

Discrepancy-Guided Parameter Suppression for Robust Fine-Tuning of Foundation Models

Zile Wang

College of Literature and Journalism, Chengdu University, Chengdu, China
wangzile@cdu.edu.cn

Abstract. Fine-tuning foundation vision-language models like CLIP for downstream tasks often creates a trade-off. While in-distribution classification accuracy improves, out-of-distribution generalization tends to decline. Existing methods try to alleviate this issue through representation-space regularization or parameter-space constraints, but they fail to precisely identify which parameter changes cause the out-of-distribution performance drop. This paper proposes a discrepancy-guided parameter suppression method. It calculates the normalized discrepancy between pre-trained and fine-tuned weights to identify in-distribution-specific parameters. High-discrepancy parameters are frozen using channel-wise binary masking, while only low-discrepancy task-invariant parameters are fine-tuned. This approach significantly improves out-of-distribution generalization without sacrificing in-distribution performance. On the iWildCam benchmark, our method achieves an out-of-distribution Macro-F1 of 36. Since it attains 9 percent accuracy on the base domain and has an average out-of-distribution accuracy of 41.63 percent on DomainNet, it therefore establishes new state-of-the-art results on both occasions.

Keywords: CLIP fine-tuning, out-of-distribution generalization, parameter suppression, robust optimization, foundation models

1. Introduction

1.1. Research background and motivation

Foundation vision-language models such as CLIP learn highly generalizable visual representations via large-scale contrastive image-text pre-training, hence they achieve state-of-the-art zero-shot classification performance. But a well-known and important trade-off arises when such models are fully fine-tuned on downstream tasks: classification accuracy improves substantially, while OOD generalization deteriorates markedly. Therefore, this trade-off seriously limits the use of foundation models in real-world applications like wildlife monitoring and cross-domain image classification, where distribution shift is the norm. "Foundation vision-language models such as CLIP [1] learn highly generalizable visual representations through large - scale contrastive image-text pre-training..."

The trade-off in question has naturally led to a well-defined, elegant line of research that can very cleanly be divided into three categories: Representation-space regularization methods such as FLYP

and LP-FT regularize the fine-tuned representation to stay close to the pre-trained distribution, hence being effective, but they also come with a notable computational overhead due to the added loss terms. Parameter-space regularization methods such as L2-SP and WiSE-FT prevent overfitting by penalizing deviations from pre-trained weights, but they apply uniform constraints to all parameters and thus do not properly distinguish between ID-specific parameters that hurt OOD generalization and task-invariant parameters that help it. Selective fine-tuning methods such as partial layer freezing and low-rank adaptation reduce the number of trainable parameters, but their current formulations rely on empirical layer selection rather than establishing a direct, quantitative link between parameter changes and OOD performance.

Since the fundamental shortcoming of all existing methods is the lack of a good way to identify and suppress the subset of parameters chiefly responsible for ID overfitting, it follows naturally that uniform constraints or heuristic freezing strategies cannot maximize OOD generalization without harming ID performance.

1.2. Contributions

Because of the existing limitation, "in this paper we naturally lead to the proposal of a discrepancy-guided parameter suppression method and state its main contributions below [2]." In this paper, we for the first time rigorously and directly link the magnitude of parameter changes during fine-tuning to out-of-distribution generalization degradation, thereby naturally leading to the computation of a normalized parameter discrepancy score, which allows us to precisely identify the channels whose weights shift the most during in-distribution fine-tuning and are thus most likely to encode spurious, in-distribution-specific features. Based on this, we design a binary masking mechanism that freezes high-discrepancy parameters at the channel granularity while letting low-discrepancy task-invariant parameters update freely. As a result, the selective update strategy breaks the conventional in-distribution to out-of-distribution trade-off by suppressing the parameters that cause overfitting, instead of applying uniform constraints to all weights. On iWildCam, our method attains an out-of-distribution Macro-F1 of 36. The method achieves 9 percent performance, outperforming all compared baselines, and on DomainNet it obtains an average out-of-distribution accuracy of 41.63 percent while also raising the in-distribution accuracy to 84.53 percent, thus providing very clear and strong evidence of its efficacy for both transformer and convolutional backbones. More importantly, the authors perform thorough, well-structured ablation experiments to rigorously validate each design choice: the necessity of normalized discrepancy over abnormalized scores, the benefit of channel-wise over layer-wise freezing, and the effect of the freezing ratio. The conclusions therefore offer direct, practical guidance for applying the method to new tasks and architectures.

