BEYOND ID BIAS: PCA-GUIDED DROPOUT FOR ROBUST FINE-TUNING

Bo Fei[1]*, **Xiaocheng Li**[2], **Zhiqi Zhang**[3], **Youchen Qing**[4], **Yancong Deng**[5]
[1]Northeastern University, Shenyang, China
[2]University of California, Irvine, USA
[3]University of Toronto, Canada
[4]Shanghai Jiao Tong University, China
[5]University of California, San Diego, USA

ABSTRACT

Fine-tuning large-scale pre-trained models often improves in-distribution (ID) performance at the cost of out-of-distribution (OOD) generalization due to overfitting to ID-specific features. To mitigate this, we propose **PCA Dropout**, a novel fine-tuning strategy that suppresses ID-specific feature dependencies by leveraging Principal Component Analysis (PCA). Our method identifies dominant feature components that contribute the most to ID variance and applies structured dropout to reduce their influence, encouraging the model to learn more generalizable representations. We evaluate PCA Dropout on DomainNet and iWildCam using CLIP-based models, demonstrating consistent improvements in OOD robustness over state-of-the-art fine-tuning methods while maintaining strong ID accuracy. Ablation studies further confirm that structured dropout at the feature level outperforms unstructured feature suppression and random dropout strategies.

1 INTRODUCTION

Fine-tuning large-scale pre-trained models has become the standard approach for adapting foundation models (FMs) to specific downstream tasks. However, a persistent challenge is balancing in-distribution (ID) performance with out-of-distribution (OOD) generalization. While fine-tuning improves accuracy on ID data, it often results in overfitting, making the model overly reliant on dataset-specific patterns that do not transfer well to unseen domains. This issue is particularly evident in vision-language models (VLMs) such as CLIP Radford et al. (2021), where fine-tuning can alter pre-trained representations and reduce robustness Kumar et al. (2022). To address this, there is a need for fine-tuning strategies that retain high ID accuracy while mitigating reliance on spurious correlations, ultimately improving OOD generalization.

Recent approaches to robust fine-tuning have focused on regularization techniques in both parameter space and representation space. Parameter-based methods such as L2-SP Xuhong et al. (2018) and FTP Tian et al. (2024) constrain parameter deviations from pre-trained weights, preventing excessive adaptation to ID-specific patterns. Representation-based methods, such as WiSE-FT Wortsman et al. (2022), improve generalization by interpolating between the zero-shot model and the fine-tuned model. Inspired by these efforts, we propose an alternative approach that directly suppresses ID-specific feature dependencies to encourage learning more transferable representations.

In this work, we introduce **PCA Dropout**, a novel fine-tuning strategy designed to improve OOD generalization by suppressing features that predominantly capture ID-specific variations. Our method is motivated by the observation that dominant principal components in feature representations are often over-aligned with the ID distribution, leading to poor generalization on OOD data. To address this, we first apply Principal Component Analysis (PCA) to quantify the contribution of each feature to the ID-specific variance. We then selectively suppress high-contribution features via a structured dropout mechanism, reducing their influence during fine-tuning. This encourages the model to rely on alternative, more generalizable feature representations.

*Corresponding author: Fei Bo 20216764@stu.neu.edu.cn

We evaluate PCA Dropout on two benchmark datasets, DomainNet Peng et al. (2019) and iWildCam Koh et al. (2021), using CLIP-based models. Our results demonstrate that PCA Dropout improves OOD generalization without sacrificing ID accuracy. On DomainNet, our method achieves an OOD accuracy of **41.40%**, surpassing prior state-of-the-art approaches such as FTP Tian et al. (2024). Similarly, on iWildCam, PCA Dropout attains the highest OOD macro F1-score of **36.8%**, outperforming competitive baselines.

Our key contributions are as follows:

- We propose **PCA Dropout**, a feature-suppression method that mitigates overfitting by selectively removing ID-specific features identified through PCA analysis.
- We introduce a novel feature importance scoring mechanism that quantifies ID-specific reliance, enabling dynamic dropout mask generation during training.
- We conduct extensive experiments on DomainNet and iWildCam, demonstrating that PCA Dropout improves OOD robustness while maintaining ID accuracy, surpassing prior state-of-the-art fine-tuning strategies.

