ImbSAM: A Closer Look at Sharpness-Aware Minimization in Class-Imbalanced Recognition

Yixuan Zhou[1] Yi Qu[1] Xing Xu[1,*] Hengtao Shen[1,2]
[1]Center for Future Media & School of Computer Science and Engineering,
University of Electronic Science and Technology of China [2]Peng Cheng Laboratory, China
yxzhou@std.uestc.edu.cn, iquyiiii@gmail.com, xing.xu@uestc.edu.cn, shenhengtao@hotmail.com

Abstract

Class imbalance is a common challenge in real-world recognition tasks, where the majority of classes have few samples, also known as tail classes. We address this challenge with the perspective of generalization and empirically find that the promising Sharpness-Aware Minimization (SAM) fails to address generalization issues under the class-imbalanced setting. Through investigating this specific type of task, we identify that its generalization bottleneck primarily lies in the severe overfitting for tail classes with limited training data. To overcome this bottleneck, we leverage class priors to restrict the generalization scope of the class-agnostic SAM and propose a class-aware smoothness optimization algorithm named Imbalanced-SAM (ImbSAM). With the guidance of class priors, our ImbSAM specifically improves generalization targeting tail classes. We also verify the efficacy of ImbSAM on two prototypical applications of class-imbalanced recognition: long-tailed classification and semi-supervised anomaly detection, where our ImbSAM demonstrates remarkable performance improvements for tail classes and anomaly. Our code implementation is available at https://github.com/cool-xuan/Imbalanced_SAM.

1. Introduction

Over the course of decades, deep neural networks have achieved remarkable success in various already-intensely-studied tasks, including classification [14, 24], segmentation [40, 50], and object detection [49, 56]. These impressive achievements are largely attributed to large-scale datasets [13, 36], which strive for a uniform distribution of categories, contrary to the data distribution in the real open world. The real-world data is usually class-imbalanced [7, 23, 32, 39, 57], following a long-tailed distribution: a small number of dominant classes (head classes) have numerous samples, while the majority of classes (tail classes) contain only a few samples. Directly applying SOTA methods [14, 24] built under balanced data distribution to the class-imbalanced setting suffers from dramatic performance degradation [39]. This critical performance reduction is primarily caused by the overwhelming presence of head classes during training, which in turn results in inadequate learning for tail classes [23, 28, 61]. This limitation motivates the theoretical research on class-imbalanced recognition, which drives lots of practical applications such as long-tailed classification [2, 31, 39, 57] and semi-supervised anomaly detection [20, 44, 51].

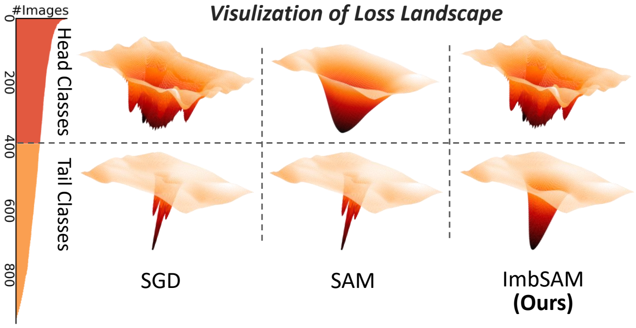


Figure 1: The visualization of separate loss landscape for *head* and *tail* classes in class-imbalanced recognition, optimized by SGD [6], SAM [17] and our ImbSAM respectively.

Many excellent methods [9, 11, 12, 15, 19, 35, 61] have been proposed to tackle the issue of class-imbalanced data, where plenty of methods re-balance the long-tailed data by re-sampling [15, 16] or assign large loss weights to the tail classes. While these methods alleviate the dominant presence of head classes over tail classes, they overexpose the limited tail class samples, increasing the risk of overfitting for these tail classes [22].

[*]Corresponding author.

Recent methods [2, 31] fine-tune the regularization to penalize the large parameters in turn avoiding overfitting. Compared with the empirical regularization, Sharpness-Aware Minimization (SAM) [17], an effective optimization algorithm, is supported by a solid theoretical foundation. SAM connects the smooth geometry of the loss landscape with generalization and captures the sharpness of the loss landscape. By simultaneously minimizing the loss value and sharpness, SAM converges the model weights to reach a smooth minimum (neighborhoods having uniformly low loss).

However, the SAM is proposed and effective in the ideal data setting (balanced distribution [13, 32]), ignoring the class imbalance in the real world. As the loss landscape visualization of SAM shown in Figure 1, SAM tends to prioritize generalization on the head classes since the heavily imbalanced data, while overlooking the tail classes in class-imbalanced recognition. Nevertheless, even without SAM, the abundant training data of head classes also prevents them from suffering overfitting.

To address this issue, we first investigate the class-imbalanced recognition and identify its generalization bottleneck primarily lying in tail classes. As for such specific tasks with long-tailed distribution, the head classes, benefiting from sufficient training samples, are less affected by generalization problems [22]. On the other hand, the tail classes, with only a few data instances (sometimes even less than 10), are highly susceptible to severe overfitting. Based on these insights, we propose a class-aware smoothness optimization algorithm named Imbalanced-SAM (*ImbSAM*) to tackle the overfitting problem with respect to (w.r.t.) tail classes. In contrast to the class-agnostic SAM, our ImbSAM introduces class priors to restrict the smoothness optimization scope to the tail classes as illustrated in Figure 1, thereby alleviating severe overfitting of inadequate training samples of tail classes.

Our ImbSAM is compatible with existing methods, demonstrating remarkable performance promotion in prototypical applications of class-imbalanced recognition: long-tailed classification (LTC) [2, 24, 31] and semi-supervised anomaly detection (SSAD) [20, 44, 51]. Notably, our ImbSAM impressively improves recognition accuracy for tail classes as illustrated in Figure 2, which firmly verifies the efficacy of ImbSAM in focusing generalization scope on classes with limited training data. Our main contributions are summarized as follows:

- We approach the challenge of class-imbalanced recognition from a generalization perspective and identify severe overfitting on tail classes as the main generalization bottleneck. A theoretical analysis is provided to reveal why the promising SAM fails to address the generalization issues in class-imbalanced recognition.

