2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Multi-Scale Contrastive Adapter for Vision-Language Model Group Robustness

Yue Cai[1], Wenqiong Zhang[1*], and Yikai Wang[1]

[1] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, 320100, China
2021070707@njupt.edu.cn

Abstract. While vision-language models (VLMs) like CLIP demonstrate strong zero-shot classification capabilities, their robustness to group shifts remains a critical challenge, as classification accuracy degrades significantly for minority groups. Existing methods to improve group robustness often require costly full-model retraining or rely on single-scale feature representations, which may inadequately capture diverse group characteristics. We propose the Multi-Scale Contrastive Adapter (MSCA) and related modules, a novel framework designed to improve group robustness in VLMs with less computational cost. MSCA employs a multi-scale feature representation strategy, leveraging contrastive learning across multiple dimensions to alleviate the group shift of the model on the dataset in multiple different dimensional spaces. A feature voting mechanism is introduced to dynamically select the most relevant feature dimensions during inference, further improving group robustness. Experiments across benchmarks (Waterbirds, CelebA, CIFAR-10.02) show that MSCA significantly improves the worst-group accuracy to 86.1% and reduces GAP from 55.2% to 4.1%, outperforming recent advanced methods like FairerCLIP. Our findings highlight that MSCA offers a practical pathway toward more robust vision-language models.

Keywords: Group Robustness, Vision-Language Models, Multi-Scale

1 Introduction

With the development of Vision-Language Models (VLMs) such as CLIP on large-scale image-text datasets [1, 2, 3, 4, 5], these models have demonstrated the ability to learn highly correlated representations between image-text pairs, exhibiting powerful zero-shot classification capabilities [1, 6]. CLIP learns from massive multimodal data to establish a close correspondence between image content and text description, allowing it to recognize and classify previously unseen categories or concepts without task-specific training [1].

However, a question worthy exploration is: Does zero-shot inference maintain group robustness—i.e. performing well on all groups—when confronted with group shift? Group shifts occur when different groups within the same classification task form

* Corresponding author

distinctly different distribution patterns in the model's feature space, causing the model to perform significantly worse on certain groups [7, 8, 9, 10, 11]. For example, in the Waterbirds dataset [12], images of the "waterbird" category may appear against either "water" or "land" backgrounds, and models tend to use the background as a basis for classification, resulting in substantially decreased classification accuracy for samples appearing in "atypical" backgrounds (such as waterbirds against land backgrounds). Additionally, in the CelebA dataset [13], the group of blond-haired men may be misclassified more frequently due to the model's reliance on gender or hair color. In such cases, "waterbird" or "landbird" and "blond hair" or "not blonde hair" are the class labels for classification in the Waterbirds and CelebA datasets, respectively. "land background" or "water background" and "male" or "female" are the group labels in the Waterbirds and CelebA datasets, respectively. Despite models demonstrating excellent overall classification accuracy, their prediction accuracy may dramatically decrease when identifying the same object under specific backgrounds or conditions, forming a significant accuracy gap. This disparity in accuracy and the low classification accuracy of the worst-performing group directly reflect poor group robustness in models.

The model's average classification accuracy on evaluation datasets often masks its severe failures on specific challenging groups, forming significant classification accuracy gaps. Zhang and Ré confirmed this phenomenon on CLIP, finding through experiments that CLIP's zero-shot classification across multiple standard benchmarks showed up to 80.7% lower accuracy in the worst-performing groups compared to average accuracy [14], and classification accuracy is as low as 6.0% in some groups. These findings reflect group robustness weakness in some VLM models. Most existing VLM zero-shot classification evaluations primarily focus on overall or average accuracy metrics but neglect to examine these models' accuracy variations across diverse groups. This evaluation limitation prevents a comprehensive understanding of these VLMs' true capabilities. Therefore, improving VLMs' group robustness under diverse data distributions, ensuring they maintain reliable accuracy rates across all relevant groups, has become a key challenge in current studies [15].

Methods dedicated to improving model group robustness typically fall into two categories: one assumes the availability of group labels during training, balancing training proportions across different groups through techniques such as sample balancing [16, 17, 18], importance weighting [19, 20], or robust optimization [21, 22]; some other methods do not rely on training group labels, but adopt a two-stage strategy: first, they train an initial model using Empirical Risk Minimization (ERM) [23, 24], then they use this model's predictions to infer group labels, and finally train a second robust model through sample balancing, importance weighting, or representation learning. Although these methods have achieved significant effectiveness in improving group robustness, they generally require training one or more complete models, which is computationally costly and less practical for large-scale VLMs.