2 METHOD

2.1 PROBLEM DEFINITION

Fine-tuning pre-trained models for downstream tasks involves adapting a model f (e.g., CLIP Radford et al. (2021)) using training data sampled from an in-distribution P_{ID}. The objective is to learn a function $f : \mathbb{R}^d \to Y$ that maps input features $x \in \mathbb{R}^d$ to corresponding labels $y \in Y$, optimizing the supervised loss:

$$L_{\text{sup}}(f, P_{\text{ID}}) = \mathbb{E}_{(x,y)\sim P_{\text{ID}}} \left[\ell(f(x), y) \right]. \tag{1}$$

However, direct fine-tuning on P_{ID} often leads to overfitting, causing the model to rely on features that do not generalize well to out-of-distribution (OOD) data P_{OOD}. This results in a significant performance drop under distribution shifts. The goal is to maintain high accuracy on P_{ID} while enhancing robustness to unseen distributions.

2.2 IDENTIFYING ID-SPECIFIC FEATURES VIA PCA-GUIDED SCORING

Neural networks often rely on spurious correlations present in the in-distribution (ID) data, leading to overfitting and degraded generalization to out-of-distribution (OOD) samples. To characterize ID-specific variations, we leverage Principal Component Analysis (PCA) to quantify the contribution of each feature to the dominant patterns in the data. The key intuition is that features highly aligned with principal components that explain the majority of variance in the data are more likely to be specific to the ID distribution.

Given an input feature matrix $\mathbf{X} \in \mathbb{R}^{N\times D}$, where N represents the number of samples and D is the feature dimension, we first perform **feature standardization** by subtracting the mean of each feature to ensure a zero-mean distribution:

$$\mathbf{X}_{\text{center}} = \mathbf{X} - \bar{\mathbf{X}}, \quad \text{where } \bar{\mathbf{X}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i. \tag{2}$$

Next, we apply Principal Component Analysis (PCA) using Singular Value Decomposition (SVD) to extract the principal directions:

$$\mathbf{X}_{\text{center}} = \mathbf{U}\mathbf{S}\mathbf{V}^\top. \tag{3}$$

Here, $\mathbf{U} \in \mathbb{R}^{N\times D}$ contains the left singular vectors, $\mathbf{S} \in \mathbb{R}^{D\times D}$ is a diagonal matrix of singular values that represent the variance captured by each principal component, and $\mathbf{V} \in \mathbb{R}^{D\times D}$ contains the right singular vectors, where each column corresponds to a principal component.

To quantify the ID-specific contribution of each feature, we compute a **feature importance score** based on its alignment with the top R principal components:

$$s_i = \sum_{j=1}^{R} V_{ij}^2, \quad i = 1, \ldots, D. \tag{4}$$

The number of retained principal components R is determined adaptively based on the explained variance. Specifically, we select the smallest R such that the cumulative variance explained satisfies:

$$\frac{\sum_{j=1}^{R} S_{jj}^2}{\sum_{j=1}^{D} S_{jj}^2} \geq 1 - \epsilon_{\text{th}}, \tag{5}$$

where ϵ_{th} is a threshold controlling the proportion of discarded variance. In our experiments, we set $\epsilon_{\text{th}} = 10^{-3}$, ensuring that components capturing less than 0.1% of the variance are discarded. This selection criterion prevents insignificant principal components from affecting feature importance estimation.

By ranking features based on their importance scores, we can analyze their relative contributions to ID-specific variations and identify patterns that predominantly characterize the training distribution.

2.2.1 Feature-Specific Dropout Based on Principal Component Contributions

To mitigate over-reliance on ID-specific features, we introduce a dropout mechanism that selectively removes highly influential features based on their contribution scores. Given the computed feature contribution scores s_i, we construct a binary dropout mask $\mathbf{M} \in \{0, 1\}^D$ to suppress features with the highest scores.

Let p be the dropout rate, which determines the proportion of features to be masked out. We first rank all features in descending order based on s_i and select the top $p \times D$ features with the highest contribution scores for suppression. The dropout mask is then defined as:

$$M_i = \begin{cases} 0, & \text{if feature } i \text{ is among the top } p \times D \text{ ranked features,} \\ 1, & \text{otherwise.} \end{cases} \tag{6}$$

During training, the mask is applied element-wise to the input feature matrix:

$$\mathbf{X}_{\text{dropout}} = \mathbf{X} \odot \mathbf{M}, \tag{7}$$

where $\odot$ denotes element-wise multiplication. This operation ensures that the most ID-specific features, as determined by their alignment with dominant principal components, are effectively removed from the learning process. By enforcing this structured dropout, the model is encouraged to rely on alternative, more generalizable features, thereby improving robustness to distributional shifts.