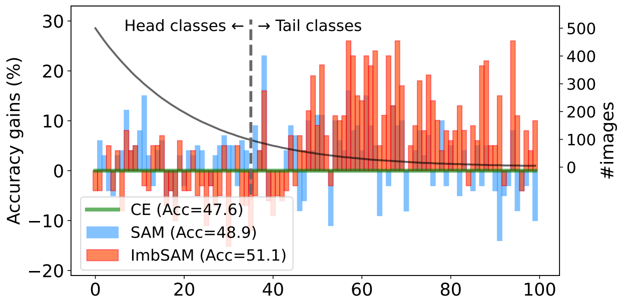


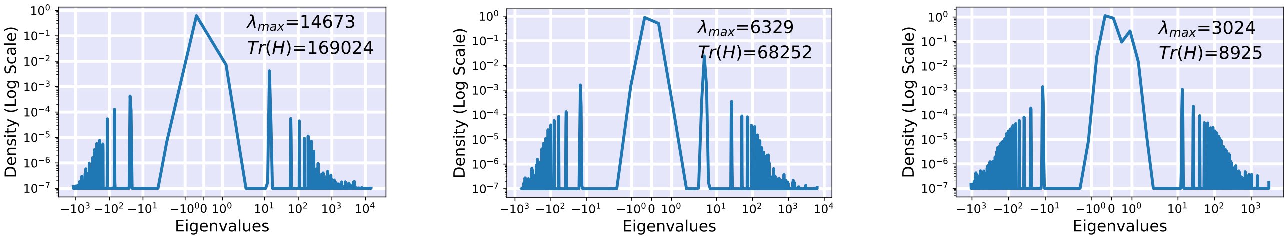
Figure 2: **Accuracy gains** (%) **of each classes on CIFAR100-LT** derived from the standard SAM [17] and our class-aware ImbSAM. With the guidance of class priors (dividing data into two splits by η), our ImbSAM successfully performs generalization targeting tail classes, which are neglected by SAM.

- To overcome the limitation of SAM, we propose the Imbalanced SAM (*ImbSAM*), which incorporates class priors into the class-agnostic SAM to specifically address the overfitting problem on tail classes.
- We evaluate the efficacy of ImbSAM on two prototypical applications of class-imbalanced recognition: long-tailed classification and semi-supervised anomaly detection, where it demonstrates remarkable performance improvements for the classes with inadequate training samples.

2. Related Work

In this section, two prototypical applications derived from class-imbalanced recognition are first introduced: long-tailed classification and semi-supervised anomaly detection, followed by the introduction of smoothness of loss landscape and its effective implementation: sharpness-aware minimization (SAM).

Long-Tailed Classification. Long-tail classification has been extensively studied [7, 18, 23, 28, 30] in recent years due to its importance in real-world applications with heavily imbalanced data distribution. Early approaches for long-tail classification focused on re-sampling [15, 16] or re-weighting [12, 35] strategies to balance the class distribution, such as over-sampling tail classes [19] and under-sampling head classes [38], or using weighted cross-entropy loss to focus on underrepresented classes [9, 52]. Recently, more flexible and robust methods are proposed involving transfer learning [37, 63], self-supervision [34], contrastive learning [11], representation learning [31], and ensemble learning with multi-experts separately recognize relatively balanced sub-groups [8, 59].



(a) SGD (b) SAM (c) Our ImbSAM

Figure 3: **Eigen Spectral Density of Hessian** for the CE [2] respectively optimized by SGD [6], SAM [17] and our ImbSAM on *Few* classes in CIFAR100-LT (IF=100). Each graph is annotated with the maximum eigenvalue (λ_{max}) and the trace of the Hessian matrix ($Tr(H)$), which are indicators of the smoothness of loss landscape (Lower λ_{max} and $Tr(H)$ reveal the smoother loss landscape).

Some existing methods [2, 58] also improve long-tailed classification with the perspective of generalization and address the generalization problem by tuning regularization. However, these methods are more empirical and less controllable compared with Sharpness-Aware Minimization [17], which is supported by a solid theoretical foundation.

Semi-Supervised Anomaly Detection. Anomaly detection plays a critical role in broad applications such as violence detection [46] and industrial manufacture [64]. In recent years, semi-supervised anomaly detection methods have shown great potential by leveraging both limited anomaly annotation [1, 42, 43, 51, 53, 69] and statistic information of abundant normal data, outperforming traditional unsupervised anomaly detection methods [20].

On the other hand, semi-supervised anomaly detection can be viewed as a special case of long-tailed classification, with only one head class (normal) and one tail class (anomaly), which poses unique challenges since the unpredictable and diverse types of anomalies. To address this problem, existing semi-supervised anomaly detection methods also borrow the re-sampling strategy [44] or ensemble learning [67] from long-tailed classification to balance the extremely long-tailed data. However, the generalization for limited but diverse anomalies remains an open and crucial problem for semi-supervised anomaly detection methods to achieve high recognition accuracy and detect unseen anomalies.

Smoothness of Loss Landscape. The issue of model generalization in deep learning has always been an essential yet challenging problem [10, 41]. A panoply of outstanding methods have been developed from the scope of model adjustment [3, 27] or data augmentation [29, 65]. In recent years, with the perspective of the connection between the geometry of the loss landscape and generalization, the Sharpness-Aware Minimization (SAM) [17] has emerged as a promising generalization improvement approach by identifying and minimizing the sharpness of the loss landscape.

SAM and its variants [33, 48] have achieved SOTA performance on various challenging tasks [13, 45], while its application in class-imbalanced recognition remains nascent [47, 58]. To tackle the generalization issues in class-imbalanced recognition, we enhance SAM with class awareness by incorporating class priors [21] and introduce the class-aware ImbSAM. Our ImbSAM successfully narrows down the generalization scope of the standard SAM, specifically targeting the tail classes, which suffer from intractable overfitting.

3. Proposed Methodology

3.1. Notation and Problem Definition

In class-imbalanced recognition tasks, the training set $\mathcal{S} = \cup_{i=1}^{n}\{(\mathbf{x}_i, y_i)\}$ is heavily imbalanced, where $y_i \in [1, ..., K]$ is labeled class k for the data sample $\mathbf{x}_i$. According to the data amount of class k, the whole set $\mathcal{S}$ is divided into two parts: $\mathcal{S}^{\text{head}}$ including data of head classes and $\mathcal{S}^{\text{tail}}$ including data of tail classes, and $\mathcal{S} = \mathcal{S}^{\text{tail}} \cup \mathcal{S}^{\text{head}}$. For convenience, $\mathcal{S}^k$ denotes the set of training samples belonging to the class k, and $|\mathcal{S}^k|$ refers to its data amount. To quantitatively measure how imbalanced the long-tailed dataset $\mathcal{S}$ is, the imbalanced factor IF $= \frac{\max_k |\mathcal{S}^k|}{\min_k |\mathcal{S}^k|}$ is defined, with IF $\gg 1$ in class-imbalanced training set $\mathcal{S}$.

