use any information of group attributes for training, (ii) *linear*: a lightweight method that trains only a single linear projection, and (iii) *calibration-based*: does not change the original classifier at all. The negative effects of the group attributes can be canceled by projecting the classification space to the orthogonal complement of the subspace spanned by the group attributes. To achieve this, we propose Spurious Subspace Mining (SSM) to discover the subspace from a random set of text embeddings without any supervision. Experimental results on two standard benchmark datasets, Waterbirds and CelebA, show that the proposed method outperforms various existing methods and improves zero-shot baseline by 35.2% in worst-group accuracy. Our code is available at https://github.com/LyricProduct/SSM.git

1 Introduction

Large-scale vision-language models like CLIP and ALIGN have demonstrated impressive zero-shot performance on various downstream tasks without any fine-tuning [9, 20]. However, due to their high learning ability, these models are known to learn not only the core features truly important for classification, but also undesirable biases contained in the training dataset [13, 16, 17, 22]. For example, vision-language models may misclassify a person's hair color based on his/her gender, due to the bias that blonde is more common in women. In this paper, we refer to the attributes (e.g., gender) that are essentially unrelated to class

Method	Knowledge -free	Linear	Calibration -based
ERM Linear Probe [12]	✓	✓	-
ERM Adapter [8]	✓	-	-
WiSE-FT [25]	✓	✓	-
DFR [11]	✓	✓	-
Orth-Cali [4]	-	✓	✓
Contrastive Adapter [29]	-	-	-
FairerCLIP [5]	-	-	✓
ROBOSHOT [1]	✓	✓	✓
Ours	✓	✓	✓

Table 1: **Properties of Group Robust Classification Methods for Vision-Language Models.** Unlike most existing methods, our method satisfies the three desirable properties of *knowledge-free*, *linear*, and *calibration-based*. The only exception is ROBOSHOT [1], but we show the significant superiority of our method in our experiments.

labels (e.g., hair color) but are spuriously correlated with them, and thus cause misclassification, as "group attributes". The goal of this work is to improve the robustness of pre-trained vision-language models to the group attributes.

In addressing the challenge of group robustness, several methods have been proposed [2, 19, 26, 27]. While some earlier methods are focused on unimodal classifiers [6, 23, 30], recent studies have explored improving robustness for pre-trained vision-language models [4, 5, 25, 29]. For example, Chuang et al. [4] propose a calibration method that mitigates their negative impact by projecting the feature space to the orthogonal complement of the subspace spanned by (the embedding vectors of) the group attributes. Zhang et al. [29] propose to learn a nonlinear adaptor through supervised contrast learning with extended positives that include samples that are in the same class but far from the anchor, as these are likely from different groups. FairerCLIP [5] propose to project image and text embeddings by a pair of nonlinear kernel mappings so that the features are insensitive to given group attributes.

In this paper, we propose a novel approach to group robust classification for improving pre-trained vision-language models. As we summarize in Table 1, unlike most of the existing methods, our method satisfies all of the following three desirable properties.

(i) **Knowledge-Free.** Most of the existing methods [4, 5, 29] assume that the group attributes are completely known for training, which cannot always be assumed in practical scenarios. Our approach does not require any knowledge of the group attributes for training.

(ii) **Linear.** Several methods learn non-linear projections for tuning [5, 29]. Compared to these methods, our method trains only a single linear projection and is thus lightweight.

(iii) **Calibration-based.** While some existing methods change the original vision-language classifiers by directly updating their parameters or appending additional adapters for tuning [11, 25, 29], our method does not change the classifier at all.

One very recent method, ROBOSHOT [1], is the only exception that satisfies all the above three properties as our method does; we show later in our experiments that our method significantly outperforms ROBOSHOT.

To achieve these desirable properties, we propose Spurious Subspace Mining (SSM), a

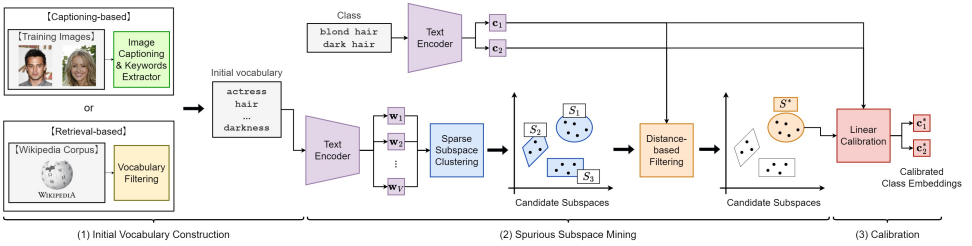


Figure 1: **Method Overview.** Our method consists of three steps. (1) Initial Vocabulary Construction: acquire a set of words likely to contain group attributes. (2) Spurious Subspace Mining (SSM): discover a set of subspaces $\{S_j\}$ of the word embeddings and specify the *spurious subspace* S^* most likely to be spanned by the group attributes. (3) Calibration: project the original class embeddings onto the orthogonal complement of the spurious subspace S^* to cancel the negative impact of the group attributes.

