Robustness Preserving Fine-tuning using Neuron Importance

Guangrui Li * Rahul Duggal** Aaditya Singh
Kaustav Kundu Bing Shuai Jonathan Wu

AWS/Amazon AI
guangrui.li@outlook.com {dugrahul, singaadi, kaustavk, bshuai, jonwu}@amazon.com

Abstract. Robust fine-tuning aims to adapt a vision-language model to downstream tasks while preserving its zero-shot capabilities on unseen data. Recent studies have introduced fine-tuning strategies to improve in-distribution (ID) performance on the downstream tasks while minimizing deterioration in out-of-distribution (OOD) performance on unseen data. This balance is achieved either by aligning the fine-tuned representations with the pre-trained ones or by constraining significant deviations in fine-tuned weights compared to the pre-trained model. In the latter approach, the regularization term is uniformly applied to all parameters. Our work proposes selectively applying the regularization term based on the importance of each neuron to the fine-tuning dataset. To this end, we have developed an importance-score metric to quantify each neurons' importance to the downstream task, which we then leverage to develop two fine-tuning strategies: importance-guided selective fine-tuning and importance-guided regularization. Our approach can be used concurrently with representation space-based methods, outperforming other approaches based on parameter space. We achieve improvements over the state-of-the-art on standard robust fine-tuning benchmarks across various datasets, in both full-shot and low-shot settings.

Keywords: Robust Finetuning · Foundation Models · Transfer Learning

1 Introduction

Foundation Models (FM) have enabled a paradigm shift in computer vision through large-scale pre-training on web-scale data [12, 19, 21]. These models not only exhibit tremendous zero-shot capabilities across a wide array of tasks [12, 19], but are also amenable to efficient adaptation via fine-tuning [11, 14, 29]. However, recent studies [14, 29] show that while naive fine-tuning enhances in-distribution (ID) accuracy, it also degrades the fine-tuned model's generalization and robustness to out-of-distribution (OOD) samples (See Fig. 1a). Indeed, robustness is an essential quality for machine learning models deployed in the real

* Work conducted during an internship with AWS AI.
** Corresponding author.

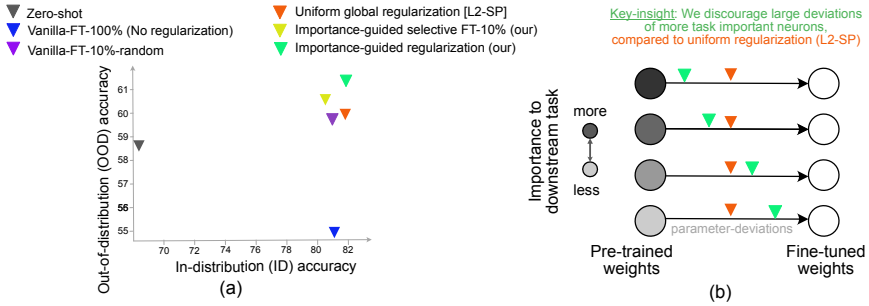

Fig. 1: Illustrating our approach by fine-tuning a CLIP pre-trained model (▼) on ImageNet (ID) and evaluating on 5 distribution shifts (OOD) [1, 9, 10, 20, 26]. Observe in part (a) that vanilla fine-tuning all neurons (▼) enhances ID accuracy, at the cost of reduced OOD robustness. In contrast, constraining parameter deviation via randomly fine-tuning only 10% neurons (▼), uniform global regularization (▼), our importance guided selective fine-tuning (▼) and our importance guided regularization (▼) improve ID without degrading OOD accuracy. Part (b) illustrates the key insight behind our importance-guided regularization – constrain deviations of the most task-important neurons while allowing larger deviations for others.

world where they commonly encounter open-set scenarios.To tackle this challenge, recent works in robust fine-tuning [8, 25, 28] aim to improve the ID task performance while preserving the innate robustness of pre-trained representations during fine-tuning on downstream tasks.

Existing works in robust fine-tuning can be categorized along two broad directions: the first aims to preserve robustness in the *representation space* by mimicking the pre-training objectives [8], representations [18] or classifier initialization [14]. The second stream operates in the *parameter space* by restricting large parameter deviation from the pre-trained weights through global weight interpolation [28], global L2 constraint on deviations [29] or by projecting gradients of individual parameters [24, 25]. Notably, the latter works provide a key theoretical insight – constraining the deviation of individual parameters in a fine-grained manner can lead to better OOD performance, but their performance on challenging benchmarks still lags the representation space paradigm methods.

Our work presents a new parameter-space fine-tuning method that aims to constrain parameter deviations at a more fine-grained (per-neuron) level and synergies with the representation-space training methods. To develop such a fine-grained constraint, we quantify each neuron's contribution towards the downstream task via an importance score. Subsequently, we leverage this importance score to guide the fine-tuning process. This leads to the following key questions:

Q1 Given a pre-trained model and the in-distribution (ID) dataset, how to evaluate the neuron-wise importance?

Q2 How to leverage the importance score during fine-tuning to achieve better ID and out-of-distribution (OOD) performances?

We answer the first question by assessing each neuron's importance and quantify it as the neuron importance score. This score essentially tracks the error change induced by updating an individual neuron, while holding the others constant. A similar importance criterion has previously been studied in the neural network pruning literature [6, 15–17] albeit for a different objective (pruning less important neurons). Our goal on the other hand, is to leverage this score to modulate gradient updates during the fine-tuning process.

