BALANCING MIXED LABELS: MIXUP MEETS NEURAL COLLAPSE IN IMBALANCED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Minority collapse, where minor classes become indistinguishable, is a key challenge in imbalanced learning, addressed by methods like Mixup with class-balanced sampling. In parallel, a simplex equiangular tight frame from Neural Collapse (NC) has emerged as an effective frame in the classifier to mitigate minority collapse. While NC has been studied in both Mixup and imbalanced learning independently, its combination remains unexplored, particularly regarding the balance of mixed labels. We investigate this overlooked factor and pose the question: *Is the mixed label balance important for alleviating minority collapse?* Our analysis reveals that (i) mixed labels should be balanced, and (ii) in this setting, interpreting mixed labels as singletons is beneficial. Building on the analysis, we propose a balanced mixed label sampler and a mixed-singleton classifier, which balance mixed labels and treat them as singleton labels. Through theoretical analysis, visualization, and ablation studies, we demonstrate the effectiveness of our approach. Experiments on standard benchmarks further confirm consistent performance gains, highlighting the importance of balancing mixed labels in imbalanced learning.

1 INTRODUCTION

In imbalanced learning, severe class imbalance often causes a significant degradation of model accuracy, particularly on the minority classes (Liu et al., 2019). One known cause of this performance drop is the phenomenon termed *minority collapse* (Fang et al., 2021), wherein the class vectors of minority classes converge and become nearly identical. To mitigate this issue, a wide range of strategies has been explored, including data augmentation (Zhang et al., 2018; Verma et al., 2019; Shi et al., 2023), calibration technique (Zhong et al., 2021), mixture-of-experts models (Cai et al., 2021; Zhang et al., 2021; Xiang et al., 2020), and class-balanced loss functions (Cao et al., 2019; Cui et al., 2019) or sampling schemes (Kang et al., 2020; Cao et al., 2019; Zhang et al., 2022; Shen & Lin, 2016). Among these approaches, Mixup, especially when combined with class-balanced sampling, has been shown to effectively improve the model performance under class-imbalanced conditions.

Meanwhile, Neural Collapse (NC) (Papyan et al., 2020) has emerged as a key framework for analyzing geometric properties of last-layer features and classifier in classification models at the terminal phase of training. Although NC has been studied in both Mixup (Fisher et al., 2024) and imbalanced learning (Liu et al., 2023; Yang et al., 2022) separately, Mixup in imbalanced settings has not been investigated in conjunction with NC. In particular, the balance of mixed labels has received little attention. The only related finding comes from M-lab NC (Li et al., 2024), which observes that even when multi-label samples are imbalanced, NC occurs at the singleton-class level as long as singleton label samples are balanced, with multi-label class emerging as combinations of singletons. However, whether the balance of input samples still hold for mixed labels under Mixup remains unclear. This motivates our central research question:

Could the balance of mixed labels be a critical factor in minority collapse?

Building on the proof approach of Fang et al. (2021), we first demonstrate that minority collapse still occurs under Mixup when the mixed-label samples are not balanced. Motivated by this theorem, we conducted label-variants experiments, which revealed that epoch-wise imbalance in mixed labels significantly impacts the model performance. To address this issue, we propose a sampler that balances mixed labels across epochs. Both theoretically and empirically, we show that aligning the balance of mixed-label samples across epochs mitigates minority collapse. Furthermore, our analysis uncovers

that minority collapse in Mixup is determined solely by the frequency of singleton and mixed-label samples, independent of the mixup ratio. Leveraging this insight, we introduce a simple but effective classifier, which treats mixed labels as singleton labels when learning class vectors. Compared with a conventional singleton classifier implemented as a fully connected layer, our approach achieves superior performance, particularly improving accuracy on minority classes.

Our contributions are formally outlined as follows.