2.3 Training Procedure with Feature-Specific Dropout

To enhance model robustness and mitigate reliance on ID-specific features, we integrate feature-specific dropout within the training pipeline. This dropout mechanism is dynamically updated at each epoch based on the principal component contributions of feature activations. The overall training procedure consists of four key stages: First, during each forward pass, feature activations are computed and stored in a feature bank. At the end of each epoch, Principal Component Analysis (PCA) is applied to these stored activations to identify the dominant feature components. Using the computed feature contribution scores, a dropout mask is generated to suppress the most influential features, ensuring that the model does not overly depend on ID-specific patterns. This mask is then applied in subsequent training iterations. Finally, standard gradient-based optimization is performed using the masked feature activations. The complete training process is formally described in Algorithm 1.

Algorithm 1 Training Procedure with Feature-Specific Dropout

Require: Pre-trained model f, training data P_{ID}, dropout rate p, number of epochs T

1: **for** epoch $t = 1$ to T **do**
2: **Forward Propagation:** Compute feature activations for each layer and store them in the feature bank.
3: **PCA Computation:** Perform PCA on stored feature activations.
4: Compute feature contribution scores $s_i = \sum_{j=1}^{R} V_{ij}^2$.
5: **Dropout Mask Generation:** Identify the top $p \times D$ features with the highest scores and mask them.
6: Apply the dropout mask to feature activations: $\mathbf{X}_{\text{dropout}} = \mathbf{X} \odot \mathbf{M}$.
7: **Backward Propagation:** Update model parameters via gradient descent.

3 EXPERIMENTS

3.1 DATASETS

To evaluate the effectiveness of our approach in both in-distribution (ID) and out-of-distribution (OOD) settings, we adopt widely used benchmarks from prior studies Goyal et al. (2022); Tian et al. (2024). These datasets provide diverse domain shifts, enabling a comprehensive assessment of model generalization.

DomainNet Peng et al. (2019) is a large-scale dataset designed for domain adaptation research. It consists of six domains: Clipart, Infograph, Painting, Quickdraw, Real, and Sketch, each exhibiting distinct visual styles. This dataset is particularly suitable for evaluating a model's ability to generalize across different visual distributions.

iWildCam (Koh et al., 2021) is a classification dataset featuring 182 animal species captured in varying environmental conditions. The ID and OOD splits are determined based on differences in camera sources and background attributes such as illumination and habitat variability.

3.2 IMPLEMENTATION DETAILS

Baselines. We compare our approach against state-of-the-art fine-tuning methods, including LP-FT Kumar et al. (2022), MARS-SP Gouk et al. (2021), WiSE-FT Wortsman et al. (2022), FLPY Goyal et al. (2022), FTP Tian et al. (2024), TPGM Tian et al. (2023), and L2-SP Xuhong et al. (2018). These baselines represent a diverse set of strategies for enhancing model robustness during fine-tuning.

Models and Optimization. Our experiments are conducted on two architectures: CLIP ResNet-50 for DomainNet and CLIP ViT-B/16 for iWildCam. For CLIP ResNet-50, training is performed using Stochastic Gradient Descent (SGD) with a learning rate of 0.01, momentum of 0.9, and a batch size of 32. For CLIP ViT-B/16, we use AdamW as the optimizer along with a cosine learning rate scheduler, setting the learning rate to 1×10^{-5}, weight decay to 0.2, and batch size to 64.