In contrast, the Contrastive Adapter method leverages contrastive learning ideas [14], using a lightweight adapter to process embeddings output by CLIP's image encoder, resulting in a model that not only maintains high group robustness but also greatly reduces computational cost. Since only the adapter needs to be trained instead of the entire CLIP model, this method requires training only about 1% of the VLM's

parameters but can improve worst-group accuracy by 8.5% to 56.0% across multiple benchmarks. However, the Contrastive Adapter still relies on a hidden assumption: the feature dimensions of embeddings output by the VLM's image encoder are always effective representations to do downstream tasks for different groups across different datasets, enabling the worst-performing groups in original zero-shot prediction to show significant improvement in CLIP with the Contrastive Adapter. But this assumption may not hold in practical applications because the key distinguishing features of different groups may be distributed across feature spaces of different dimensions. Therefore, it is necessary to combine feature representations of different granularities in the network structure.

To verify this conjecture, we designed a simple experiment, observing model worst-group accuracy by varying the hidden layer dimensions of the Contrastive Adapter. Experiment shows that when the hidden layer dimensions change, the worst-group accuracy of the model exhibits significant fluctuations. This fluctuation phenomenon indicates that the selection of feature space dimensionality indeed has an important impact on the model's group robustness, and there is no single optimal dimension that can simultaneously satisfy the representational needs of all groups.

This finding led us to consider that different groups may behave differently depending on the feature dimensions. Based on this, can we design a mechanism that simultaneously exploits representations of multiple granularities, contrastively learns in subspaces of different dimensions, and finally provides the most appropriate feature representation for the downstream tasks of VLM? Based on this reasoning, we propose the Multi-Scale Contrastive Adapter (MSCA).

The Multi-Scale Contrastive Adapter (MSCA) approach begins by extracting the original embedding vectors from the Vision-Language Model (VLM) image encoder. These embeddings are then processed using a matryoshka embedding strategy [25], which generates nested multi-scale representations. Through contrastive learning in these multi-dimensional spaces, the model ouuputs multiple new embeddings, allowing features at various scales to be compared within their respective feature spaces. The following up-projection layer ensures that each low-dimensional feature is mapped back to the original input dimension. This restoration allows all processed features to engage in similarity calculations within the same space as the original class query vectors. During training, the model assigns weights to the logits from embeddings of different dimensions and queries, with these weights being learned by a feature voter. During inference, the feature voter assesses the importance of features at each dimension and selects the embedding that will be used for the final dot product with the query, thus generating logits for classification. Our method constitutes a complete framework that both brings same-class samples closer together and adapts to group-specific representations, thereby significantly enhancing the group robustness of VLMs.

In this work, we introduce several key contributions aimed at improving the efficiency and group robustness of the CLIP model through a novel framework design.

Contribution 1: To the best of our knowledge, we first applied Matryoshka embedding within the CLIP adapter as a more efficient and more suitable dimensionality reduction method for MSCA.

Contribution 2: Our Multi-Scale Contrastive Adapter pulls the positive anchor samples closer and pulls the anchor samples and hard negative samples farther away in multiple dimensional spaces instead of only one original dimensional space. This can alleviate the group shift and improve group robustness of the model on the dataset in multiple dimensional spaces. Crucially, MSCA achieves these improvements without requiring full model retraining or additional group labels, making it a computationally efficient and scalable solution for addressing group shifts.

Contribution 3: Our up-projection layer restores all embeddings output by MSCA to the original dimension, and then uses feature voters to obtain the most appropriate feature representation for samples from different groups, further improving group robustness.

2 Related Work

Traditional Group Robustness Methods via Full Model Training: Traditional Group Robustness Methods via Full Model Training: Various studies have focused on enhancing group robustness. When group annotations are available during training, existing techniques frequently employ strategies such as group distribution balancing [16, 17, 18], weighted importance sampling [19, 20], or optimization for robustness [21, 22]. Under such conditions, a typical methodology includes approaches like ERM Linear Probe [26] - a linear classifier trained on frozen foundation model embeddings using Empirical Risk Minimization, and ERM Adapter [27]- small bottleneck MLPs trained with ERM that transform foundation model embeddings to improve downstream task accuracy metrics. Other methods include DFR variants that balance group distributions through subsampling instances before training a classifier on frozen features [28], or upsampling minority group samples [28]. Despite their effectiveness in improving group robustness, these approaches necessitate the training of at least one complete model, and often multiple models. This requirement renders them computationally prohibitive when applied to large-scale foundation models.

Enhancing group robustness in the CLIP: Many works aim to improve group robustness and debias in the CLIP. Debiasing refers to the process of mitigating unwanted correlations and prejudices in model representations and outputs that can lead to unfair or biased predictions across different demographic groups. Zhang & Ré employed a contrastive adapter training strategy to enhance group robustness in foundation models [14]. Berg et al. proposed a method that reduces bias in vision-language models by using adversarial debiasing with contrastive loss [29]. Seth et al. introduced DeAR, which uses additive residuals to offset image representations and ensure fairer outputs in vision-language models [30]. Chuang et al. proposed a method to debias vision-language models by projecting out biased directions in text embeddings [31]. Chen et al. found that adapter-based methods outperform full fine-tuning in vision-language model robustness [32]. Dehdashtian et al. FairerCLIP debiases CLIP's zero-shot predictions using functions in RKHSs, enhancing fairness and robustness [33]. These methods may lead to a large increase in computational overhead and poor adaptability to specific data.