- **Proposed methods: Balanced Mixed Label Sampler and Mixed-Singleton Classifier** (§4). We propose the Balanced Mixed Label Sampler, which balances the frequency of mixed-label samples during training, and the Mixed-Singleton Classifier, which learns class vectors by treating mixed labels as singletons under a mixed-label balanced condition.
- **Finding 1: The balance of mixed-label samples is crucial** (§5.2). Using a Layer-Peeled Model with Mixup, we show that ensuring balanced frequencies of mixed-label samples in a singleton-class imbalanced setting mitigates minority collapse.
- **Finding 2: Interpreting mixed labels as singletons further alleviates minority collapse** (§5.3). From our analysis, we find that minority collapse in Mixup depends on not only the frequency of singleton label but also that of mixed-label samples. From our analysis, we find that minority collapse in Mixup depends not only on the frequency of singleton labels but also on that of mixed-label samples. Based on this observation, we demonstrate that treating mixed labels as singletons within the classifier further reduces minority collapse beyond balancing frequencies alone.
- **Empirical validation on our analysis and proposed methods** (§6). Extensive visualization and ablation studies confirm the effectiveness of our methods, and experiments on standard imbalanced learning benchmarks demonstrate consistent performance improvements.

2 Related Work

Due to the paper's length constraints, this section has been moved to the Appendix A, except for the discussion of our work's novelty.

Mixup-based Method. Many attempts have been made to address the challenges of imbalanced learning environments using Mixup (Zhang et al., 2018), which increases the diversity of sampled data and alleviates risk of overfitting on tail classes (Zhang et al., 2018), including data augmentation, architecture improvements, and calibration methods. However, no research has specifically studied on the balance of mixed labels in minority collapse.

Class-balanced Methods. Various class-balanced samplers have been proposed, yet no work has mainly focused on the balance of mixed labels. Additionally, while Logit Adjustment (Menon et al., 2021) and UniMix (Xu et al., 2021) have concentrated on the effect of the class vectors of singleton labels, they did not interpret mixed labels as singletons.

Neural Collapse in Mixup and Imbalanced Learning. NC in imbalanced learning has been studied in Fang et al. (2021). To alleviate the minority collapse, Yang et al. (2022) assumed that the classifier is fixed to the K-simplex ETF and proved that LPM with the classifier satisfies NC properties. Also, the fixed ETF classifier with Mixup has improved the model performance in imbalanced learning. Building on the theorems, Fisher et al. (2024) proved Mixup also satisfies NC properties for both same class and different class. However, Yang et al. (2022) and Fisher et al. (2024) did not consider the minority collapse from the mixed label balance in the LPM with learnable classifiers.

3 Preliminary

3.1 Notations

Let $\mathcal{X}$ be the dataset with N samples where the number of singleton label classes is K and $\mathbb{S}$ be the set of their feature vectors $\boldsymbol{h}$. Then, we formulate them as $\mathcal{X} := [(\boldsymbol{x}_i, c_i)]_{i=1}^N$ where c_i is the class label of the i-th sample x_i and $\mathbb{S} := \{\boldsymbol{h}_i\}_{i=1}^N$. As a result, we define $\boldsymbol{y}_i = \boldsymbol{e}^{(c_i)}$ as the one-hot vector of x_i. Then, we denote the subset of $\mathbb{S}$ which has only k-th class feature vectors $\boldsymbol{h}_{k,i}$ as $\mathbb{S}_k := \{\boldsymbol{h}_{k,i}\}_{i=1}^{n_k}$ where n_k is the number of k-th class samples and $k \in [K]$. Thus, $N = \sum_{k=1}^K n_k$.

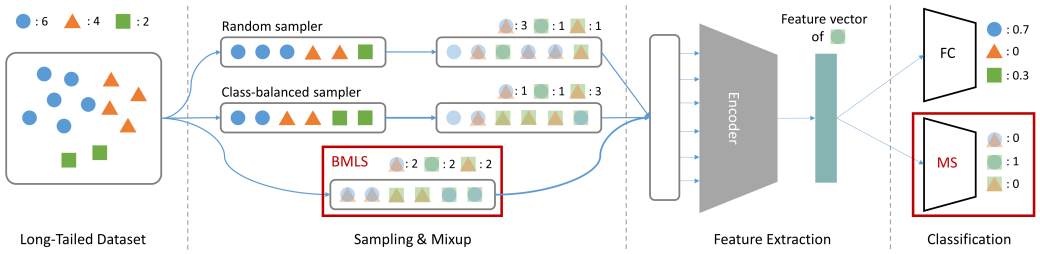


Figure 1: Overview of Balanced Mixed Label Sampler (BMLS) and Mixed-Singleton Classifier (MS)