3.3 RESULTS

DomainNet. On the DomainNet dataset, the proposed PCA Dropout method demonstrates strong performance in both in-distribution (ID) and out-of-distribution (OOD) settings, as shown in Table 1. PCA Dropout achieves an ID accuracy of **84.48%**, surpassing prior methods such as FTP (84.22%) and L2-SP (82.07%). In terms of OOD generalization, PCA Dropout consistently outperforms baseline approaches, particularly in challenging domains such as Clipart (**48.94%**) and Painting (**47.17%**), leading to an overall OOD average of **41.40%**. This represents a notable improvement over FTP (39.94%) and TPGM (39.65%), demonstrating PCA Dropout's ability to mitigate overfitting to ID-specific features while improving robustness across diverse visual distributions.

iWildCam. On the iWildCam dataset, PCA Dropout also achieves competitive results. The method attains an ID macro F1-score of **49.5%**, performing on par with leading methods such as FLPY (51.4%) and LP-FT (49.7%). More importantly, PCA Dropout exhibits superior OOD robustness, achieving the highest OOD macro F1-score at **36.8%**, outperforming baselines like FTP (35.8%)

Table 1: DomainNet Results using CLIP pre-trained ResNet50. Note that the results of baselines are adopted from FTP Tian et al. (2024).

Methods	ID	OOD				Statistics
	Real	Sketch	Painting	Infograph	Clipart	OOD Avg.
Vanilla FT	80.93	31.81	41.02	20.29	43.59	34.18
Linear Prob.	52.56	20.05	24.92	19.18	21.15	21.33
L2-SP Xuhong et al. (2018)	82.07	36.67	45.62	22.97	47.78	38.26
MARS-SP Gouk et al. (2021)	77.19	25.33	33.43	14.81	39.20	28.19
LP-FT Kumar et al. (2022)	80.82	34.85	44.03	22.23	46.13	36.81
TPGM Tian et al. (2023)	83.64	38.78	43.11	28.70	48.01	39.65
FTP Tian et al. (2024)	84.22	37.66	46.11	28.33	47.67	39.94
PCA Dropout	**84.48**	**39.63**	**47.17**	**29.85**	**48.94**	**41.40**

Table 2: iWildCam Results using CLIP pre-trained ViT B/16. Following common practice, we report macro F1-score. Note that the results of baselines are adopted from FLPY Goyal et al. (2022). †indicates our reproduced result.

Methods	iWildCam	
	ID	OOD
Zeroshot	8.7	11.0
Linear Prob.	44.5	31.1
Vanilla FT	48.1	35.0
L2-SP Xuhong et al. (2018)	48.6	35.3
LP-FT Kumar et al. (2022)	49.7	34.7
WiSE-FT Wortsman et al. (2022)	48.1	35.0
FTP Tian et al. (2024) †	47.3	35.8
FLPY Goyal et al. (2022) †	**51.4**	35.2
PCA Dropout	49.5	**36.8**

and WiSE-FT (35.0%). These results highlight PCA Dropout's effectiveness in handling real-world distribution shifts, particularly those caused by variations in environmental conditions and camera perspectives, reinforcing its potential for robust fine-tuning in natural image classification tasks.

3.4 Ablation Study

To better understand the effect of PCA Dropout, we conduct a series of ablation experiments on the iWildCam dataset using CLIP ViT-B/16. Specifically, we analyze the impact of dropout on different architectural components, the choice of layers for dropout, the effect of varying dropout rates, and the influence of randomly selecting layers for dropout. The results are summarized in Table 3.

Effect of Dropout on Attention and MLP Layers. We first examine how dropout applied to the *Attention module*, *MLP module*, or both affects performance when applied to layer 9 with a dropout rate of 0.3. Table 3 shows that dropping the MLP or Attention module alone yields similar results, with MLP achieving a slightly higher ID F1-score (47.6% vs. 46.9%). Further, applying dropout to both improves ID performance to 49.3% and OOD generalization to 36.0%, demonstrating a combined benefit.

Effect of Dropout on Different Layers. Next, we analyze the impact of *dropout at different layers* while keeping the dropout rate at 0.3 and applying it to both Attention and MLP modules. When applied to *layer 3*, the model achieves *ID: 48.4%* and *OOD: 35.1%*, whereas dropout at *layer 6* yields *ID: 47.4%*, but the highest *OOD performance of 36.5%*. Dropout at *layer 9* leads to *ID: 49.3%* and *OOD: 36.0%*, while applying dropout at *layer 12* results in the highest ID score of *50.1%* but the lowest OOD generalization at *34.2%*. These results indicate that *dropout at intermediate layers (e.g.,*

Table 3: Abation Experiment in iWildCam using CLIP pre-trained ViT B/16. We report macro F1-score.