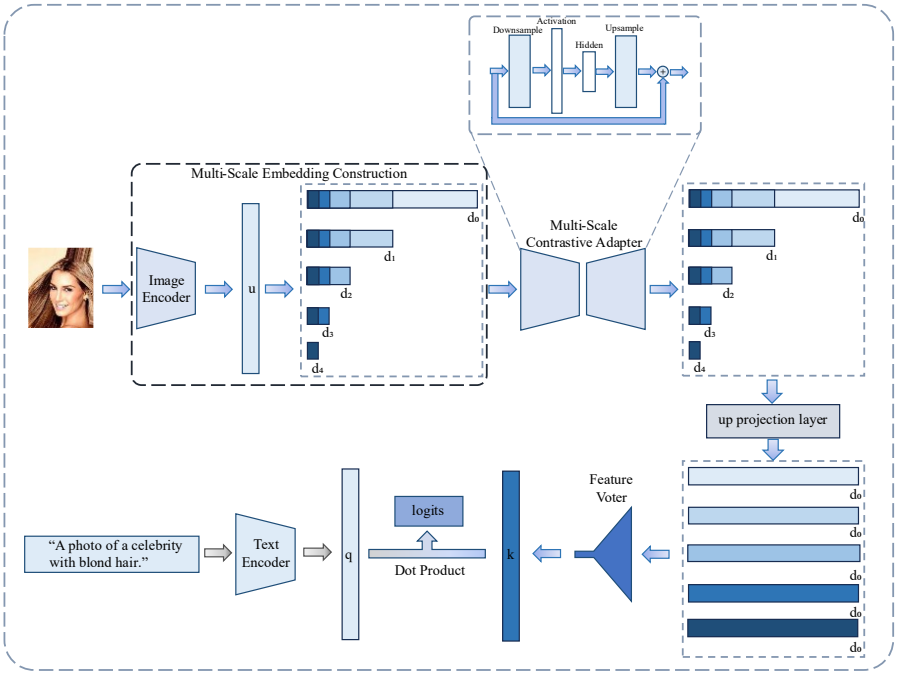

Fig. 1. The overall architecture of our approach.

3 Method

The core idea of MSCA is that different groups in the same dataset may be better represented under different feature dimensions, thus, multiple dimensional feature representations should be utilized simultaneously. Unlike simply selecting a single optimal dimension, MSCA—small bottleneck MLPs—conducts contrastive learning across multiple selected dimensional spaces: bringing the embeddings of anchor samples closer to same-class positive samples while pushing away distances from negative samples across various dimensions, and choosing the best dimension after learning the optimal combination of these multi-scale representations. This nested multi-scale feature learning strategy not only avoids the uncertainty of dimension selection but also simultaneously captures global consistency features and group-specific features, fundamentally enhancing the model's capability to address group shifts and thereby improving group robustness.

We acquire the original embedding vectors from the VLM's image encoder and construct nested multi-scale representations with the matryoshka embedding strategy. Each embedding undergoes adapter, contrastive learning in the selected multi-dimensional space, producing n new embeddings, thus ensuring that features at different scales can be compared in their feature space. To ensure that all features processed at different dimensions can interact and be compared in a unified semantic space, we introduce an up-projection layer. Specifically, for each low-dimensional feature (such as 512, 256,

128, and 64 dimensions), we apply a linear projection layer to map it back to the original input dimension (such as 1024 dimensions). This dimension restoration mechanism ensures that all feature vectors processed through different scale MSCAs can perform similarity calculations in the same space as the original class query vectors. During training, the logits of different dimensions of embedding and query dot product will have a weight, and this weight is trained by feature voter network training. During inference, feature voter evaluates the importance of features at different dimensions and selects embedding to dot product with query to get logits for classification. The overall architecture of our approach is shown in Fig. 1.

3.1 Multi-Scale Embedding Construction

We use a pretrained Vision-Language Model (VLM) such as CLIP to extract embeddings for both the training samples and class descriptions. The process begins with obtaining image embeddings by passing the images through the model's image encoder. For the class embeddings, we generate natural language prompts (e.g., "a picture of [class name]") and pass them through the text encoder of the VLM. This ensures that the zero-shot classification capability of the model is preserved.