Class-imbalanced recognition intrinsically follows the common classification framework: training a neural network $f(\cdot; \boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$. Given a sample $\mathbf{x}_i$, the neural network f predicts a label $y_i'{=}f(\mathbf{x}_i; \boldsymbol{\theta})$. The classic cross-entropy (CE) loss function [4] or its variants [12, 35] is chosen as the criterion $\ell(y_i', y_i)$ to supervise the optimization of parameters $\boldsymbol{\theta}$. The training optimization can be formulated as follows:

$$
\begin{aligned}
\boldsymbol{\theta}^* &= \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{\mathcal{S}} \ell(f(\mathbf{x}_i;\boldsymbol{\theta}), y_i), \\
&= \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{k} \sum_{\mathcal{S}^k} \ell(f(\mathbf{x}_i;\boldsymbol{\theta}), y_i), \\
&= \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{\mathcal{S}^{\text{head}}} \ell(f(\mathbf{x}_i;\boldsymbol{\theta}), y_i) + \sum_{\mathcal{S}^{\text{tail}}} \ell(f(\mathbf{x}_i;\boldsymbol{\theta}), y_i),
\end{aligned} \tag{1}
$$

where $\boldsymbol{\theta}^*$ is the theoretical optimum of $\boldsymbol{\theta}$, and we also provide a class-wise equivalence. Furthermore, we simplify the class-wise equivalence by dividing all K classes into 2 types: *head* classes with ample samples and *tail* classes with restricted samples. These two parts of losses are respectively denoted as follows:

$$
\mathcal{L}_{\mathcal{S}^{\text{head}}}(\boldsymbol{\theta}) = \sum_{\mathcal{S}^{\text{head}}} \ell(f(\mathbf{x}_i;\boldsymbol{\theta}), y_i), \tag{2}
$$

$$
\mathcal{L}_{\mathcal{S}^{\text{tail}}}(\boldsymbol{\theta}) = \sum_{\mathcal{S}^{\text{tail}}} \ell(f(\mathbf{x}_i;\boldsymbol{\theta}), y_i). \tag{3}
$$

The summarized loss $\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta})$ is rewritten as

$$
\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}) = \sum_{\mathcal{S}} \ell(f(\mathbf{x}_i;\boldsymbol{\theta}), y_i) = \mathcal{L}_{\mathcal{S}^{\text{head}}}(\boldsymbol{\theta}) + \mathcal{L}_{\mathcal{S}^{\text{tail}}}(\boldsymbol{\theta}). \tag{4}
$$

3.2. Preliminaries: Sharpness-Aware Minimization

From the perspective of the connection of smooth loss landscape and generalization, Sharpness-Aware Minimization (SAM) not only minimizes the single point in the loss landscape of criterion $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$ w.r.t. data distribution $\mathcal{D}$ but also consistently brings its neighborhoods down. As a result, SAM turns to minimize the following PAC-Bayesian error upper bound:

$$
\begin{aligned}
\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) \leq & \left[\max_{\|\boldsymbol{\epsilon}\| \leq \rho} \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}+\boldsymbol{\epsilon}) - \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}) \right] \\
& + \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}) + h(\|\boldsymbol{\theta}\|_2^2/\rho^2),
\end{aligned} \tag{5}
$$

with some strictly increasing function h. Compared with standard loss, the additional term in square brackets measures the loss sharpness by capturing the loss increasing rate when perturbing $\boldsymbol{\theta}$ with noise $\boldsymbol{\epsilon}$ in the neighborhood of ρ. Since the monotonicity of h, it can be theoretically replaced by the L2 regularization term $\lambda\|\boldsymbol{\theta}\|_2^2$ with weight decay coefficient λ. Thus, the optimization target of SAM, the right-hand side of the inequality Eq. 5, is rewritten as

$$
\min_{\boldsymbol{\theta}} \max_{\|\boldsymbol{\epsilon}\| \leq \rho} \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}+\boldsymbol{\epsilon}) + \lambda\|\boldsymbol{\theta}\|_2^2, \tag{6}
$$

which is transferred as a minimax optimization.

To solve this minimax problem, SAM first tackles the max problem by seeking the maximum perturbation $\boldsymbol{\epsilon}_t$ in the range of ρ at training step t. This inner maximization problem can be calculated via a first-order Taylor approximation w.r.t. $\boldsymbol{\epsilon} \to 0$ and dual norm as follows:

$$
\begin{aligned}
\boldsymbol{\epsilon}_t &= \operatorname*{argmax}_{\|\boldsymbol{\epsilon}\| \leq \rho} \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}_t+\boldsymbol{\epsilon}) \\
&\approx \operatorname*{argmax}_{\|\boldsymbol{\epsilon}\| \leq \rho} \boldsymbol{\epsilon}^\top \nabla\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}_t) \\
&= \rho \operatorname{sign}(\nabla\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}_t)) \frac{|\nabla\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}_t)|^{q-1}}{\|\nabla\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}_t)\|_q^{q/p}},
\end{aligned} \tag{7}
$$

where $|\cdot|^{q-1}$ refers to element-wise absolute value and power, sign$(\cdot)$ is the signum function, and $1/p + 1/q = 1$. Secondly, the outer minimization problem can be solved as

$$
\begin{aligned}
\boldsymbol{\theta}_t &= \operatorname*{argmin}_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_t) + \lambda\|\boldsymbol{\theta}\|_2^2 \\
&\approx \boldsymbol{\theta}_t - \alpha_t \left[\nabla\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}_t+\boldsymbol{\epsilon}_t) + \lambda\boldsymbol{\theta}_t \right],
\end{aligned} \tag{8}
$$

where α_t is the learning rate at training step t. It is empirically confirmed that the above 2-step optimization yields the best performance when $p = 2$, resulting in $\boldsymbol{\epsilon}_t$ formulated as

$$
\boldsymbol{\epsilon}_t = \rho \frac{\nabla\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}_t)}{\|\nabla\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}_t)\|_2}. \tag{9}
$$

In summary, SAM converges $\boldsymbol{\theta}$ to a smooth minimum with uniformly low loss by iteratively solving Eq. 9 and Eq. 8.

3.3. The Limitation of Class-Agnostic SAM

While SAM is effective and supported by a solid theoretical foundation, the class-agnostic SAM forfeits its impressive generalization power when confront with class-imbalanced data. In long-tailed datasets, the SAM optimization can be re-formulated by introducing the split loss function (Eq. 4) as follows:

$$
\begin{aligned}
\min_{\boldsymbol{\theta}} \max_{\|\boldsymbol{\epsilon}\| \leq \rho} & \left[\mathcal{L}_{\mathcal{S}^{\text{head}}}(\boldsymbol{\theta}+\boldsymbol{\epsilon}) \right. \\
& \left. + \mathcal{L}_{\mathcal{S}^{\text{tail}}}(\boldsymbol{\theta}+\boldsymbol{\epsilon}) \right] + \lambda\|\boldsymbol{\theta}\|_2^2.
\end{aligned} \tag{10}
$$