To answer the second question, we devise two fine-tuning strategies with seemingly opposing goals: *importance-guided selective fine-tuning* to encourage the most important neurons to deviate from pre-trained weights while freezing the rest; and our *importance-guided regularization* to discourage the most important neurons to deviate from the pre-trained weights, while allowing the remaining to deviate. We reconcile the seemingly opposing goals with an empirically driven hypothesis (see Fig. 2): if only a few parameters are allowed deviation, then prioritizing further optimization of the task-important neurons performs better empirically. On the other hand, if more parameters are allowed to deviate, then preserving the task-important neurons performs better empirically as the ID accuracy improves further by optimizing the task-agnostic neurons. However, the parameter deviation in this case requires a more careful and fine-grained treatment to preserve OOD robustness thus motivating our regularization-based approach. Indeed Fig. 1a demonstrates that while both our fine-tuning strategies surpass vanilla fine-tuning, the regularization approach is superior in enhancing both ID and OOD accuracy. We illustrate the key insight behind our regularization strategy in Fig. 1b.

In principle, our two fine-tuning strategies relate to prior work [24, 25, 28, 29] that minimize the parameter deviation from the pre-trained weights, but at a more fine-grained level. Further, our strategies are orthogonal to recent representation based strategies [8, 18] and can be combined to improve upon the state-of-art. Specifically, combined with the state-of-the-art FLYP, we improve absolute OOD accuracy by 2.5% for ImageNet, 0.8% for iWildCam, and 1.4% for FMoW with while maintaining ID performance. Additionally in low-shot scenarios, we achieve 1-2% ID/OD accuracy improvements on ImageNet and iWildCam across 3 different low-shot regimes. Interestingly, compared to FLYP [8] that also fine-tunes the language encoder; and FTP [25], that adds extra projection parameters, our approach adds negligible computational overhead and saves up to 10% training time compared to these approaches.
To summarize, the overall contributions of our work are:

- We develop an importance-guided fine-tuning framework that presents a more fine-grained perspective for robust fine-tuning.
- We propose a novel importance-score metric to rank neurons' importance to downstream task and leverage this score through two strategies: importance-guided selective fine-tuning; and importance-guided regularization.
- We demonstrate state-of-the-art results on standard benchmarks by combining our regularizer with recent methods, without introducing any computational overhead.

2 Related Works

Robustness benchmarks and vision-language models (VLMs). Owing to concerns of generalizability of models trained and evaluated only on ImageNet [4, 7], researchers have developed several challenging benchmarks to measure robustness to natural distribution shifts [1, 9, 10, 13, 20, 26]. These can occur due to the natural variations in images, and in particular [23] have demonstrated that robustness to synthetic distribution shifts (i.e. programmatically deterministic) does not transfer well to such shifts. The advent of vision-language models [12, 19] has lead to unprecedented robustness gains over several of these shifts. However, naive fine-tuning of such models to improve in-distribution (ID) performance is known to deteriorate out-of-distribution (OOD) performance [8, 14, 28]. Hence, a large number of works focus on robust fine-tuning of such VLMs.

Robust fine-tuning of VLMs. While many recent works propose different methods to achieve better downstream task (i.e ID) performance than the zero-shot VLM while preserving or improving its robustness to distribution shifts, they can be broadly grouped into explorations in *representation space* and *parameter space*. The first line of works attempt to mitigate the distortion of pre-trained features during fine-tuning via multi-stage training or alternate objective functions. LP-FT [14] identifies such feature distortion as the cause for degraded OOD robustness. To counter this, they perform a two-stage fine-tuning process of first training a linear head and then the entire vision encoder. Most recently, FLYP [8] adopts the contrastive learning objective instead of cross-entropy for fine-tuning both the vision and language encoders, and achieves state-of-the-art performance on many challenging benchmarks.

The second line of works attempt to constraint the deviation of fine-tuned model parameters from the pre-trained weights. WiSE-FT [28] first naively fine-tunes the vision encoder with zero-shot head weights and then linearly interpolates the weights of the zero-shot vision encoder. Model Soups [27] scales this kind of weight-space ensembling further by fine-tuning several models with different kinds of augmentations and optimization objectives. L2-SP [29] applies a global regularization based on deviation from zero-shot weights during fine-tuning. However, the works mentioned thus far do not discriminate between the individual model parameters while fine-tuning which could be suboptimal, as argued by [24, 25]. These works perform gradient projection to restrict the deviation of individual model parameters from the pre-trained weights, and provide theoretical insights on why that might lead to a better OOD performance. Our work extends their intuition of constraining parameter deviation during fine-tuning by directly applying a per-neuron regularization scheme. We obtain the regularization weights via gradient-based estimation of neuron importance which is inspired by pruning literature [16, 17].

3 Methodology

To transfer to downstream tasks, the fine-tuning process typically starts with the pre-trained model, that is fine-tuned on the downstream data $\mathcal{D}$ using relevant

task objectives. We begin by outlining our fine-tuning task in Sec. 3.1 and then describe the two key elements of our approach – Estimating the neuron-wise importance score as discussed in Sec. 3.2; and leveraging this score in the fine-tuning process itself as discussed in Sec. 3.3.

3.1 Preliminaries

Setup. Consider an image classification task that aims to associate an input image x with a label y using a neural network or encoder f_W parameterized by weights W. The encoder is typically pre-trained on a large image-text corpus such as LAION [21] via contrastive learning objectives [12,19]. The outcome of this pre-training phase is a pre-trained model f_{W^0} that is parameterized by W^0.

