

Focal-SAM: Focal Sharpness-Aware Minimization for Long-Tailed Classification

Sicong Li^{1,2} Qianqian Xu³ Zhiyong Yang⁴ Zitai Wang³
 Linchao Zhang⁵ Xiaochun Cao⁶ Qingming Huang^{4,7,3}

Abstract

Real-world datasets often follow a long-tailed distribution, making generalization to tail classes difficult. Recent methods resorted to long-tail variants of Sharpness-Aware Minimization (SAM), such as ImbSAM and CC-SAM, to improve generalization by flattening the loss landscape. However, these attempts face a trade-off between computational efficiency and control over the loss landscape. On the one hand, ImbSAM is efficient but offers only coarse control as it excludes head classes from the SAM process. On the other hand, CC-SAM provides fine-grained control through class-dependent perturbations but at the cost of efficiency due to multiple backpropagations. Seeing this dilemma, we introduce Focal-SAM, which assigns different penalties to class-wise sharpness, achieving fine-grained control without extra backpropagations, thus maintaining efficiency. Furthermore, we theoretically analyze Focal-SAM’s generalization ability and derive a sharper generalization bound. Extensive experiments on both traditional and foundation models validate the effectiveness of Focal-SAM.

1. Introduction

In the past decades, deep learning has achieved remarkable success in various fields, including image classification (He et al., 2016), medical image processing (Ronneberger et al., 2015), and object detection (Ren et al., 2015). However, this

¹Institute of Information Engineering, CAS ²School of Cyber Security, University of Chinese Academy of Sciences ³Key Lab. of Intelligent Information Processing, Institute of Computing Tech., CAS ⁴School of Computer Science and Tech., University of Chinese Academy of Sciences ⁵Artificial Intelligence Institute of China Electronics Technology Group Corporation, ⁶School of Cyber Science and Tech., Shenzhen Campus of Sun Yat-sen University ⁷BDKM, University of Chinese Academy of Sciences. Correspondence to: Qianqian Xu <xuqianqian@ict.ac.cn>, Qingming Huang <qmhuang@ucas.ac.cn>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

success often relies on carefully curated, balanced datasets. In real-world scenarios, data often exhibits a *long-tailed* distribution, where a few categories have abundant samples while most categories contain only a small number of examples. Long-tailed learning focuses on effectively training models on such imbalanced datasets (Zhang et al., 2023; 2024a). Numerous approaches have been proposed to address this challenge, including re-sampling (Buda et al., 2018), re-balancing (Cui et al., 2019; Ren et al., 2020; Wang et al., 2023), representation learning (Zhu et al., 2022; Cui et al., 2024), ensemble learning (Wang et al., 2021; Zhang et al., 2022), and fine-tuning foundation models (Dong et al., 2023; Shi et al., 2024).

Recently, Rangwani et al. (2022) visualized the loss landscape of different classes and observed that tail classes often suffer from saddle points. Since the loss landscape is closely related to the generalization of modern neural networks (Keskar et al., 2017; Jiang et al., 2020), they apply Sharpness-Aware Minimization (SAM) (Foret et al., 2021) to help tail classes escape from saddle points. Later, since the original SAM operates on all classes, ImbSAM (Zhou et al., 2023a) excludes the head classes to better focus on flattening the landscape of the tail classes. However, when combined with popular re-balancing methods (Cao et al., 2019; Kini et al., 2021; Menon et al., 2021), this coarse-grained approach often overemphasizes the tail classes, leading to poor head and overall performance. To achieve fine-grained control, CC-SAM (Zhou et al., 2023b) uses class-dependent perturbation. However, the per-class perturbation requires at least C additional backpropagations, where C denotes the number of classes, making it rather computationally expensive. This raises a natural question: *Can we design a method that achieves both fine-grained control and computational efficiency?*

Targeting this goal, we integrate the focal mechanism (Lin et al., 2017) with SAM, inducing a novel approach named Focal-SAM. Specifically, we introduce the focal sharpness term, which is defined as the weighted sum of class-wise sharpness, where the weights decrease in a focal-like manner from head to tail classes. On the one hand, Focal-SAM controls the flatness of different classes in a fine-grained way, better balancing the performance between head and tail classes than ImbSAM, as shown in Fig. 1(a). On the other

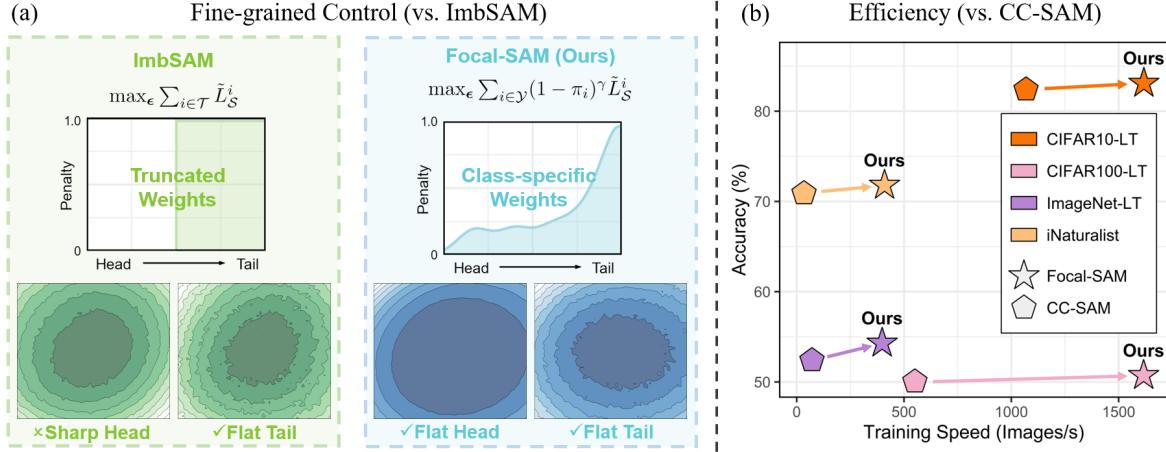


Figure 1: (a) ImbSAM applies the sharpness penalty only to tail classes, leading to a sharp loss landscape for head classes. In contrast, Focal-SAM assigns class-specific weights to the sharpness penalty, resulting in smooth loss landscapes for both head and tail classes. (b) Focal-SAM replaces per-class perturbations in CC-SAM with class-specific sharpness penalties, significantly enhancing computational efficiency while achieving better performance.

hand, Focal-SAM replaces the per-class perturbations in CC-SAM with per-class sharpness penalties, making it much more efficient than CC-SAM, as illustrated in Fig.1(b). Furthermore, we provide an informative generalization bound based on the PAC-Bayesian theory. This bound not only decreases at a faster rate than those of SAM and CC-SAM ($\tilde{\mathcal{O}}(1/n)$ vs. $\mathcal{O}(1/\sqrt{n})$, where n is the number of training samples) but also demonstrates the influence of the hyperparameters and trace of the Hessian.

Finally, we conduct extensive experiments on various benchmark datasets to validate the effectiveness of Focal-SAM, including training ResNet models from scratch and fine-tuning the foundation model CLIP (Radford et al., 2021). The results show that Focal-SAM consistently outperforms other SAM-based methods across multiple datasets and models in long-tailed recognition tasks. Prior arts (Zhou et al., 2022; Khattak et al., 2023; Park et al., 2024) have demonstrated that fine-tuning CLIP often performs well on the target domain but struggles with domain shifts. Therefore, we also assess model performance on OOD test sets when fine-tuning foundation models, referred to as long-tailed domain generalization tasks. The results indicate that Focal-SAM improves performance by approximately 0.5%~4.3% when combined with baselines on OOD test sets. These further suggest that Focal-SAM can enhance generalization, leading to better performance under domain shifts.

In summary, our key contributions are as follows:

- Systematic studies illustrate the limitations of ImbSAM and CC-SAM. ImbSAM fails to flatten the loss landscape for head classes, while CC-SAM is highly computationally expensive.

- We propose Focal-SAM, a simple yet effective method that provides fine-grained control of loss landscape and maintains computational efficiency. Theoretical analysis further offers a sharp generalization bound of Focal-SAM.
- Extensive experiments validate the effectiveness of the proposed Focal-SAM, ranging from training ResNet models from scratch to fine-tuning foundation models.

2. Related Work

2.1. Long-Tailed Learning

Several approaches address long-tailed learning challenges, such as re-sampling (Buda et al., 2018; Wang et al., 2019b; Liu et al., 2022), re-balancing (Cui et al., 2019; Ren et al., 2020; Wang et al., 2023; 2022; Han et al., 2024; Hou et al., 2022; Lyu et al., 2025; Yang et al., 2023b;a; 2022; Zhao et al., 2024a; Dai et al., 2023; Shao et al., 2023; Hong et al., 2024), data augmentation (Kim et al., 2020; Hong et al., 2022; Ahn et al., 2023; Wang et al., 2024b;a), representation learning (Cui et al., 2021; Zhu et al., 2022; Cui et al., 2024; Gao et al., 2023; Zhang et al., 2024b), ensemble learning (Wang et al., 2021; Zhang et al., 2022; Li et al., 2022; Aimar et al., 2023; Yang et al., 2024; Zhao et al., 2024b), and fine-tuning foundation models (Dong et al., 2023; Shi et al., 2024). This paper focuses on loss modification, a technique that modifies the loss function to guide the model's attention towards tail classes, consequently improving their performance. Various methods have been proposed, such as LDAM (Cao et al., 2019), which enlarges the margin for tail classes to enhance their generalization performance. Cao et al. (2019) further introduce a training scheme called Deferred Re-weighting (DRW) used in conjunction with

LDAM to improve model performance. However, Menon et al. (2021) argue that previous loss modification techniques sacrifice consistency in minimizing the balanced error. They propose the LA (Menon et al., 2021) loss, which introduces adjustments to the standard cross-entropy loss to ensure Fisher consistency for balanced error minimization. Building on this work, the VS (Kini et al., 2021) loss further improves upon the LA loss by incorporating both additive and multiplicative adjustments, beneficial during the initial and terminal phases of training respectively. Most recently, Wang et al. (2023) provide a comprehensive generalization analysis of these losses.

In this paper, we leverage these loss functions while aiming to specifically improve their generalization ability for long-tailed classification tasks.

2.2. Sharpness of Loss Landscape

Generalization in deep neural networks has always been a crucial focus in machine learning research. Recent studies (Keskar et al., 2017; Jiang et al., 2020) have empirically and theoretically demonstrated that flatter minima in the loss landscape typically lead to better generalization. Inspired by this, Sharpness-Aware Minimization (SAM) (Foret et al., 2021) is developed to find flatter minima, achieving superior performance across various tasks.

In the context of long-tailed learning, Rangwani et al. (Rangwani et al., 2022) suggest combining SAM with re-balancing techniques to help the model escape saddle points and improve generalization. Imbalanced SAM (ImbSAM) (Zhou et al., 2023a) incorporates class priors into SAM by dividing classes into head and tail groups. It applies SAM exclusively to the tail classes while maintaining standard optimization for head classes, aiming to specifically enhance the generalization of tail classes. Class-Conditional SAM (CC-SAM) (Zhou et al., 2023b) applies SAM to each class individually, using class-specific perturbation radii. These radii increase from head to tail classes, enabling fine-grained control over the loss landscape for each class.

This work also extends the SAM framework for long-tailed classification. Our method aims to achieve fine-grained control over the loss landscape while maintaining computational efficiency.

3. Motivation

3.1. Problem Setup

We define the sample space as \mathcal{X} and the label space as $\mathcal{Y} = \{1, 2, \dots, C\}$. In the long-tailed recognition task, the training set follows an imbalanced distribution \mathcal{D} and consists of data pairs denoted as $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $y_i \in \mathcal{Y}$ is the label for sample $\mathbf{x}_i \in \mathcal{X}$, and n is the total number of training samples. Let \mathcal{D}_{bal} denote the uniform test distribution. Following prior work (Cao et al., 2019; Hong et al., 2021), given a class y , \mathcal{D} and \mathcal{D}_{bal} share the same class-conditional distribution, denoted as $\mathcal{D}_y \triangleq P(\mathbf{x}|y)$. We use n_y to represent the number of samples in the y -th class and $\pi_y = n_y/n$ to denote the ratio of the y -th class in the training set. Without loss of generality, we assume $n_1 \geq n_2 \geq \dots \geq n_C$, with $n_1 \gg n_C$.

The model parameters are denoted by \mathbf{w} , with a total of k parameters. The loss for sample (\mathbf{x}, y) is defined as $\ell(\mathbf{w}; \mathbf{x}, y)$. The training loss over dataset \mathcal{S} is given by $L_S(\mathbf{w}) \triangleq \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{w}; \mathbf{x}_i, y_i)$. Similarly, the loss specifically for samples from the y -th class within \mathcal{S} is defined as $L_S^y(\mathbf{w}) \triangleq \frac{1}{n_y} \sum_{y_i=y} \ell(\mathbf{w}; \mathbf{x}_i, y_i)$. We further define the expected loss over \mathcal{D} , \mathcal{D}_{bal} and \mathcal{D}_y as $L_{\mathcal{D}}(\mathbf{w}) \triangleq \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} [\ell(\mathbf{w}; \mathbf{x}, y)]$, $L_{\mathcal{D}_{bal}}(\mathbf{w}) \triangleq \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_{bal}} [\ell(\mathbf{w}; \mathbf{x}, y)]$ and $L_{\mathcal{D}_y}(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_y} [\ell(\mathbf{w}; \mathbf{x}, y)]$, respectively. Our goal is to optimize parameters \mathbf{w} on dataset \mathcal{S} such that $L_{\mathcal{D}_{bal}}(\mathbf{w})$ is minimized, leading to good performance on the balanced test set.

3.2. Limitations in ImbSAM and CC-SAM

ImbSAM. ImbSAM divides classes into head and tail groups, denoted as \mathcal{H} and \mathcal{T} . It applies SAM only to the tail group to focus on flattening loss landscape for these classes. Its objective function is:

$$L_S^{IS}(\mathbf{w}) \triangleq L_S^{\mathcal{H}}(\mathbf{w}) + \max_{\|\epsilon\|_2 \leq \rho} L_S^{\mathcal{T}}(\mathbf{w} + \epsilon) \quad (1)$$

From Eq.(1), ImbSAM excludes all head classes from SAM. As a result, **the loss landscape for head classes becomes sharper**, which may reduce their generalization performance. To validate this, we analyze the spectral density of the Hessian H (Ghorbani et al., 2019), a common measure for the flatness of the loss landscape. We also consider two key metrics: the largest eigenvalue λ_{max} and the trace $Tr(H)$. Higher values of λ_{max} and $Tr(H)$ generally indicate a sharper loss landscape. Following prior work (Rangwani et al., 2022), we compute the eigen spectral density of the Hessian for head and tail classes on the CIFAR-10 LT dataset using the VS loss function. The results are shown in Fig.2.

A comparison between Fig.2(e) and Fig.2(f) reveals that ImbSAM effectively reduces $Tr(H)$ and λ_{max} for the tail classes, suggesting a flatter loss landscape. However, when comparing Fig.2(a) and Fig.2(b), we observe that with ImbSAM, the values of $Tr(H)$ and λ_{max} for head classes are significantly higher. This indicates that ImbSAM's exclusion of head classes from SAM sharpens their loss landscape, potentially degrading their generalization performance.