simple and effective method for improving worst-group accuracy of pre-trained CLIP in situations where group attributes are totally unknown. The classification by CLIP is performed based on the inner product of the image and class embeddings. If the group attributes are known, undesirable effects of the group attributes can be canceled by linearly projecting the embeddings to the orthogonal complement of the subspace spanned by the group attributes [4]. This suggests that even if the group attributes are unknown, we do not need to identify the group attributes themselves – **we can cancel out the negative impact of the group attributes as long as the subspace spanned by the group attributes can be estimated.** Based on this idea, in our SSM, we first discover a set of candidate subspaces from a random vocabulary in an unsupervised manner through sparse subspace clustering [7, 28] and then find the most promising subspace based on the distance to the class embeddings. Experimental results on two standard benchmark datasets, Waterbirds and CelebA, demonstrate that the proposed method improves worst-group accuracy in zero-shot CLIP by up to 35.2%.

The key contributions of this paper can be summarized as follows:

- We propose a knowledge-free, linear, and calibration-based approach to group robust classification.
- We propose Spurious Subspace Mining (SSM) to efficiently find the subspace likely to be spanned by group attributes without using any knowledge of group attributes.
- Our method outperforms the state-of-the-art group robust classification methods and is even highly competitive with the methods assuming knowledge of group attributes is available for training.

2 Related Work

Group robustness in pre-trained vision-language models has gained significant attention. For example, Chuang et al. [4] proposed to mitigate the undesirable impacts caused by group attributes by projecting the feature embeddings into the orthogonal compliment of the subspace spanned by the group attributes. Contrastive Adapter [29] trains a nonlinear adapter to pull

together distant embeddings within the same class, while to push apart nearby embeddings in different classes in a supervised contrastive learning manner. FairerCLIP [5] introduces additional image encoder and text encoder, and fine-tunes these two encoders. Including these examples, most existing methods have been developed under the assumption that group attributes are known for training the models (or tuning the hyperparameters), which may not always be true in practice.

Motivated by this idea, several recent methods have been designed to avoid using any knowledge of group attributes in training. WiSE-FT [25] takes an ensemble of the parameters of the zero-shot and fine-tuned vision-language models to improve the group robustness. Deep Feature Reweighting (DFR) [11] fine-tunes the top classification head by using group-balanced data. These methods are knowledge-free but directly update the parameters of the pre-trained models, and thus require sufficient computational resources for training.

Unlike these existing methods, our method is knowledge-free and calibration-based, which does not change the original model at all. The only exception is ROBOSHOT [1], which utilizes Large Language Models such as ChatGPT [18] and LLaMA [24] to identify beneficial and harmful components for group robust classification. However, our method outperforms ROBOSHOT by significant margins.

3 Method

The overview of the proposed method is illustrated in Fig. 1. Given a training dataset and a pre-trained CLIP classifier, our method performs the following three major steps to obtain a linear projection for calibrating the class embeddings of the CLIP classifier.

(1) **Initial Vocabulary Construction**: We obtain an initial vocabulary (i.e., a set of words) that is likely to contain group attributes by using the training dataset. In this paper, we consider two different approaches: Captioning-based and Retrieval-based.

(2) **Spurious Subspace Mining (SSM)**: Under the assumption that the embeddings of the initial vocabulary are distributed over a union of subspaces, we first discover those subspaces in an unsupervised manner by utilizing sparse subspace clustering, and then specify a single subspace most likely to have the group attributes (which we call *spurious subspace*) through filtering based on point-to-subspace distances from the class embeddings.

(3) **Calibration**: We find the linear projection to the orthogonal complement of the specified spurious subspace and project the original class embeddings with the found projection.

We hereafter describe the details of each step. The core of the proposed method is in (2) SSM.