3.2 Overview of Mixup

Mixup randomly permutes input samples and blends them with the ones before permutation, respectively. Let $\mathcal{I} := [i]_{i=1}^N$ be the indices of $\mathcal{X}$ and $\pi(\mathcal{I}) := [\pi(i)]_{i=1}^N$ be the permuted one where $\pi(i)$ represents the index number corresponding to i-th element of $\mathcal{I}$. Therefore, the index pairs of mixed samples $\mathcal{I}^\lambda$ is denoted as $\mathcal{I}^\lambda := [(i, \pi(i))]_{i\in\mathcal{I}}$. In this case, we denote $\mathcal{I}^\lambda_{(a,b)}$ as the index pairs of $(c_i, c_{\pi(i)}) = (a, b)$, and $\mathbb{S}^\lambda_{(a,b)}$ as the mixed feature set of (a, b)-label samples. Therefore, $\mathbb{S}^\lambda_{(a,b)} := \{\lambda \boldsymbol{h}_{a,i} + (1-\lambda)\boldsymbol{h}_{b,j} \,|(i,j) \in \mathcal{I}^\lambda_{(a,b)}\} = \{\boldsymbol{h}^\lambda_{(a,b),i}\}_{i=1}^{n_{(a,b)}}$ where $(a, b) \in \mathbb{K}^2$, $n_k = |\mathcal{I}^\lambda_{(a,b)}|$, and $\mathbb{K}^2 = \{(a, b)|1 \le a \le K,\ 1 \le b \le K\}$. Thus, $N = \sum_{(a,b)\in\mathbb{K}^2} n_{(a,b)}$.

Based on the notations, we perform mixup on each pair defined by $\mathcal{I}^\lambda$ to create mixed-label samples by linearly interpolating them:

$$\boldsymbol{x}_i^\lambda = \lambda \boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_{\pi(i)}, \boldsymbol{y}_i^\lambda = \lambda \boldsymbol{y}_{c_i} + (1-\lambda)\boldsymbol{y}_{c_{\pi(i)}}, \forall (i, \pi(i)) \in \mathcal{I}^\lambda, \tag{1}$$

where the mixup ratio $\lambda \in (0, 1)$ is sampled from the beta distribution D_λ, i.e., $\lambda \sim D_\lambda(\alpha, \alpha)$ and α is a hyperparameter.

4 Proposed Method

4.1 Balanced Mixed Label Sampler

We propose the Balanced Mixed Label Sampler (BMLS), where the frequency of all mixed-label samples is equal in each epoch as shown in Figure 1. When using BMLS, the probability of sampling of a (a, b)-label sample is

$$P_{(i,\pi(i))|(i,\pi(i))\in\tilde{\mathcal{I}}^\lambda} = \frac{1}{N}. \tag{2}$$

$\tilde{\mathcal{I}}^\lambda$ is the index pairs of samples where their mixed-labels are balanced, i.e., $n_{(a,b)} = n$ for all $(a, b) \in \mathbb{K}^2$. As done in the class-aware sampler (Shen & Lin, 2016), we remove the randomness by pre-defining $\tilde{\mathcal{I}}^\lambda$ for every epoch. After generating $\tilde{\mathcal{I}}^\lambda$, we simply replace $\mathcal{I}^\lambda$ to $\tilde{\mathcal{I}}^\lambda$ in Eq. 1