CC-SAM. CC-SAM applies SAM to each class individually, using class-specific perturbation radii. The objective

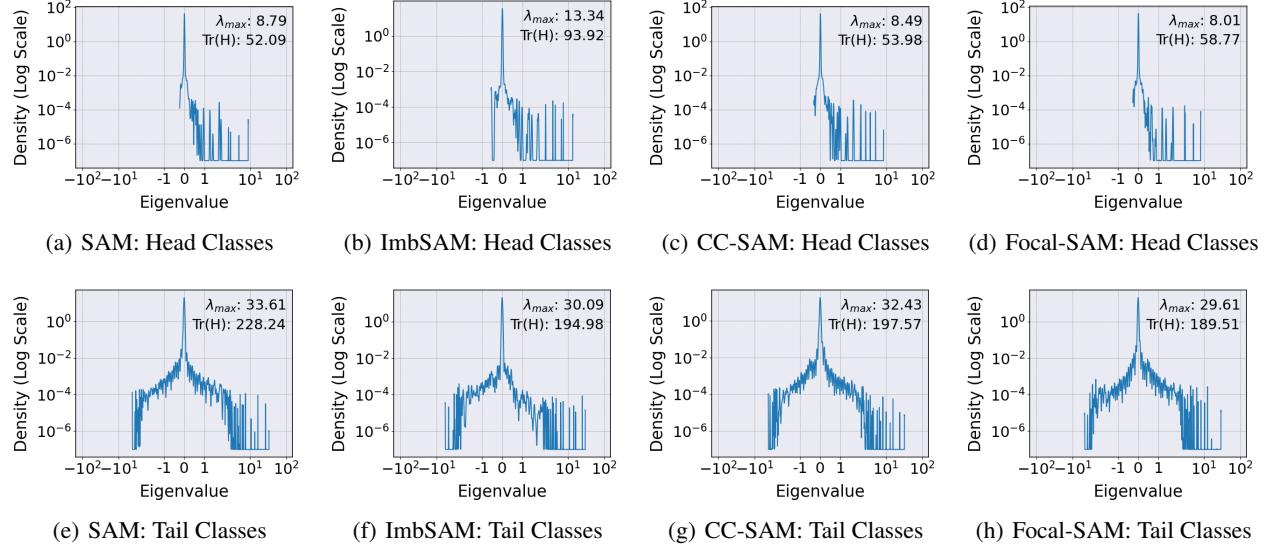


Figure 2: Eigen Spectral Density of Hessian for head and tail classes of ResNet models trained with various SAM variants on CIFAR-10 LT using VS loss. A smaller λ_{\max} and $Tr(H)$ generally indicate a flatter loss landscape.

Table 1: Average training time per epoch (in seconds) for different SAM variants across four long-tailed datasets using ResNet models. For CC-SAM, we follow its protocol by perturbing only the last few layers to improve its efficiency.

Methods	CIFAR-10 LT	CIFAR-100 LT	ImageNet-LT	iNaturalist
SAM	5.66s (1.00×)	4.81s (1.00×)	170.04s (1.00×)	831.67s (1.00×)
ImbSAM	7.80s (1.37×)	6.68s (1.39×)	293.11s (1.72×)	1088.61s (1.31×)
CC-SAM	11.61s (2.05×)	19.70s (4.10×)	1626.54s (9.57×)	12869.89s (15.47×)
Focal-SAM (Ours)	7.67s (1.36×)	6.71s (1.40×)	291.05s (1.71×)	1068.92s (1.29×)

function is defined as:

$$L_S^{CS}(\mathbf{w}) \triangleq \sum_{i=1}^C \max_{\|\epsilon\|_2 \leq \rho_i^*} \frac{1}{\pi_i} \cdot L_S^i(\mathbf{w} + \epsilon) \quad (2)$$

The optimal perturbation $\hat{\epsilon}_i(\mathbf{w})$ for each class i is also class-wise and estimated as $\rho_i^* \nabla_{\mathbf{w}} L_S^i(\mathbf{w}) / \|\nabla_{\mathbf{w}} L_S^i(\mathbf{w})\|_2$. The model parameters are updated with the learning rate η as:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \sum_{i=1}^C \frac{1}{\pi_i} \cdot \nabla_{\mathbf{w}} L_S^i(\mathbf{w})|_{\mathbf{w}+\hat{\epsilon}_i(\mathbf{w})} \quad (3)$$

This fine-grained method flattens the loss landscape of head and tail classes more effectively, as shown in Fig.2(c) and Fig.2(g). However, **CC-SAM is much more computationally demanding than SAM**. According to Eq.(3), per parameters update requires computing the gradient for each class i 's loss at $\mathbf{w} + \hat{\epsilon}_i(\mathbf{w})$, i.e., $\nabla_{\mathbf{w}} L_S^i(\mathbf{w})|_{\mathbf{w}+\hat{\epsilon}_i(\mathbf{w})}$. Therefore, CC-SAM requires at least C backpropagations per update, whereas SAM only needs two. Thus, CC-SAM has a much higher computational cost than SAM. For details on the backpropagation requirements for SAM and ImbSAM, please see App.B.

To confirm this, we measure the average training time per epoch for various SAM variants across four datasets using ResNet models. For CC-SAM, we follow its protocol by perturbing only the last few layers to enhance efficiency. As shown in Tab.1, despite perturbing fewer parameters, CC-SAM takes about $2\sim 15\times$ more time than SAM, depending on the dataset. The training time ratio of CC-SAM to SAM grows with the number of classes in the batch. These high computational costs make CC-SAM particularly impractical for large-scale datasets or fine-tuning foundation models.

4. Methodology

4.1. Focal Sharpness-Aware Minimization

Motivated by the analysis in Sec.3, we develop a new method called Focal-SAM. This approach achieves fine-grained control over the flatness between head and tail classes while maintaining computational efficiency, as shown in Tab.1 and Fig.2.

To this end, we first introduce the concept of **class-wise**

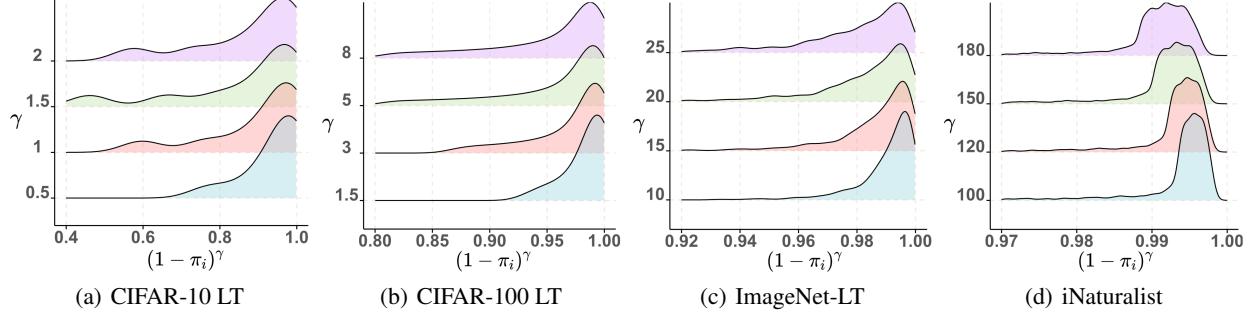


Figure 3: The probability density distributions of $(1 - \pi_i)^\gamma$ for various γ values on CIFAR-10 LT, CIFAR-100 LT, ImageNet-LT, and iNaturalist.

sharpness, defined as the loss difference between the original model parameters \mathbf{w} and the perturbed ones, to quantify the sharpness of loss landscapes across different classes:

$$\tilde{L}_S^i(\mathbf{w}, \epsilon) \triangleq L_S^i(\mathbf{w} + \epsilon) - L_S^i(\mathbf{w}), i \in \mathcal{Y}. \quad (4)$$

Next, we propose a new sharpness term called **focal sharpness**:

$$\tilde{L}_S^{FS}(\mathbf{w}) = \max_{\|\epsilon\|_2 \leq \rho} \sum_{i=1}^C (1 - \pi_i)^\gamma \tilde{L}_S^i(\mathbf{w}, \epsilon), \quad (5)$$

where $(1 - \pi_i)^\gamma$ is the focal weight that provides fine-grained control over class-wise sharpness, and γ is a tunable hyperparameter. When γ increases, the distribution of focal weight $(1 - \pi_i)^\gamma$ will skew more to tail classes. Fig.3 illustrates how the probability density distributions of $(1 - \pi_i)^\gamma$ varies with respect to γ on various long-tailed datasets.

Then, the objective of Focal-SAM is defined by the combination of the training loss and the focal sharpness term:

$$L_S^{FS}(\mathbf{w}) = L_S(\mathbf{w}) + \lambda \cdot \tilde{L}_S^{FS}(\mathbf{w}), \quad (6)$$

where λ is a hyperparameter controlling the importance of focal sharpness. This formulation highlights how Focal-SAM overcomes ImbSAM’s limitations. When $\gamma = 0$ and $\lambda = 1$, Eq.(5) penalizes the sharpness of each class equally, reverting to standard SAM. Conversely, when γ is sufficiently large, focal weights for head classes rapidly approach 0, while the weights for tail classes remain relatively large. In this scenario, Focal-SAM approximates ImbSAM. Typically, we select a moderate γ , such that the focal weights increase smoothly from head to tail classes. This fine-grained control over loss landscape improves the flatness of tail classes while maintaining that of head classes, ultimately enhancing generalization for both traditional and foundation models.

4.2. Optimizing the Focal-SAM Objective Function

In this section, we discuss how to optimize the Focal-SAM objective $L_S^{FS}(\mathbf{w})$. Let $L_S^\gamma(\mathbf{w}) \triangleq \sum_{i=1}^C (1 - \pi_i)^\gamma L_S^i(\mathbf{w})$.

Using a first-order Taylor expansion, we approximate the solution of the inner maximization problem for $\tilde{L}_S^{FS}(\mathbf{w})$:

$$\hat{\epsilon}(\mathbf{w}) \approx \underset{\|\epsilon\|_2 \leq \rho}{\operatorname{argmax}} \epsilon^T \nabla_{\mathbf{w}} L_S^\gamma(\mathbf{w}) = \rho \frac{\nabla_{\mathbf{w}} L_S^\gamma(\mathbf{w})}{\|\nabla_{\mathbf{w}} L_S^\gamma(\mathbf{w})\|_2} \quad (7)$$

Then, we can substitute ϵ and compute the gradients of $L_S^{FS}(\mathbf{w})$ to solve the outer minimization problem:

$$\begin{aligned} \nabla_{\mathbf{w}} L_S^{FS}(\mathbf{w}) &\approx \nabla_{\mathbf{w}} (L_S(\mathbf{w}) + \lambda [L_S^\gamma(\mathbf{w} + \hat{\epsilon}(\mathbf{w})) - L_S^\gamma(\mathbf{w})]) \\ &\approx \nabla_{\mathbf{w}} (L_S(\mathbf{w}) - \lambda L_S^\gamma(\mathbf{w}))|_{\mathbf{w}} + \lambda \nabla_{\mathbf{w}} L_S^\gamma(\mathbf{w})|_{\mathbf{w} + \hat{\epsilon}(\mathbf{w})} \end{aligned} \quad (8)$$

From Eq.(7) and Eq.(8), computing $\nabla_{\mathbf{w}} L_S^{FS}(\mathbf{w})$ to update model parameters requires only three backpropagations: one for $\nabla_{\mathbf{w}} L_S^\gamma(\mathbf{w})$, one for $\nabla_{\mathbf{w}} (L_S(\mathbf{w}) - \lambda L_S^\gamma(\mathbf{w}))|_{\mathbf{w}}$, and one for $\nabla_{\mathbf{w}} L_S^\gamma(\mathbf{w})|_{\mathbf{w} + \hat{\epsilon}(\mathbf{w})}$. Therefore, Focal-SAM is more computationally efficient than CC-SAM, making it more suitable for large-scale datasets or fine-tuning foundation models.

Overall, Alg.1 gives the pseudo-code to optimize the Focal-SAM objective, using SGD as the base optimizer.

Algorithm 1 Focal-SAM algorithm

Input: Training set S , perturbation radius ρ , hyperparameter λ , γ , learning rate η

Output: Model trained with Focal-SAM

- 1: Initialize weights \mathbf{w}_0 , $t = 0$;
 - 2: **while** not converged **do**
 - 3: Sample batch $B = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_b, y_b)\}$;
 - 4: Compute $L_B^\gamma(\mathbf{w})$;
 - 5: Compute $\nabla_{\mathbf{w}} L_B^\gamma(\mathbf{w})$ and $\hat{\epsilon}(\mathbf{w})$ according to Eq.(7);
 - 6: Perturb \mathbf{w} with $\hat{\epsilon}(\mathbf{w})$, and compute gradient $\mathbf{g}_1 = \nabla_{\mathbf{w}} L_B^\gamma(\mathbf{w})|_{\mathbf{w} + \hat{\epsilon}(\mathbf{w})}$;
 - 7: Compute gradient $\mathbf{g}_2 = \nabla_{\mathbf{w}} [L_B(\mathbf{w}) - \lambda \cdot L_B^\gamma(\mathbf{w})]|_{\mathbf{w}}$;
 - 8: Update weights: $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta(\lambda \mathbf{g}_1 + \mathbf{g}_2)$;
 - 9: $t = t + 1$;
 - 10: **end while**
-

4.3. Generalization Ability of Focal-SAM

Previous works have established the generalization bound for SAM (Foret et al., 2021) and CC-SAM (Zhou et al., 2023b). However, these bounds are relatively loose (with an order of $1/\sqrt{n}$) and could bias the training process. For example, the perturbation radius of CC-SAM (*i.e.*, ρ_i in Eq.(2)) is set as the solution to minimizing its PAC-Bayesian bound. Since the generalization is not sharp enough, the estimated perturbation radius ρ_i^* could deviate from the optimal one, thus leading to inferior performance. In this section, we develop a sharper generalization bound with an order of $1/n$ for Focal-SAM.

We assume the loss function $\ell(\mathbf{w}; \mathbf{x}, y)$ has an upper bound of B , which is a common and practical assumption. Then, we derive the following generalization bound based on the PAC-Bayesian theorem proposed in (Tolstikhin & Seldin, 2013). For conciseness, we present an informal formulation in the main content, leaving the formal one and the corresponding proof in App.A.

Theorem 4.1 (Informal). Assume that $\forall (\mathbf{x}, y) \in \mathcal{D}, 0 \leq \ell(\mathbf{w}; \mathbf{x}, y) \leq B$. For any $\rho > 0$, any uniform distribution \mathcal{D}_{bal} and any distribution \mathcal{D} , with high probability over the choice of the training set $S \sim \mathcal{D}$,

$$L_{\mathcal{D}_{bal}}(\mathbf{w}) \leq \underbrace{\frac{2L_S^{FS}(\mathbf{w})}{C\pi_C}}_{(I)} - \underbrace{\mathcal{O}\left(\frac{\lambda\rho^2}{k + \ln(n)} \cdot \text{tr}(H(\mathbf{w}))\right)}_{(II)} + \underbrace{\tilde{\mathcal{O}}\left(\frac{\lambda[k \log(\|\mathbf{w}\|_2^2/\rho^2) + \Psi]}{n}\right)}_{(III)}. \quad (9)$$

where $n = |S|$, $\Psi \triangleq \sum_{i=1}^C (1 - \pi_i)^\gamma \pi_i$, k is the number of parameters, $H(\mathbf{w})$ represents the Hessian matrix of $L_{\mathcal{D}}^\gamma(\mathbf{w})$ at point \mathbf{w} and $\text{tr}(\cdot)$ represents the matrix trace.

From the theorem, we have the following insights:

- The generalization bound consists of three components. Specifically, (I) is the empirical loss on the training set $L_S^{FS}(\mathbf{w})$, which can be minimized via large-scale models. (II) reveals how the generalization performance is affected by multiple factors, including $\lambda, \rho, \text{tr}(H(\mathbf{w}))$. (III) decreases at a faster rate of $\tilde{\mathcal{O}}(1/n)$.
- The hyperparameters λ and γ play a crucial role. On the one hand, a larger λ can increase both components (II) and (III) of the bound. Therefore, careful tuning of λ can induce a tighter bound. On the other hand, a larger γ leads to a smaller Ψ , also leading to a tighter bound. This suggests that assigning greater weights to the sharpness of the tail classes can improve the overall generalization ability.