Correspondingly, the formula of perturbation $\boldsymbol{\epsilon}_t$ is rewritten as follows:

$$
\boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}_t^{\text{head}} + \boldsymbol{\epsilon}_t^{\text{tail}}, \tag{11}
$$

and

$$
\boldsymbol{\epsilon}_t^{\text{head}} = \rho \frac{\nabla\mathcal{L}_{\mathcal{S}^{\text{head}}}(\boldsymbol{\theta}_t)}{\|\nabla\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}_t)\|_2}, \quad \boldsymbol{\epsilon}_t^{\text{tail}} = \rho \frac{\nabla\mathcal{L}_{\mathcal{S}^{\text{tail}}}(\boldsymbol{\theta}_t)}{\|\nabla\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}_t)\|_2}, \tag{12}
$$

where $\boldsymbol{\epsilon}^{\text{head}}$ and $\boldsymbol{\epsilon}^{\text{tail}}$ denotes the perturbations added for the head class set $\mathcal{S}^{\text{head}}$ and tail class set $\mathcal{S}^{\text{tail}}$, respectively. Due to the overwhelming data amount of head classes over tail classes ($|\mathcal{S}^{\text{head}}| \gg |\mathcal{S}^{\text{tail}}|$), the magnitude of the gradient for head classes $|\nabla\mathcal{L}_{\mathcal{S}^{\text{head}}}|$ also crushes $|\nabla\mathcal{L}_{\mathcal{S}^{\text{tail}}}|$ derived from tail classes, resulting in $|\boldsymbol{\epsilon}_t^{\text{head}}| \gg |\boldsymbol{\epsilon}_t^{\text{tail}}|$. On the other hand, $\left\|\boldsymbol{\epsilon}_t^{\text{head}} + \boldsymbol{\epsilon}_t^{\text{tail}}\right\|_2$ equals to ρ, which is a constant during training and set 0.05 for most cases [17]. Therefore, the perturbation $\boldsymbol{\epsilon}_t^{\text{tail}}$ calculated for tail classes is negligible and can be ignored, which leads to the following approximation of Eq. 10:

$$\min_{\boldsymbol{\theta}} \max_{\|\boldsymbol{\epsilon}^{\text{head}}\| \leq \rho} \left[\mathcal{L}_{\mathcal{S}^{\text{head}}}(\boldsymbol{\theta} + \boldsymbol{\epsilon}^{\text{head}}) + \mathcal{L}_{\mathcal{S}^{\text{tail}}}(\boldsymbol{\theta} + \boldsymbol{\epsilon}^{\text{head}})\right] + \lambda\|\boldsymbol{\theta}\|_2^2. \tag{13}$$

Accordingly, the gradient update formula is also approximated as

$$\boldsymbol{\theta}_t \approx \boldsymbol{\theta}_t - \alpha_t\left[\nabla\mathcal{L}_{\mathcal{S}^{\text{head}}}(\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t^{\text{head}}) + \nabla\mathcal{L}_{\mathcal{S}^{\text{tail}}}(\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t^{\text{head}}) + \lambda\boldsymbol{\theta}_t\right], \tag{14}$$

where the SAM optimization on head classes is persevered, while that on tail classes is misleading by the overwhelming perturbation $\boldsymbol{\epsilon}_t^{\text{head}}$. In other words, the class-agnostic SAM only prioritizes generalization for the head classes, while confusing the optimization of tail classes, which are prone to overfitting.

Introducing re-sampling [19, 16] or re-weighting [35, 12] methods forces the magnitude of $\boldsymbol{\epsilon}^{\text{tail}}$ to be larger and even comparable with $\boldsymbol{\epsilon}^{\text{head}}$. At the same time, these two categories of methods also increase the risk of overfitting with limited tail class instances that are over-exposed or over-focused during training. Notably, SAM is impressive for its generalization improvement, which is theoretically contradictory with re-sampling and re-weighting. The empirical evaluation (Table 3) of naively combining SAM with large loss re-weights for tail classes also validates our analysis, where the recognition accuracy of tail classes is barely unaffected even assigned with large re-weight up to 20.

3.4. Class-Aware Imbalanced SAM (ImbSAM)

Before making some adjustments in the standard SAM to adapt it to class-imbalanced recognition, we first thoroughly investigate this specific task and try to diagnose its generalization bottleneck. In the class-imbalanced dataset with a long-tailed distribution, a small number of head classes carve up the main body of training data, leaving a limited number of samples for the vast majority of tail classes. Benefiting from the abundance of training data, there are no severe generalization issues for the minority of head classes [22]. However, since tail classes have access to only a handful of training instances, with some classes having fewer than ten samples, deep neural networks are susceptible to overfitting these few training instances and failing to generalize to unseen data. This results in a generalization bottleneck that is particularly pronounced in tail classes.

Inspired by the above insights, we introduce the class priors to the standard SAM to take full advantage of its efficiently improving generalization and propose a class-aware Imbalanced SAM (*ImbSAM*) to address the generalization bottleneck on the side of tail classes in class-imbalanced recognition. With the guidance of class priors, our ImbSAM successfully shifts the focus from dominant head classes to vulnerable tail classes and significantly avoids overfitting.

Algorithm 1: Our *ImbSAM* Algorithm (p=2)

```
Input: Training dataset $\mathcal{S} = \cup_{i=1}^{n}\{(\mathbf{x}_i, y_i)\}$, neural
       network $f(\cdot)$ with parameters $\boldsymbol{\theta}$, loss
       function $\ell$, mini-batch size $b$, learning rate $\alpha$,
       weight decay coefficient $\lambda$, neighborhood
       size $\rho$, class split threshold $\eta$.
Output: Trained parameters $\boldsymbol{\theta}^*$
Initialize parameters $\boldsymbol{\theta}_0$, $t = 0$;
while not converged do
    Sample batch $\mathcal{B} = \{(\mathbf{x}_i, y_1), ..., (\mathbf{x}_b, y_b)\}$;
    Divide $\mathcal{B}$ into $\mathcal{B}^{\text{head}}$ and $\mathcal{B}^{\text{tail}}$ with $\eta$;    // Eq. 15
    $\mathcal{L}_{\mathcal{B}^{\text{head}}}(\boldsymbol{\theta_t}) = \sum_{\mathcal{B}^{\text{head}}} \ell(f(\mathbf{x}_i; \boldsymbol{\theta}_t), y_i)$;
    $\mathcal{L}_{\mathcal{B}^{\text{tail}}}(\boldsymbol{\theta_t}) = \sum_{\mathcal{B}^{\text{tail}}} \ell(f(\mathbf{x}_i; \boldsymbol{\theta}_t), y_i)$;
    $\nabla\mathcal{L}_{\mathcal{B}^{\text{head}}}(\boldsymbol{\theta}_t) = \text{Backward}(\mathcal{L}_{\mathcal{B}^{\text{head}}}, f(\cdot))$;
    $\nabla\mathcal{L}_{\mathcal{B}^{\text{tail}}}(\boldsymbol{\theta}_t) = \text{Backward}(\mathcal{L}_{\mathcal{B}^{\text{tail}}}, f(\cdot))$;
    $\boldsymbol{\epsilon}^{\text{tail}} = \rho\frac{\nabla\mathcal{L}_{\mathcal{B}^{\text{tail}}}(\boldsymbol{\theta}_t)}{\|\nabla\mathcal{L}_{\mathcal{B}^{\text{tail}}}(\boldsymbol{\theta}_t)\|_2}$;
    $\boldsymbol{\theta}_t = \boldsymbol{\theta}_t - \alpha_t\left[\nabla\mathcal{L}_{\mathcal{B}^{\text{tail}}}(\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t^{\text{tail}}) + \nabla\mathcal{L}_{\mathcal{B}^{\text{head}}}(\boldsymbol{\theta}_t) + \lambda\boldsymbol{\theta}_t\right]$.
end
```