Objective. Given an image, label pairs $(x, y) \sim \mathcal{D}$ from a downstream dataset, we aim to adapt the pre-trained model f_{W^0} to f_{W^t} such that $f_{W^t}(x) \rightarrow y$. Such adaptation can typically be achieved by minimizing the empirical loss:

$$W^t = \underset{W}{argmin}\ \{\mathcal{L}_{sup}(f_{W \leftarrow W^0}; \mathcal{D}) + \lambda \mathcal{R}(W, W^0)\}. \tag{1}$$

Here $W \leftarrow W^0$ denotes the initialization from pre-trained weights. $\mathcal{L}_{sup}$ corresponds to a supervised loss term that can be cross-entropy, or a contrastive loss [19] while $\mathcal{R}$ is a regularization term that operates on the weights W independent from the downstream dataset.

The loss formulation in Eq. (1)can be applied to fine-tune either the linear classifier atop a frozen pre-trained backbone, known as the linear probing setting, or both the classifier and the backbone in a full fine-tuning setting. For the purpose of this paper, we consider the full fine-tuning setting as it often leads to higher in-distribution accuracy [14]. Furthermore, we consider both variations of the supervised loss (cross-entropy and contrastive) with our proposed regularizer.

3.2 Neuron-wise Importance

Recall from Sec. 1 that we aim to devise a more direct and fine-grained approach of constraining deviation of individual parameters which achieves synergy with the state-of-the-art training methodologies and further improves their in-distribution (ID) and out-of-distribution (OOD) performances. This section answers the first overarching question: given a pre-trained model and ID dataset, how to assess the importance of each neuron for ID performance?

To this end, we propose a metric – importance score – that is inspired from pruning literature [16,17] which studies the same question to determine the least important neurons that can ultimately be pruned from the network. Instead of pruning weights, we use the importance-score to enable a more fine-grained modulation over the weights during the fine-tuning process.

Our importance-score approximates the contribution of each neuron by modeling the change in empirical risk that is attributable to the individual neuron. Specifically, given network parameters $W = \{w_0, w_1, ..., w_m\}$ and task dataset

$\mathcal{D}$, a neuron i's contribution, r_i towards the downstream task can be estimated by the change in empirical risk E cause by updating the parameter as:

$$r_i = E(\mathcal{D}, w_i) - E(\mathcal{D}, w_i | w_i = w_i + \triangle_{w_i}). \tag{2}$$

where $\triangle_{w_i}$ denotes the change in weights during the fine-tuning process.

In practice, calculating r_i for each neuron is intractable for modern neural networks which can contain millions of parameters. Therefore, we approximate r in the vicinity of W by its first-order Taylor expansion similar to [17]:

$$r_i^{(1)} = -g_i \triangle_{w_i} \tag{3}$$

where $g = \partial E / \partial w_i$ is the gradient. The above approximation suggests that gradient magnitude is a viable proxy to importance score since even a small change in weights for important neurons will result in large change in empirical risk. With this insight, we define the importance score as:

$$r_i = \frac{1}{n_t} \sum_{j=1}^{n_t} |g_i(x_j)|, (x_j, y_j) \in \mathcal{D}. \tag{4}$$

Intuitively, our importance-score averages the absolute gradient over all n_t samples from the downstream task dataset. This metric encapsulates the insight that neurons with large gradients contribute more towards the downstream task. Note that while both Eq. 4 and [17] leverage the taylor expansion of Eq. 3 we focus on the first-order approximation as it suffices for our use-case of modulating gradient updates via neuron importance during fine-tuning. In the next section, we discuss how to leverage our importance metric to guide the parameters updates during the fine-tuning process.

3.3 Importance-guide Modulation

Given the importance score for each neuron determined via Eq. 4, how to modulate parameter updates with importance guidance? To answer this question, we devise two fine-tuning strategies with seemingly opposing objectives:

- **Importance-guided Selective Fine-tuning**: This strategy encourages the optimization of neurons with higher ID importance, prioritizing their role in handling ID data effectively. Thus the top-k% important neurons are updated with their gradient while the remaining are frozen.
- **Important-guided Regularization**: This strategy discourages the optimization of neurons with higher ID importance while allowing others to update freely. Furthermore, instead of masking neurons we devise a novel regularization term that allows updating all neurons.

Despite the seeming contrarian objectives, our two fine-tuning strategies share the high-level idea of recent works [24, 25, 28, 29] that minimize parameter deviation from pre-trained representations as a means to the preserve innate robustness gained during pre-training. The difference lies in *which* neurons' weights to preserve and *how* to preserve them. The next sub-section delves into these details.