- Focal-SAM enables a more effective optimization of $L_{\mathcal{D}_{bal}}(\mathbf{w})$. Specifically, we can reformulate Eq.(9) to

$$L_{\mathcal{D}_{bal}}(\mathbf{w}) + (II) \leq (I) + (III). \quad (10)$$

Typically, (II) tends to be large without SAM-based techniques. As a result, minimizing the right-hand side (RHS) of Eq.(10) in such cases may not induce a small $L_{\mathcal{D}_{bal}}(\mathbf{w})$. In contrast, Focal-SAM reduces the trace $\text{tr}(H(\mathbf{w}))$ by effectively flattening the loss landscape, leading to a small (II). This makes it more effective to minimize $L_{\mathcal{D}_{bal}}(\mathbf{w})$ when we optimize the RHS of Eq.(10). This insight again validates the necessity of Focal-SAM.

5. Experiments

This section evaluates the effectiveness of Focal-SAM through a series of experiments. **Detailed experimental settings and additional results are provided in App.C and App.D due to space constraints.**

5.1. Experiment Protocols

Datasets. We use four widely adopted long-tailed datasets for long-tailed recognition tasks: CIFAR-10 LT (Cao et al., 2019), CIFAR-100 LT (Cao et al., 2019), ImageNet-LT (Liu et al., 2019) and iNaturalist (Horn et al., 2018). The CIFAR-LT datasets include variants with imbalance ratios of $\{200, 100, 50, 10\}$. In addition to evaluating model performance on ID test sets, we also assess it on OOD test sets, referred to as long-tailed domain generalization tasks. Specifically, we train the model on ImageNet-LT and evaluate it on three OOD datasets: ImageNet-Sketch (Wang et al., 2019a), ImageNetV2 (Recht et al., 2019), and ImageNet-C (Hendrycks & Dietterich, 2019). For more details, see App.C.1.

Competitors. When training ResNet models on the CIFAR-LT dataset, we assess several loss functions. These methods are further combined with SAM (Foret et al., 2021), ImbSAM (Zhou et al., 2023a), and CC-SAM (Zhou et al., 2023b) as baselines. For the ImageNet-LT and iNaturalist datasets, we employ a range of representative methods as baseline methods. When fine-tuning the foundation model CLIP (Radford et al., 2021), we evaluate both full fine-tuning with LA loss (denoted as FFT) and parameter-efficient fine-tuning using the LIFT method (Shi et al., 2024), along with their performance when combined with different SAM variants. For more details, please refer to App.C.2.

Evaluation Protocol. For long-tailed recognition tasks, we assess model performance using balanced accuracy (Menon et al., 2021). To provide deeper insights, we split the classes into three groups: Head, Medium, and Tail, and report accuracy for each group individually. For long-tailed domain generalization tasks, we evaluate performance on OOD balanced test sets, including top-1 accuracy and accuracy for

Table 2: Performance comparison on CIFAR-100 LT datasets with various imbalance ratios (IR). FFT denotes fully fine-tuning the foundation model with LA loss. **Due to space limitations, additional CIFAR-100 LT results combining more methods, as well as the CIFAR-10 LT results, are shown in Tab.6 and Tab.5.**

Method	IR100				IR200 All	IR50 All	IR10 All
	Head	Med	Tail	All			
Training from scratch							
CE	69.2	41.6	9.0	41.5	37.5	45.6	58.1
CE+SAM	72.7	41.8	7.0	42.2	38.9	46.8	59.7
CE+ImbSAM	68.5	46.0	9.6	43.0	38.7	47.8	60.1
CE+CC-SAM	70.1	44.2	9.0	42.7	39.1	47.4	60.0
CE+Focal-SAM	73.8	44.2	8.9	44.0	39.6	48.1	60.9
LA (Menon et al., 2021)	61.3	42.3	28.6	44.9	41.8	50.3	59.4
LA+SAM	63.1	52.2	32.2	50.0	45.5	52.8	62.6
LA+ImbSAM	57.4	51.1	31.0	47.3	43.4	52.2	62.4
LA+CC-SAM	63.7	51.9	32.3	50.1	45.6	53.0	63.0
LA+Focal-SAM	63.9	53.0	32.5	50.7	46.0	54.5	63.8
Fine-tuning foundation model							
FFT	88.2	79.3	66.1	78.5	76.3	81.2	85.5
FFT+SAM	87.9	82.5	70.8	80.9	77.7	83.4	86.8
FFT+ImbSAM	87.5	82.0	70.2	80.4	77.2	81.9	86.7
FFT+CC-SAM	87.8	82.9	70.9	81.0	78.2	83.5	87.0
FFT+Focal-SAM	88.1	82.8	72.4	81.6	79.0	83.9	87.3
LIFT (Shi et al., 2024)	85.3	81.1	79.2	82.0	79.6	82.8	85.0
LIFT+SAM	85.0	81.5	79.4	82.1	79.6	83.0	85.1
LIFT+ImbSAM	84.7	81.9	78.9	82.0	79.8	83.1	85.2
LIFT+CC-SAM	84.8	81.8	79.0	82.0	79.7	83.1	85.2
LIFT+Focal-SAM	85.4	81.9	79.4	82.4	80.0	83.2	85.4

each class group. For more details of the evaluation protocol, please refer to App.C.3.

Implementation Details. For CIFAR-LT datasets, we train ResNet models using ResNet-32 (He et al., 2016) as the backbone. For ImageNet-LT and iNaturalist datasets, we employ ResNet-50 (He et al., 2016). Training is conducted for 200 epochs. For fine-tuning foundation models, we follow the protocols outlined in LIFT (Shi et al., 2024). Specifically, we fine-tune the image encoder of CLIP (Radford et al., 2021) with a ViT-B/16 (Dosovitskiy et al., 2021) backbone. The training lasts for 20 epochs. For further implementation details, please refer to App.C.4.

5.2. Performance Comparison

Tab.2 summarizes the experimental results on the CIFAR-LT datasets with different imbalance ratios. From these results, we have the following observations: 1) Focal-SAM consistently performs better than SAM, ImbSAM, and CC-SAM across various loss functions. 2) Focal-SAM significantly outperforms ImbSAM on head classes, while maintaining

or surpassing ImbSAM on tail classes. Additionally, Focal-SAM generally outperforms CC-SAM on both head and tail classes, showing its ability to achieve a finer balance between head and tail classes performance.

Tab.3 presents results on the larger ImageNet-LT and iNaturalist datasets. Combining the baseline LA with Focal-SAM improves performance by approximately 1.9%~2.3% when training ResNet models. Similarly, pairing the baseline FFT or LIFT with Focal-SAM yields a performance gain of 0.3%~2.4% when fine-tuning foundation models, outperforming several competitors.

5.3. Long-tailed Domain Generalization

In Tab.4, we evaluate the model trained on the ImageNet-LT dataset across three OOD datasets. The results show the following: 1) SAM-based methods, when combined with FFT or LIFT, generally achieve more performance gain on OOD datasets than on the ID dataset (ImageNet-LT). This observation aligns with prior studies (Zhou et al., 2022; Khattak et al., 2023; Park et al., 2024), which suggest that

Table 3: Performance comparison on ImageNet-LT and iNaturalist. The results for methods marked with \dagger are taken from the original paper. “-” indicates that the original paper didn’t report the corresponding results.

Method	ImageNet-LT				iNaturalist			
	Head	Med	Tail	All	Head	Med	Tail	All
Training from scratch								
CB (Cui et al., 2019) \dagger	39.6	32.7	16.8	33.2	53.4	54.8	53.2	54.0
cRT (Kang et al., 2020) \dagger	61.8	46.2	27.3	49.6	69.0	66.0	63.2	65.2
DiVE (He et al., 2021) \dagger	64.1	50.4	30.7	49.4	70.6	70.0	67.6	69.1
DRO-LT (Samuel & Chechik, 2021) \dagger	64.0	49.8	33.1	53.5	-	-	-	69.7
DisAlign (Zhang et al., 2021) \dagger	61.3	52.2	31.4	52.9	69.0	71.1	70.2	70.6
WB (Alshammari et al., 2022) \dagger	62.5	50.4	41.5	53.9	71.2	70.4	69.7	70.2
CC-SAM (Zhou et al., 2023b) \dagger	61.4	49.5	37.1	52.4	65.4	70.9	72.2	70.9
LA (Menon et al., 2021)	62.8	49.0	31.8	52.0	68.4	69.4	69.2	69.2
LA+SAM	63.1	51.7	33.1	53.6	68.3	70.8	71.9	71.0
LA+ImbSAM	62.6	50.3	32.6	52.6	68.0	70.2	70.2	69.9
LA+Focal-SAM	63.9	52.2	34.4	54.3	68.4	72.0	72.5	71.8
Fine-tuning foundation model								
Decoder (Wang et al., 2024c) \dagger	-	-	-	73.2	-	-	-	59.2
LPT (Dong et al., 2023) \dagger	-	-	-	-	-	-	79.3	76.1
FFT	79.9	70.5	51.0	71.5	69.7	71.9	71.7	71.6
FFT+SAM	80.9	72.9	54.3	73.5	69.5	74.4	74.4	73.8
FFT+ImbSAM	80.6	72.6	52.2	72.9	68.5	73.4	73.8	73.1
FFT+CC-SAM	80.6	73.6	54.2	73.6	69.2	74.1	74.2	73.6
FFT+Focal-SAM	80.8	73.9	54.4	73.9	69.1	74.7	74.3	74.0
LIFT (Shi et al., 2024)	79.7	76.2	72.8	77.1	74.1	79.4	81.5	79.7
LIFT+SAM	79.9	76.4	72.7	77.2	73.5	79.7	81.6	79.8
LIFT+ImbSAM	79.8	76.4	72.5	77.2	73.2	79.5	81.4	79.6
LIFT+CC-SAM	79.8	76.4	73.3	77.3	74.0	79.4	81.5	79.7
LIFT+Focal-SAM	79.7	76.6	73.6	77.4	73.9	79.8	81.7	80.0

fine-tuning foundation models often perform well on target (ID) datasets but struggles with unseen (OOD) datasets. 2) Focal-SAM achieves a performance improvement of 0.5% to 4.3%, surpassing SAM, ImbSAM, and CC-SAM. This is because Focal-SAM effectively enhances the model’s generalization ability by flattening the loss landscape, which mitigates performance issues on OOD test sets.

5.4. Training Speed of Focal-SAM

To assess the computational efficiency of Focal-SAM, we evaluate the training time per epoch across various long-tailed datasets, as shown in Tab.1. Focal-SAM requires about 50% more running time than SAM and has a similar running time to ImbSAM. Given that our method consistently outperforms SAM and ImbSAM, thus the computational cost is acceptable for the performance gain. Furthermore, Focal-SAM is significantly faster than CC-SAM while delivering better performance, aligning with our goal

of improving CC-SAM’s efficiency.

5.5. Sharpness of Loss Landscape for Focal-SAM

To examine the impact of Focal-SAM on the loss landscape, Fig.2(d) and Fig.2(h) show the eigenvalue spectrum of Hessian for head and tail classes of models trained with Focal-SAM on CIFAR-10 LT using the VS loss function. Comparing Fig.2(e) and Fig.2(h), the trace $Tr(H)$ and the maximum eigenvalue λ_{max} for tail classes in Focal-SAM are significantly lower than those in SAM. Similarly, Fig.2(b) and Fig.2(d) reveal that $Tr(H)$ and λ_{max} for head classes in Focal-SAM are much smaller than those in ImbSAM. These results suggest that Focal-SAM achieves a fine-grained balance in the flatness between head and tail classes.

5.6. Ablation Study About γ and λ

We analyze the influence of hyperparameters γ and λ to Focal-SAM on the CIFAR-LT datasets.

Table 4: Performance comparison for domain generalization. The source models are trained on the ImageNet-LT dataset and evaluated on out-of-distribution datasets, including ImageNet-Sketch, ImageNetV2, and ImageNet-C.

Method	ImageNet-Sketch				ImageNetV2				ImageNet-C			
	Head	Med	Tail	All	Head	Med	Tail	All	Head	Med	Tail	All
FFT	42.9	35.5	21.4	36.4	70.1	60.2	45.2	62.0	50.3	41.4	26.1	42.8
FFT+SAM	44.9	39.3	26.1	39.6	71.2	62.6	48.0	63.9	52.5	44.6	29.3	45.6
FFT+ImbSAM	45.2	39.5	24.8	39.7	71.0	62.2	46.5	63.5	52.0	44.7	28.3	45.2
FFT+CC-SAM	45.0	41.0	26.8	40.6	71.3	63.2	48.4	64.3	52.0	45.1	29.4	45.6
FFT+Focal-SAM	45.5	41.2	27.3	41.0	71.8	63.6	48.8	64.8	52.6	45.5	29.8	46.1
LIFT (Shi et al., 2024)	46.4	43.3	45.7	44.8	70.4	65.9	64.7	67.5	52.6	48.7	47.3	50.0
LIFT+SAM	46.9	43.5	46.4	45.2	70.4	66.0	65.5	67.6	52.9	49.2	48.1	50.5
LIFT+ImbSAM	46.4	43.5	46.0	44.9	70.0	66.2	65.5	67.6	52.6	49.0	47.7	50.2
LIFT+CC-SAM	46.8	44.1	47.6	45.6	70.4	66.2	65.4	67.7	53.0	49.7	49.2	50.9
LIFT+Focal-SAM	46.9	44.7	49.4	46.2	70.0	66.8	66.9	68.0	53.1	49.9	49.8	51.1

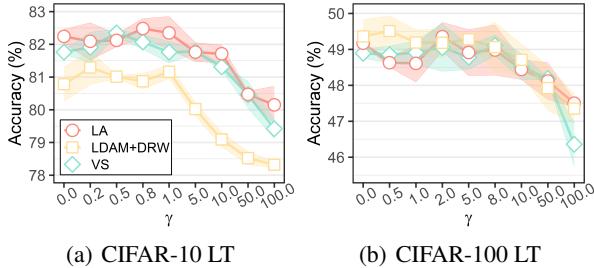


Figure 4: Ablation Study of Focal-SAM w.r.t. γ

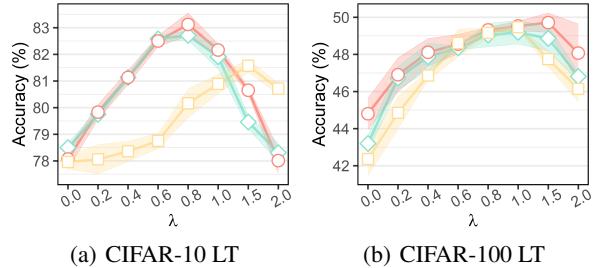


Figure 5: Ablation Study of Focal-SAM w.r.t. λ

Impact of γ : Fig.4 explores the effect of γ . As γ increases, performance initially improves, suggesting that assigning greater weight to the class-wise sharpness of tail classes benefits performance. However, a further increase in γ leads to declining accuracy, indicating that assigning excessive weight to the class-wise sharpness of tail classes can harm performance.

Impact of λ : Fig.5 investigates the effect of λ . As λ increases, accuracy initially improves but subsequently decreases. This indicates a trade-off between minimizing the training loss and minimizing the sharpness of the loss landscape.

6. Conclusion

This paper examines the limitations of ImbSAM and CC-SAM in long-tailed learning. ImbSAM excludes all head classes from SAM, often overemphasizing tail classes when combined with rebalancing methods. CC-SAM’s per-class perturbation strategy provides fine-grained control over the loss landscape but is computationally costly. To address these issues, we propose Focal-SAM, a method that efficiently balances loss landscape flatness between head and tail classes. Additionally, we offer a theoretical analysis of

Focal-SAM’s generalization ability, deriving a tighter bound. Extensive experiments validate Focal-SAM’s effectiveness.