We acquire the original embedding vectors from the VLM's image encoder and construct nested multi-scale representations. Although PCA is a commonly used method for dimensionality reduction, PCA's orthogonality properties may limit its reliability for dimensionality reduction for vectors with highly non-linear relationships, especially at high dimensionality [34]. Drawing inspiration from matryoshka embedding in our classification task [25], we utilize different leading dimensions of the same embedding to perform classification tasks with varying dimensional emphasis, achieving significant efficiency improvements while maintaining dimensional characteristics through the use of dimensionally decreasing nested embeddings. We can use only initial embedding for downstream tasks. Specifically, as shown in Fig. 1, we employ a nested list $[d_1, d_2, ..., d_n]$, where $d_0 > d_1 > d_2 > ... > d_n$ (d_0 is the original embedding dimension). We first use the embeddings from the VLM's image encoder as input, then truncate each to the leading d_i dimensions, obtaining n different embeddings. Each embedding passes through MSCA, undergoing contrastive learning across selected multiple dimensional spaces, resulting in n new embeddings, thus ensuring that features of different scales can be compared in their feature space.

3.2 Multi-Scale Contrastive Adapter

A crucial component of our method involves optimizing the model's embeddings through contrastive learning. We define three categories of samples for contrastive learning:

- **Anchor Samples:** Misclassified samples during zero-shot classification.
- **Positive Samples:** Correctly classified samples from the same class as the anchor samples.
- **Hard Negative Samples:** Samples from different classes that are closest to the anchor samples in the embedding space.

The contrastive sampling process is shown in Fig. 2.

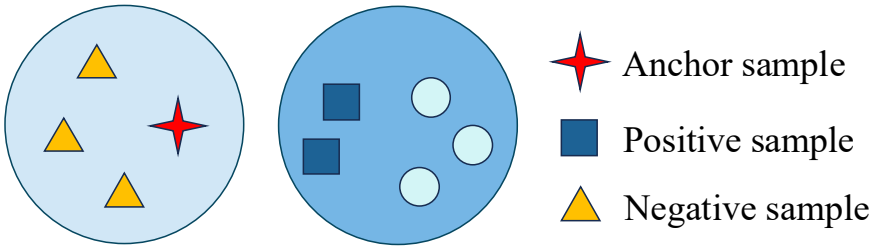

Fig. 2. The contrastive sampling.

The reason for this choice is to simultaneously bring the same type of misclassified samples closer to the correctly classified samples and push away the easily confused samples of different categories, thereby enhancing the model's ability to distinguish between groups. This contrastive learning strategy explicitly brings different groups within the same class (particularly samples that are easily confused in zero-shot classification) closer in the feature space while maintaining discriminability between samples of different classes, effectively addressing the group robustness issue.

To optimize sample representations simultaneously across multiple feature dimensional spaces, we propose a multi-scale contrastive loss. This loss function independently calculates contrastive loss at each dimension in the nested list, then uses hyperparameter weights to create a weighted combination of contrastive losses across all dimensional spaces, which is backpropagated during adapter training to achieve multi-scale collaborative optimization. That is to say, not only in the dimensional space of d_0, but also in the dimensional space of d_1, d_2, ..., d_n, the anchor sample is getting closer to the positive samples and farther away from the hard negative samples. As shown in Fig. 1, the original input embeddings of n different dimensions will get new embeddings of n original dimensions after adapter. This multi-scale collaborative optimization ensures that the model maintains good discrimination ability at different dimensional spaces. It effectively solves the limitation that a single-dimensional representation is difficult to adapt to all groups, and further enhances the model's ability to cope with group shifts. To formalize the multi-scale contrastive loss. This loss function is specifically designed for the MSCA to optimize the sample representations across multiple dimensional spaces. We define the following objective function:

$$\mathcal{L} = \sum_{i=1}^{n} \lambda_i \left[\frac{-1}{P} \sum_{p \in \mathcal{P}} \log \frac{\exp\left(f_\theta\left(u^{[d_i]}\right)^\top f_\theta\left(p^{[d_i]}\right)/\tau\right)}{\exp(f_\theta(u^{[d_i]})^\top f_\theta(p^{[d_i]})/\tau) + \sum_{m \in \mathcal{M}} \exp(f_\theta(u^{[d_i]})^\top f_\theta(m^{[d_i]})/\tau)} \right] \quad (1)$$

where $u^{[d_i]}$, $p^{[d_i]}$, $m^{[d_i]}$ are respectively representing the anchor sample, positive sample, and negative sample embeddings that are truncated to dimension d_i. For each class, we are identifying an "anchor" sample that is incorrectly predicted by zero-shot, $\mathcal{P}$ "positive" samples that are correctly classified by zero-shot, and $\mathcal{M}$ hard "negative" samples. λ_i is representing the hyperparameter weight for the i-th sample. n is the number of feature dimensions that is being considered.