3.1 Initial Vocabulary Construction

Given a training dataset and other possible sources of vocabulary or models, we want to obtain an initial set of V words that is likely to cover the target group attributes. While there are several possible approaches, we use the following two approaches in our method:

1. **Captioning-based**: This approach acquires the initial vocabulary by applying an external pre-trained image captioning model to the images in the training dataset. Following Kim et al. [10], we adopt ClipCap [15] as our image captioning model to obtain the text captions of the training images, and then extract the V keywords from the captions by using YAKE [3], an unsupervised keyword extraction method. While the B2T approach filters the obtained keywords using a proprietary score denoted as CLIP score, our method uses all the V keywords to construct our initial vocabulary.

2. **Retrieval-based**: This approach uses a Wikipedia corpus to build the initial vocabulary. We first extract only nouns from the corpus and then filter only those highly correlated with the training images. More specifically, we first compute the text embeddings of all the nouns using the prompt "This is a picture of a [noun]." After evaluating the similarities (dot products) of all the pairs of image and text embeddings, we sort the similarities in descending order and filter only the top V frequent nouns for constructing the initial vocabulary.

3.2 Spurious Subspace Mining (SSM)

Note that our calibration is done by linearly projecting the class embeddings of the original CLIP classifier. Hence, we do not need to directly identify the group attributes themselves from the words in the initial vocabulary, but we only need to find the *spurious subspace*, i.e., the subspace spanned by the group attributes. To identify the spurious subspace, we perform (i) discovery of candidate subspaces by sparse subspace clustering and (ii) identification of the spurious subspace by Distance-based subspace filtering.

Sparse Subspace Clustering. Let us assume that the V text embeddings in the initial vocabulary are distributed over a union of low-dimensional subspaces in an ambient high-dimensional embedding space. Under this assumption, Sparse Subspace Clustering (SSC) [7, 28] can be employed to discover the set of subspaces in which V text embeddings live. Based on the fact that a sample in a low-dimensional subspace can be efficiently reconstructed by a linear combination of a small number of samples in the same subspace, SSC discovers each subspace by solving a sparse L_1 reconstruction problem. We apply SSC to the V text embeddings to discover K candidate subspaces.

Distance-based Subspace Filtering. We then specify the spurious subspace from the K subspaces. Since the group attributes are spuriously correlated with the true class labels, the spurious subspace is likely to be close to the original class embeddings. Based on this idea, we identify the subspace closest to the class embeddings as the spurious subspace.

Let $\{S_j\}_{j=1}^K$ be the set of K candidate subspaces obtained by SSC and $\{\mathbf{v}_k^j\}_{k=1}^n$ be the orthonormal basis obtained by Gram-Schmidt's orthogonalization for embeddings of all the words in S_j. Then the subspace-to-point distance of the subspace S_j to the i-th class embedding $\mathbf{c}_i$ is computed as the length of the perpendicular line between the class embedding $\mathbf{c}_i$ and its point projected into the subspace $\sum_k (\mathbf{c}_i^\top \mathbf{v}_k^j)\mathbf{v}_k^j$. By taking its average over all the C class embeddings $\{\mathbf{c}_i\}_{i=1}^C$, the final spurious subspace S^* is determined by:

$$S^* = \operatorname*{argmin}_{S_j} \frac{1}{C} \sum_i \|\mathbf{c}_i - \sum_k (\mathbf{c}_i^T \mathbf{v}_k^j)\mathbf{v}_k^j\|_2. \tag{1}$$

3.3 Calibration

We calibrate the original class embeddings by linearly projecting them onto the orthogonal complement of the spurious subspace S^*. Given a matrix A whose columns are the embeddings of all the words (group attributes) in S^*, the projection matrix P_0 to the orthogonal complement of S^* is obtained as:

$$P_0 = I - A(A^T A)^{-1} A^T. \tag{2}$$

However, this simple projection matrix P_0 is not sufficiently reliable, because the complete set of group attributes cannot always be accurately identified, and the number of available group attributes may be smaller than the number of dimensions of the original embedding space, which makes the problem underdetermined. To improve the reliability of the projection matrix, instead of using the simple orthogonal projection P_0, we follow [4] and use a set of positive pairs $\mathcal{Q}$ to regularize the projection matrix. Formally, the problem is written as:

$$\min_P \|P - P_0\|^2 + \frac{\eta}{|\mathcal{Q}|} \sum_{(i,j)\in\mathcal{Q}} \|P\mathbf{z}_i - P\mathbf{z}_j\|^2, \tag{3}$$

where $(\mathbf{z}_i, \mathbf{z}_j)$ is the embedding of pair (i, j) in $\mathcal{Q}$ and (i, j) are prompts that describe the same class but different group attributes in S^*. η is a hyperparameter. This problem has a closed-form solution P^* as:

$$P^* = P_0 \left(I + \frac{\eta}{|\mathcal{Q}|} \sum_{(i,j)\in\mathcal{Q}} (\mathbf{z}_i - \mathbf{z}_j)(\mathbf{z}_i - \mathbf{z}_j)^T \right)^{-1}. \tag{4}$$