4.2 Mixed-Singleton Classifier

Let $\boldsymbol{W} \in \mathbb{R}^{K\times p}$ be a classifier of singleton labels, which is a fully-connected layer. We define the Mixed-Singleton Classifier (MS) as

$$\boldsymbol{W}^\lambda = [\lambda \boldsymbol{w}_a + (1-\lambda)\boldsymbol{w}_b]_{(a,b)\in\mathbb{K}^2}, \tag{3}$$

where p is the last-layer feature dimension, as shown in Figure 1. We replace the singleton classifier with MS and perform Mixup with BMLS, where mixed-label samples $\tilde{\boldsymbol{x}}_i^\lambda$ and their one-hot vectors $\tilde{\boldsymbol{y}}_i^\lambda$ are defined as:

$$\tilde{\boldsymbol{x}}_i^\lambda = \lambda \boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_{\pi(i)}, \tilde{\boldsymbol{y}}_i^\lambda = \boldsymbol{e}^{\mathcal{I}^2(c_i, c_{\pi(i)})}, \forall (i, \pi(i)) \in \tilde{\mathcal{I}}^\lambda, \tag{4}$$

where $\mathcal{I}^2$ denotes the index pairs of $\mathbb{K}^2$, and $\mathcal{I}^2(a, b)$ gives the index number of $(a, b) \in \mathbb{K}^2$.

Building on these methods, we generated mixed labels (a, b) only for the case where $a < b$, ensuring that the existing theorem and proposition still hold, thereby mitigating the limitations of both methods. The limitation and proof are described in §8 and Appendix C.3.

5 THEORETICAL ANALYSIS

5.1 PROOF SKETCH

We first present a proof sketch that outlines the approach we followed to propose and prove our theorems. Fang et al. (2021) proved that oversampling mitigates minority collapse when singleton label samples are imbalanced, following the sequence outlined below. (Gray indicates the part as defined in Fang et al. (2021).)

(1) Define the Layer-Peeled Model. (Eq. 7)
(2) Prove that NC properties are satisfied when the LPM has global optimality in the case where singleton label samples are balanced. (Theorem 1)
(3) Demonstrate that the LPM suffers from minority collapse in the case where singleton label samples are imbalanced. (Lemma 1 and Theorem 5)
(4) Show that oversampling alleviates minority collapse in the imbalanced case. (Proposition 1)

Our theorem and proof leverages strategies similar to those in Fang et al. (2021), but we extend these concepts to Mixup focusing on the balance of mixed label samples.

In §5.2, (1) we define the Layer-Peeled Model with Mixup (LPM$_\lambda$) and omit step (2), which holds true according to the theorem of Fisher et al. (2024); (3) we prove that in the imbalanced case the LPM$_\lambda$ also suffers from minority collapse; and in closing, (4) we show that the Balanced Mixed Label Sampler (BMLS) alleviates the minority collapse. In §5.3, we extend the LPM$_\lambda$ by modifying the classifier: (1) we newly define the Layer-Peeled Model with Mixup and Mixed-Singleton Classifier (LPM$_\lambda$-MS); (2) we prove that when this model achieves global optimality, it also satisfies the NC properties; and finally, following the same reasoning as in §5.2, (3–4) we show that in the imbalanced case the LPM$_\lambda$-MS suffers from minority collapse, and that BMLS is effective to the minority collapse even in this setting.

5.2 LABEL IMBALANCE IN MIXUP

Remark 1. According to Theorem 1, which will be described in this section, Mixup also experiences the minority collapse. Additionally, even when using class-balanced samplers to alleviate label suppression and learn an unbiased classifier, minority collapse is partially mitigated but not fully resolved. This is because Mixup blends input samples with a random permutation of them, and as a result, the balance of samples is disrupted, even when class-balanced samplers are employed. For this reason, when using Mixup in imbalanced learning, not only singleton labels but also mixed ones should be balanced, as proven in Proposition 1.