3.3 Feature Voter

As shown in Fig. 1. after the embeddings of different dimensions pass through MSCA, all embeddings undergo a simple yet effective up-projection layer to restore them to the original embedding dimensionality, while preserving the feature and semantic information inherent to their original dimensions.

To dynamically balance the contributions of feature dimensions at different granularities, we introduce a feature voting mechanism from ensemble learning—a lightweight neural network that receives the original embedding vector and outputs a multi-scale feature weight w_i distribution. This voter independently calculates weight w_i for each embedding, which are used during the training phase to weight the classification logits between embeddings of various granularities and the query, and are optimized end-to-end through classification cross-entropy loss with respect to the labels.

We employ different strategies during training and inference phases: during training, we use a batch-based weighted combination of logits from all scale features to comprehensively learn the weights of representations at each scale in the classification task. During inference, we select the single-scale embedding with the highest weight in the process of training to dot product with query and get logits for classification. This adaptive weight allocation mechanism enables the model to select the most appropriate feature representation for samples from different groups, thereby enhancing group robustness.

The formula for calculating classification logits during training is:

$$\text{logits} = \sum_{i=1}^{n} w_i \text{sim}(z[i], \text{query}) \tag{2}$$

where $z[i]$ is the embeddings of various dimensions after passing through the up-projection layer, $\text{sim}(z[i], \text{query})$ is the temperature-scaled cosine similarity between embedding $z[i]$ and class embedding query, w_i is the weights trained by the feature voter.

4 Experiments

4.1 Datasets

We conducted a series of classification task evaluations for MSCA across multiple datasets. These include Waterbirds [12], which contains two classes of land birds and waterbirds, exhibiting significant group imbalance where minority groups (such as waterbirds on land) have extremely limited samples. We also used different configurations of CelebA [13], which contains over 200,000 facial images of celebrities in the wild, annotated with 40 binary attributes, where minority groups (such as "dark-haired females" and "blonde males") have notably fewer samples. Additionally, we utilized CIFAR-10.02, which combines CIFAR-10 and CIFAR-10.2 datasets that are similar but with slightly different visual styles [35], containing images across 10 categories, including airplanes, cars, etc., demonstrating distribution shifts within the same class

but from different data sources. So in CIFAR10.02, whether the sample is from CIFAR10 or CIFAR10.02 is the group label.

4.2 Empirical Results

We report the results of MSCA and compare them with related baseline methods: ERM Linear Probe (2022), ERM Adapter (2021), DFR (Subsample) (2022), DFR (Upsample) (2022), Contrastive Adapter (2022), FairerCLIP (2024).

Implementing Details. In this paper, the experiments of the baseline model are reproduced with their default parameters. To optimize model worst group accuracy, we configured the *learning rate* to 1e–3. The *multi-scale list* is set to [64,128,256,512]. The hyperparameter λ_i are set to 0.25, 0.2, 0.1, 0.2, 0.25, the *batch_size* is configured at 128.

4.2.1 The Impact of Dimensionality on the Group Robustness of CA

To comprehensively understand the impact of feature dimension selection on group robustness, we systematically evaluated the original Contrastive Adapter (CA) model's worst-group accuracy across various hidden layer dimensions. We conducted three tests on the Waterbirds dataset using different random seeds. When random seeds differ, significant variations occur in the random sampling of anchor samples, positive samples, and negative samples. Additionally, the model may prioritize learning different group feature patterns during representation learning, causing changes in sensitivity to different dimensional representations. We experimented with four distinct hidden layer dimension configurations: 512, 256, 128, and 64.

Table 1. Effects of different hidden layer dimensions on the robustness of CA

Method	Waterbirds-1st		Waterbirds-2nd		Waterbirds-3rd	
	WG	GAP	WG	GAP	WG	GAP
CA-512	82.9	7.2	**84.6**	**5.7**	83.4	7.4
CA-256	83.5	7.1	82.5	7.2	**84.3**	**6.7**
CA-128	84.1	6.3	84.2	6.2	82.5	7.9
CA-64	**84.3**	**5.5**	83.8	6.3	83.1	**6.7**

Table 1 demonstrates two phenomena: First, variations in hidden layer dimensions indeed affect the model's group robustness. For instance, in the first experiment (waterbirds-1st), as the hidden layer dimension decreases from 512 to 64, the worst-group accuracy (WG) improves from 82.9% to 84.3%, while the GAP decreases from 7.2% to 5.5%. Second, the optimal dimension setting differs across experimental runs, indicating that no single "best" dimension consistently performs optimally across all scenarios. These observations support our hypothesis that different groups may require feature representations of varying granularity, and adapters with a single dimension may not simultaneously satisfy the representational needs of all groups. This finding also provides empirical foundation for our proposed nested multi-scale feature learning strategy, suggesting that combining features from multiple scales may more effectively improve model group robustness than selecting a single "best" dimension.