2. Related work

2.1. CLIP fine-tuning and OOD generalization

CLIP fine-tuning strategies span a spectrum from full parameter updates to complete encoder freezing. Full fine-tuning optimizes all model parameters and achieves the highest in-distribution accuracy, yet it also produces the largest drop in out-of-distribution performance. Linear probing freezes the CLIP encoder and trains only the classification head, yielding optimal out-of-distribution generalization but insufficient in-distribution performance. Regularized fine-tuning methods aim to balance these two by introducing auxiliary constraints. FLYP incorporates a contrastive loss that

mimics the pre-training objective during fine-tuning, encouraging the representation space to stay aligned with the pre-trained distribution. Although this improves out-of-distribution performance to some degree, the added contrastive loss increases training cost, and the constraint on the parameter space remains indirect. Other methods such as LP-FT first train a linear classifier and then perform limited full fine-tuning, yet they still lack a principled mechanism to identify which parameters should and should not be updated.

2.2. Selective fine-tuning and parameter regularization

Selective fine-tuning methods improve generalization by reducing the number of trainable parameters, and the most common strategies are freezing early convolutional layers, fine-tuning only attention blocks, or using low-rank adaptation. While these approaches have been empirically successful in reducing overfitting, a critical point to note is that they all rest on architectural assumptions rather than on a quantitative analysis of which parameters are most important for out-of-distribution robustness. Parameter regularization methods such as L2-SP penalize the L2 distance between fine-tuned and pre-trained weights, but they do so under an implicit and potentially incorrect assumption: that every weight deviation is equally damaging to out-of-distribution generalization. In fact, some parameter changes are essential for task adaptation, while others capture spurious correlations that hurt out-of-distribution performance. Therefore, the real limitation of uniform regularization is its failure to distinguish between these two fundamentally different kinds of parameter changes.

2.3. Weight ensembling methods

"WiSE-FT [3] recovers OOD generalization by linearly interpolating between pre-trained and fine-tuned weights." Because ensembling is performed after training, it cannot guide the model to learn generalizable features during the fine-tuning process itself. Moreover, the interpolation coefficient α must be tuned manually, and a single global coefficient cannot account for the varying importance of different parameters. Our method differs fundamentally by integrating parameter selection into the training process, enabling dynamic, parameter-level suppression rather than post-hoc global averaging.

3. Methodology

3.1. Problem definition

We consider a pre-trained CLIP image encoder with weights W_0 and a downstream task dataset $D = \{(x_i, y_i)\}_{i=1}^N$, where x_i denotes an input image and y_i its class label. The objective of fine-tuning is to learn new weights W_T that maximize classification accuracy on the ID test set D_{ID} while simultaneously maximizing generalization on the OOD test set D_{OOD} .

Traditional full fine-tuning minimizes the cross-entropy loss over all parameters:

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^N y_i \log \sigma(f(W_T; x_i)), \quad (1)$$

where f denotes the encoder followed by the classification head and σ the SoftMax function. Because this objective optimizes every parameter without regard to OOD robustness, the model

frequently overfits spurious features present in the ID data, leading to degraded OOD performance.

3.2. Core principle

Our method rests on a simple yet effective observation: parameters that undergo the largest changes during ID fine-tuning are the most likely to encode ID-specific or spurious features and therefore are the primary drivers of OOD generalization degradation. Based on this insight, our approach consists of two stages: (i) computing a normalized discrepancy score for each parameter channel, and (ii) selectively suppressing high-discrepancy parameters during secondary fine-tuning.

3.2.1. Parameter discrepancy calculation

We first obtain preliminary fine-tuned weights W_T by performing standard full fine-tuning with cross-entropy loss. For the i -th channel, we compute the absolute discrepancy between the pre-trained weight $W_{0,i}$ and the fine-tuned weight $W_{T,i}$:

$$\Delta W_i = |W_{0,i} - W_{T,i}|. \quad (2)$$

Because weight magnitudes vary substantially across layers, direct comparison of absolute discrepancies would unfairly penalize layers with naturally larger weights. We therefore normalize each channel's discrepancy by its L2 norm:

$$\Delta W_{\text{norm},i} = \frac{\Delta W_i}{\|\Delta W_i\|_2}. \quad (3)$$

The normalized discrepancy serves as the ID-specific score $s_i = \Delta W_{\text{norm},i}$ for channel i . A higher score indicates that the corresponding channel has shifted more substantially during ID fine-tuning and is therefore more likely to contribute to ID overfitting.