(1) Problem Settings. The Layer-Peeled Model (LPM) (Fang et al., 2021) is the optimization program of simplified neural network, modeled by only last-layer features and classifier. Following the definition of LPM, we obtain the Layer-Peeled Model with Mixup (LPM$_\lambda$):

$$\min_{\boldsymbol{W},\boldsymbol{H}^\lambda} \mathbb{E}_\lambda \frac{1}{N} \sum_{k\in\mathbb{K}^2} \sum_{i=1}^{n_k} \mathcal{L}(\boldsymbol{W}\boldsymbol{h}^\lambda_{k,i}, \boldsymbol{y}^\lambda_k) \text{ s.t. } \frac{1}{K}\sum_{k=1}^{K} \|\boldsymbol{w}_k\|^2 \le E_W,\ \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}^\lambda_{k,i}\right\|^2 \le E_H, \tag{5}$$

where $\boldsymbol{y}^\lambda_{(a,b)} = \lambda \boldsymbol{e}^{(a)} + (1-\lambda)\boldsymbol{e}^{(b)}$. For simplicity, we hereafter denote $\boldsymbol{W} = [\boldsymbol{w}_k]_{k=1}^K \in \mathbb{R}^{K\times p}$ for the weights of the classifier and the positive thresholds $E_W \propto 1/K$ and $E_H \propto 1/K$.

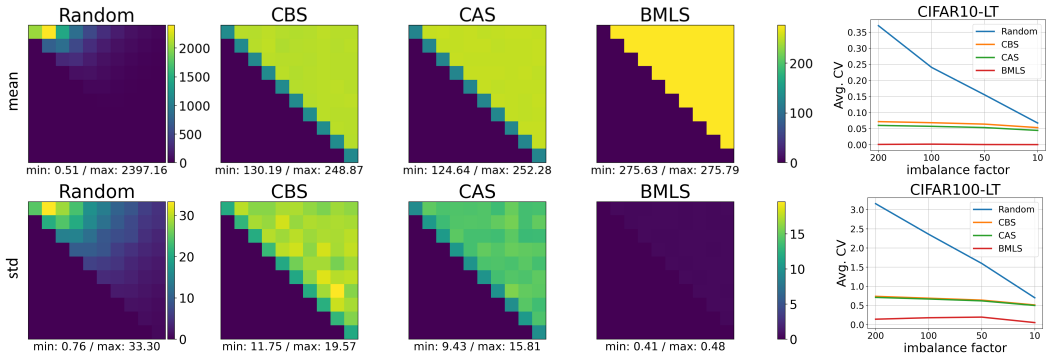
We present a convex optimization program that serves as a relaxation of the non-convex LPM$_\lambda$ (Eq. 5), leveraging the established result that a quadratically constrained quadratic program can be transformed into a semidefinite program. This formulation is provided as Eq. 11 in Appendix B.

(2) Satisfying NC properties. As proven in Fisher et al. (2024), when LPM$_\lambda$ (Eq. 5) has the global optimality, NC properties are satisfied. We omit this step.

(3) Minority collapse occurs in LPM$_\lambda$. Now, we are ready for proving that LPM$_\lambda$ also suffers from minority collapse. Lemma 1 below relates the solutions of Eq. 11 to that of Eq. 5.

Lemma 1. *Assume $p \ge K^2 + K$ and the loss function $\mathcal{L}$ is convex in its first argument. Let $\boldsymbol{X}^\star$ be a minimizer of the convex program (Eq. 11). Define $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ as*

$$\left[\boldsymbol{h}^\star_{(1,1)}, \boldsymbol{h}^\star_{(1,2)}, \ldots, \boldsymbol{h}^\star_{(K,K)},\ (\boldsymbol{W}^\star)^\top\right] = \boldsymbol{P}(\boldsymbol{X}^\star)^{1/2}, \tag{6}$$