4.2.2 Evaluation Metrics Comparison of MSCA and Baseline Methods

To assess the effectiveness of our MSCA in improving group robustness, we conducted experiments on multiple group shift benchmark datasets, including Waterbirds, CelebA, and CIFAR-10.02. As our method primarily focuses on effectively addressing classification accuracy disparities across different groups, we compared it with various existing group robustness methods. For comprehensive metrics evaluation, we employed two key metrics: 1) Worst Group Accuracy (WG), defined as the lowest accuracy among all groups, reflecting the model's performance on the most challenging groups; and 2) Gap, the difference between average accuracy and worst group accuracy, where a smaller gap indicates more balanced cross-group performance.

Table 2. The worst group accuracy (WG) and the GAP between the previous methods and MLCA are compared on two different CLIP variants, CLIP ResNet-50 and CLIP ViT-L/14, on the WaterBirds, CelebA and CIFAR10.02 datasets.

Method	Waterbirds		CelebA		CIFAR-10.02	
	WG	GAP	WG	GAP	WG	GAP
CLIP RN50						
Zero-shot(ZS)	36.9	55.2	75.3	6.2	39.8	29.7
ERM Linear Probe	14.5	79.4	14.2	80.5	52.1	25.8
ERM Adapter	63.1	32.2	42.8	51.8	<u>65.9</u>	20.2
DFR(Subsample)	66.7	26.2	81.3	10.4	45.5	29.7
DFR(Upsample)	54.4	35.8	**88.9**	**2.5**	38.7	39.4
Contrastive Adapter	<u>84.5</u>	<u>5.9</u>	87.4	<u>3.1</u>	60.8	<u>19.7</u>
FairerCLIP	75.4	8.9	81.5	3.5	N/A	N/A
MSCA(Ours)	**86.1**	**4.1**	<u>87.5</u>	3.5	**67.6**	**16.8**
CLIP ViT L/14						
Zero-shot(ZS)	26.5	61.1	62.8	9.3	71.8	21.4
ERM Linear Probe	65.6	31.9	30.9	63.2	87.0	<u>9.3</u>
ERM Adapter	76.6	21.3	40.5	53.5	<u>87.1</u>	9.8
DFR(Subsample)	59.5	36.2	79.1	12.2	85.6	11.1
DFR(Upsample)	66.8	29.5	<u>83.9</u>	7.4	72.4	21.4
Contrastive Adapter	85.3	9.2	83.8	6.9	82.2	13.9
FairerCLIP	<u>86.0</u>	<u>6.2</u>	**85.2**	**2.6**	N/A	N/A
MSCA(Ours)	**87.1**	**6.0**	83.5	<u>6.7</u>	**87.4**	**9.0**

Table 2 clearly demonstrates the significant advantages of our MSCA method in terms of group robustness. Our method, MSCA, achieved first place in 8 categories and second place in 2 categories. Compared to existing methods, MSCA brings substantial improvements across all tested datasets. On CIFAR-10.02, our method shows the most significant enhancement, increasing the worst-group accuracy to 87.4% with ViT-L/14 while reducing the GAP to 9.0%, the highest WG and lowest GAP among all methods. On the WaterBirds dataset, the improvements are equally effective, with MSCA raising the worst group accuracy from the zero-shot's 36.9% to 86.1% on CLIP RN50, while significantly reducing the GAP from 55.2% to just 4.1%. For CelebA, our method maintains competitive performance on worst-group accuracy comparable to the

strongest baselines. These results demonstrate MSCA's group robustness across different types of group shift scenarios.

4.2.3 Ablation Experiment of Multi-Scale Feature Combination list

Table 3. Impact of Different Multi-Scale Lists of MSCA on Group Robustness

Method	Waterbirds		CelebA		CIFAR-10.02	
	WG	GAP	WG	GAP	WG	GAP
CLIP RN50						
Contrastive Adapter	84.5	5.9	**87.7**	3.1	60.8	19.7
[64,128,256,512]	**86.1**	**4.1**	87.5	**2.5**	**67.6**	**16.8**
[64,512]	85.7	**4.1**	**87.7**	3.2	66.8	18.3
[64]	85.9	4.7	86.2	3.4	65.4	17.8
CLIP ViT L/14						
Contrastive Adapter	85.3	9.2	**83.8**	**6.6**	82.2	13.9
[64,128,256,512]	**87.1**	**6.0**	83.5	6.7	**87.4**	9.0
[64,512]	86.7	7.2	83.1	7.0	87.2	**8.9**
[64]	86.7	6.9	82.5	7.5	86.1	9.8