Acknowledgements

This work was supported in part by the Fundamental Research Funds for the Central Universities, in part by the National Key R&D Program of China under Grant 2018AAA0102000, in part by National Natural Science Foundation of China: 62236008, 62441232, U21B2038, U23B2051, 62122075, 62025604, 62441619, 62206264 and 92370102, in part by Youth Innovation Promotion Association CAS, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No.XDB0680201, in part by the China National Postdoctoral Program for Innovative Talents under Grant BX20240384.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

- Ahn, S., Ko, J., and Yun, S. CUDA: curriculum of data augmentation for long-tailed recognition. In *International Conference on Learning Representations*, 2023.
- Aimar, E. S., Jonnarth, A., Felsberg, M., and Kuhlmann, M. Balanced product of calibrated experts for long-tailed recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19967–19977, 2023.
- Alshammari, S., Wang, Y., Ramanan, D., and Kong, S. Long-tailed recognition via weight balancing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6887–6897, 2022.
- Bernstein, S. On a modification of chebyshev’s inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, pp. 38–49, 1924.
- Buda, M., Maki, A., and Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, pp. 249–259, 2018.
- Cao, K., Wei, C., Gaidon, A., Aréchiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Annual Conference on Neural Information Processing Systems*, pp. 1565–1576, 2019.
- Cui, J., Zhong, Z., Liu, S., Yu, B., and Jia, J. Parametric contrastive learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 695–704, 2021.
- Cui, J., Zhong, Z., Tian, Z., Liu, S., Yu, B., and Jia, J. Generalized parametric contrastive learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 7463–7474, 2024.
- Cui, Y., Jia, M., Lin, T., Song, Y., and Belongie, S. J. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019.
- Dai, S., Xu, Q., Yang, Z., Cao, X., and Huang, Q. DRAUC: an instance-wise distributionally robust AUC optimization framework. In *Annual Conference on Neural Information Processing Systems*, 2023.
- Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.
- Dong, B., Zhou, P., Yan, S., and Zuo, W. LPT: long-tailed prompt tuning for image classification. In *International Conference on Learning Representations*, 2023.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.
- Gao, P., Xu, Q., Wen, P., Yang, Z., Shao, H., and Huang, Q. Feature directions matter: Long-tailed learning via rotated balanced representation. In *International Conference on Machine Learning*, pp. 27542–27563, 2023.
- Ghorbani, B., Krishnan, S., and Xiao, Y. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pp. 2232–2241, 2019.
- Han, B., Xu, Q., Yang, Z., Bao, S., Wen, P., Jiang, Y., and Huang, Q. Aucseg: Auc-oriented pixel-level long-tail semantic segmentation. In *Annual Conference on Neural Information Processing Systems*, 2024.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.
- He, Y., Wu, J., and Wei, X. Distilling virtual examples for long-tailed recognition. In *IEEE/CVF International Conference on Computer Vision*, pp. 235–244, 2021.
- Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.
- Hong, F., Yao, J., Lyu, Y., Zhou, Z., Tsang, I. W., Zhang, Y., and Wang, Y. On harmonizing implicit subpopulations. In *International Conference on Learning Representations*, 2024.
- Hong, Y., Han, S., Choi, K., Seo, S., Kim, B., and Chang, B. Disentangling label distribution for long-tailed visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6626–6636, 2021.
- Hong, Y., Zhang, J., Sun, Z., and Yan, K. Safa: Sample-adaptive feature augmentation for long-tailed image classification. In *European Conference on Computer Vision*, pp. 587–603, 2022.
- Horn, G. V., Aodha, O. M., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. J. The inaturalist species classification and detection dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

- Hou, W., Xu, Q., Yang, Z., Bao, S., He, Y., and Huang, Q. Adauc: End-to-end adversarial AUC optimization against long-tail problems. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8903–8925, 2022.
- Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020.
- Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.
- Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.
- Khattak, M. U., Wasim, S. T., Naseer, M., Khan, S., Yang, M., and Khan, F. S. Self-regulating prompts: Foundational model adaptation without forgetting. In *IEEE/CVF International Conference on Computer Vision*, pp. 15144–15154, 2023.
- Kim, J., Jeong, J., and Shin, J. M2m: Imbalanced classification via major-to-minor translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13893–13902, 2020.
- Kini, G. R., Paraskevas, O., Oymak, S., and Thrampoulidis, C. Label-imbalanced and group-sensitive classification under overparameterization. In *Annual Conference on Neural Information Processing Systems*, pp. 18970–18983, 2021.
- Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Master’s thesis, Department of Computer Science, University of Toronto, 2009.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Annual Conference on Neural Information Processing Systems*, pp. 1106–1114, 2012.
- Li, J., Tan, Z., Wan, J., Lei, Z., and Guo, G. Nested collaborative learning for long-tailed visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6939–6948, 2022.
- Lin, T., Goyal, P., Girshick, R. B., He, K., and Dollár, P. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pp. 2999–3007, 2017.
- Liu, B., Li, H., Kang, H., Hua, G., and Vasconcelos, N. Breadcrumbs: Adversarial class-balanced sampling for long-tailed recognition. In *European Conference on Computer Vision*, pp. 637–653, 2022.
- Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2537–2546, 2019.
- Lyu, X., Xu, Q., Yang, Z., Lyu, S., and Huang, Q. SSE-SAM: balancing head and tail classes gradually through stage-wise SAM. In *Association for the Advancement of Artificial Intelligence*, pp. 19278–19286, 2025.
- Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.
- Park, J., Ko, J., and Kim, H. J. Prompt learning via meta-regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26930–26940, 2024.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.
- Rangwani, H., Aithal, S. K., Mishra, M., and R., V. B. Escaping saddle points for effective generalization on class-imbalanced data. In *Annual Conference on Neural Information Processing Systems*, pp. 22791–22805, 2022.
- Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400, 2019.
- Ren, J., Yu, C., Sheng, S., Ma, X., Zhao, H., Yi, S., and Li, H. Balanced meta-softmax for long-tailed visual recognition. In *Annual Conference on Neural Information Processing Systems*, pp. 4175–4186, 2020.
- Ren, S., He, K., Girshick, R. B., and Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. In *Annual Conference on Neural Information Processing Systems*, pp. 91–99, 2015.
- Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9351, pp. 234–241, 2015.
- Samuel, D. and Chechik, G. Distributional robustness loss for long-tail learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 9475–9484, 2021.

- Shao, H., Xu, Q., Yang, Z., Wen, P., Gao, P., and Huang, Q. Weighted ROC curve in cost space: Extending AUC to cost-sensitive learning. In *Annual Conference on Neural Information Processing Systems*, 2023.
- Shi, J., Wei, T., Zhou, Z., Shao, J., Han, X., and Li, Y. Long-tail learning with foundation model: Heavy fine-tuning hurts. In *International Conference on Machine Learning*, pp. 45014–45039, 2024.
- Tolstikhin, I. O. and Seldin, Y. Pac-bayes-empirical-bernstein inequality. In *Annual Conference on Neural Information Processing Systems*, pp. 109–117, 2013.
- Wang, B., Wang, P., Xu, W., Wang, X., Zhang, Y., Wang, K., and Wang, Y. Kill two birds with one stone: Rethinking data augmentation for deep long-tailed learning. In *International Conference on Learning Representations*, 2024a.
- Wang, H., Ge, S., Lipton, Z. C., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In *Annual Conference on Neural Information Processing Systems*, 2019a.
- Wang, P., Zhao, Z., Wen, H., Wang, F., Wang, B., Zhang, Q., and Wang, Y. Llm-autoda: Large language model-driven automatic data augmentation for long-tailed problems. In *Annual Conference on Neural Information Processing Systems*, 2024b.
- Wang, X., Lian, L., Miao, Z., Liu, Z., and Yu, S. X. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021.
- Wang, Y., Gan, W., Yang, J., Wu, W., and Yan, J. Dynamic curriculum learning for imbalanced data classification. In *IEEE/CVF International Conference on Computer Vision*, pp. 5016–5025, 2019b.
- Wang, Y., Yu, Z., Wang, J., Heng, Q., Chen, H., Ye, W., Xie, R., Xie, X., and Zhang, S. Exploring vision-language models for imbalanced learning. *Int. J. Comput. Vis.*, pp. 224–237, 2024c.
- Wang, Z., Xu, Q., Yang, Z., He, Y., Cao, X., and Huang, Q. Openauc: Towards auc-oriented open-set recognition. In *Annual Conference on Neural Information Processing Systems*, 2022.
- Wang, Z., Xu, Q., Yang, Z., He, Y., Cao, X., and Huang, Q. A unified generalization analysis of re-weighting and logit-adjustment for imbalanced learning. In *Annual Conference on Neural Information Processing Systems*, pp. 48417–48430, 2023.
- Yang, Z., Xu, Q., Bao, S., Cao, X., and Huang, Q. Learning with multiclass AUC: theory and algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):7747–7763, 2022.
- Yang, Z., Xu, Q., Bao, S., Wen, P., He, Y., Cao, X., and Huang, Q. Auc-oriented domain adaptation: From theory to algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12):14161–14174, 2023a.
- Yang, Z., Xu, Q., Hou, W., Bao, S., He, Y., Cao, X., and Huang, Q. Revisiting auc-oriented adversarial training with loss-agnostic perturbations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12):15494–15511, 2023b.
- Yang, Z., Xu, Q., Wang, Z., Li, S., Han, B., Bao, S., Cao, X., and Huang, Q. Harnessing hierarchical label distribution variations in test agnostic long-tail recognition. In *International Conference on Machine Learning*, pp. 56624–56664, 2024.
- Zhang, C., Almanidis, G., Fan, G., Deng, B., Zhang, Y., Liu, J., Kamel, A., Soda, P., and Gama, J. A systematic review on long-tailed learning. *CoRR*, abs/2408.00483, 2024a.
- Zhang, S., Li, Z., Yan, S., He, X., and Sun, J. Distribution alignment: A unified framework for long-tail visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2361–2370, 2021.
- Zhang, T., Zheng, H., Yao, J., Wang, X., Zhou, M., Zhang, Y., and Wang, Y. Long-tailed diffusion models with oriented calibration. In *International Conference on Learning Representations*, 2024b.
- Zhang, Y., Hooi, B., Hong, L., and Feng, J. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. In *Annual Conference on Neural Information Processing Systems*, pp. 34077–34090, 2022.
- Zhang, Y., Kang, B., Hooi, B., Yan, S., and Feng, J. Deep long-tailed learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 10795–10816, 2023.
- Zhao, Z., Wang, P., Wen, H., Xu, W., Lai, S., Zhang, Q., and Wang, Y. Two fists, one heart: Multi-objective optimization based strategy fusion for long-tailed learning. In *International Conference on Machine Learning*, 2024a.
- Zhao, Z., Wen, H., Wang, Z., Wang, P., Wang, F., Lai, S., Zhang, Q., and Wang, Y. Breaking long-tailed learning bottlenecks: A controllable paradigm with hypernetwork-generated diverse experts. In *Annual Conference on Neural Information Processing Systems*, 2024b.
- Zhong, Z., Cui, J., Liu, S., and Jia, J. Improving calibration for long-tailed recognition. In *IEEE Conference on*

Computer Vision and Pattern Recognition, pp. 16489–16498, 2021.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16795–16804, 2022.

Zhou, Y., Qu, Y., Xu, X., and Shen, H. Imbsam: A closer look at sharpness-aware minimization in class-imbalanced recognition. In IEEE/CVF International Conference on Computer Vision, pp. 11311–11321, 2023a.

Zhou, Z., Li, L., Zhao, P., Heng, P., and Gong, W. Class-conditional sharpness-aware minimization for deep long-tailed recognition. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3499–3509, 2023b.

Zhu, J., Wang, Z., Chen, J., Chen, Y. P., and Jiang, Y. Balanced contrastive learning for long-tailed visual recognition. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6898–6907, 2022.

Appendix

Contents

A Proof of Theorem	15
A.1 Framework of the Proof	15
A.2 Proof of Lem.A.2	15
A.3 Proof of Lem.A.3 and Lem.A.4	16
A.4 Proof of Thm.4.1	20
B Analysis of Backpropagation Requirements for SAM and ImbSAM	20
B.1 Backpropagation Requirements for SAM	20
B.2 Backpropagation Requirements for ImbSAM	21
C More Experiment Protocols	21
C.1 Datasets	21
C.2 Competitors	21
C.3 Evaluation Protocol	22
C.4 Implementation Details	22
C.5 Experimental Hardware Setup	22
D More Experiment Results	23
D.1 Additional Results on the CIFAR-LT Datasets	23
D.2 CE and mCE Metrics on ImageNet-C for Long-tailed Domain Generalization	24
D.3 Results for Aligning Computational Cost	24
D.4 Visualization of Loss Landscape	25
D.5 Ablation Study About Perturbation Radius ρ	26
D.6 Additional Results for Eigen Spectral Density of Hessian	27

A. Proof of Theorem

A.1. Framework of the Proof

Goal. To bound the balanced loss $L_{\mathcal{D}_{bal}}(\mathbf{w})$ using our objective loss:

$$L_S^{FS}(\mathbf{w}) \triangleq \underbrace{[L_S(\mathbf{w}) - \lambda \cdot L_S^\gamma(\mathbf{w})]}_{(a)} + \lambda \cdot \underbrace{\max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} L_S^\gamma(\mathbf{w} + \boldsymbol{\epsilon})}_{(b)} \quad (11)$$

Framework of the proof.

- Essentially, the generalization bound describes how empirical values ($L_S^{FS}(\mathbf{w})$) deviate from the expected one ($L_{\mathcal{D}_{bal}}(\mathbf{w})$). To bound such deviations, Bernstein's inequality (Bernstein, 1924) and PAC-Bayesian theorem (Tolstikhin & Seldin, 2013) are convenient tools. Notice that these tools **require the empirical values to be sampled i.i.d. from the distribution on which the expectation is based**. Since the training set $S \sim D$, we first transform the distribution from \mathcal{D}_{bal} to D building on the work of Wang et al. (2023), i.e.,

$$L_{D_{bal}}(\mathbf{w}) \lesssim L_D(\mathbf{w}) \stackrel{\text{split into}}{=} \underbrace{[L_D(\mathbf{w}) - \lambda \cdot L_D^\gamma(\mathbf{w})]}_{(c)} + \underbrace{\lambda \cdot L_D^\gamma(\mathbf{w})}_{(d)} \quad (12)$$

- Get (c) \lesssim (a) via Bernstein's inequality (Lem.A.2).
- Get (d) \lesssim (b):
 - Bound (d) using a intermediate value $\mathbb{E}_\epsilon[L_D^\gamma(\mathbf{w} + \boldsymbol{\epsilon})]$ via Taylor expansion (Lem.A.4).
 - Bound $\mathbb{E}_\epsilon[L_D^\gamma(\mathbf{w} + \boldsymbol{\epsilon})]$ by (b) via PAC Bayesian bound (Lem.A.3).

Combine all, we get Thm.A.5 as follow:

$$L_{D_{bal}}(\mathbf{w}) \lesssim (c) + (d) \lesssim (a) + (b) = L_S^{FS}(\mathbf{w}) \quad (13)$$

A.2. Proof of Lem.A.2

We begin by introducing Bernstein's inequality to prove the first part.