(a) Heatmap of label frequencies on CIFAR10-LT (b) Avg. CV Curve

Figure 2: Mean and standard deviation of label frequencies including mixed label across epochs (higher imbalance factor means higher imbalanced)(The closer Avg. CV is to 0, the more evenly the labels appear across epochs)

$$\boldsymbol{h}^\star_{k,i} = \boldsymbol{h}^\star_k, \ \ \textit{for all } i \in \mathcal{I}^\lambda_k, k \in \mathbb{K}^2,$$

where $(\boldsymbol{X}^\star)^{1/2}$ *denotes the positive square root of* $\boldsymbol{X}^\star$ *and* $\boldsymbol{P} \in \mathbb{R}^{p\times(K^2+K)}$ *is any partial orthogonal matrix such that* $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}_{K^2+K}$*. Then,* $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ *is a minimizer of Eq. 5. Moreover, if all* $\boldsymbol{X}^\star$*'s satisfy* $\frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}^\star(k,k) = E_H$*, then all the solutions of Eq. 5 are in the form of Eq. 6.*

Proof. See Appendix C.1 □

Theorem 1. *Assume* $p \geq K$ *and* $n_A/n_B \to \infty$*, and fix* K_A *and* K_B*. Let* $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ *be any global minimizer of the LPM*$_\lambda$ *(Eq. 5). As the imbalance factor* $R \equiv n_A/n_B \to \infty$*, we have*

$$\lim \boldsymbol{w}^\star_k - \boldsymbol{w}^\star_{k'} = \boldsymbol{0}_p, \ \ \textit{for all } K_A < k < k' \leq K.$$

Proof. See Appendix C.2 □

From Lemma 1 and Theorem 1, we demonstrate that LPM$_\lambda$ also exhibits minority collapse.

(4) Balancing mixed labels mitigates minority collapse in LPM$_\lambda$. To formalize the behavior of a neural network trained by minimizing a new program with balanced samples including mixed-label ones through BMLS, we propose that it may perform as if it were trained on a larger dataset containing n_A examples in the majority class and $w_r n_B$ examples in the minority class. We begin by analyzing the LPM$_\lambda$ in the context of BMLS:

$$\min_{\boldsymbol{W},\boldsymbol{H}^\lambda} \frac{1}{N'}\left[\sum_{k\in\mathbb{K}^2_A}\sum_{i=1}^{n_A}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}^\lambda_{k,i}, \boldsymbol{y}^\lambda_k) + w_r\sum_{k\in\mathbb{K}^2_B}\sum_{i=1}^{n_B}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}^\lambda_{k,i}, \boldsymbol{y}^\lambda_k)\right] \tag{7}$$

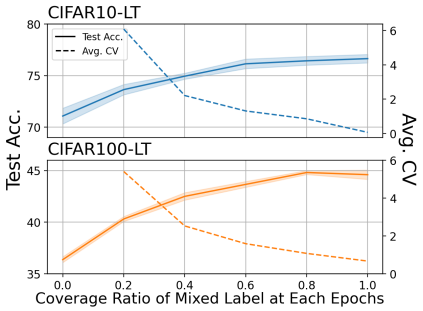
$$\text{s.t. } \frac{1}{K}\sum_{k=1}^{K}\|\boldsymbol{w}_k\|^2 \leq E_W, \ \frac{1}{|\mathbb{K}^2_A|}\sum_{k\in\mathbb{K}^2_A}\frac{1}{n_A}\sum_{i=1}^{n_A}\left\|\boldsymbol{h}^\lambda_{k,i}\right\|^2 + \frac{1}{|\mathbb{K}^2_B|}\sum_{k\in\mathbb{K}^2_B}\frac{1}{n_B}\sum_{i=1}^{n_B}\left\|\boldsymbol{h}^\lambda_{k,i}\right\|^2 \leq E_H,$$

where $N' = n_A|\mathbb{K}^2_A| + w_r n_B|\mathbb{K}^2_B|$

The following result supports the intuition that BMLS enhances the size of the minority classes in the LPM$_\lambda$. For simplicity, we omit the superscript λ in Proposition 1.