The calibration is performed by applying P^* to the original class embeddings; the calibrated confidence values are computed in the orthogonal compliment of the subspace spanned by potential group attributes S^*, which makes the final classification result insensitive to the group attributes.

4 Experiments

We conduct experiments to evaluate the effectiveness of the proposed method.

Datasets. We use the following two common benchmark datasets for evaluating group robust classification:

- **CelebA** [14] is a set of 162,770 images of celebrities. The task is to classify the hair color of a celebrity as "blonde" or "not blonde". The class label is spuriously correlated with the celebrity's gender, which can be either of "male" or "female".
- **Waterbirds** [22] consists of 4,795 images constructed by cropping birds from images in Caltech-UCSD Birds-200-2011 and transferring them to the background of Places. The task is to classify images of birds as "waterbird" or "landbird", and the class label is spuriously correlated with the background, which is either of "land background" or "water background."

Metrics. Following prior work [6, 10], we use the following three major evaluation metrics for group robust classification. (i) WG: Both datasets have two classes and two group

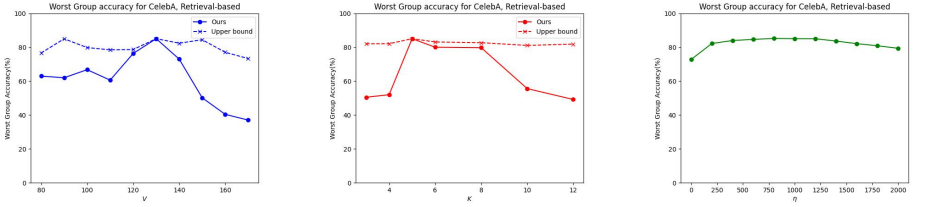
Method	CelebA			Waterbirds		
	WG↑	Avg↑	Gap↓	WG↑	Avg↑	Gap↓
methods without group attributes knowledge						
Zero-shot CLIP [20]	72.8	87.7	14.9	44.2	90.4	46.2
ERM Linear Probe [12]	28.3	94.7	66.4	65.9	97.6	31.7
ERM Adapter [8]	42.8	93.6	50.8	<u>77.6</u>	97.8	<u>20.2</u>
WiSE-FT [25]	80.0	87.4	7.4	65.9	97.6	31.7
DFR (Subsample) [11]	76.3	92.1	15.8	51.9	95.7	43.8
DFR (Upsample) [11]	<u>83.7</u>	91.2	7.5	65.9	96.1	30.2
B2T [10]	73.3	88.0	14.7	61.2	84.9	23.7
ROBOSHOT [1]	82.6	85.5	2.9	45.2	79.9	34.7
Ours (Captioning-based)	82.2	84.2	<u>2.0</u>	**79.4**	88.5	**9.1**
Ours (Retrieval-based)	**85.1**	85.6	**0.5**	69.0	91.2	22.2
methods using group attributes knowledge						
Orth-Cali [4]	76.1	86.2	10.1	67.1	83.6	16.4
Contrastive Adapter [29]	84.6	90.4	5.8	86.9	96.2	9.3
FairerCLIP [5]	85.2	87.8	2.5	78.1	85.1	7.1

Table 2: **Comparison with Existing Methods.** The best and the second best in WG and Gap are **bolded** and <u>underlined</u>, respectively. Compared with the existing knowledge-free methods (which do not require knowledge of group attributes), our method outperforms all the baselines in WG and Gap values on both datasets. Furthermore, our method is highly competitive to the knowledge-based methods in some cases.

attributes, resulting in all the samples being divided into one of four groups (2 classes $\times$ 2 group attributes = 4 groups). WG is the average classification accuracy within the lowest accuracy group out of the four. Higher WG means higher group robustness, (ii) Avg: The weighted average accuracy with the weights corresponding to the relative proportions of each group in the training data, and (iii) Gap: The gap between WG and Avg.