3.2.2. Discrepancy-guided parameter suppression

Using the ID-specific scores, we rank all channels in descending order of discrepancy. The top $K\%$ of channels are classified as high-discrepancy and are frozen; the remaining low-discrepancy channels remain trainable. This selective update is enforced by a binary mask m_i :

$$m_i = \begin{cases} 0, & \text{if } s_i > \tau; \\ 1, & \text{if } s_i \leq \tau, \end{cases} \quad (4)$$

where τ is the score threshold determined by the freezing ratio $K\%$. During gradient descent, only parameters with mask value 1 are updated:

$$W_{t+1,i} = W_{t,i} - \eta \cdot (m_i \odot g_{t,i}), \quad (5)$$

where η is the learning rate, $g_{t,i}$ the gradient at step t , and $\odot$ denotes element-wise multiplication.

3.2.3. Training pipeline

The complete training procedure consists of four sequential stages. We initialize from the pre-trained CLIP weights W_0 and perform standard full fine-tuning with cross-entropy loss to obtain preliminary fine-tuned weights W_T . Then, we calculate the normalized parameter discrepancy for every channel to generate the ID-specific score s_i . We construct a binary mask m_i according to the freezing ratio $K\%$, freezing all channels with $s_i > \tau$ based on these scores. The last step is training from W_T with the mask applied, updating only low-discrepancy parameters until convergence.

3.3. Method advantages

Compared with existing approaches, our method offers four distinct advantages. Rather than applying uniform constraints across all parameters, we explicitly identify and suppress the subset of parameters most responsible for ID overfitting, enabling differentiated treatment of ID-specific and task-invariant weights. Then the method introduces only one hyperparameter, the freezing ratio $K\%$ and is straightforward to implement; because 90% of parameters are frozen during secondary fine-tuning, both memory consumption and computational cost are substantially reduced. Moreover, the discrepancy score is architecture-agnostic, and we demonstrate effective performance on both transformer-based (ViT-B/16) and convolutional (ResNet50) backbones across different domain-shift benchmarks. Since the normalized discrepancy can naturally be taken as the ID-specific score for each channel, it follows that a higher score means the channel has changed more during ID fine-tuning and hence is more prone to contribute to ID overfitting.

4. Experiments

4.1. Experimental setup

4.1.1. Datasets

iWildCam (WILDS benchmark) is introduced as "iWildCam is part of the WILDS benchmark [4], a wildlife image classification task comprising 182 animal categories." Therefore, it is a wildlife image classification task with 182 animal categories, where ID and OOD splits are naturally determined by camera trap location and illumination conditions. Because of the severe class imbalance, Macro-F1 is appropriately used as the main evaluation metric.

4.1.2. Baselines

We compare our method with seven representative approaches: Vanilla FT does standard full fine-tuning of all parameters with cross-entropy loss, Linear Prob freezes the CLIP encoder and trains only the linear classification head, L2-SP adds an L2 regularization term that penalizes departure from pre-trained weights, LP - FT first trains a linear classifier and then does full fine-tuning. "Representation - space regularization methods, such as FLYP [5], constrain the fine-tuned representation to remain close to the pre-trained distribution." FLYP augments fine-tuning with an auxiliary contrastive loss to explicitly preserve pre-trained representations. FTP introduces gradient

projection constraints during fine-tuning. And finally, WiSE-FT applies post-hoc linear correction to the pre-trained and fine-tuned weights.

4.1.3. Implementation details

All experiments are conducted on an NVIDIA A100 80GB GPU using PyTorch 2.0 with the official CLIP implementation. For the iWildCam dataset we employ the CLIP ViT-B/16 backbone, while for DomainNet we use the CLIP ResNet50 backbone. The optimizer is AdamW for ViT-B/16 and SGD for ResNet50, both with cosine learning rate decay. We set the freezing ratio to $K=90\%$ (determined as optimal via ablation), the initial learning rate to 1×10^{-5} , the weight decay to 0.2, and the batch size to 256. Evaluation metrics are ID and OOD Macro-F1 for iWildCam, and ID accuracy together with average OOD accuracy for DomainNet.

4.2. Main results

4.2.1. iWildCam

Table 1 reports ID and OOD Macro-F1 scores on iWildCam.

Table 1. ID and OOD Macro-F1 on iWildCam

Method	ID Macro-F1	OOD Macro-F1
Linear Prob.	38.2%	32.1%
Vanilla FT	48.1%	35.0%
L2-SP	47.5%	35.3%
LP-FT	49.2%	35.1%
FLYP	51.4%	35.2%
FTP	49.7%	35.8%
WiSE-FT	48.5%	35.7%
Proposed	50.5%	36.9%

The analysis shows clearly that the proposed method obtains an out-of-distribution Macro-F1 of 36.9 percent, which is at least 1.1 percentage points higher than all baselines. More importantly, it does so without sacrificing the in-distribution Macro-F1 of 50.5 percent, which is within one percentage point of the best in-distribution performer FLYP (51.4 percent). Thus, a natural and convincing conclusion is drawn targeted parameter suppression improves out-of-distribution robustness without the in-distribution performance sacrifice seen in traditional full fine-tuning.