How to Build Class Priors. Due to the lack of class prior that is essential in recognition of tail classes [8, 12, 18, 35, 58], SAM fails to optimize a smooth minimum for tail classes. Thus, how to construct the class priors for SAM should be first solved. In our ImbSAM, the class priors are simply built by introducing a class split threshold η to divide the entire training set into two sub-sets according to their training data amount: head sub-set $\mathcal{S}^{\text{head}}$ and tail sub-set $\mathcal{S}^{\text{tail}}$. In particular, head and tail sub-sets comprise all data samples categorized into the class with data amount more or not more than η respectively, formulated as

$$\begin{cases} (\mathbf{x}_i, y_i) \in \mathcal{S}^{\text{head}} & \text{if } |\mathcal{S}^{y_i}| > \eta \\ (\mathbf{x}_j, y_j) \in \mathcal{S}^{\text{tail}} & \text{if } |\mathcal{S}^{y_j}| \leqslant \eta\,, \end{cases} \tag{15}$$

where the class split threshold η is a hyperparameter to control the training set splitting and can be set 100 for the most long-tailed datasets [12, 39, 57]. Although our class priors only roughly split the entire training set $\mathcal{S}$ into two parts without considering the specific data amount of each class, they still play a crucial role in endowing our ImbSAM with class awareness.

How to Utilize Class Priors in ImbSAM. To incorporate class awareness into the standard SAM, we treat the losses derived from the two sub-sets divided according to our class priors, $\mathcal{L}_{\mathcal{S}^{\text{head}}}(\boldsymbol{\theta})$ and $\mathcal{L}_{\mathcal{S}^{\text{tail}}}(\boldsymbol{\theta})$, differently. Particularly, the optimization target in our ImbSAM is adapted from Eq. 10 to the following formula by ignoring the SAM optimization term for head classes:

$$\min_{\boldsymbol{\theta}} \max_{\|\boldsymbol{\epsilon}\| \le \rho} \mathcal{L}_{\mathcal{S}^{\text{tail}}}(\boldsymbol{\theta} + \boldsymbol{\epsilon}^{\text{tail}}) + \mathcal{L}_{\mathcal{S}^{\text{head}}}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_2^2. \tag{16}$$

To make explicit our sharpness-aware term, the above optimization target can be rewritten as follows:

$$\begin{aligned} \min_{\boldsymbol{\theta}} & \overbrace{\left[\max_{\|\boldsymbol{\epsilon}\| \le \rho} \mathcal{L}_{\mathcal{S}^{\text{tail}}}(\boldsymbol{\theta} + \boldsymbol{\epsilon}^{\text{tail}}) - \mathcal{L}_{\mathcal{S}^{\text{tail}}}(\boldsymbol{\theta}) \right] + \mathcal{L}_{\mathcal{S}^{\text{tail}}}(\boldsymbol{\theta})}^{\text{optimization term for } tail \text{ classes}} \\ & + \underbrace{\mathcal{L}_{\mathcal{S}^{\text{head}}}(\boldsymbol{\theta})}_{\text{optimization term for } head \text{ classes}} + \lambda \|\boldsymbol{\theta}\|_2^2, \end{aligned} \tag{17}$$

where the term in square brackets specifically captures the sharpness for the loss derived from tail classes. Unlike the class-agnostic SAM treating all classes equally, our ImbSAM leverage class priors to focus the sharpness-aware minimization on tail classes specifically, and maintain the standard optimization for head classes.

The gradient update is also changed accordingly as

$$\begin{aligned} \boldsymbol{\theta}_t = \boldsymbol{\theta}_t - \alpha_t \big[& \nabla \mathcal{L}_{\mathcal{S}^{\text{tail}}}(\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t^{\text{tail}}) \\ & + \nabla \mathcal{L}_{\mathcal{S}^{\text{head}}}(\boldsymbol{\theta}_t) + \lambda \boldsymbol{\theta}_t \big] \end{aligned} \tag{18}$$

with $\boldsymbol{\epsilon}_t^{\text{tail}}$ calculated as

$$\boldsymbol{\epsilon}^{\text{tail}} = \rho \frac{\nabla \mathcal{L}_{\mathcal{S}^{\text{tail}}}(\boldsymbol{\theta}_t)}{\|\nabla \mathcal{L}_{\mathcal{S}^{\text{tail}}}(\boldsymbol{\theta}_t)\|_2}. \tag{19}$$

As demonstrated in the above gradient update, the proposed ImbSAM suspends the sharpness-aware minimization for head classes, which are less prone to overfitting with the support of sufficient training data.

By incorporating class priors into the class-agnostic SAM, our ImbSAM efficiently restricts its uncontrollable generalization scope from all classes, which are typically dominated by head classes with overwhelming data, to the overlooked tail classes that are plagued with overfitting problems. Algorithm 1 outlines the full ImbSAM algorithm, with SGD as the base gradient optimizer. Additionally, the pseudo-code in `PyTorch` style of our ImbSAM is displayed in the supplementary, which only requires a few changes in the implementation of the standard SAM. We also provide the other implementation with the Huawei MindSpore toolkit at `https://github.com/cool-xuan/Imbalanced_SAM`.

4. Experiments

Comprehensive empirical experiments are conducted to verify the generalization improvement efficacy of our ImbSAM when confront with class-imbalanced data. Our experiments encompass two prototypical applications: Long-Tailed Classification (LTC) and Semi-Supervised Anomaly Detection (SSAD), demonstrating the broad applicability of the proposed ImbSAM. In all experiments, we evaluate the effectiveness of our ImbSAM by simply replacing the original optimization algorithm (SGD [6] with momentum [54]) used in existing methods with our class-aware ImbSAM, without any other hyperparameter changing.

4.1. Long-Tailed Classification (LTC)

Datasets. We conduct experiments on three mainstream long-tailed datasets including CIFAR100-LT, ImageNet-LT, and iNaturalist. CIFAR100-LT [12] and ImageNet-LT [39] are artificially truncated from the balanced CIFAR100 [32] and ImageNet [13] datasets, while iNaturalist [57] is a large-scale naturally imbalanced dataset comprising $8,142$ species with the number of samples per class ranges from $1,000$ to 2. ImageNet-LT also has 1000 classes like the balanced version and the number of samples per class ranges from $1,280$ to 5 images. Particularly, there are three sub-versions of CIFAR100-LT by varying the imbalanced factor (IF) in $[100, 50, 10]$.

Evaluation Protocol. In long-tailed classification, all classes with long-tailed distribution are treated equally during testing. The overall accuracy is calculated as $\text{acc}_{\text{All}}{=}\frac{1}{K}\sum \text{acc}_k$ including *All* classes, where acc_k is the Top-1 recognition accuracy for class k. Following [39], we also report accuracy on three splits of classes according to the number of training data: *Many* classes (>100), *Medium* classes (20∼100), and *Few* classes (<20).