Lemma A.1 (Bernstein's Inequality (Bernstein, 1924)). *Let X_1, \dots, X_n be i.i.d. random variables, $\mu = \mathbb{E}[X_1]$ and $\forall i, |X_i - \mu| \leq b$. Let $\sigma^2 = \text{Var}(X_i)$. With probability at least $1 - \delta$,*

$$|\bar{X}_n - \mu| \leq \sqrt{\frac{4\sigma^2 \log(\frac{2}{\delta})}{n}} + \frac{4b \log(\frac{2}{\delta})}{3n} \quad (14)$$

where $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

Employing Lem.A.1, we can derive the following lemma to bound $L_{\mathcal{D}}(\mathbf{w}) - \lambda \cdot L_{\mathcal{D}}^\gamma(\mathbf{w})$ by $L_S(\mathbf{w}) - \lambda \cdot L_S^\gamma(\mathbf{w})$.

Lemma A.2. *Assume that $\forall (\mathbf{x}, y) \in \mathcal{D}, 0 \leq \ell(\mathbf{w}; \mathbf{x}, y) \leq B$. With probability at least $1 - \delta$ over the choice of the training set $S \sim \mathcal{D}$*

$$\Phi_{\mathcal{D}}^\lambda(\mathbf{w}) \leq 2 \cdot \Phi_S^\lambda(\mathbf{w}) + \frac{40 \cdot (B + \lambda B') \cdot \log(\frac{2}{\delta})}{3n} \quad (15)$$

where $B' \triangleq \sum_{i=1}^C (1 - \pi_i)^\gamma \pi_i B$, $\Phi_{\mathcal{D}}^\lambda(\mathbf{w}) \triangleq L_{\mathcal{D}}(\mathbf{w}) - \lambda \cdot L_{\mathcal{D}}^\gamma(\mathbf{w})$ and $\Phi_S^\lambda(\mathbf{w}) \triangleq L_S(\mathbf{w}) - \lambda \cdot L_S^\gamma(\mathbf{w})$.

Proof. Since $\forall (\mathbf{x}, y) \in \mathcal{D}, \forall \mathbf{w} \in \mathcal{W}, \ell(\mathbf{w}; \mathbf{x}, y) \leq B$, we have

$$0 \leq L_S(\mathbf{w}) \leq B, 0 \leq L_{\mathcal{D}}(\mathbf{w}) \leq B \quad (16)$$

and

$$\begin{aligned} 0 \leq L_S^\gamma(\mathbf{w}) &= \sum_{i=1}^C (1 - \pi_i)^\gamma L_S^i(\mathbf{w}) \leq \sum_{i=1}^C (1 - \pi_i)^\gamma \pi_i B \triangleq B' \\ 0 \leq L_D^\gamma(\mathbf{w}) &= \mathbb{E}_S[L_S^\gamma(\mathbf{w})] \leq \sum_{i=1}^C (1 - \pi_i)^\gamma \pi_i B \triangleq B' \end{aligned} \quad (17)$$

By the above two inequalities, we can obtain

$$\begin{aligned} |\Phi_D^\lambda(\mathbf{w})| &\leq B + \lambda B' \\ |\Phi_S^\lambda(\mathbf{w})| &\leq B + \lambda B' \end{aligned} \quad (18)$$

Thus, we have

$$|\Phi_S^\lambda(\mathbf{w}) - \Phi_D^\lambda(\mathbf{w})| \leq 2 \cdot (B + \lambda B') \quad (19)$$

To simplify the analysis, we assume $\Phi_D^\lambda(\mathbf{w}) \geq 0$. This assumption is reasonable because our experiments in Sec.5.6 typically show that the best value for λ is slightly less than 1, where this assumption holds true. With this assumption, the variance of $\Phi_S^\lambda(\mathbf{w})$ can be bounded as:

$$\text{Var}(\Phi_S^\lambda(\mathbf{w})) \leq \mathbb{E}[(\Phi_S^\lambda(\mathbf{w}))^2] \leq 2 \cdot (B + \lambda B') \cdot \Phi_D^\lambda(\mathbf{w}) \quad (20)$$

Using Lem.A.1, with probability at least $1 - \delta$, we have

$$\begin{aligned} \Phi_D^\lambda(\mathbf{w}) &\leq \Phi_S^\lambda(\mathbf{w}) + \sqrt{\frac{8 \cdot (B + \lambda B') \cdot \Phi_D^\lambda(\mathbf{w}) \cdot \log(\frac{2}{\delta})}{n}} \\ &\quad + \frac{8 \cdot (B + \lambda B') \cdot \log(\frac{2}{\delta})}{3n} \\ &\leq \Phi_S^\lambda(\mathbf{w}) + \frac{1}{2} \cdot \Phi_D^\lambda(\mathbf{w}) + \frac{20 \cdot (B + \lambda B') \cdot \log(\frac{2}{\delta})}{3n} \end{aligned} \quad (21)$$

where the last inequality leverages the property that for any positive numbers a and b , $\sqrt{ab} \leq \frac{a}{2} + \frac{b}{2}$.

Reformulate the inequality, we can obtain that with probability at least $1 - \delta$,

$$\Phi_D^\lambda(\mathbf{w}) \leq 2 \cdot \Phi_S^\lambda(\mathbf{w}) + \frac{40 \cdot (B + \lambda B') \cdot \log(\frac{2}{\delta})}{3n} \quad (22)$$

□

A.3. Proof of Lem.A.3 and Lem.A.4

The following lemmas utilize the PAC-Bayesian theorem to prove the second part. We first derive an intermediate result in the following lemma.

Lemma A.3. Assume that $\forall (\mathbf{x}, y) \in \mathcal{D}, 0 \leq \ell(\mathbf{w}; \mathbf{x}, y) \leq B$. Then, for any $\rho > 0$ and any distribution \mathcal{D} , with probability $1 - \delta$ over the choice of the training set $S \sim \mathcal{D}$

$$\begin{aligned} \mathbb{E}_{\epsilon_i \sim \mathcal{N}(0, \sigma_Q)}[L_D^\gamma(\mathbf{w} + \boldsymbol{\epsilon})] &\leq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} 2L_S^\gamma(\mathbf{w} + \boldsymbol{\epsilon}) \\ &\quad + \frac{2 + 2B' + 2k \log\left(1 + \frac{\|\mathbf{w}\|_2^2}{k\rho^2}\right) + 4k \log\left(\sqrt{k} + \sqrt{2 \ln n}\right) + 4 \log \frac{\pi^2 \sqrt{n}(nB' + 1)^2}{3\delta}}{n} \end{aligned} \quad (23)$$

where $n = |S|$, k is the number of parameters, $B' \triangleq \sum_{i=1}^C (1 - \pi_i)^\gamma \pi_i B$ and $\sigma_Q \triangleq \frac{\rho}{\sqrt{k} + \sqrt{2 \ln(n)}}$.

Proof. Inspired by the proof technique in SAM (Foret et al., 2021), we provide the following proof.

Since $\forall (\mathbf{x}, y) \in \mathcal{D}, \forall \mathbf{w} \in \mathcal{W}, \ell(\mathbf{w}; \mathbf{x}, y) \leq B$, we have:

$$L_S^\gamma(\mathbf{w}) = \sum_{i=1}^C (1 - \pi_i)^\gamma L_S^i(\mathbf{w}) \leq \sum_{i=1}^C (1 - \pi_i)^\gamma \pi_i B = B' \quad (24)$$

$$L_D^\gamma(\mathbf{w}) = \mathbb{E}_S[L_S^\gamma(\mathbf{w})] \leq \sum_{i=1}^C (1 - \pi_i)^\gamma \pi_i B = B' \quad (25)$$

Thereby, the right-hand side of the bound in the theorem is lower bounded by $\frac{k}{n} \log(1 + \frac{\|\mathbf{w}\|_2^2}{k\rho^2})$ which is greater than B' when $\|\mathbf{w}\|_2^2 > k\rho^2[\exp(nB'/k) - 1]$ and in this case the inequality holds trivially. Thereby, we only consider the case when $\|\mathbf{w}\|_2^2 \leq k\rho^2[\exp(nB'/k) - 1]$ in the rest of the proof.

Using PAC-Bayesian generalization bound in (Tolstikhin & Seldin, 2013), for any fixed prior \mathcal{P} over parameters, with probability $1 - \delta$ over training set S , for any posterior \mathcal{Q} over parameters, the following generalization bound holds:

$$\begin{aligned} \mathbb{E}_{\mathbf{w} \sim \mathcal{Q}}[L_D^\gamma(\mathbf{w})] &\leq \mathbb{E}_{\mathbf{w} \sim \mathcal{Q}}[L_S^\gamma(\mathbf{w})] + \sqrt{2\mathbb{E}_{\mathbf{w} \sim \mathcal{Q}}[L_S^\gamma(\mathbf{w})] \frac{KL(\mathcal{Q}||\mathcal{P}) + \log \frac{2\sqrt{n}}{\delta}}{n}} \\ &\quad + 2 \frac{KL(\mathcal{Q}||\mathcal{P}) + \log \frac{2\sqrt{n}}{\delta}}{n} \\ &\leq 2\mathbb{E}_{\mathbf{w} \sim \mathcal{Q}}[L_S^\gamma(\mathbf{w})] + 4 \frac{KL(\mathcal{Q}||\mathcal{P}) + \log \frac{2\sqrt{n}}{\delta}}{n} \end{aligned} \quad (26)$$

where the last inequality leverages the property that for any positive numbers a and b , $\sqrt{ab} \leq a + b$.

Following SAM (Foret et al., 2021), we assume $\mathcal{P} = \mathcal{N}(\boldsymbol{\mu}_P, \sigma_P^2 \mathbf{I})$ and $\mathcal{Q} = \mathcal{N}(\boldsymbol{\mu}_Q, \sigma_Q^2 \mathbf{I})$, then the KL divergence can be written as:

$$KL(\mathcal{Q}||\mathcal{P}) = \frac{1}{2} \left[\frac{k\sigma_Q^2 + \|\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q\|_2^2}{\sigma_P^2} - k + k \log \left(\frac{\sigma_P^2}{\sigma_Q^2} \right) \right] \quad (27)$$

Let $T = \{c \exp((1-j)/k) | j \in \mathbb{N}\}$ be the predefined set of values for σ_P^2 . If for any $j \in \mathbb{N}$, the bounds holds with probability $1 - \delta_j$ with $\delta_j = \frac{6\delta}{\pi^2 j^2}$, then by the union bound, all above bounds hold simultaneously with probability $1 - \sum_{j=1}^{\infty} \frac{6\delta}{\pi^2 j^2} = 1 - \delta$.

Let $\sigma_Q = \frac{\rho}{\sqrt{k} + \sqrt{2 \ln(n)}}$, $\boldsymbol{\mu}_Q = \mathbf{w}$ and $\boldsymbol{\mu}_P = \mathbf{0}$. We have:

$$\sigma_Q^2 + \frac{\|\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q\|_2^2}{k} \leq \rho^2 + \frac{\|\mathbf{w}\|_2^2}{k} \leq \rho^2 \exp\left(\frac{nB'}{k}\right) \quad (28)$$

Let $j = \lfloor 1 - k \log((\rho^2 + \|\mathbf{w}\|_2^2/k)/c) \rfloor$. We can ensure $j \in \mathbb{N}$ by setting $c = \rho^2 \exp(nB'/k)$. For $\sigma_P^2 = c \exp((1-j)/k)$, we have:

$$\rho^2 + \frac{\|\mathbf{w}\|_2^2}{k} \leq \sigma_P^2 \leq \exp\left(\frac{1}{k}\right)(\rho^2 + \frac{\|\mathbf{w}\|_2^2}{k}) \quad (29)$$

Building on Eq.(28) and Eq.(29), we can obtain an upper bound for the KL divergence:

$$KL(\mathcal{Q} \parallel \mathcal{P}) = \frac{1}{2} \left[\frac{k\sigma_Q^2 + \|\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q\|_2^2}{\sigma_P^2} - k + k \log \left(\frac{\sigma_P^2}{\sigma_Q^2} \right) \right] \quad (30)$$

$$\leq \frac{1}{2} \left[\frac{k(\rho^2 + \frac{\|\boldsymbol{w}\|_2^2}{k})}{\rho^2 + \frac{\|\boldsymbol{w}\|_2^2}{k}} - k + k \log \left(\frac{\exp(\frac{1}{k})(\rho^2 + \frac{\|\boldsymbol{w}\|_2^2}{k})}{\sigma_Q^2} \right) \right] \quad (31)$$

$$= \frac{1}{2} \left[k \log \left(\frac{\exp(\frac{1}{k})(\rho^2 + \frac{\|\boldsymbol{w}\|_2^2}{k})}{\sigma_Q^2} \right) \right] \quad (32)$$

$$= \frac{1}{2} \left[k \log \left(\frac{\exp(\frac{1}{k})(\rho^2 + \frac{\|\boldsymbol{w}\|_2^2}{k})(\sqrt{k} + \sqrt{2 \ln n})^2}{\rho^2} \right) \right] \quad (33)$$

$$= \frac{1}{2} \left[1 + k \log \left(1 + \frac{\|\boldsymbol{w}\|_2^2}{k\rho^2} \right) + 2k \log \left(\sqrt{k} + \sqrt{2 \ln n} \right) \right] \quad (34)$$

Given the bound that corresponds to j holds with probability $1 - \delta_j$ for $\delta_j = \frac{6\delta}{\pi^2 j^2}$, the log term can be bounded as:

$$\log \frac{2\sqrt{n}}{\delta_j} = \log \frac{2\sqrt{n}}{\delta} + \log \frac{\pi^2 j^2}{6} \quad (35)$$

$$\leq \log \frac{2\sqrt{n}}{\delta} + \log \frac{\pi^2 (1 + \log(c/\rho^2))^2}{6} \quad (36)$$

$$\leq \log \frac{2\sqrt{n}}{\delta} + \log \frac{\pi^2 (1 + k \log(\exp(nB'/k)))^2}{6} \quad (37)$$

$$= \log \frac{2\sqrt{n}}{\delta} + \log \frac{\pi^2 (1 + nB')^2}{6} \quad (38)$$

$$= \log \frac{\pi^2 \sqrt{n} (1 + nB')^2}{3\delta} \quad (39)$$

where the first inequality is derived from the fact that $j \leq 1 + k \log(c/(\rho^2 + \|\boldsymbol{w}\|_2^2/k)) \leq 1 + k \log(c/\rho^2)$.

Therefore, the generalization bound can be written as:

$$\begin{aligned} \mathbb{E}_{\epsilon_i \sim \mathcal{N}(0, \sigma_Q)} [L_D^\gamma(\boldsymbol{w} + \boldsymbol{\epsilon})] &\leq 2\mathbb{E}_{\epsilon_i \sim \mathcal{N}(0, \sigma_Q)} [L_S^\gamma(\boldsymbol{w} + \boldsymbol{\epsilon})] \\ &+ \frac{2 + 2k \log \left(1 + \frac{\|\boldsymbol{w}\|_2^2}{k\rho^2} \right) + 4k \log \left(\sqrt{k} + \sqrt{2 \ln n} \right) + 4 \log \frac{\pi^2 \sqrt{n} (1 + nB')^2}{3\delta}}{n} \end{aligned} \quad (40)$$

Since $\|\boldsymbol{\epsilon}\|_2^2$ has chi-square distribution, for any positive t , we have:

$$P(\|\boldsymbol{\epsilon}\|_2^2 - k\sigma_Q^2 \geq 2\sigma_Q^2 \sqrt{kt} + 2t\sigma_Q^2) \leq \exp(-t) \quad (41)$$

Therefore, with probability $1 - 1/n$, we have:

$$\|\boldsymbol{\epsilon}\|_2^2 \leq \sigma_Q^2 \left[k + 2\sqrt{k \ln(n)} + 2 \ln(n) \right] \quad (42)$$

$$\leq \sigma_Q^2 \left[\sqrt{k} + \sqrt{2 \ln(n)} \right]^2 \quad (43)$$

$$\leq \rho^2 \quad (44)$$

Therefore, we have:

$$\begin{aligned}
 & \mathbb{E}_{\epsilon_i \sim \mathcal{N}(0, \sigma_Q)} [L_{\mathcal{D}}^{\gamma}(\mathbf{w} + \boldsymbol{\epsilon})] \leq 2(1 - 1/n) \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} L_S^{\gamma}(\mathbf{w} + \boldsymbol{\epsilon}) + \frac{2B'}{n} \\
 & + \frac{2 + 2k \log \left(1 + \frac{\|\mathbf{w}\|_2^2}{k\rho^2}\right) + 4k \log \left(\sqrt{k} + \sqrt{2 \ln n}\right) + 4 \log \frac{\pi^2 \sqrt{n}(nB' + 1)^2}{3\delta}}{n} \\
 & \leq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} 2L_S^{\gamma}(\mathbf{w} + \boldsymbol{\epsilon}) \\
 & + \frac{2 + 2B' + 2k \log \left(1 + \frac{\|\mathbf{w}\|_2^2}{k\rho^2}\right) + 4k \log \left(\sqrt{k} + \sqrt{2 \ln n}\right) + 4 \log \frac{\pi^2 \sqrt{n}(nB' + 1)^2}{3\delta}}{n}
 \end{aligned} \tag{45}$$

□

Combining the above lemma with the Taylor expansion, we can derive the following lemma to bound $\lambda \cdot L_{\mathcal{D}}^{\gamma}(\mathbf{w})$ by $\lambda \cdot \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} L_S^{\gamma}(\mathbf{w} + \boldsymbol{\epsilon})$.