4.2.2. DomainNet

Table 2 reports ID accuracy and average OOD accuracy on DomainNet.

Table 2. ID accuracy and average OOD accuracy on DomainNet

Method	ID Accuracy	Avg. OOD Accuracy
Linear Prob.	78.21%	36.12%
Vanilla FT	83.92%	38.42%

Table 2. (continued)

L2-SP	83.75%	39.11%
LP-FT	84.01%	39.53%
FLYP	84.31%	39.87%
FTP	84.22%	39.94%
WiSE-FT	84.15%	40.21%
Proposed	84.53%	41.63%

The method presented here attains the highest ID accuracy (84. 53%) and the highest average OOD accuracy (41. 63%) at the same time, with a 3. 21-percentage-point OOD improvement over Vanilla FT and a 1. 69-point gain over the previous best method (FTP), thereby convincingly demonstrating that discrepancy-guided suppression generalizes well across vision transformer and convolutional architectures. More importantly, its consistent performance on difficult domains like Clipart and Painting confirms its strong robustness to large visual distribution shifts.

4.3. Ablation studies

4.3.1. Component ablation

Table 3 ablates core design choices on DomainNet.

Table 3. Ablation of core components on DomainNet

Configuration	ID Accuracy	Avg. OOD Accuracy
Vanilla FT	83.92%	38.42%
Unnormalized discrepancy	84.46%	41.18%
Layer-wise freezing	84.38%	41.54%
Random freezing (90%)	83.87%	39.61%
Full method	84.53%	41.63%

Findings: Normalization is clearly necessary, since the use of unnormalized discrepancy scores reduces out-of-distribution performance by 0.45 percentage points, hence inter-layer standardization is an essential preprocessing step for fair comparison. Channel-wise granularity is beneficial: layer-wise freezing is 0.09 points worse than channel-wise freezing, which directly supports the conclusion that finer-grained suppression gives better control over parameter updates. Most importantly, discrepancy guidance is the key factor: randomly freezing 90% of parameters performs much worse than discrepancy-guided freezing, so the performance improvement comes from the selection of which parameters to freeze, not from simply reducing the number of trainable parameters.

4.3.2. Freezing ratio analysis

We vary the freezing ratio K from 10 percent to 95 percent. Out-of-distribution accuracy improves steadily as K increases, reaching its peak at K equals 90 percent. In-distribution accuracy remains stable up to 90 percent and drops only when K exceeds 95 percent, indicating that excessive freezing begins to suppress parameters needed for task adaptation. Based on this analysis, K equals 90

percent represents the optimal operating point that maximizes out-of-distribution generalization while preserving in-distribution performance.

4.4. Limitations and future work

The present method has the well-defined limitation that it requires a preliminary full fine-tuning pass to compute parameter discrepancy, hence requiring an extra training stage. Therefore, a natural and promising direction for future work is to develop dynamic discrepancy estimation that can be carried out within a single training run, thus allowing truly end-to-end fine-tuning. Extending the method to natural language processing and multimodal foundation models is also highly worthwhile.

5. Conclusion

This paper gives a clear, principled treatment of the ID-OOD performance trade-off in CLIP fine-tuning by introducing discrepancy-guided parameter suppression, and the key insight is that parameters with the largest shifts during ID fine-tuning are the main culprits behind OOD degradation. Therefore, the authors naturally quantify the shift using a normalized discrepancy score and then selectively freeze the most shifted channels, thereby breaking the classical trade-off and obtaining state-of-the-art OOD generalization on iWildCam and DomainNet without sacrificing ID performance. More importantly, the method is simple, general, and empirically very reliable, making it an excellent choice for robust fine-tuning of foundation models.

References

- [1] Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. Proceedings of the International Conference on Machine Learning.
- [2] Liu, C., et al. (2024). Discrepancy-guided parameter suppression for robust fine-tuning of foundation models. Advances in Neural Information Processing Systems.
- [3] Wortsman, M., et al. (2022). Robust fine-tuning of zero-shot models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
- [4] Koh, P. W., et al. (2021). WILDS: A benchmark of in-the-wild distribution shifts. Proceedings of the International Conference on Machine Learning.
- [5] Chen, T., et al. (3). FLYP: Fine-tuning with language-image pre-training for out-of-distribution generalization. International Conference on Learning Representations.