Implementation. Our ImbSAM is combined with existing long-tailed classification methods to demonstrate its efficacy, including the baseline trained by cross-entropy loss (CE) with fine-tuned weight decay [2] and two strong SOTA methods: weight balancing (WB) [2] for CIFAR100-LT and learnable weight scaling (LWS) [31] for the other two large-scale datasets. For a fair comparison to prior methods, we use ResNet32 [12] on CIFAR100-LT, ResNeXt50 [60] on ImageNet-LT, and ResNet50 [24] on iNaturalist2018. We set SGD optimizer with momentum 0.9 as the base optimizer and train all models for 200 epochs, with a batch size of 64 for CIFAR100-LT and ImageNet-LT, and 512 for iNaturalist. For all experiments, if not specified, the hyperparameters ρ=0.05 and η is set as 100, which equals the upper limit of *Medium* classes.

Comparison on CIFAR100-LT. As reported in Table 1, our ImbSAM demonstrates consistently significant accuracy improvement for the naive baseline (CE) or strong SOTA (WB [2]) on the CIFAR100-LT dataset. When combined with the prior SOTA WB, our ImbSAM further achieves a novel SOTA performance with overall accuracy 54.8%, 59.3% and 69.7% respectively, on all three sub-versions (IF=[100, 50, 10]).

Furthermore, we report the detailed accuracy of *Many*, *Medium*, and *Few* on the most imbalanced CIFAR100-LT

Table 1: **Comparison of overall accuracy** (%) **on CIFAR100-LT** with IF=[100, 50, 10]. The reported accuracy of CE and WB [2] is implemented by ourselves. '*' refers to the SOTA with bells and whistles.

Imbalance Factor	100	50	10
CB [12]	39.6	45.2	58.0
KD [26]	40.4	45.5	59.2
LDAM-DRW [9]	42.0	46.6	58.7
BBN [68]	42.6	47.0	58.7
De-confound [55]	44.1	50.3	59.6
τ-norm [31]	47.7	52.5	63.8
DiVE [25]	45.4	51.1	62.0
DRO-LT [52]	47.3	57.6	63.4
SSD* [34]	46.0	50.5	62.3
PaCO* [11]	52.0	56.0	64.2
ACE* [8]	49.6	51.9	–
CE [2]	47.6	52.8	66.9
CE+**ImbSAM**	**51.1**	**56.4**	**69.2**
WB [2]	51.9	56.7	68.9
WB+**ImbSAM**	**54.8**	**59.3**	**69.7**

Table 2: **Comparison of overall accuracy and split accuracy** (%) **on large-scale ImageNet-LT and iNaturalist**. The reported accuracy of CE [2] and LWS [31] is implemented by ourselves. '*' refers to the SOTA with bells and whistles. The unreported accuracy in [52] and [34] is replaced with '-'. 'Med.' denotes 'Medium' classes.

	ImageNet-LT [39]				iNaturaList [57]			
	Many	Med.	Few	All	Many	Med.	Few	All
CB [12]	39.6	32.7	16.8	33.2	53.4	54.8	53.2	54.0
τ-norm [31]	59.1	46.9	30.7	49.4	65.6	65.3	65.5	65.6
DiVE [25]	64.1	50.4	31.5	53.1	70.6	70.0	67.7	69.1
DRO-LT [52]	64.0	49.8	33.1	53.5	–	–	–	69.7
DisAlign [66]	61.3	52.2	31.4	52.9	69.0	71.1	70.2	70.6
WB [2]	62.5	50.4	41.5	53.9	71.2	70.4	69.7	70.2
PaCO* [11]	63.2	51.6	39.2	54.4	69.5	72.3	73.1	72.3
SSD* [34]	66.8	53.1	35.4	56.0	–	–	–	71.5
RIDE* [59]	67.9	52.3	36.0	56.1	66.5	72.1	71.5	71.3
CE [2]	69.3	41.7	10.3	48.2	75.4	66.9	61.7	65.7
CE+SAM [17]	**70.0**	41.1	10.2	48.2	**75.6**	66.7	61.8	65.7
CE+**ImbSAM**	68.5	**47.5**	**21.6**	**52.2**	73.5	**69.2**	**67.9**	**69.1**
LWS [31]	**64.1**	49.1	31.2	52.5	**71.7**	69.4	68.7	69.4
LWS+SAM [17]	64.0	48.8	30.5	52.3	**71.7**	69.6	68.8	69.5
LWS+**ImbSAM**	63.2	**53.7**	**38.3**	**55.3**	68.2	**72.5**	**72.9**	**71.1**

with IF=100 in Table 3, where we also apply the standard SAM ($\rho = 0.05$ [17]) to these methods and coordinate it with large weights for tail classes. As we claimed in Section 3.3, the class-agnostic SAM only performs its powerful generalization improvement on the dominant head classes while neglecting the severe overfitting lying in tail classes. Assigning SAM with large loss weights up to 20 does not yield an obvious corrective effect. However, with the class priors, our ImbSAM efficiently addresses the overfitting issues for tail classes, resulting in a significant accuracy improvement (>4%) for the *Medium* and *Few* sub-sets without heavily sacrificing the performance on *Many* classes.

Detailed performance improvements for each class. We also visualize the accuracy gains of SAM and our ImbSAM for each class in Figure 2, which intuitively displays our ImbSAM's efficacy in avoiding overfitting for the tail classes with barely a handful of training samples. Besides, we calculate the Eigen Spectral Density [62] for CE respectively optimized by SGD [6], SAM and the proposed ImbSAM on *Few* classes in CIFAR100-LT, as illustrated in Figure 3. Particularly, the maximum eigenvalue (λ_{max}) and the trace of the Hessian matrix ($Tr(H)$) derived from the model optimized by ImbSAM are the smallest among the model trained by three optimizers, empirically demonstrating the efficacy of our ImbSAM in performing smoothness optimization targeting tail classes.

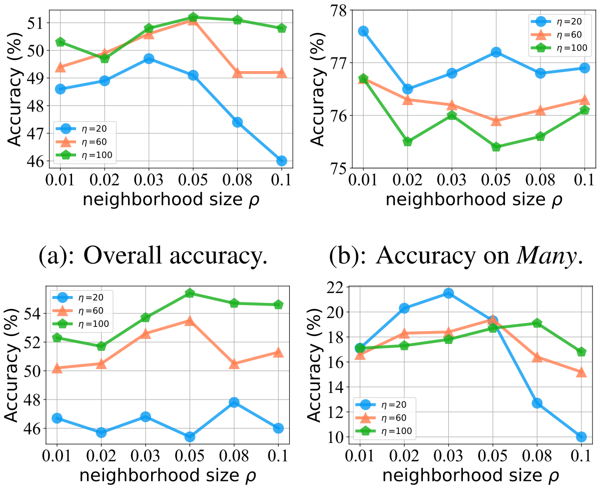
Comparison on ImageNet-LT and iNaturalist. On these two large-scale datasets following irregular and complex data distribution, our ImbSAM also exhibits superior accuracy gains. Specifically, Table 2 shows that when combined with the selected baseline (CE [2]) and SOTA (LWS [31]), the proposed ImbSAM yields significant accuracy improvements of 3% $\sim$ 11% on the *Medium* and *Few* class splits. Although our ImbSAM slightly affects the performance on *Few* classes, it still leads to overall accuracy increases of 1.7% $\sim$ 4%. Without bells and whistles, our ImbSAM promotes the prior method LWS [31] to be comparable with the prior SOTAs that commonly employ ensemble learning (RIDE [59]), self-pretraining (PaCO [11] or SSD [34]), and achieves a novel SOTA performance of the trade-off on *Medium* and *Few* classes.