Lemma A.4. Assume that $\forall (\mathbf{x}, y) \in \mathcal{D}, 0 \leq \ell(\mathbf{w}; \mathbf{x}, y) \leq B$. Then, for any $\rho > 0$ and any distribution \mathcal{D} , with probability $1 - \delta$ over the choice of the training set $S \sim \mathcal{D}$

$$\begin{aligned}
 L_{\mathcal{D}}^{\gamma}(\mathbf{w}) & \leq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} 2L_S^{\gamma}(\mathbf{w} + \boldsymbol{\epsilon}) - \frac{\rho^2}{2(\sqrt{k} + \sqrt{2 \ln(n)})^2} \cdot \text{tr}(H(\mathbf{w})) - o\left(\frac{k\rho^2}{(\sqrt{k} + \sqrt{2 \ln(n)})^2}\right) \\
 & + \frac{2 + 2B' + 2k \log \left(1 + \frac{\|\mathbf{w}\|_2^2}{k\rho^2}\right) + 4k \log \left(\sqrt{k} + \sqrt{2 \ln n}\right) + 4 \log \frac{\pi^2 \sqrt{n}(nB' + 1)^2}{3\delta}}{n}
 \end{aligned} \tag{46}$$

where $n = |S|$, k is the number of parameters, $B' \triangleq \sum_{i=1}^C (1 - \pi_i)^{\gamma} \pi_i B$, $H(\mathbf{w})$ represents the Hessian matrix of $L_{\mathcal{D}}^{\gamma}(\mathbf{w})$ at point \mathbf{w} and $\text{tr}(\cdot)$ represent the matrix trace.

Proof. By expanding $\mathbb{E}_{\epsilon_i \sim \mathcal{N}(0, \sigma_Q)} [L_{\mathcal{D}}^{\gamma}(\mathbf{w} + \boldsymbol{\epsilon})]$ around \mathbf{w} using a second-order Taylor Series expansion, we can obtain

$$\begin{aligned}
 \mathbb{E}_{\epsilon_i \sim \mathcal{N}(0, \sigma_Q)} [L_{\mathcal{D}}^{\gamma}(\mathbf{w} + \boldsymbol{\epsilon})] & = \mathbb{E}_{\epsilon_i \sim \mathcal{N}(0, \sigma_Q)} [L_{\mathcal{D}}^{\gamma}(\mathbf{w}) + \boldsymbol{\epsilon}^T \nabla L_{\mathcal{D}}^{\gamma}(\mathbf{w}) + \frac{1}{2} \boldsymbol{\epsilon}^T H(\mathbf{w}) \boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|_2^2)] \\
 & = L_{\mathcal{D}}^{\gamma}(\mathbf{w}) + \frac{1}{2} \mathbb{E}_{\epsilon_i \sim \mathcal{N}(0, \sigma_Q)} [\boldsymbol{\epsilon}^T H(\mathbf{w}) \boldsymbol{\epsilon}] + \mathbb{E}_{\epsilon_i \sim \mathcal{N}(0, \sigma_Q)} [o(\|\boldsymbol{\epsilon}\|_2^2)]
 \end{aligned} \tag{47}$$

where $\sigma_Q \triangleq \frac{\rho}{\sqrt{k} + \sqrt{2 \ln(n)}}$ and $H(\mathbf{w})$ represents the Hessian matrix of $L_{\mathcal{D}}^{\gamma}(\mathbf{w})$ at point \mathbf{w} .

Thereby, we have:

$$\begin{aligned}
 \mathbb{E}_{\epsilon_i \sim \mathcal{N}(0, \sigma_Q)} [L_{\mathcal{D}}^{\gamma}(\mathbf{w} + \boldsymbol{\epsilon})] & = L_{\mathcal{D}}^{\gamma}(\mathbf{w}) + \frac{1}{2} \mathbb{E}_{\epsilon_i \sim \mathcal{N}(0, \sigma_Q)} [\boldsymbol{\epsilon}^T H(\mathbf{w}) \boldsymbol{\epsilon}] + \mathbb{E}_{\epsilon_i \sim \mathcal{N}(0, \sigma_Q)} [o(\|\boldsymbol{\epsilon}\|_2^2)] \\
 & = L_{\mathcal{D}}^{\gamma}(\mathbf{w}) + \frac{\text{tr}(H(\mathbf{w}))}{2} \cdot \mathbb{E}_{\epsilon_1 \sim \mathcal{N}(0, \sigma_Q)} [\boldsymbol{\epsilon}_1^2] + o(k \cdot \mathbb{E}_{\epsilon_1 \sim \mathcal{N}(0, \sigma_Q)} [\boldsymbol{\epsilon}_1^2]) \\
 & = L_{\mathcal{D}}^{\gamma}(\mathbf{w}) + \frac{\rho^2}{2(\sqrt{k} + \sqrt{2 \ln(n)})^2} \cdot \text{tr}(H(\mathbf{w})) + o\left(\frac{k\rho^2}{(\sqrt{k} + \sqrt{2 \ln(n)})^2}\right)
 \end{aligned} \tag{48}$$

Combining Eq.(48) with Lem.A.3, with probability $1 - \delta$, we have

$$\begin{aligned}
 L_{\mathcal{D}}^{\gamma}(\mathbf{w}) & \leq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} 2L_S^{\gamma}(\mathbf{w} + \boldsymbol{\epsilon}) - \frac{\rho^2}{2(\sqrt{k} + \sqrt{2 \ln(n)})^2} \cdot \text{tr}(H(\mathbf{w})) - o\left(\frac{k\rho^2}{(\sqrt{k} + \sqrt{2 \ln(n)})^2}\right) \\
 & + \frac{2 + 2B' + 2k \log \left(1 + \frac{\|\mathbf{w}\|_2^2}{k\rho^2}\right) + 4k \log \left(\sqrt{k} + \sqrt{2 \ln n}\right) + 4 \log \frac{\pi^2 \sqrt{n}(nB' + 1)^2}{3\delta}}{n}
 \end{aligned} \tag{49}$$

□

A.4. Proof of Thm.4.1

Combining the above two parts, we can finally derive the following theorem.

Theorem A.5 (Restate of Thm.4.1). *Assume that $\forall (\mathbf{x}, y) \in \mathcal{D}, 0 \leq \ell(\mathbf{w}; \mathbf{x}, y) \leq B$. For any $\rho > 0$, any uniform distribution \mathcal{D}_{bal} and any distribution \mathcal{D} , with probability $1 - \delta$ over the choice of the training set $S \sim \mathcal{D}$,*

$$\begin{aligned} L_{\mathcal{D}_{bal}}(\mathbf{w}) &\leq \frac{2L_S^{FS}(\mathbf{w})}{C\pi_C} + \frac{40 \cdot (B + \lambda B') \cdot \log(\frac{4}{\delta})}{3n \cdot C\pi_C} - \frac{\lambda\rho^2}{2(\sqrt{k} + \sqrt{2\ln(n)})^2 \cdot C\pi_C} \cdot \text{tr}(H(\mathbf{w})) \\ &+ \lambda \cdot \frac{2 + 2B' + 2k \log\left(1 + \frac{\|\mathbf{w}\|_2^2}{k\rho^2}\right) + 4k \log\left(\sqrt{k} + \sqrt{2\ln n}\right) + 4 \log \frac{2\pi^2 \sqrt{n}(nB'+1)^2}{3\delta}}{n \cdot C\pi_C} \\ &- o\left(\frac{\lambda k \rho^2}{(\sqrt{k} + \sqrt{2\ln(n)})^2 \cdot C\pi_C}\right) \end{aligned} \quad (50)$$

where $n = |S|$, k is the number of parameters, $B' \triangleq \sum_{i=1}^C (1 - \pi_i)^\gamma \pi_i B$, $H(\mathbf{w})$ represents the Hessian matrix of $L_{\mathcal{D}}^\gamma(\mathbf{w})$ at point \mathbf{w} and $\text{tr}(\cdot)$ represent the matrix trace.

Proof. Combining Lem.A.2 and Lem.A.4 and using union bound, with probability at least $1 - \delta$, we have

$$\begin{aligned} L_{\mathcal{D}}(\mathbf{w}) &\leq 2L_S^{FS}(\mathbf{w}) + \frac{40 \cdot (B + \lambda B') \cdot \log(\frac{4}{\delta})}{3n} - \frac{\lambda\rho^2}{2(\sqrt{k} + \sqrt{2\ln(n)})^2} \cdot \text{tr}(H(\mathbf{w})) \\ &+ \lambda \cdot \frac{2 + 2B' + 2k \log\left(1 + \frac{\|\mathbf{w}\|_2^2}{k\rho^2}\right) + 4k \log\left(\sqrt{k} + \sqrt{2\ln n}\right) + 4 \log \frac{2\pi^2 \sqrt{n}(nB'+1)^2}{3\delta}}{n} \\ &- o\left(\frac{\lambda k \rho^2}{(\sqrt{k} + \sqrt{2\ln(n)})^2}\right) \end{aligned} \quad (51)$$

We further recognize that:

$$L_{\mathcal{D}}(\mathbf{w}) = \sum_{i=1}^C \pi_i L_{\mathcal{D}_i}(\mathbf{w}) \geq \sum_{i=1}^C \pi_C L_{\mathcal{D}_i}(\mathbf{w}) = C\pi_C \cdot L_{\mathcal{D}_{bal}}(\mathbf{w}) \quad (52)$$

Substituting Eq.(52) into Eq.(51) leads to Thm.A.5. □

B. Analysis of Backpropagation Requirements for SAM and ImbSAM

B.1. Backpropagation Requirements for SAM

SAM aims to find flatter minima, ensuring the entire neighborhood around the model parameters has consistently low training loss. The objective loss function is defined as:

$$L_S^{SAM}(\mathbf{w}) \triangleq \max_{\|\epsilon\|_2 \leq \rho} L_S(\mathbf{w} + \epsilon) \quad (53)$$

The optimal perturbation $\hat{\epsilon}_{SAM}(\mathbf{w})$ for the inner maximization problem is estimated as follow:

$$\hat{\epsilon}_{SAM}(\mathbf{w}) \approx \rho \frac{\nabla_{\mathbf{w}} L_S(\mathbf{w})}{\|\nabla_{\mathbf{w}} L_S(\mathbf{w})\|_2} \quad (54)$$

Thus, the gradient of $L_S^{SAM}(\mathbf{w})$ can be approximated as:

$$\nabla_{\mathbf{w}} L_S^{SAM}(\mathbf{w}) \approx \nabla_{\mathbf{w}} L_S(\mathbf{w})|_{\mathbf{w} + \hat{\epsilon}(\mathbf{w})} \quad (55)$$

To update parameters once using SAM, **two** backpropagations are required: one for $\nabla_{\mathbf{w}} L_S(\mathbf{w})$, and another for $\nabla_{\mathbf{w}} L_S(\mathbf{w})|_{\mathbf{w} + \hat{\epsilon}(\mathbf{w})}$

B.2. Backpropagation Requirements for ImbSAM

ImbSAM divides classes into head and tail groups, denoted as \mathcal{H} and \mathcal{T} , and applies SAM only to the tail group. Its objective function is:

$$L_S^{IS}(\mathbf{w}) \triangleq L_S^{\mathcal{H}}(\mathbf{w}) + \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} L_S^{\mathcal{T}}(\mathbf{w} + \boldsymbol{\epsilon}) \quad (56)$$

The optimal perturbation $\hat{\boldsymbol{\epsilon}}_{IS}(\mathbf{w})$ for the inner maximization problem is estimated as follow:

$$\hat{\boldsymbol{\epsilon}}_{IS}(\mathbf{w}) \approx \rho \frac{\nabla_{\mathbf{w}} L_S^{\mathcal{T}}(\mathbf{w})}{\|\nabla_{\mathbf{w}} L_S^{\mathcal{T}}(\mathbf{w})\|_2} \quad (57)$$

Thus, the gradient of $L_S^{IS}(\mathbf{w})$ can be approximated as:

$$\nabla_{\mathbf{w}} L_S^{IS}(\mathbf{w}) \approx \nabla_{\mathbf{w}} L_S^{\mathcal{H}}(\mathbf{w}) + \nabla_{\mathbf{w}} L_S^{\mathcal{T}}(\mathbf{w})|_{\mathbf{w} + \hat{\boldsymbol{\epsilon}}(\mathbf{w})} \quad (58)$$

To update parameters once using ImbSAM, **three** backpropagations are required: one for $\nabla_{\mathbf{w}} L_S^{\mathcal{T}}(\mathbf{w})$, one for $\nabla_{\mathbf{w}} L_S^{\mathcal{H}}(\mathbf{w})$, and another for $\nabla_{\mathbf{w}} L_S^{\mathcal{T}}(\mathbf{w})|_{\mathbf{w} + \hat{\boldsymbol{\epsilon}}(\mathbf{w})}$

C. More Experiment Protocols

C.1. Datasets

For long-tailed recognition tasks, we conduct experiments on four widely used long-tailed datasets: CIFAR-10 LT, CIFAR-100 LT, ImageNet-LT, and iNaturalist. For long-tailed domain generalization tasks, we train the model on ImageNet-LT and evaluate it on three OOD datasets: ImageNet-Sketch, ImageNetV2, and ImageNet-C. Below is a detailed description of these datasets:

- **CIFAR-100 LT and CIFAR-10 LT** (Cao et al., 2019). The original CIFAR-100 (Krizhevsky & Hinton, 2009) and CIFAR-10 (Krizhevsky & Hinton, 2009) datasets contain 50,000 training images and 10,000 testing images for 100 and 10 classes, respectively. We utilize their various long-tailed versions with different imbalance ratios of {100, 50, 10}.
- **ImageNet-LT** (Liu et al., 2019). The ImageNet-LT dataset is derived from the ImageNet (Deng et al., 2009) dataset according to a Pareto distribution, containing 1000 categories. The dataset includes 115,846 training images and 50,000 test images. The dataset has an imbalance ratio of 256.
- **iNaturalist** (Horn et al., 2018). The iNaturalist dataset is a real-world large-scale dataset, consisting of 8142 categories. The training set contains approximately 430,000 images, while the test set contains about 24,000 images. The dataset's imbalance ratio is 500.
- **ImageNet-Sketch** (Wang et al., 2019a). The ImageNet-Sketch dataset is an OOD test set derived from the ImageNet (Deng et al., 2009) dataset, comprising 50,000 images across 1000 classes. Each image is a sketch, introducing a domain shift relative to ImageNet.
- **ImageNetV2** (Recht et al., 2019). The ImageNetV2 dataset consists of 10,000 images spanning the same 1000 classes as ImageNet. The images are sourced differently from the original ImageNet (Deng et al., 2009), resulting in a slight domain shift.
- **ImageNet-C** (Hendrycks & Dietterich, 2019). The ImageNet-C dataset includes the same 1000 classes as ImageNet (Deng et al., 2009) but features corrupted versions of the original validation set. Each image undergoes one of 15 corruption types at 5 severity levels, resulting in 75 dataset variations.