Algorithm 1 Importance-guided Selective Fine-Tuning

Input: Set of parameters W, learning rate η, top k% neurons T_{ID}, loss function $\mathcal{L}_{sup}$
for each parameter W_i^t in W^t **do**
  if $W_i^t \in T_{ID}$ **then**
    Compute gradient g^i for W_i^t with respect to $\mathcal{L}_{sup}$
    Update $W_i^{t+1} = W_i^t - \eta \cdot g^i$
  else
    $W_i^{t+1} = W_i^t$ ▷ No update for parameters outside T_{ID}
  end if
end for
return W^{t+1}

Importance-guided Selective Optimization The overall approach is outlined in Algorithm. 1. We start by determining the importance score via Eq. 4 for each neuron. Then, we partition all the pre-trained weights in the network W^0 into, either the set $\mathcal{T}_{ID}$ containing the top k percent of neurons representing the highest ID importance; or the remainder $W^0 - \mathcal{T}_{ID}$. Given the i^{th} neuron at the t-th iteration w_i^t, its gradient update is performed as:

$$w_{t+1}^i = w_t^i - \eta \cdot m^i \cdot g^i, \tag{5}$$

where η is the learning rate and m is a binary optimization mask:

$$m^i = \begin{cases} 1 & \text{if } w^i \in \mathcal{T}_{ID} \\ 0 & \text{if } w^i \notin \mathcal{T}_{ID}. \end{cases} \tag{6}$$

The overall outcome of our importance-guided selective fine-tuning is that only the top-k most-important neurons are updated with the supervised loss term, *i.e.*, $\mathcal{L}_{sup}$ in Eq. 1, while the others retain the pre-trained weights. We present the impact of importance guidance in Fig. 2 on both the ID and OOD accuracy while fine-tuning an CLIP [19] pre-trained ViT-B/16 model on ImageNet. Observe that with vanilla fine-tuning all neurons improves ID accuracy while degrading the OOD accuracy compared to zero-shot accuracy. Interestingly, fine-tuning a randomly chosen subset of neurons improves OOD accuracy. Most notably, tuning the most important subset of neurons improves both the ID and OOD accuracy, thereby validating the significance of importance guidance. Note that reversing the recommended direction of importance-guidance *i.e.* tuning the least important neurons degrades both ID and OOD accuracy.

Importance-guided Regularization. The previous section demonstrates the significance of importance-guided modulation. However, the hard top-k cutoff can still be sub-optimal as (1) fine-tuning only a subset of model parameters reduces its learning capacity and imposes a strong inductive bias (2) experiments from prior works [25,28] in *parameter space* (see Sec. 1) suggest that for optimal

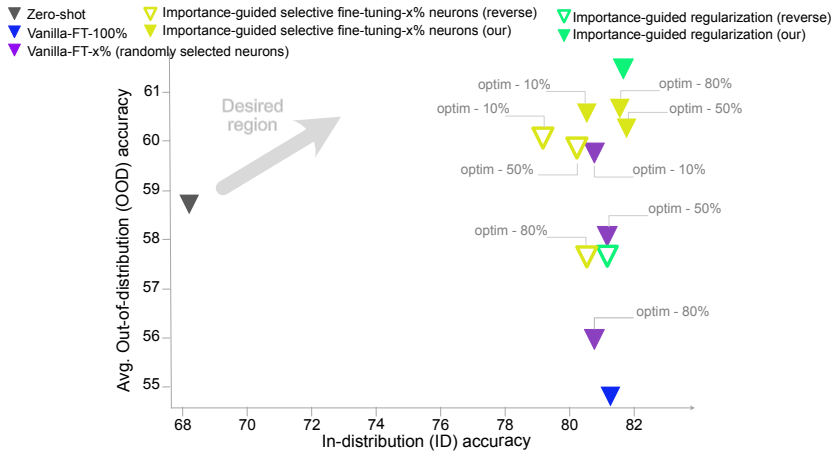

Fig. 2: Comparing optimization schemes while fine-tuning a CLIP pre-trained ViT-B/16 on ImageNet and evaluating on 5 natural distribution shifts (see Sec. 4). Observe that vanilla fine-tuning 100% parameters improves ID while sacrificing OOD accuracy compared to the zero-shot model. In comparison fine-tuning a random subset of parameters improves OOD, while optimizing important neurons only via importance-guided selective fine-tuning (see Sec. 3.3) further improves performance. The best overall ID/OOD accuracy is achieved via our importance-guided regularization (see Sec. 3.3) demonstrating the impact of importance guided modulation. Importantly, reversing the recommended importance-guidance directions degrades accuracy.

performance individual model parameters should be kept in the vicinity of the pre-trained weights but also be allowed to fine-tune.

Hence, this sub-section presents a novel regularizer that enables importance-guided fine-tuning for *all* neurons in the network. Our key motivation to develop the regularizer is to penalize large deviations of the most important neurons from the pre-trained weights. Specifically, given a parameter i at iteration t, w_i^t, and it's importance score r_i, we first normalize the importance-score against a cohort set S_i as,

$$\bar{r_i} = \|r_i\|_\infty, \quad w_i^0 \in S_i, \tag{7}$$

where the set S_i contains all neurons in the same transformer block [5] as neuron i. Then, we use the normalized importance score to constrain parameter deviations through our regularizer:

$$\mathcal{R}(w_i^t, w_i^0) = \bar{r}_i |w_i^t - w_i^0|. \tag{8}$$

We observe that our regularizer's performance is sensitive to the selection of the cohort set S_i used for normalizing the importance score. In Sec. 4.4, we explore different cohort sets S_i for normalizing the importance score of such neurons in the same transformer layer (or layer-wise), the entire network-wide (or global) and find that the block-wise cohort set leads to the best perfor-

Algorithm 2 Importance-Guided Regularization

Input: Set of parameters W^t at iteration t, learning rate η, initial weights W^0, importance scores r, supervised loss $\mathcal{L}_{sup}$
Output: Regularized weights W^{t+1}
Normalize importance scores in each parameters set (S_k), $\bar{r}_i = \|r_i\|_\infty, w_i \in S_k$.
Define regularization term: $\mathcal{R}(w_i^t, w_i^0) = \sum_i \bar{r}_i |w^t - w^0|$.
Compute total loss: $\mathcal{L}_{total} = \mathcal{L}_{sup}(W^t) + \lambda \mathcal{R}(W^t, W^0)$ ▷ λ is the regularization strength
for each parameter W_i^t in W^t **do**
    Compute gradient of $\mathcal{L}_{total}$ with respect to W_i^t: $g^i = \nabla_{W_i^t} \mathcal{L}_{total}$
    Update $W_i^{t+1} = W_i^t - \eta \cdot g^i$
end for
return W^{t+1}

mances. The overall importance-guided regularization approach is outlined in Algorithm 2.