Training Configuration					iWildCam	
Attention	MLP	Dropout Rate	Layer	Random	ID (%)	OOD (%)
✓		0.3	9		46.9	34.5
	✓	0.3	9		47.6	34.4
✓	✓	0.3	3		48.4	35.1
✓	✓	0.3	6		47.4	36.5
✓	✓	0.3	9		49.3	36.0
✓	✓	0.3	12		**50.1**	34.2
✓	✓	0.1	9		49.5	**36.8**
✓	✓	0.1	6,9		48.5	35.6
✓	✓	0.1	3,6,9		48.7	34.3
✓	✓	0.1	3,6,9,12		45.2	34.1
✓	✓	0.1	6,9	✓	49.9	36.0
✓	✓	0.1	3,6,9	✓	49.7	35.8

layer 6 and 9) enhances OOD robustness, whereas dropout at deeper layers primarily benefits ID accuracy.

Effect of Dropout Rate. To understand the influence of the *dropout rate*, we compare results when applying dropout at *layer 9* with rates of *0.1* and *0.3*. At a *dropout rate of 0.3*, the model attains *ID: 49.3%* and *OOD: 36.0%*, whereas lowering the dropout rate to *0.1* yields *ID: 49.5%* and *OOD: 36.8%*. The results suggest that *reducing the dropout rate slightly improves OOD generalization while maintaining similar ID performance*.

Effect of Dropout Across Multiple Layers. We further investigate whether *dropping multiple layers simultaneously* improves generalization by applying dropout at *layers 3, 6, 9, and 12* with a rate of *0.1*. Compared to dropout at only layer 9 (*ID: 49.5%, OOD: 36.8%*), applying dropout to *all four layers degrades ID accuracy to 45.2% and OOD performance to 34.1%*. Similarly, applying dropout to *layers 3, 6, and 9* leads to *ID: 48.7% and OOD: 34.3%*, and at *layers 6 and 9*, the model achieves *ID: 48.5% and OOD: 35.6%*. These results indicate that *aggressive dropout across multiple layers negatively impacts both ID and OOD performance*, likely due to excessive feature suppression.

Effect of Randomly Selecting Dropout Layers. Finally, we evaluate whether *randomly selecting a layer for dropout* instead of applying dropout to predefined layers improves generalization. With dropout at *layers 3, 6, and 9*, selecting a layer randomly achieves *ID: 49.7% and OOD: 35.8%*, which is *slightly better than manually selecting these layers (ID: 48.7%, OOD: 34.3%)*. Similarly, for dropout at *layers 6 and 9*, random selection results in *ID: 49.9% and OOD: 36.0%*, compared to *ID: 48.5% and OOD: 35.6%* for fixed layer selection. These results suggest that *introducing randomness in layer selection can improve generalization by preventing the model from adapting too strongly to specific dropout patterns*.

4 Conclusion

We propose **PCA Dropout**, a fine-tuning strategy that suppresses ID-specific features to enhance out-of-distribution (OOD) generalization while maintaining strong in-distribution (ID) performance. By leveraging Principal Component Analysis (PCA) to identify dominant ID-specific features, our method applies structured dropout to encourage reliance on more transferable representations. Unlike parameter-based regularization, PCA Dropout operates at the feature level, offering a complementary approach to improving robustness. Experiments on DomainNet and iWildCam show that PCA Dropout consistently outperforms state-of-the-art fine-tuning methods in OOD generalization. Suppressing dominant ID-specific features effectively mitigates overfitting, leading to stronger generalization.

Limitations and Future Work While PCA Dropout improves OOD robustness, its effectiveness should be validated on larger-scale datasets and across different modalities beyond vision. Our current study focuses on CLIP ViT-B/16, and further investigation is needed to assess its adaptability to other model architectures. Additionally, the choice of hyperparameters, such as which layer to dropout, is model-dependent. Future work will explore more adaptive and hyperparameter-light strategies to enhance the general applicability of PCA Dropout across diverse architectures and tasks.

REFERENCES

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

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016.

Henry Gouk, Timothy M Hospedales, and Massimiliano Pontil. Distance-based regularisation of deep networks for fine-tuning. *ICLR*, 2021.

Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. *arXiv preprint arXiv:2212.00638*, 2022.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Ananya Kumar et al. Fine-tuning can distort pretrained features and underperform out-of-distribution. *ICLR*, 2022.

Kai Li, Chang Liu, Handong Zhao, Yulun Zhang, and Yun Fu. Ecacl: A holistic framework for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8578–8587, 2021.