Proposition 1. *Assume* $p \geq K^2 + K$ *and the loss function* $\mathcal{L}$ *is convex in the first argument. Let* $\boldsymbol{X}^\star$ *be any minimizer of the convex program (11) with* $n_{(1,1)} = n_{(1,2)} = \cdots = n_{(K_A,K_A)} = n_A$ *and* $n_{(K_A+1,K_A+1)} = n_{(K_A+1,K_A+2)} = \cdots = n_{(K,K)} = w_r n_B$*. Define* $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ *as*

$$\left[\boldsymbol{h}^\star_{(1,1)}, \boldsymbol{h}^\star_{(1,2)}, \ldots, \boldsymbol{h}^\star_{(K,K)}, \ (\boldsymbol{W}^\star)^\top\right] = \boldsymbol{P}(\boldsymbol{X}^\star)^{1/2}, \tag{8}$$



$\mathcal{D}$	Ctgy.	Coverage ratio					
		0.0	0.2	0.4	0.6	0.8	1.0
C10	few	58.74	58.56	58.50	58.51	60.06	65.60
	med	68.46	71.09	74.63	77.85	77.10	75.21
	many	86.93	92.13	91.79	91.49	91.91	89.59
	all	71.08	73.64	74.94	76.14	76.43	76.64
C100	few	7.72	10.28	12.17	12.66	13.16	13.20
	med	34.49	39.13	42.95	44.47	46.21	46.97
	many	61.26	65.64	66.49	67.85	68.93	67.60
	all	36.37	40.31	42.50	43.66	44.81	44.60

Figure 3: Mixed-label frequency control experiments on CIFAR10/100 LT datasets. Coverage ratio represents the proportion of mixed labels used in training during one epoch compared to the total number of mixed labels. (e.g., when coverage ratio is 0.6 in CIFAR100-LT, the model trains on mixed labels consisting of combinations of 60 different classes, which change with each epoch.) (figure) Test Acc. (%) and Avg. CV over coverage ratio (table) Comparison of test accuracies

$$\boldsymbol{h}^\star_{k,i} = \boldsymbol{h}^\star_k, \;\; \textit{for all } i \in \mathcal{I}^\lambda_k, k \in \mathbb{K}^2_A, \;\; \boldsymbol{h}^\star_{k,i} = \boldsymbol{h}^\star_k, \;\; \textit{for all } i \in \mathcal{I}^\lambda_k, k \in \mathbb{K}^2_B,$$

where $\boldsymbol{P} \in \mathbb{R}^{p\times(K^2+K)}$ *is any partial orthogonal matrix such that* $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}_{K^2+K}$*. Then,* $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ *is a global minimizer of the mixed-label balanced LPM*$_\lambda$ *(Eq. 7). Moreover, if all* $\boldsymbol{X}^\star$*'s satisfy* $\frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\boldsymbol{X}^\star(k,k) = E_H$*, then all the solutions of Eq. 7 are in the form of Eq. 8.*

In conjunction with Lemma 1, Proposition 1 demonstrates that the number of training examples in each minority class is effectively $w_r n_B$ instead of n_B in the LPM$_\lambda$. In the special case where $w_r = n_A/n_B \equiv R$, the results indicate that the angles between any pair of last-layer classifiers are equal, regardless of whether they belong to the majority or minority classes.

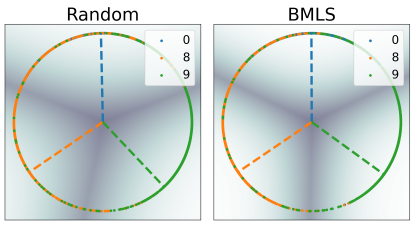
To demonstrate the empirical evidence of *Remark* 1, we examine the mean and standard deviation of label frequencies from various sampler: random sampler, class-balanced sampler (CBS) (Kang et al., 2020), class-aware sampler (CAS) (Shen & Lin, 2016), and ours (BMLS), as shown in Figure 2. We use the average of Coefficient of Variation ($\overline{\text{CV}}$) (Dodge, 2008) as the metric to measure the dispersion of each label frequency distributions: $\overline{\text{CV}} = \frac{1}{C}\sum_{c=1}^{C}\frac{\sigma_c}{\mu_c}$, where the lower $\overline{\text{CV}}$, the less dispersion, which means labels evenly appear across epochs. After training, the mean of label frequencies is almost balanced across all samplers, but epoch-wise balance is not. To empirically validate that the epoch-wise label imbalance is a problem in imbalanced learning, we do mixed-label frequency control experiments. As shown in Figure 3, the more imbalanced mixed label appears from epoch to epoch, the lower the performance of models.