We choose the lists of [64], [64,512], [64,128,256,512] dimensions based on the strategy of gradually increasing the feature dimensions, starting from smaller dimensions and gradually expanding. Table 3 illustrates the impact of different multi-scale feature combinations on group robustness. The results indicate that the choice of feature combinations significantly affects WG, and GAP. For CLIP RN50, the combination [64,128,256,512] achieves the best worst-group accuracy on CIFAR-10.02 (67.6%) and competitive results on CelebA (87.5%), while showing the lowest GAP on Waterbirds (4.1%). For CLIP ViT L/14, the same combination [64,128,256,512] achieves the best worst-group accuracy on CIFAR-10.02 (87.4%) and high worst-group accuracy on Waterbirds (87.1%), with the lowest GAP on CIFAR-10.02 (9.0%) and one of the lowest on Waterbirds (6.0%). Notably, even simpler feature combinations like [64,512] perform well across datasets, sometimes matching the worst-group accuracy of more complex combinations. The baseline Contrastive Adapter shows competitive evaluation results on Waterbirds with both model architectures but is outperformed by multi-scale feature combinations on other datasets. Overall, these results confirm our hypothesis: appropriate selection of multi-scale feature combinations can significantly improve model robustness across different groups, while the optimal combination may vary depending on the specific task and model architecture.

4.2.4 The Impact of Feature Voting

Table 4. Impact of Feature Voter of MSCA on Group Robustness

Method	Waterbirds		CelebA		CIFAR-10.02	
	WG	GAP	WG	GAP	WG	GAP
CLIP RN50						

Contrastive Adapter	84.5	5.9	**87.7**	3.1	60.8	19.7
MSCA-input dim	84.2	6.6	87.1	3.4	62.4	19.0
MSCA-512	85.3	4.5	86.9	3.3	63.2	19.1
MSCA-128	86.0	**4.1**	86.8	**3.0**	65.6	**16.3**
MSCA-random	85.6	4.3	87.2	3.6	66.1	17.8
MSCA-voting	**86.1**	**4.1**	87.5	3.5	**67.6**	16.8
CLIP ViT L/14						
Contrastive Adapter	85.3	9.2	**83.8**	**6.6**	82.2	13.9
MSCA-input dim	85.9	8.0	83.5	7.4	87.0	9.7
MSCA-512	85.2	8.1	82.6	7.1	87.1	9.2
MSCA-128	86.7	7.4	83.7	7.8	86.8	9.3
MSCA-random	85.7	7.9	82.5	8.0	86.1	9.8
MSCA-voting	**87.1**	**6.0**	83.5	6.7	**87.4**	**9.0**

Table 4 demonstrates the impact of different feature selection strategies on the group robustness of the MSCA model. Results indicate that our proposed feature voting mechanism (MSCA-feature voting) has optimal metrics in most scenarios. On CLIP RN50, the feature voting mechanism achieves 86.1% worst-group accuracy and 4.1% GAP on the Waterbirds dataset, while improving worst-group accuracy to 67.6% on CIFAR-10.02, substantially outperforming methods using fixed input dimensions. On CLIP ViT-L/14, this mechanism also performs excellently, reaching 87.1% worst-group accuracy on Waterbirds. Notably, fixed dimension strategies can also perform well on specific datasets, such as MSCA-128 achieving 86.8% worst-group accuracy and 3.0% GAP on the CelebA dataset with CLIP RN50. This suggests varying sensitivity to feature dimensions across different datasets, with certain types of group shifts potentially better suited to feature representations at specific scales. However, considering overall evaluation results across datasets and model architectures, the feature voting mechanism provides the most group robustness. By learning to dynamically allocate weights to features at different scales, this mechanism better adapts to the characteristics of different datasets, confirming our hypothesis: Different groups require feature representations of varying granularity, and adaptively combining multi-scale features more effectively captures features of different groups, enhancing the model's overall group robustness.

5 Conclusion

In this paper, we introduced Multi-Scale Contrastive Adapter (MSCA), which significantly enhances the group robustness of vision-language models. By leveraging multi-scale feature representations, multi-scale contrastive learning strategy and a dynamic feature voting mechanism, MSCA effectively addresses group shifts, significantly improving worst-group accuracy without requiring full model retraining or group labels. Our experiments demonstrate that MSCA outperforms existing methods across multiple benchmark datasets, offering a robust approach to tackling group-specific challenges in zero-shot classification tasks. Importantly, multi-scale feature representations and a dynamic feature voting mechanism could also offer valuable insights for

improving other vision-language models, such as ALBEF and BLIP [3, 4], potentially extending the benefits of our approach to a broader range of VLMs without requiring full model retraining.