Chang Liu, Lichen Wang, Kai Li, and Yun Fu. Domain generalization via feature variation decorrelation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1683–1691, 2021.

Chang Liu, Xiang Yu, Yi-Hsuan Tsai, Masoud Faraki, Ramin Moslemi, Manmohan Chandraker, and Yun Fu. Learning to learn across diverse data biases in deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4072–4082, 2022.

Chang Liu, Gaurav Mittal, Nikolaos Karianakis, Victor Fragoso, Ye Yu, Yun Fu, and Mei Chen. Hyperstar: Task-aware hyperparameter recommendation for training and compression. *International Journal of Computer Vision*, pp. 1–15, 2023. doi: 10.1007/s11263-023-01961-0. URL `https://doi.org/10.1007/s11263-023-01961-0`.

Gaurav Mittal, Chang Liu, Nikolaos Karianakis, Victor Fragoso, Mei Chen, and Yun Fu. Hyperstar: Task-aware hyperparameters for deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8736–8745, 2020.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019.

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 *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. *ECCV*, 2016.

Junjiao Tian, Xiaoliang Dai, Chih-Yao Ma, Zecheng He, Yen-Cheng Liu, and Zsolt Kira. Trainable projected gradient method for robust fine-tuning. *arXiv preprint arXiv:2303.10720*, 2023.

Junjiao Tian, Yen-Cheng Liu, James S Smith, and Zsolt Kira. Fast trainable projection for robust fine-tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

Vladimir Vapnik. Statistical learning theory wiley. *New York*, 1998.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971, 2022.

Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. *arXiv*, 2019.

LI Xuhong, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning*, pp. 2825–2834. PMLR, 2018.

A Appendix

B Related Works

B.1 Robust Fine-tuning

Ensuring both in-distribution (ID) accuracy and out-of-distribution (OOD) robustness when fine-tuning large-scale models remains a key challenge, as excessive adaptation to ID data often leads to overfitting. A prominent strategy to mitigate this issue is ensembling, as demonstrated by WiSE-FT Wortsman et al. (2022), which blends predictions from the original zero-shot model with those from a fine-tuned model. This approach balances generalization and task-specific adaptation by preserving pre-trained knowledge while allowing controlled fine-tuning adjustments.

Another line of work constrains the degree of fine-tuning to maintain the model's pre-trained generalization properties. For instance, FLYP Goyal et al. (2022) encourages a training process that mimics pre-training dynamics to avoid excessive specialization. Similarly, LP-FT Kumar et al. (2022) restricts updates to the classifier layer, ensuring that the feature extractor remains unchanged to preserve the model's transferability.

Beyond structural constraints, explicit regularization has been employed to limit divergence from pre-trained parameters. L2-SP Xuhong et al. (2018) applies an L2 penalty to prevent drastic shifts in weight space, while approaches like FTP Tian et al. (2024) and TPGM Tian et al. (2023) impose gradient projections to control updates within constrained subspaces. These techniques regulate model adaptation to improve generalization without compromising ID performance. Our method builds upon this paradigm by introducing a feature-level regularization mechanism that selectively suppresses ID-specific components, mitigating over-reliance on spurious correlations and enhancing robustness under distributional shifts.

B.2 Domain Generalization

Domain generalization (DG) aims to train models that can generalize to unseen distributions without access to target domain data. Traditional Empirical Risk Minimization (ERM) (Vapnik, 1998)

minimizes the expected loss across training domains but does not explicitly encourage robustness to domain shifts. To improve generalization, various methods have been proposed, including invariant representation learning (Arjovsky et al., 2019; Liu et al., 2021), adversarial perturbations (Xu et al., 2019; Li et al., 2021), domain adaptation via adversarial training (Ganin et al., 2016), feature distribution alignment (Sun & Saenko, 2016), and meta-learning (Liu et al., 2022; 2023; Mittal et al., 2020). Additionally, SWAD (Cha et al., 2021) demonstrated that flatter loss landscapes lead to improved domain generalization, ensuring more stable performance across shifts.

Our approach connects to DG by focusing on suppressing ID-specific feature dependencies during fine-tuning, a complementary perspective to domain-invariant learning. By introducing a selective dropout mechanism based on principal component analysis, we reduce reliance on features that are overly adapted to the training distribution, improving OOD generalization while preserving ID performance.