To evaluate the significance of our regularizer, we plot a model fine-tuned via importance-guided regularization in Fig. 2. Observe that the ID and OOD performance improves further beyond our importance-guided selective fine-tuning approach. We attribute this improvement to fine-tuning of *all* parameters in the network. Additionally, our contribution lies in developing the novel regularizer which can be fitted along any supervised loss term L_{sup} in Eq. 1. Indeed, while using the contrastive loss of FLYP [8] we achieve state-of-the-art accuracy in most settings as demonstrated by the results in Sec. 4. Our evaluations in the remainder of the paper focus on importance-guided regularization due to it's superiority with respect to selective fine-tuning.

4 Experiments

We begin by describing our evaluation benchmarks in Sec. 4.1, baselines in Sec. 4.2 and then draw a comparison in Sec. 4.3 against the recent state-of-the-art across several datasets, and low shot regimes. To dive deeper into the underlying reasons of performance gains and evaluate robustness to hyper-parameters we present detailed ablations in Sec. 4.4.

4.1 Datasets and Data Regimes.

We follow prior works and consider the following benchmarks for measuring in-distribution (ID) and out-of-distribution (OOD) performance.

– **ImageNet & Distribution Shifts** ImageNet (IN-1k) [4] is a widely used object-centric dataset containing 1.2M training and 50k validation images from 1000 classes. We consider ImageNet-R (IN-R) [9], ObjectNet (ON) [1], ImageNet-S (IN-S) [26], ImageNet-A (IN-A) [10], ImageNet-v2 (IN-v2) [20] as the OOD shifts following prior works [8, 19, 28]. We report the top-1

accuracy on the validation set for ID performance and the averaged top-1 accuracy across these 5 shifts for OOD performance.
- **iWildCam** [2] dataset consists of images belonging to 182 animals species captured by different camera traps, that serve as distribution shifts. We use the WILDS benchmark [13] and the standard splits for training, validation, and ID and OOD testing that contain O(10k-100k) images. As the dataset has label imbalance, we follow prior works [8,28] and report macro F1-score.
- **FMoW** [3] consists of satellite images of 62 classes for different land and building types, where the year and continents in which they were taken lead to distribution shifts. We use WILDS benchmark and standard training, validation, and ID and OOD splits containing O(10k-100k) images. We report worst-case region accuracy following prior works [8, 28].

Additionally, to measure the effectiveness of the method in data scarce scenarios, we use the low-shot training splits curated by [22] for IN-1k and iWildCam datasets. For IN-1k, the splits contain 1, 5, and ∼13 images per class, *i.e.* 1k, 5k, and 13k images respectively. For iWildCam, the splits contain images in 1%, 10%, 20% ratios per-class, *i.e.* 1.3k, 12.9k, and 25.9k images respectively.

4.2 Baselines and Implementation Details.

We consider five most recent and well performing methods namely, WiSE-FT [28], LP-FT [14], FLYP [8], FTP [25] and L2-SP [29], for robust fine-tuning of vision-language models. For benchmarking, we use the CLIP ViT-B/16 model pre-trained on LAION-2B [5,19] as the base model and report average results across 3 seeds. We re-use the codebase and most of the training hyper-parameters from [8] including the AdamW optimizer with a cosine learning rate scheduler and a batch size of 512 for ImageNet and 256 for other datasets. We fix the regularization strength λ to 0.01 as discussed in Sec. 4.4 and provide other details in Appendix.

4.3 Results

We present our extensive empirical evaluations in this section. For our method, we mainly use importance-guided regularization as it outperforms importance-guided selective fine-tuning (see Sec. 3.3). The following sections show the results for ImageNet, iWildCam and FMoW in the full-shot regime and on ImageNet and iWildCam in 3 different low-shot regimes.

Full-shot results on ImageNet. We report the results on ImageNet on both in-distribution (ID) and averaged on 5 out-of-distribution shifts (OOD-Avg.) in table 1 from which we can draw the following conclusions.

- Incorporating our importance-guided regularizer with the conventional cross-entropy objective leads to a 0.2% improvement (from 81.3 → 81.5) in ID shift and a significant 6.8% improvement (from 54.9 → 61.7) on OOD shifts.

Table 1: Experimental Results ImageNet benchmark using CLIP ViT-B/16. Here we report the results of ID dataset ImageNet (IN), and five OOD datasets along with their average. † denotes our re-run with the publicly available codebase.