References

1. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning, PMLR, vol. 2021, pp. 8748–8763 (2021)
2. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: Proceedings of the International Conference on Machine Learning, PMLR, vol. 2021, pp. 5583–5594 (2021)
3. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hong Hoi, S.: Align before fuse: Vision and language representation learning with momentum distillation. In: Advances in Neural Information Processing Systems, vol. 34, pp. 9694–9705 (2021)
4. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of the International Conference on Machine Learning, PMLR, pp. 12888–12900 (2022)
5. Liu, H., Li, C., Wu, Q., Lee, Y. J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems, vol. 36, pp. 34892–34916 (2023)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)
7. Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European Conference on Computer Vision, vol. 2018, pp. 456–473 (2018)
8. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, PMLR, vol. 2018, pp. 77–91 (2018)
9. Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., et al.: Wilds: A benchmark of in-the-wild distribution shifts. In: Proceedings of the International Conference on Machine Learning, PMLR, vol. 2021, pp. 5637–5664 (2021)
10. Nam, J., Cha, H., Ahn, S., Lee, J., Shin, J.: Learning from failure: De-biasing classifier from biased classifier. In: Advances in Neural Information Processing Systems, vol. 33, pp. 20673–20684 (2020)
11. Sohoni, N., Dunnmon, J., Angus, G., Gu, A., Ré, C.: No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In: Advances in Neural Information Processing Systems, vol. 33, pp. 19339–19352 (2020)
12. Sagawa, S., Koh, P. W., Hashimoto, T. B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
13. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3730–3738 (2015)
14. Zhang, M., Ré, C.: Contrastive adapters for foundation model group robustness. In: Advances in Neural Information Processing Systems, vol. 35, pp. 21682–21697 (2022)
15. Rezaei, A., Liu, A., Memarrast, O., Ziebart, B. D.: Robust fairness under covariate shift. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 11, pp. 9419–9427 (2021)

16. Cui, Y., Jia, M., Lin, T.-Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9268–9277 (2019)
17. He, H., Garcia, E. A.: Learning from imbalanced data. In: IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284 (2009)
18. Idrissi, B. Y., Arjovsky, M., Pezeshki, M., Lopez-Paz, D.: Simple data balancing achieves competitive worst-group-accuracy. In: Proceedings of the Conference on Causal Learning and Reasoning, PMLR, pp. 336–351 (2022)
19. Byrd, J., Lipton, Z.: What is the effect of importance weighting in deep learning?. In: Proceedings of the International Conference on Machine Learning, PMLR, pp. 872–881 (2019)
20. Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. In: Journal of Statistical Planning and Inference, vol. 90, no. 2, pp. 227–244 (2000)
21. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
22. Sagawa, S., Koh, P. W., Hashimoto, T. B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
23. Addepalli, S., Asokan, A. R., Sharma, L., Babu, R. V.: Leveraging vision-language models for improving domain generalization in image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23922–23932 (2024)
24. Varma, M., Delbrouck, J.-B., Chen, Z., Chaudhari, A., Langlotz, C.: Ravl: Discovering and mitigating spurious correlations in fine-tuned vision-language models. In: Advances in Neural Information Processing Systems, vol. 37, pp. 82235–82264 (2024)
25. Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., et al.: Matryoshka representation learning. In: Advances in Neural Information Processing Systems, vol. 35, pp. 30233–30249 (2022)
26. Kumar, A., Raghunathan, A., Jones, R., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. arXiv preprint arXiv:2202.10054 (2022)
27. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. In: International Journal of Computer Vision, vol. 132, no. 2, pp. 581–595 (2024)
28. Kirichenko, P., Izmailov, P., Wilson, A. G.: Last layer re-training is sufficient for robustness to spurious correlations. arXiv preprint arXiv:2204.02937 (2022)
29. Berg, H., Hall, S. M., Bhalgat, Y., Yang, W., Kirk, H. R., Shtedritski, A., Bain, M.: A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. arXiv preprint arXiv:2203.11933 (2022)
30. Seth, A., Hemani, M., Agarwal, C.: Dear: Debiasing vision-language models with additive residuals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6820–6829 (2023)
31. Chuang, C.-Y., Jampani, V., Li, Y., Torralba, A., Jegelka, S.: Debiasing vision-language models via biased prompts. arXiv preprint arXiv:2302.00070 (2023)
32. Chen, S., Gu, J., Han, Z., Ma, Y., Torr, P., Tresp, V.: Benchmarking robustness of adaptation methods on pre-trained vision-language models. In: Advances in Neural Information Processing Systems, vol. 36, pp. 51758–51777 (2023)
33. Dehdashtian, S., Wang, L., Boddeti, V. N.: Fairerclip: Debiasing clip's zero-shot predictions using functions in rkhss. arXiv preprint arXiv:2403.15593 (2024)

34. Jolliffe, I. T., Cadima, J.: Principal component analysis: a review and recent developments. In: Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065, pp. 20150202 (2016)
35. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. In: Technical Report, University of Toronto, pp. 7 (2009)