Implementation Details. We use a pre-trained CLIP models with ViT-L/14 visual backbones [21] as the classifier. The size of the initial vocabulary V and the number of subspaces K in SSC are set to $V = 40, K = 2$ for Waterbirds and $V = 200, K = 7$ for CelebA in Captioning-based approach, and $V = 60, K = 3$ for Waterbirds and $V = 130, K = 5$ for CelebA in Retrieval-based approach. Another hyperparameter inside SSC, i.e., the weight of the L_1 regularization term, is set to 1.0. The coefficient of the regularization term in Eq. (4), denoted as η is set to 1000, following the default value established by Chuang et al [4]. We do not tune this parameter.

Baselines. We compare our method with a variety of existing knowledge-free group robust classification methods (i.e., methods that assume that group attributes are unknown) for vision-language models. Specifically, zero-shot CLIP [20], empirical risk minimization (ERM) with linear probing [12], ERM with non-linear adapter [8], WiSE-FT [25], DFR [11], B2T [10], and ROBOSHOT [1]. We also compare our method with three "knowledge-based" methods, i.e., methods that assume that group attributes are known for training, namely, Orth-Cali [4], Contrastive Adapter [29] and FairerCLIP [5].



(a) Sensitivity to V. (b) Sensitivity to K. (c) Sensitivity to η.

Figure 2: **Sensitivity to Hyperparameters.** Our method is affected by several hyperparameters. In particular, the number of words included in the initial vocabulary (V), the number of clusters preset in sparse subspace clustering (K) and the value of the coefficient η for the regularization term in Eq. (4) have a significant impact. We also indicate the upper bounds that the subspace filtering could achieve. The results of the sensitivity analysis indicate that our method is somewhat sensitive to V, K and η, but it demonstrates favorable performance within reasonable ranges.

4.1 Main Results

Table 2 shows the results on CelebA and Waterbirds. Our method outperforms all the existing knowledge-free methods in WG and Gap for both datasets. Ours improves Zero-shot CLIP in WG by 12.3% on the CelebA and by 35.2% on Waterbirds, which proves the significant effectiveness of our method. DFR and ERM Adapter show excellent WG accuracy on one of the datasets but are severely degraded on the other dataset. In contrast, our method yields satisfactory performance on both datasets, demonstrating its stability and versatility. When ours is compared with B2T, the initial vocabulary of our method (Captioning-based) is constructed in a similar way to B2T. However, ours significantly surpasses B2T in the WG accuracy. ROBOSHOT, the most recent knowledge-free method having the closest property to ours, has a distinct disadvantage of significantly lower WG values on Waterbirds. Our approach surpasses ROBOSHOT, underscoring the superiority of our method.

Furthermore, our method is highly competitive with the existing knowledge-based methods, despite being knowledge-free. In particular, ours is superior or comparable to all the knowledge-based methods in WG accuracy; the only exception is Contrastive Adapter on Waterbirds. Although our method follows Orth-Cali in calibrating the embeddings by seeking linear projections to a healthy subspace, our method significantly outperforms Orth-Cali in WG on both of the two datasets. This highlights the remarkable advantage of our method.

Finally, one limitation of our method is that it shows a slight degradation from Zero-shot CLIP in Avg. However, this is not unique to our method, and most of the methods show a similar trend; especially the ones that show stable improvements in WG tend to exhibit lower Avg. Exploring group robust classification methods that perform well in both WG and Avg would be a major challenge in this field today.

4.2 Analysis

Sensitivity to Hyperparameters. Fig. 2 shows the sensitivity of our method to the hyperparameters, the size of the initial vocabulary V, the number of subspaces K and the value of the coefficient η for the regularization term in Eq. (4). Our method is somewhat sensi-

Method	Distance -based	Similarity -based	CelebA WG↑	CelebA Avg↑	CelebA Gap↓	Waterbirds WG↑	Waterbirds Avg↑	Waterbirds Gap↓
Ours (Captioning-based)	✓		82.2	84.2	2.0	**79.4**	88.5	9.1
Ours (Captioning-based)		✓	80.5	83.2	2.7	77.4	84.2	**6.8**
Ours (Retrieval-based)	✓		**85.1**	85.6	**0.5**	69.0	91.2	22.2
Ours (Retrieval-based)		✓	84.1	84.6	**0.5**	69.0	91.2	22.2

Table 3: **Analysis of Method Configuration.** The best and the second best in WG and Gap are **bolded** and underlined, respectively. The results demonstrate that the Distance-based subspace filtering approach is better than Similarity-based approach in WG and Avg.