C.2. Competitors

When training ResNet models from scratch, we evaluate several competitive methods on different datasets. For the CIFAR-LT dataset, we assess multiple loss functions, including CE loss, LDAM+DRW (Cao et al., 2019), LA loss (Menon et al., 2021), and VS loss (Kini et al., 2021). These methods are further combined with SAM (Foret et al., 2021), ImbSAM (Zhou et al., 2023a), and CC-SAM (Zhou et al., 2023b) as baseline comparisons. For the ImageNet-LT and iNaturalist datasets, we

employ a range of representative methods, including CB (Cui et al., 2019) for class re-balancing, cRT (Kang et al., 2020) for decoupled training, DiVE (He et al., 2021) for transfer learning, DRO-LT (Samuel & Chechik, 2021) for representation learning, DisAlign (Zhang et al., 2021) for class re-balancing, and WB (Alshammari et al., 2022) for regularization. When fine-tuning the foundation model CLIP (Radford et al., 2021), we use Decoder (Wang et al., 2024c) and LPT (Dong et al., 2023) as baselines. We also evaluate both fully fine-tuning the models with LA loss (denoted as FFT), and parameter-efficient fine-tuning using the LIFT method (Shi et al., 2024), as well as their performance when combined with different SAM variants.

C.3. Evaluation Protocol

For long-tailed recognition tasks, we assess model performance using top-1 accuracy on balanced test sets. This ensures all classes contribute equally to the evaluation. To provide a more detailed analysis, we follow the approach in (Zhong et al., 2021; Liu et al., 2019) by splitting the classes into three subsets: Head, Medium, and Tail. Accuracy is then reported for each subset individually. For CIFAR-10 LT (IR = 100), the Head classes contain more than 1000 samples, the Medium classes have 200~1000 samples, and the Tail classes have less than 200 samples. For CIFAR-100 LT (IR = 100), ImageNet-LT, and iNaturalist, the Head classes contain more than 100 samples, the Medium classes have 20~100 samples, and the Tail classes have less than 20 samples. Prior arts (Zhou et al., 2022; Khattak et al., 2023; Park et al., 2024) have demonstrated that fine-tuning CLIP (Radford et al., 2021) often performs well on the target domain but struggles with domain shifts. Therefore, when fine-tuning the foundation models, we also assess model performance on OOD test sets, referred to as long-tailed domain generalization tasks. Specifically, models are trained on the ImageNet-LT dataset and evaluated on out-of-distribution datasets, including ImageNet-Sketch (Wang et al., 2019a), ImageNetV2 (Recht et al., 2019), and ImageNet-C (Hendrycks & Dietterich, 2019). We evaluate model performance on these OOD balanced test sets, including top-1 accuracy and accuracy for each class subset.

C.4. Implementation Details

We follow the procedures described below to train ResNet models from scratch. For the CIFAR-LT datasets, we use ResNet-32 (He et al., 2016) as the backbone. We employ stochastic gradient descent (SGD) as the base optimizer, with an initial learning rate of 0.1, a batch size of 64, and a momentum of 0.9. Training spans 200 epochs, using a cosine annealing scheduler to reduce the learning rate from 0.1 to 0 gradually. For the larger-scale ImageNet-LT and iNaturalist datasets, we employ ResNet-50 (He et al., 2016) as the backbone. SGD is again used as the base optimizer with a momentum of 0.9. For ImageNet-LT, the initial learning rate is set to 0.1, with a batch size of 256, while for iNaturalist, the initial learning rate is 0.2, and the batch size is increased to 512. Training for these datasets also lasts 200 epochs with a cosine annealing scheduler. Additionally, We employ a step scheduler for ρ , following the approach of Rangwani et al. (2022). This scheduler initializes ρ until the 160th epoch and then increases its value towards the end of training.

For fine-tuning foundation models, we follow the protocols outlined in LIFT (Shi et al., 2024). A cosine classifier is added after the image encoder of CLIP (Radford et al., 2021), with its weights initialized using the text encoder, which is then discarded. We fine-tune the image encoder of CLIP with a ViT-B/16 (Dosovitskiy et al., 2021) backbone. Stochastic gradient descent (SGD) is used as the base optimizer, with a batch size of 128 and momentum of 0.9. The initial learning rate is 0.01 for parameter-efficient fine-tuning and 0.001 for full fine-tuning. Unlike LIFT (Shi et al., 2024), all models in our experiments are fine-tuned for 20 epochs across datasets and methods. In LIFT, models are trained for 10 epochs on the CIFAR-LT and the ImageNet-LT datasets, and 20 epochs on the iNaturalist dataset. We extend the training to 20 epochs because the models do not fully converge under the original settings.

C.5. Experimental Hardware Setup

All the experiments are conducted on Ubuntu servers equipped with Nvidia(R) RTX 3090 GPUs and RTX 4090 GPUs. Fine-tuning the foundation models is performed using a single GPU for all datasets. The number of GPUs used for training the ResNet models from scratch varies based on dataset size: a single GPU for the CIFAT-LT datasets, 2 GPUs for the ImageNet-LT dataset, and 4 GPUs for the iNaturalist dataset.

D. More Experiment Results

D.1. Additional Results on the CIFAR-LT Datasets

In this section, we show additional results on the CIFAR-LT datasets. Specifically, Tab.6 presents additional experimental results on the CIFAR-100 LT dataset with more combined methods. Tab.5 provides the experimental results on the CIFAR-10 LT dataset. The results suggest that Focal-SAM consistently outperforms SAM, ImbSAM, and CC-SAM across all methods and datasets, regardless of whether ResNet models are trained from scratch or foundation models are fine-tuned. This indicates that Focal-SAM offers better fine-grained control over the loss landscape for both head and tail classes, leading to improved overall performance. This further highlights the effectiveness of Focal-SAM.

Table 5: Performance comparison on CIFAR-10 LT datasets with various imbalance ratios (IR). FFT denotes fully fine-tuning the foundation model with LA loss.

Method	IR100				IR200 All	IR50 All	IR10 All
	Head	Med	Tail	All			
Training from scratch							
CE	87.0	73.6	54.0	73.1	68.6	78.3	87.4
CE+SAM	89.5	73.9	56.7	75.0	69.8	79.6	88.8
CE+ImbSAM	88.0	79.0	60.1	76.9	72.6	81.1	89.3
CE+CC-SAM	88.9	74.1	61.3	76.2	71.3	80.0	89.2
CE+Focal-SAM	89.3	75.4	62.9	77.2	71.7	82.0	90.0
LDAM+DRW (Cao et al., 2019)	85.5	74.6	69.0	77.3	73.8	80.8	87.3
LDAM+DRW+SAM	88.9	78.3	73.2	81.0	78.6	84.5	89.4
LDAM+DRW+ImbSAM	86.5	79.7	73.7	80.6	77.3	84.0	88.9
LDAM+DRW+CC-SAM	88.4	79.2	73.3	81.1	78.9	84.4	89.4
LDAM+DRW+Focal-SAM	88.7	79.5	74.2	81.6	79.2	84.5	89.5
LA (Menon et al., 2021)	87.6	72.6	70.1	77.9	74.3	81.6	87.8
LA+SAM	86.7	80.6	78.2	82.3	78.9	85.4	90.2
LA+ImbSAM	84.1	81.6	80.1	82.2	78.6	84.7	89.5
LA+CC-SAM	86.6	80.8	78.5	82.5	79.1	85.5	90.2
LA+Focal-SAM	86.9	81.2	79.2	82.9	79.6	85.5	90.5
VS (Kini et al., 2021)	88.1	77.1	68.4	78.9	74.7	81.5	88.3
VS+SAM	85.6	82.7	76.6	82.0	79.0	85.4	90.3
VS+ImbSAM	85.3	82.1	77.3	81.9	78.7	84.8	90.0
VS+CC-SAM	85.6	82.0	78.2	82.3	79.3	85.5	90.4
VS+Focal-SAM	87.7	80.6	78.8	82.9	79.5	85.8	90.7
Fine-tuning foundation model							
FFT	97.9	95.8	95.9	96.7	95.7	97.1	97.9
FFT+SAM	97.5	96.5	97.0	97.0	96.6	97.5	98.0
FFT+ImbSAM	97.0	97.0	97.4	97.1	96.5	97.7	97.9
FFT+CC-SAM	97.6	96.2	97.0	97.0	96.6	97.6	98.0
FFT+Focal-SAM	97.5	96.4	97.5	97.2	96.6	97.5	98.2
LIFT (Shi et al., 2024)	96.6	95.7	97.4	96.6	96.3	96.8	97.2
LIFT+SAM	96.6	95.6	97.8	96.7	96.4	96.7	97.0
LIFT+ImbSAM	96.5	95.9	97.7	96.7	96.4	96.7	97.2
LIFT+CC-SAM	96.5	95.6	97.9	96.6	96.4	96.7	97.3
LIFT+Focal-SAM	96.6	95.6	98.1	96.8	96.4	96.9	97.3

Table 6: Performance comparison on CIFAR-100 LT with more combined methods

Method	IR100				IR200	IR50	IR10
	Head	Med	Tail	All	All	All	All
Training from scratch							
LDAM+DRW (Cao et al., 2019)	63.1	44.4	18.6	43.2	40.3	46.1	57.3
LDAM+DRW+SAM	67.6	51.7	25.9	49.5	45.8	52.6	61.1
LDAM+DRW+ImbSAM	62.5	48.8	26.4	46.9	42.5	51.3	59.8
LDAM+DRW+CC-SAM	66.5	52.2	26.2	49.4	45.7	52.3	61.0
LDAM+DRW+Focal-SAM	67.9	52.7	26.9	50.3	46.2	53.8	62.3
VS (Kini et al., 2021)	58.3	43.8	31.1	45.1	41.6	49.3	59.4
VS+SAM	62.7	52.0	29.3	49.0	45.5	53.5	62.5
VS+ImbSAM	56.1	53.3	29.9	47.2	44.7	52.6	62.6
VS+CC-SAM	62.2	52.2	30.3	49.1	45.2	53.7	62.9
VS+Focal-SAM	62.7	52.6	31.0	49.7	45.8	54.5	63.7

D.2. CE and mCE Metrics on ImageNet-C for Long-tailed Domain Generalization

ImageNet-C (Hendrycks & Dietterich, 2019) contains the corrupted versions of ImageNet (Deng et al., 2009) dataset, with 15 corruption types applied at 5 severity levels, resulting in 75 dataset variations. In addition to the model’s average accuracy across these 75 datasets, as shown in Tab.4, ImageNet-C introduces two additional metrics: Corruption Error (CE) and Mean Corruption Error (mCE). These metrics systematically assess the robustness of models against image corruption. CE measures the accuracy drop of a model on a specific type and severity of corruption compared to a baseline model, typically AlexNet (Krizhevsky et al., 2012). mCE aggregates the CE values across all corruption types and severity levels, providing a single robustness score for the model. Tab.7 presents the CE and mCE results on the ImageNet-C dataset when fine-tuning the foundation models. The results show that Focal-SAM generally achieves significantly lower CE and mCE values across the entire dataset and for each corruption type. This further demonstrates Focal-SAM’s effectiveness in improving generalization.

Table 7: The CE and mCE values for different methods on the ImageNet-C dataset. The source models are trained on ImageNet-LT and evaluated on ImageNet-C. Lower values indicate better performance.

Method	mCE \downarrow	Blur				Noise			Digital				Weather			
		Motion	Defoc	Glass	Zoom	Gauss	Impul	Shot	Contr	Elast	JPEG	Pixel	Bright	Snow	Fog	Frost
FFT	72.6	72.8	74.6	78.2	79.4	73.6	74.1	75.1	66.7	80.0	78.4	71.3	63.7	66.6	65.7	68.5
+SAM	69.0	69.2	72.2	76.9	76.9	68.9	69.7	70.7	63.6	77.5	74.5	66.8	59.2	62.6	60.6	65.7
+ImbSAM	69.5	69.4	72.2	76.8	77.1	70.2	70.7	72.0	62.9	77.6	76.4	68.5	59.6	62.5	60.2	65.7
+CC-SAM	69.0	69.0	74.1	76.6	77.8	69.5	69.3	71.2	64.0	76.6	75.2	67.0	58.1	61.3	59.6	65.0
+Focal-SAM	68.3	68.2	72.7	76.6	76.6	68.3	68.8	70.1	63.2	76.2	74.2	66.0	58.0	61.5	59.7	65.1
LIFT	63.6	61.7	67.6	75.9	72.4	61.0	61.2	62.7	54.1	80.8	72.7	60.3	52.1	57.0	52.2	62.1
+SAM	63.0	61.1	67.2	75.7	71.5	60.4	60.6	62.0	53.6	80.6	72.1	59.4	51.7	56.3	51.7	61.7
+ImbSAM	63.4	61.6	67.5	75.8	72.1	60.9	60.9	62.6	53.8	80.6	72.5	59.9	51.9	56.7	51.9	61.8
+CC-SAM	62.5	61.0	66.5	75.3	70.9	59.5	59.9	61.3	53.7	79.9	71.7	58.2	51.3	55.5	51.6	61.2
+Focal-SAM	62.2	60.6	66.3	75.2	71.1	59.3	59.9	61.0	53.1	79.6	71.0	58.2	50.7	55.3	51.1	61.0

D.3. Results for Aligning Computational Cost

Focal-SAM requires about 50% more training time than SAM. To fairly evaluate the benefit of Focal-SAM, we conduct experiments where we extend the training epochs of SAM to match Focal-SAM’s total computational cost. Specifically, we increase the training epochs to 300 or 30 ($1.5 \times$ the original 200 or 20) for SAM, while keeping Focal-SAM at 200 or 20 epochs. In this setting, the total computational cost of SAM and Focal-SAM becomes comparable. We conduct these experiments on CIFAR-100 LT, ImageNet-LT, and iNaturalist datasets. The results are shown in Tab.8, Tab.9, and Tab.10.

Table 8: Performance comparison on CIFAR-100 LT with aligned computational cost.

Method	Epoch	IR100	IR200	IR50	IR10
		All	All	All	All
Training from scratch					
CE+SAM	300	43.0	39.2	46.9	60.0
CE+Focal-SAM	200	44.0	39.6	48.1	60.9
LDAM+DRW+SAM	300	50.4	46.4	53.0	61.2
LDAM+DRW+Focal-SAM	200	50.3	46.2	53.8	62.3
VS+SAM	300	49.2	45.5	53.0	63.3
VS+Focal-SAM	200	49.7	45.8	54.5	63.7
LA+SAM	300	50.1	45.5	53.8	63.0
LA+Focal-SAM	200	50.7	46.0	54.5	63.8
Fine-tuning foundation model					
FFT+SAM	30	81.2	78.3	83.6	86.9
FFT+Focal-SAM	20	81.6	79.0	83.9	87.3
LIFT+SAM	30	82.1	80.2	83.1	85.2
LIFT+Focal-SAM	20	82.4	80.0	83.2	85.4

Table 9: Performance comparison on ImageNet-LT with aligned computational cost.