Table 3: **Split accuracy** (%) **comparison on CIFAR100-LT (IF=100)** between SAM [17] and our ImbSAM. In particular, we assign the class-agnostic SAM with large weights (2, 3, 5, 10, and 20) for tail classes while no re-weighting for our ImbSAM. 'Med.' denotes 'Medium'.

Models	Many	Med.	Few	All
CE [2]	77.8	46.6	13.5	47.6
CE+SAM [17]	79.7	49.3	12.4	48.9
CE+SAM (reweight=2)	79.4	49.1	12.3	48.7
CE+SAM (reweight=3)	79.9	48.9	13.4	49.1
CE+SAM (reweight=5)	79.5	49.2	13.2	49.0
CE+SAM (reweight=10)	79.7	48.6	13.1	48.8
CE+SAM (reweight=20)	**80.1**	48.7	13.5	49.1
CE+**ImbSAM**	75.9	**53.5**	**19.4**	**51.1**
LTR [2]	68.5	49.1	35.9	51.9
LTR+SAM [17]	**76.1**	54.6	25.3	53.3
LTR+**ImbSAM**	64.1	**58.6**	**39.4**	**54.8**

In particular, on these two large-scale datasets containing much more classes ($1,000$ classes for ImageNet-LT and $8,142$ classes for iNaturalist) than the artificial CIFAR100-

LT, the standard SAM loses its impressive efficacy in generalization improving even for *Many* classes. This is because the head classes in these datasets have sufficient training data, which inherently guarantees their generalization [22]. In this case, the classification performance is mainly limited by the model capacity [5].



(a): Overall accuracy.

(b): Accuracy on *Many*.

(c): Accuracy on *Medium*.

(d): Accuracy on *Few*.

Figure 4: **Detailed accuracy** (%) **variation on CIFAR100-LT (IF=100)** by varying the neighborhood size ρ and class split threshold η in our ImbSAM, based on the baseline naive trained with cross-entropy loss (CE) [2].

Ablation study about ρ and η in ImbSAM. We conduct the ablation studies by varying neighborhood size $\rho \in$ [0.01, 0.02, 0.03, 0.05, 0.08, 0.1] and $\eta \in$ [20, 60, 100] on CIFAR100-LT with IF=100, based on the naive baseline (CE) optimized with our ImbSAM. According to the overall accuracy variation tendency as shown in Figure 4(a), the performance achieves the best with ρ=0.05 and class split threshold η=100, which exactly equals to the upper bound of *Medium* classes. Particularly, for *Few* classes, our ImbSAM boosts CE to achieve a superior accuracy up to 21.5% when η is set as the threshold of *Few* classes, which further verifies the controllable generalization scope of our ImbSAM.

Similar to SAM [17], ρ=0.05 is the best hyperparameter setting on *Medium* and *Few* classes, when η is set 60 or 100 (Figure 4(b) and Figure 4(c)). However, the best setting for ρ is 0.03 under η=20 as shown in Figure 4(d), and the accuracy on *Few* classes is dramatically reduced with ρ increasing. Since the harsh data limitation ($<$ 20), the gradient supervision derived from neighborhoods of the large range is noisy, which contributes little to model training.

Table 4: **Comparison of AUCROC score** (%) **on five AD datasets** with anomaly ratio γ_l=25%. '*' refers to the SOTA with ensembling.

Model	CIFAR	F-MNIST	MNIST-C	MVTec	SVHN	Avg
GANomaly [1]	67.8	79.4	75.1	76.0	56.9	71.0
REPEN [42]	67.9	87.1	80.9	75.9	58.8	74.1
PReNet [43]	87.5	96.1	94.4	90.2	78.7	89.4
FEAWAD [69]	85.2	95.1	95.4	96.2	77.4	89.9
XGBOD* [67]	87.8	96.4	95.8	99.1	81.2	92.1
DeepSAD [51]	86.5	96.3	**96.4**	93.1	80.9	90.6
+SAM [17]	86.9	96.7	96.1	91.3	81.0	90.4
+ImbSAM	**87.9**	**97.0**	96.2	**95.3**	**82.4**	**91.8**
DevNet [44]	88.4	96.4	95.6	95.6	82.1	91.6
+SAM [17]	88.2	96.2	95.3	94.2	81.0	91.0
+ImbSAM	**88.7**	**96.7**	**96.0**	**96.2**	**83.7**	**92.2**

4.2. Semi-Supervised Anomaly Detection (SSAD)

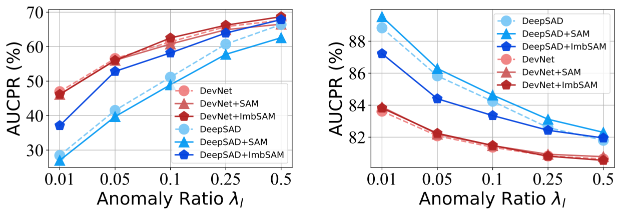
Compared with long-tailed classification, although there are only one head class (normal) and one tail class (anomaly) in SSAD, the diversity and uncertainty of anomalies make this task challenging.

Datasets. Following the impressive ADBench [20], we evaluate our ImbSAM on five image datasets: CIFAR10, SVHN, FashionMNIST, MNIST-C, and MVTec-AD. The former 4 datasets respectively contain 10 sub-sets with one class as normal and other classes as abnormal. For MNIST-C, original MNIST samples are set as normal and corrupted images as abnormal. In MVTec-AD, 15 types of industrial products are collected with accepts as normal and defects as abnormal. The number of accessible anomalies during training is controlled by the anomaly ratio γ_l following [20].

Evaluation protocol. We calculate the widely-used AUCROC (Area Under Receiver Operating Characteristic Curve) and AUCPR (Area Under Precision-Recall Curve) to evaluate the detection accuracy following [20]. In particular, we report the AUCPR w.r.t. both normal and anomaly, which can be viewed as the detection performance for normal and anomaly, respectively. Notably, we only report the dataset-wise performance and the average performance of five datasets in the text, detailed performance in each dataset is displayed in our supplementary.

Implementation. The SOTA methods DeepSAD [51] and DevNet [44] are selected as two strong baselines to combine with our ImbSAM. For all SSAD datasets, the frozen ResNet18 [24] is adopted to extract features. All models are trained for 50 epochs with a batch size of 128, and ρ in SAM and ImbSAM is set as 0.05 for all experiments.