Methods	IN (ID)	ImageNet Distribution Shifts (OOD)					OOD Avg.
		IN-V2	IN-R	IN-S	ObjectNet	IN-A	
ZeroShot [19]	68.3	61.9	77.7	50.0	48.3	55.4	58.7
Linear Probing	79.9	69.8	70.8	46.4	46.9	52.1	57.2
L2-SP [29] [ICML 2019]	81.7	71.8	70.0	42.5	48.5	56.2	57.8
LP-FT [14] [ICLR 2022]	81.7	72.1	73.5	47.6	50.3	58.2	60.3
FTP † [25] [NeurIPS 2023]	81.6	72.1	74.9	51.6	56.6	51.0	61.2
WiSE-FT [28] ($\alpha = 0.5$) [CVPR 2022]	81.7	72.8	78.7	53.9	57.3	52.2	63.0
+ Imp.-guided-reg. (Ours)	**82.0** (↑)	73.3	77.9	53.0	58.1	55.8	**63.6** (↑)
Fine-tuning	81.3	71.2	66.1	37.8	46.1	53.3	54.9
+ Imp.-guided-reg. (Ours)	**81.5** (↑)	72.0	75.9	51.7	56.6	52.2	**61.7** (↑)
FLYP † [8] [CVPR 2023]	82.5	73.0	71.5	48.6	49.9	54.6	59.5
+ Imp.-guided-reg. (Ours)	**82.5** (–)	73.2	75.0	51.1	56.7	53.8	**62.0** (↑)

Table 2: Experimental Results on iWildCam and FMOW. Following common practice, we report macro F1-score for iWildCam and worst-case region accuracy for FMoW.

Method	iWildCam		FMoW	
	ID	OOD	ID	OOD
ZeroShot	8.7	11.0	20.4	18.7
Linear Probing	44.5	31.1	48.2	30.5
LP-FT [14]	49.7	34.7	68.4	40.4
L2-SP [29]	48.6	35.3	68.6	39.4
WiSE-FT [28]	48.1	35.0	66.8	42.3
FTP [25]	47.1	35.9	66.3	39.2
Fine tuning (FT)	48.1	35.0	66.8	40.9
+ Imp.-guided-reg. (Ours)	49.4 (↑)	36.2 (↑)	68.8 (↑)	42.0 (↑)
FLYP† [8]	51.2	35.3	67.7	40.8
+ Imp.-guided-reg. (Ours)	51.1 (↓)	36.1 (↑)	67.6 (↓)	42.2 (↑)

Also, OOD-Avg. performance improves by 3% (61.7 *vs*. 58.7) compared to the zero-shot model.

– Compared to other works sharing the same underlying intuition of restricting parameter deviation from zero-shot weights, *i.e*. L2-SP and FTP, we observe comparable (within 0.3%) ID performance but improvement in OOD-Avg. performance by 3.9% and 0.7% respectively which demonstrates the effectiveness of the regularization strategy.
– As our importance-guided regularization is orthogonal to the works in feature learning paradigm, it can be added to further boost their performance. For the state-of-the-art FLYP, our regularizer improves its OOD-Avg. performance by 2.5% (from 59.5 → 63.0). Additionally, it also demonstrates synergy with weight-space ensembling (*i.e*. WiSE-FT) and leads to 0.3% and 0.6% gains in ID and OOD-Avg. respectively.

Full-shot results on iWildCam. We report the results on iWildCam in table 2. Combined with conventional cross-entropy, our importance-guided regularization improves ID performance (*i.e*. macro F1-score) by 1.3% and OOD performance by 1.2% which outperforms all the methods that constrain parameter deviations, *i.e*. L2-SP, FTP, and the strong baseline of WiSE-FT by 1.3% ID and 1.2% OOD. Note that these improvements are obtained even though the zero-shot performance (row 1) is considerably lower. Combined with FLYP,

Table 3: Experimental results in low-shot data regimes on ImageNet and iWildCam.

Methods	ImageNet						iWildsCam					
	1-shot		5-shot		1% subset		1% subset		5% subset		20% subset	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
Zeroshot	68.3	58.4	68.3	58.4	68.3	58.4	8.7	11.0	8.7	11.0	8.7	11.0
FLYP	66.6	55.6	69.5	56.8	72.1	56.4	16.7	15.2	21.0	17.7	31.5	26.0
+Imp.-gd reg. (Ours)	70.3 (↑)	59.3 (↑)	70.6 (↑)	59.1 (↑)	72.4 (↑)	58.7 (↑)	16.8 (↑)	15.3 (↑)	22.5 (↑)	19.4 (↑)	32.6 (↑)	27.2 (↑)
FT	68.4	56.8	68.3	58.4	68.3	58.4	19.7	17.5	21.7	20.5	31.3	26.8
+Imp.-gd reg. (Ours)	69.4 (↑)	58.1 (↑)	70.2 (↑)	58.2 (↑)	71.6 (↑)	58.7 (↑)	20.7 (↑)	17.5(−)	22.5 (↑)	20.9 (↑)	31.4 (↑)	24.7(↓)

our approach leads to comparable (within 0.1%) ID performance with a 0.8% improvement (from 35.3 → 36.1%) on OOD shift, improving the state-of-the-art.

Full-shot results on FMoW. We report the results on FMoW dataset in table 2. With conventional cross-entropy, our importance-guided regularization improves ID performance (*i.e.* worst-case region accuracy) by 2.0% and OOD performance by 1.1% which outperforms L2-SP and FTP with at least 0.2 % on ID shift and ∼1.6% on OOD shift. Combined with FLYP, it leads to comparable (within 0.3%) ID and a significant 1.4% OOD performance improvement.

Results on low-shot regimes. To test the effectiveness of importance-guided regularization in low-shot data regimes, we follow [22] and conduct experiments on ImageNet and iWildCam datasets and their associated OOD shifts as described in Sec. 4. We consider vanilla fine-tuning (which uses conventional cross-entropy loss) and FLYP (which uses contrastive loss) as baselines and show the results of this experiment in table 3.