Method	Epoch	ImageNet-LT			
		Head	Medium	Tail	All
Training from scratch					
LA+SAM	300	63.2	51.6	34.8	53.8
LA+Focal-SAM	200	63.9	52.2	34.4	54.3
Fine-tuning foundation model					
FFT+SAM	30	80.6	73.1	56.1	73.6
FFT+Focal-SAM	20	80.8	73.9	54.4	73.9
LIFT+SAM	30	79.8	76.1	73.5	77.2
LIFT+Focal-SAM	20	79.7	76.6	73.6	77.4

Table 10: Performance comparison on iNaturalist with aligned computational cost.

Method	Epoch	iNaturalist			
		Head	Medium	Tail	All
Training from scratch					
LA+SAM	300	68.0	71.4	72.4	71.5
LA+Focal-SAM	200	68.4	72.0	72.5	71.8

D.4. Visualization of Loss Landscape

Fig.6 and Fig.7 visualize the loss landscape for head and tail classes of ResNet models trained with SAM, ImbSAM, CC-SAM, and Focal-SAM on the CIFAR-100 LT and CIFAR-10 LT datasets using VS loss respectively. From the results, we can observe that the loss landscape for tail classes with ImbSAM generally appears flatter and smoother than with SAM, suggesting that ImbSAM better flattens the loss landscape for tail classes. However, the head class loss landscape with

ImbSAM is generally sharper than with SAM, indicating that ImbSAM’s exclusion of all head classes from the SAM term can sharpen the loss landscape for head classes, which might reduce their generalization performance. In contrast, CC-SAM and Focal-SAM provide fine-grained class-wise control, leading to a flatter loss landscape for both head and tail classes.

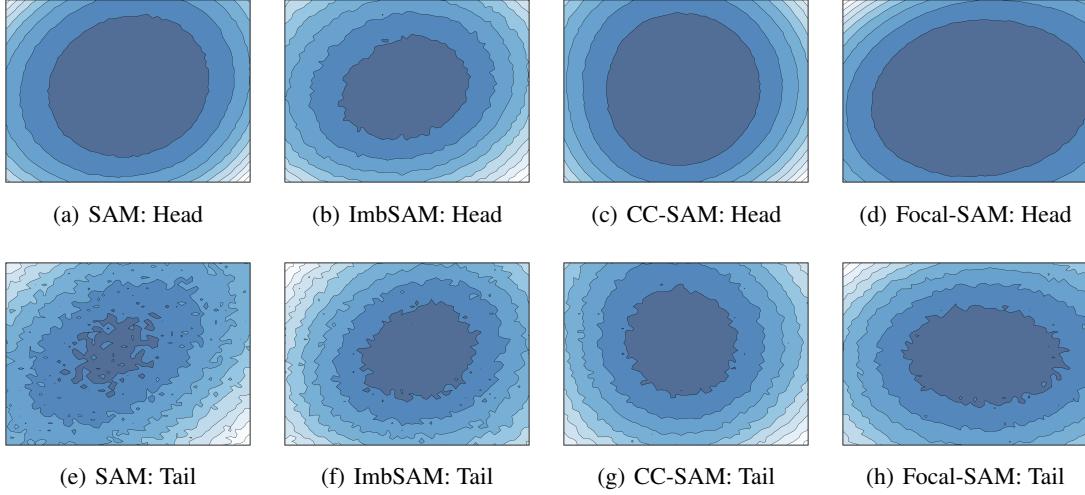


Figure 6: Visualization of loss landscape for head and tail classes of ResNet models trained with SAM, ImbSAM, CC-SAM, and Focal-SAM on CIFAR-100 LT using VS loss respectively.

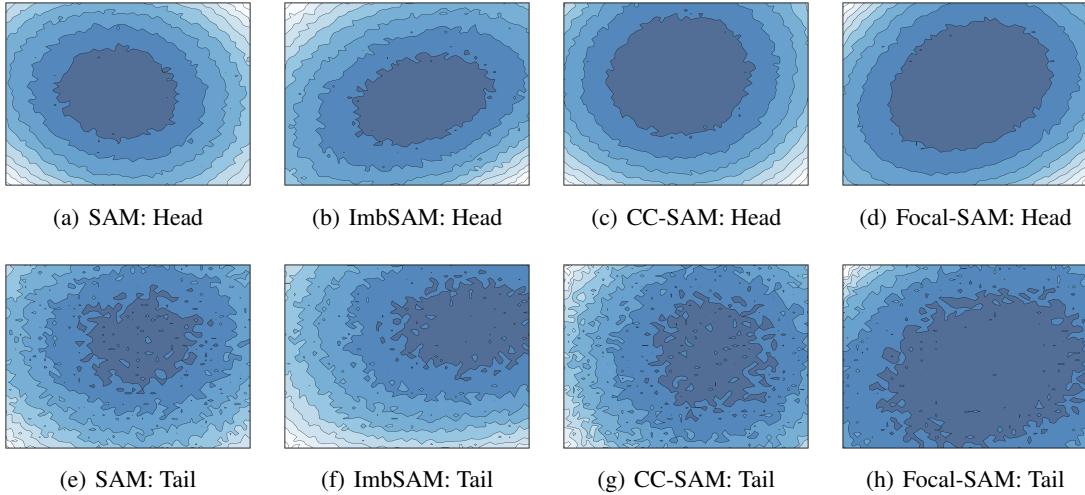


Figure 7: Visualization of loss landscape for head and tail classes of ResNet models trained with SAM, ImbSAM, CC-SAM, and Focal-SAM on CIFAR-10 LT using VS loss respectively.

D.5. Ablation Study About Perturbation Radius ρ

Fig.8 illustrates the impact of the hyperparameter ρ on the performance of Focal-SAM when combined with LDAM+DRW, LA, and VS methods on the CIFAR-LT datasets during ResNet models training. As ρ increases, Focal-SAM’s performance initially improves but then declines. This indicates a trade-off between achieving flatter minima and reducing training loss. The optimal value of ρ for Focal-SAM is approximately 0.3, which is higher than the commonly optimal value for SAM on balanced training datasets, as reported by Foret et al. (2021). This observation is consistent with Rangwani et al. (2022), who suggest that a larger ρ can enhance performance in long-tailed learning.

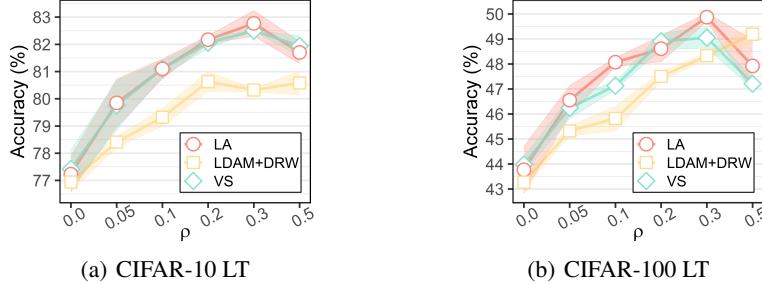


Figure 8: Ablation Study of Focal-SAM w.r.t. ρ

D.6. Additional Results for Eigen Spectral Density of Hessian

This section presents additional results of the spectral density of hessian for ResNet models trained with SAM, ImbSAM, CC-SAM, and Focal-SAM. We analyze models trained on CIFAR-10 LT and CIFAR-100 LT datasets using VS and CE loss functions. The results are visualized in Fig.9, Fig.10 and Fig.11.

The results indicate that the largest eigenvalue λ_{max} and the trace $tr(H)$ of the Hessian for tail classes are generally smaller with ImbSAM than with SAM. This suggests that ImbSAM flattens the loss landscape for tail classes more effectively. However, λ_{max} and $tr(H)$ for head classes are typically larger with ImbSAM than with SAM, indicating that ImbSAM's coarse-grained strategy of excluding all head classes from SAM terms sharpens the loss landscape for those classes. In contrast, CC-SAM applies finer control over the loss landscape by using class-dependent perturbation radii, generally achieving lower λ_{max} and $tr(H)$ for head and tail classes. Overall, both λ_{max} and $tr(H)$ for head and tail classes are relatively lower with Focal-SAM than other SAM-based methods. This further suggests that Focal-SAM provides fine-grained control over the loss landscape, leading to a flatter landscape for both head and tail classes.

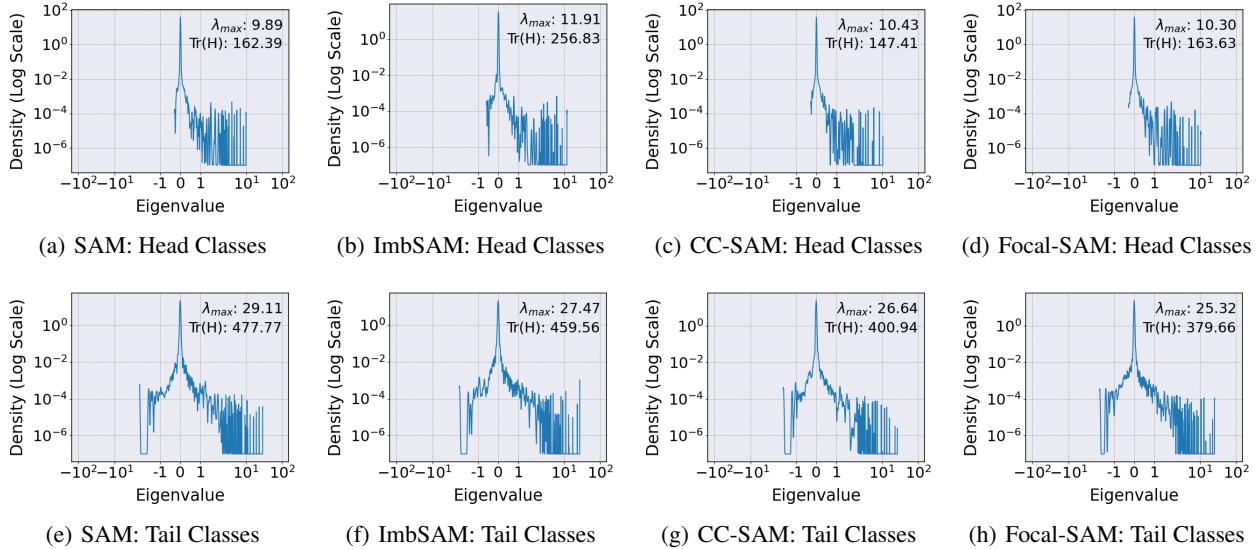


Figure 9: Eigen Spectral Density of Hessian for head and tail classes of ResNet models trained with SAM, ImbSAM, CC-SAM, and Focal-SAM on CIFAR-100 LT using VS loss respectively. A smaller λ_{max} and $Tr(H)$ generally indicate a flatter loss landscape.

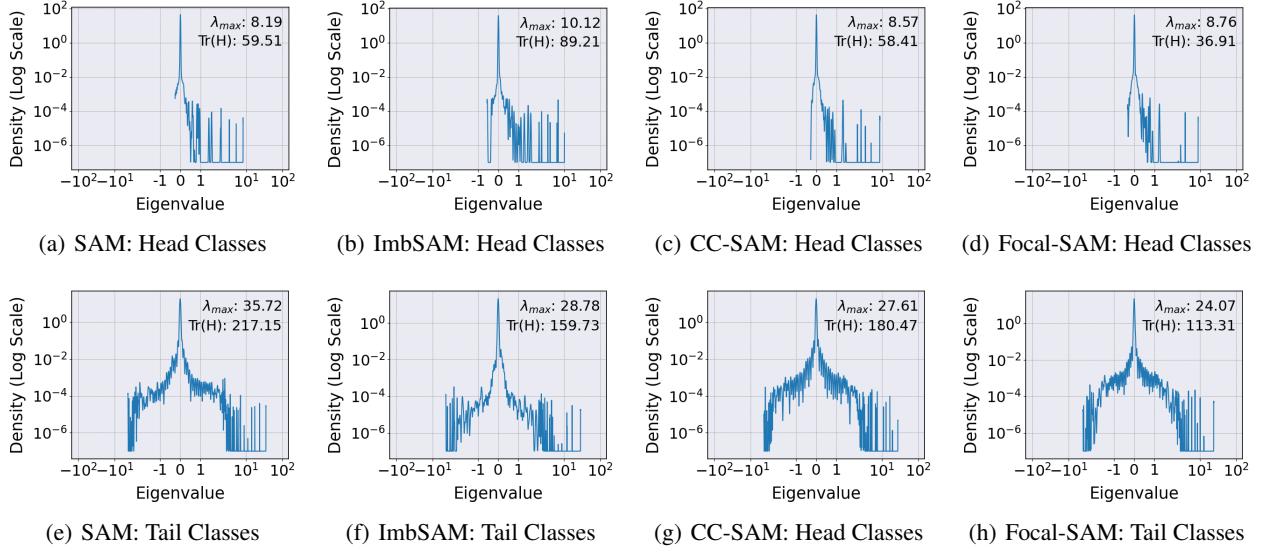


Figure 10: Eigen Spectral Density of Hessian for head and tail classes of ResNet models trained with SAM, ImbSAM, CC-SAM, and Focal-SAM on CIFAR-10 LT using CE loss respectively. A smaller λ_{max} and $Tr(H)$ generally indicate a flatter loss landscape.

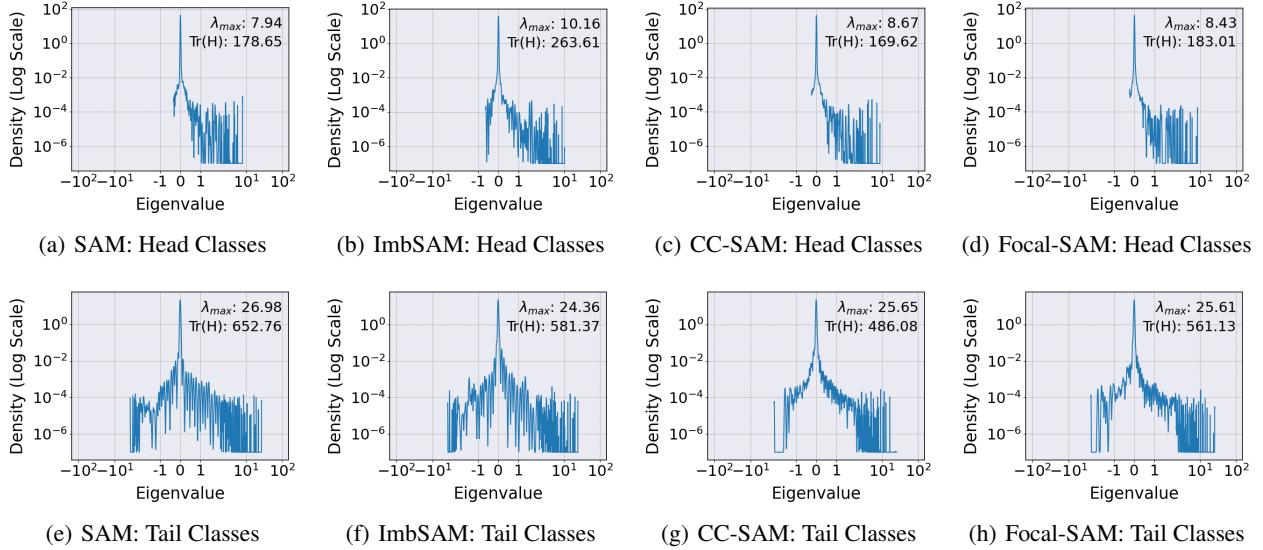


Figure 11: Eigen Spectral Density of Hessian for head and tail classes of ResNet models trained with SAM, ImbSAM, CC-SAM, and Focal-SAM on CIFAR-100 LT using CE loss respectively. A smaller λ_{max} and $Tr(H)$ generally indicate a flatter loss landscape.