Comparison on SSAD datasets. When applied to SSAD, our ImbSAM simply treats the anomalies as the only tail class and adopts the generalization capacity of SAM targeting the optimization for anomaly perception. As shown in Table 4, ImbSAM further boosts the selected two strong baselines, DeepSAD [51] and DevNet [44], to



(a) Anomaly

(b) Normal

Figure 5: **AUCPR of anomaly and normal** for the DeepSAD [51] and DevNet [44] optimized by SGD [6], SAM [17] and our ImbSAM, averaged on five SSAD datasets.

achieve higher AUCROC scores, outperforming the prior ensembling SOTA (92.1% of XGBOD [67] v.s. 92.2% of DevNet optimized with our ImbSAM). In particular, DevNet incorporated with our ImbSAM achieves superior performance over XGBOD on all datasets except for MVTec AD, an industrial defect detection dataset. As for this specific dataset, XGBOD benefits from the prediction of ensemble unsupervised methods only trained by normal samples, thoroughly capturing the characteristics of defect-free data, leading to the best performance.

To further confirm the generalization scope of our ImbSAM, we report the AUCPR for normal and anomaly respectively, with an increasing anomaly ratio $\gamma_l \in$ [1%, 5%, 10%, 25%, 50%]. Notably, the imbalanced factor $IF \gg 1$ even when $\gamma_l = 50\%$. As displayed in Figure 5, our ImbSAM significantly improves the AUCPR of anomaly for DeepSAD by about 10%, comparable with the DevNet that introduces re-sampling (Figure 5(a)), while SAM only slightly increases AUCPR on the side of normal (Figure 5(b)). In particular, with γ_l <0.1, the anomaly AUCPR of DevNet slightly drops when optimized with our ImbSAM. This also verifies our analysis in Section 3.3 that overexpose of limited data conflicts with SAM.

5. Conclusion

In order to adapt the promising SAM to tackle the overfitting issues in class-imbalanced recognition, we leverage class priors to control the generalization scope of SAM to focus tail classes and propose a class-aware ImbSAM. Our ImbSAM demonstrates remarkable performance improvement especially for the classes with limited training data, achieving novel SOTA in both long-tailed classification and semi-supervised anomaly detection.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62222203, 61976049) and CAAI-Huawei MindSpore Open Fund.

References

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *ACCV*. Springer, 2019. 3, 8

[2] Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *CVPR*, 2022. 1, 2, 3, 6, 7, 8

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[4] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006. 3

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 8

[6] Léon Bottou. Stochastic gradient descent tricks. *Neural Networks: Tricks of the Trade: Second Edition*, 2012. 1, 3, 6, 7, 9

[7] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 2018. 1, 2

[8] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *ICCV*, 2021. 2, 5, 7

[9] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS*, 2019. 1, 2, 7

[10] Rich Caruana, Steve Lawrence, and C Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *NeurIPS*, 2000. 3

[11] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *ICCV*, 2021. 1, 2, 7

[12] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 1, 2, 3, 5, 6, 7

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2, 3, 6

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020. 1

[15] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, 2003. 1, 2

[16] Chengjian Feng, Yujie Zhong, and Weilin Huang. Exploring classification equilibrium in long-tailed object detection. In *ICCV*, 2021. 1, 2, 5

[17] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *ICLR*, 2021. 1, 2, 3, 5, 7, 8, 9

[18] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 2017. 2, 5

[19] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *ICIC*. Springer, 2005. 1, 2, 5

[20] Songqiao Han, Xiyang Hu, Hailiang Huang, Mingqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. *NeurIPS*, 2022. 1, 2, 3, 8

[21] David J Hand and Robert J Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 2001. 3

[22] Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 2004. 1, 2, 5, 8

[23] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *TKDE*, 2009. 1, 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2, 6, 8

[25] Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. Distilling virtual examples for long-tailed recognition. In *ICCV*, 2021. 7

[26] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NeurIPS*, 2015. 7

[27] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 3

[28] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 2002. 1, 2

[29] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *ICLR*, 2019. 3

[30] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 2019. 2

[31] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *ICLR*, 2019. 1, 2, 6, 7

[32] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 2, 6

[33] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *ICML*. PMLR, 2021. 3

[34] Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *ICCV*, 2021. 2, 7

[35] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 1, 2, 3, 5

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[37] Si Liu, Risheek Garrepalli, Thomas Dietterich, Alan Fern, and Dan Hendrycks. Open category detection with pac guarantees. In *ICML*, 2018. 2

[38] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE TSMC*, 2008. 2

[39] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019. 1, 5, 6, 7

[40] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[41] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *NeurIPS*, 2017. 3

[42] Guansong Pang, Longbing Cao, Ling Chen, and Huan Liu. Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *ACM SIGKDD*, 2018. 3, 8

[43] Guansong Pang, Chunhua Shen, Huidong Jin, and Anton van den Hengel. Deep weakly-supervised anomaly detection. *arXiv preprint arXiv:1910.13601*, 2019. 3, 8

[44] Guansong Pang, Chunhua Shen, and Anton van den Hengel. Deep anomaly detection with deviation networks. In *ACM SIGKDD*, 2019. 1, 2, 3, 8, 9

[45] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 3

[46] Didik Purwanto, Yie-Tarng Chen, and Wen-Hsien Fang. Dance with self-attention: A new look of conditional random fields on anomaly detection in videos. In *ICCV*, 2021. 3

[47] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, and R Venkatesh Babu. Escaping saddle points for effective generalization on class-imbalanced data. *NeurIPS*, 2022. 3

[48] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, Arihant Jain, and Venkatesh Babu Radhakrishnan. A closer look at smoothness in domain adversarial training. In *ICML*. PMLR, 2022. 3

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 1

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 2015. 1

[51] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *ICLR*, 2019. 1, 2, 3, 8, 9

[52] Dvir Samuel and Gal Chechik. Distributional robustness loss for long-tail learning. In *ICCV*, 2021. 2, 7

[53] Shelly Sheynin, Sagie Benaim, and Lior Wolf. A hierarchical transformation-discriminating generative model for few shot anomaly detection. In *ICCV*, 2021. 3

[54] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*. PMLR, 2013. 6

[55] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *NeurIPS*, 2020. 7

[56] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019. 1

[57] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 1, 5, 6, 7

[58] Dong Wang, Yicheng Liu, Liangji Fang, Fanhua Shang, Yuanyuan Liu, and Hongying Liu. Balanced gradient penalty improves deep long-tailed learning. In *ACM MM*, 2022. 3, 5

[59] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2020. 2, 7

[60] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 6

[61] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. *NeurIPS*, 2020. 1

[62] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *Big data*. IEEE, 2020. 7

[63] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *CVPR*, 2019. 2

[64] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem - a discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, 2021. 3

[65] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ICLR*, 2017. 3

[66] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, 2021. 7

[67] Yue Zhao and Maciej K Hryniewicki. Xgbod: improving supervised outlier detection with unsupervised representation learning. In *IJCNN*. IEEE, 2018. 3, 8, 9

[68] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, 2020. 7

[69] Yingjie Zhou, Xucheng Song, Yanru Zhang, Fanxing Liu, Ce Zhu, and Lingqiao Liu. Feature encoding with autoencoders for weakly supervised anomaly detection. *IEEE TNNLS*, 2021. 3, 8