Across both datasets and low-shot regimes, ID and OOD performances consistently improve with our regularization for both vanilla fine-tuning and FLYP. Specifically compared to FLYP, adding our regularization improves ImageNet ID performance by 0.3-3.7% and OOD performance by 2.3-3.7% across different shot regimes. Similarly for iWildCam ID performance improves by 0.1-1.1% and OOD performance improves 0.1-1.7% respectively. With vanilla fine-tuning we observe similar gains of 1-3.3% ID, 0.3-1.3% OD on ImageNet and 0.1-1% ID, 0-0.4% OD on iWildsCam. Note that even though FLYP performs worse than vanilla fine-tuning in these low-shot regimes, our regularization technique consistently brings improvements to both types of fine-tuning.

4.4 Ablation Studies and Analysis

This section ablates on different design and hyper-parameter choices of our proposed method. We conduct evaluations on ImageNet by fitting our importance-guided regularization onto the FLYP [8] loss, unless mentinoned otherwise.

Analysis on cohort sets for importance-score normalization. Here we investigate the influence of different cohort sets for normalizing the importance score in Eq. 7. Rows #1-3 of Table 4 compares three different cohort sets, *i.e.*, layer-wise, and block-wise and network-wide (global norm) to normalize the importance scores. We find that block-wise normalization leads to the best results.

Table 4: Ablation Studies regarding normalization technique (1-3), weighting choice (4-7) on ImageNet and iWildCam.

Modification	ImageNet		iWildCam	
	ID	OOD	ID	OOD
#0 FLYP + imp.-guided reg. (our)	82.5	62.0	51.1	36.1
#1 Layer-wise norm	78.9	61.1	44.7	29.7
#3 Block-wise norm	82.5	62.0	51.1	36.1
#2 Global norm	82.6	58.9	50.7	34.3
#4 Uniform	81.7	57.8	48.6	35.3
#5 Reversed ID relevance	81.5	57.6	37.1	25.0
#6 Unsupervised Estimation of r	82.8	61.6	52.3	35.6

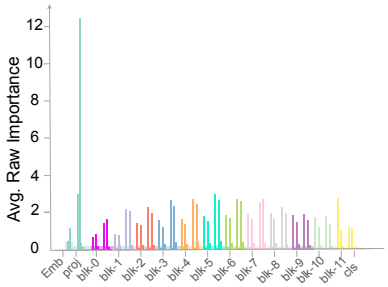

Fig. 3: Visualizing the raw importance score, across layers and blocks highlighting its variance and the need for normalization.

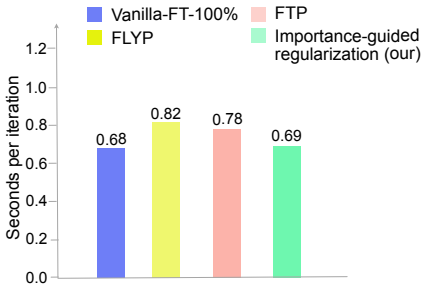

Fig. 4: Comparison on training time per iteration. Compared with recent robust-FT methods, our method adds negligible computation overhead.

The reasoning for selecting block-wise normalization is further strengthened by the illustration in Fig. 3 that plots the un-normalized importance score for each layer (bar) and block (shade). Observe the large variance in the importance scores across different blocks which illustrates the need for block-wise normalization.

Analysis on weighting strategies. A critical part of our contribution lies in answering how to determine the neuron-importance scores *i.e.* Eq. 4 and how to leverage the importance for fine-tuning *i.e.* Eq. 8. We investigate different choices for these in rows #4-6 of Table 4. Firstly, in row #4, we present the choice of uniformly weighting all neurons, which is equivalent to L2-SP [29]. As expected our method easily outperforms this on both ID and OOD due to the more fine-grained importance guidance. Second, in row #5, we reverse importance score by changing $\bar{r}$ to $1-\bar{r}$ in Eq. 8 thereby encouraging the most important neurons to deviate further during fine-tuning. This strategy degrades ID (-1%/-2.5%) and OOD (-4.4%/-0.8%) on both benchmarks further validating our choice of penalizing deviation of the most important neurons. Finally, in row # 6, we use the entropy loss (instead of cross-entropy) to get the gradient signals necessary for importance estimation. Interestingly, this leads to competitive performance

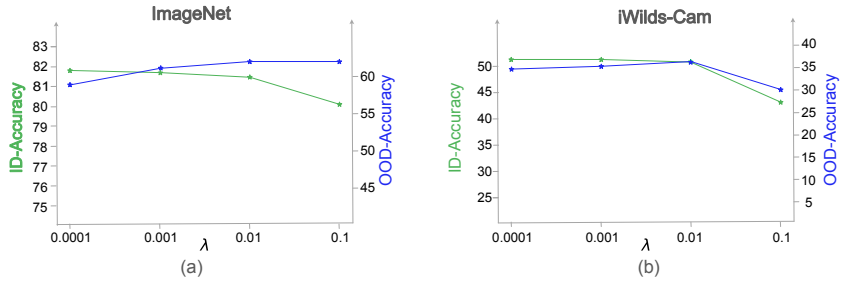

Fig. 5: Sensitivity Analysis to λ in regularizer. We present the analysis on two benchmarks, ImageNet (left), and iWild-CAM (right). With a wide range of λ, both ID and OOD performance suffers from limited fluctuations, indicating the robustness of the proposed method to the selection of λ.

with our proposed approach thereby neglecting the need of downstream label supervision to estimate neuron importance.