Method	Regular -ization	CelebA WG↑	CelebA Avg↑	CelebA Gap↓	Waterbirds WG↑	Waterbirds Avg↑	Waterbirds Gap↓
Ours (Captioning-based)	✓	82.2	84.2	2.0	**79.4**	88.5	9.1
Ours (Captioning-based)		**84.2**	85.6	**1.4**	78.0	83.1	**5.1**
Ours (Retrieval-based)	✓	**85.1**	85.6	**0.5**	**69.0**	91.2	**22.2**
Ours (Retrieval-based)		84.1	85.4	1.3	66.2	90.9	24.7

Table 4: **Analysis of Regularization in Calibration.** The best values in WG and Gap are **bolded**. The analysis results demonstrate that in most cases, regularization leads to an improvement in WG values over simple projection. Furthermore, the values of Avg are also improved by regularization in most cases.

tive to V, K and η but achieves reasonable accuracy over wide ranges of their values. The upper bounds of performance for different V and K are also shown by dashed lines, which are obtained by using the subspace that gives the best WG values from among the K subspaces, instead of identifying it by our Distance-based subspace filtering based on Eq. (1). Our Distance-based filtering achieves performance fairly close to the upper bounds within reasonable ranges ($120 \leq V \leq 140$ and $5 \leq K \leq 8$), which confirms its strong effectiveness.

Analysis of Method Configuration. Besides Distance-based filtering based on Eq. (1), we test Similarity-based filtering that finds the subspace with the highest average cosine similarity between the class embeddings and embeddings of all the words in S_j. Table 3 shows the results. While the Similarity-based filtering performs reasonably well, the Distance-based filtering is more favorable in WG accuracy, regardless of the initial vocabulary construction method. This highlights the advantage of the proposed configuration.

Analysis of Regularization in Calibration. In sec. 3.3, we describe two types of calibration: simple projection and projection with regularization. Here, we present the experimental results of these two types of calibration in Table 4. Calibration based on P_0 calculated by Eq. (2), is shown without regularization, while calibration based on P^*, computed by Eq. (4), is presented with regularization. These results indicate that the introduction of regularization improves WG and Avg in most cases.

5 Conclusions

We proposed a knowledge-free, linear, and calibration-based approach to group robust classification for pre-trained vision-language models. The core is Spurious Subspace Mining (SSM), which finds the subspace spanned by the group attributes in an unsupervised manner and eliminates their negative influences. Results demonstrated its remarkable superiority to the state-of-the-art methods.

References

[1] Dyah Adila, Changho Shin, Linrong Cai, and Frederic Sala. Zero-shot robustification of zero-shot models with foundation models. In *Proc. ICLR*, 2024.

[2] Bang An, Sicheng Zhu, Michael-Andrei Panaitescu-Liess, Chaithanya Kumar Mummadi, and Furong Huang. PerceptionCLIP: Visual classification by inferring and conditioning on contexts. In *Proc. ICLR*, 2024.

[3] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020.

[4] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

[5] L Dehdashtian, Lan Wang, and V Boddeti. Fairerclip: Debiasing zero-shot predictions of clip in rkhss. In *Proc. ICLR*, 2024.

[6] Yihe Deng, Yu Yang, Baharan Mirzasoleiman, and Quanquan Gu. Robust learning with progressive data expansion against spurious correlation. In *Proc. NeurIPS*, 2023.

[7] Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *TPAMI*, 35(11):2765–2781, 2013.

[8] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, 132(2):581–595, 2024.

[9] 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 *Proc. ICML*, 2021.

[10] Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and mitigating visual biases through keyword explanation. In *Proc. CVPR*, 2024.

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

[12] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *Proc. ICLR*, 2022.

[13] 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 *Proc. ICML*, 2021.

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

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

[16] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Proc. NeurIPS*, 2020.

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

[18] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proc. NeurIPS*, 2022.

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

[20] 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 *Proc. ICML*, 2021.

[21] 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 *Proc. ICML*, 2021.

[22] 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 *Proc. ICLR*, 2020.

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

[24] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[25] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proc. CVPR*, 2022.

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

[27] Chenyu You, Yifei Min, Weicheng Dai, Jasjeet S Sekhon, Lawrence Staib, and James S Duncan. Calibrating multi-modal representations: A pursuit of group robustness without annotations. In *Proc. CVPR*, 2024.

[28] Chong You, Daniel Robinson, and René Vidal. Scalable sparse subspace clustering by orthogonal matching pursuit. In *Proc. CVPR*, 2016.

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

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