Analysis on training efficiency. Fig. 4,compares the training iteration time of different methods. Compared with vanilla-FT, our method brings negligible computation overhead and is, *i.e.*, 16% faster than FLYP and 11% faster than FTP. The is because FLYP and FTP introduces extra parameters to optimization, *i.e.*, FLYP tunes both the visual encoder and text encoder, and FTP introduces additional projection parameters. Instead, our method realizes a more fine-grained modulation without any extra parameters.

Sensitivity to hyper-parameter. We analyze the sensitivity of the hyper-parameter λ, that controls regularization strength in Eq. 1. Fig. 5 plots the ID & OOD accuracy for FLYP + importance-guided regularization for four values of $\lambda \in \{0.1, 0.01, 0.001, 0.0001\}$. Expectantly, increasing λ degrades ID while bettering OOD. Thus we select the largest $\lambda = 0.01$ leading to 1% drop from the maximum ID accuracy.

5 Conclusion, Limitations and Future Work

This work tackles the challenge of fine-tuning a pre-trained model to improve in-domain task accuracy while preserving the innate robustness of the pre-trained representations. We propose an importance-guided fine-tuning method that leverages a novel importance score to modulate neurons through two strategies: selective fine-tuning; and fine-grained regularization. We demonstrate significant improvements in both in-distribution, and out-of-distribution performances post fine-tuning that furthers the state-of-the-art on standard benchmarks. There are some limitations of our work: (1) we consider fine-tuning of vision-language pre-trained models. Although a broad class in itself, future work can explore self-supervised pre-trained models; (2) We consider image classification as the main downstream task. These limitations show that there is scope for improving our fine-tuning framework which can be tackled by future work.

References

1. Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., Katz, B.: Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. Advances in neural information processing systems **32** (2019)
2. Beery, S., Cole, E., Gjoka, A.: The iwildcam 2020 competition dataset. arXiv preprint arXiv:2004.10340 (2020)
3. Christie, G., Fendley, N., Wilson, J., Mukherjee, R.: Functional map of the world. In: Computer Vision and Pattern Recognition (CVPR) (2018)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Duggal, R., Xiao, C., Vuduc, R., Chau, D.H., Sun, J.: Cup: Cluster pruning for compressing deep neural networks. In: 2021 IEEE International Conference on Big Data (Big Data). pp. 5102–5106. IEEE (2021)
7. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
8. Goyal, S., Kumar, A., Garg, S., Kolter, Z., Raghunathan, A.: Finetune like you pretrain: Improved finetuning of zero-shot vision models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19338–19347 (June 2023)
9. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8349 (2021)
10. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15262–15271 (2021)
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), `https://openreview.net/forum?id=nZeVKeeFYf9`
12. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). `https://doi.org/10.5281/zenodo.5143773`, `https://doi.org/10.5281/zenodo.5143773`, if you use this software, please cite it as below.
13. Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., et al.: Wilds: A benchmark of in-the-wild distribution shifts. In: International Conference on Machine Learning. pp. 5637–5664. PMLR (2021)
14. Kumar, A., Raghunathan, A., Jones, R.M., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022), `https://openreview.net/forum?id=UYneFzXSJWh`

15. Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H.P.: Pruning filters for efficient convnets. arXiv preprint arXiv:1608.08710 (2016)
16. Liu, H., Simonyan, K., Yang, Y.: Darts: Differentiable architecture search. arXiv preprint arXiv:1806.09055 (2018)
17. Molchanov, P., Mallya, A., Tyree, S., Frosio, I., Kautz, J.: Importance estimation for neural network pruning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11264–11272 (2019)
18. Mukhoti, J., Gal, Y., Torr, P.H., Dokania, P.K.: Fine-tuning can cripple your foundation model; preserving features may be the solution. arXiv preprint arXiv:2308.13320 (2023)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
20. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400. PMLR (2019)
21. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5b: An open large-scale dataset for training next generation image-text models. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022), https://openreview.net/forum?id=M3Y74vmsMcY
22. Singh, A., Sarangmath, K., Chattopadhyay, P., Hoffman, J.: Benchmarking low-shot robustness to natural distribution shifts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16232–16242 (October 2023)
23. Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., Schmidt, L.: Measuring robustness to natural distribution shifts in image classification. Advances in Neural Information Processing Systems **33**, 18583–18599 (2020)
24. Tian, J., He, Z., Dai, X., Ma, C.Y., Liu, Y.C., Kira, Z.: Trainable projected gradient method for robust fine-tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7836–7845 (2023)
25. Tian, J., Liu, Y.C., Smith, J.S., Kira, Z.: Fast trainable projection for robust fine-tuning. Advances in Neural Information Processing Systems (2023)
26. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in Neural Information Processing Systems **32** (2019)
27. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., Schmidt, L.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, PMLR (2022), https://proceedings.mlr.press/v162/wortsman22a.html
28. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Gontijo-Lopes, R., Hajishirzi, H., Farhadi, A., Namkoong, H., Schmidt, L.: Robust fine-tuning of zero-shot models. arXiv preprint arXiv:2109.01903 (2021), https://arxiv.org/abs/2109.01903

29. Xuhong, L., Grandvalet, Y., Davoine, F.: Explicit inductive bias for transfer learning with convolutional networks. In: International Conference on Machine Learning. pp. 2825–2834. PMLR (2018)