5.3 Interpreting Mixed Labels as Singletons

Remark 2. As proven in Theorem 2, balancing mixed labels and interpreting them as singletons allows the LPM$_\lambda$-MS to operate in the same manner of the LPM. At the same time, it is expected to preserve the strong feature learning effect of Mixup while potentially reducing its negligible influence on classifier learning by maintaining mixed-label samples but removing the mixup loss.

Building on Theorem 1 and Proposition 1, we raise a conjecture: *If mixed labels are interpreted as singletons, then the mitigation of minority collapse will be enhanced.*

The rationale for the conjecture can be summarized as follows: *(i) Feature-based differences*. In Proposition 1, minority collapse occurs regardless of the mixup ratio λ. This is because the total loss derived from features is equivalent to that obtained without Mixup. However, the behavior of features differs: while the loss is divided between classes according to the mixup ratio λ, the mixed features are not generally decomposed in this way due to the non-linearity of the model; *(ii) Similarity in terms of sample frequency*. In addition, the minority collapse of LPM$_\lambda$ depends not only on the number of singleton label samples but also on that of mixed-label samples, as if the mixed labels were singletons; *(iii) Harmful influence on classifier learning*. Furthermore, MiSLAS (Zhong et al., 2021) reports that Mixup primarily facilitates representation learning while exerting a minimal or adverse effect on classifier learning. For this reason, *would it not be more effective in alleviating minority collapse to interpret mixed labels as singletons, as this reduces the adverse effect of Mixup?*



Sampler	Test Acc. (%) ↑	U_G ↑	U ↑
Random	72.91	14.6404	4.2325
CBS	75.86	15.2521	4.4848
CAS	76.60	15.3319	4.4999
BMLS	78.71	15.3379	4.5651

Figure 4: Experiments on CIFAR10-LT dataset for the effectiveness of BMLS to minority collapse. (figure) Visualization of 2D-projection of class vectors about Many class $\{0\}$ and Few classes $\{8, 9\}$. Dashed line indicates each class vector and contrast of background means the confidence value, i.e., a confidence close to 0.5 indicates that the model is confused between the two classes for the given sample, and this is represented by darker colors in the figure. (table) Quantitative comparison results. (U_G: Uniformity of all classes, U: Uniformity of $\{0, 8, 9\}$ classes)

(1) Problem Settings. By replacing the classifier as Mixed-Singleton Classifier defined in §4, we obtain the LPM$_\lambda$ with Mixed-Singleton Classifier (LPM$_\lambda$-MS):

$$\begin{aligned} &\min_{\boldsymbol{W}^\lambda, \boldsymbol{H}^\lambda} \mathbb{E}_\lambda \frac{1}{N} \sum_{k\in\mathbb{K}^2} \sum_{i=1}^{n_k} \mathcal{L}(\boldsymbol{W}^\lambda \boldsymbol{h}_{k,i}^\lambda, \boldsymbol{y}_k^\lambda) \\ &\text{s.t.} \quad \frac{1}{|\mathbb{K}^2|} \sum_{k\in\mathbb{K}^2} \left\|\boldsymbol{w}_k^\lambda\right\|^2 \le E_{W^\lambda}, \ \frac{1}{K^2} \sum_{k\in\mathbb{K}^2} \frac{1}{n_k} \sum_{i=1}^{n_k} \left\|\boldsymbol{h}_{k,i}^\lambda\right\|^2 \le E_H, \end{aligned} \tag{9}$$

where the only differences are $\boldsymbol{W}^\lambda = [\lambda \boldsymbol{w}_a + (1-\lambda)\boldsymbol{w}_b]_{(a,b)\in\mathbb{K}^2}$ and $E_{W^\lambda} \propto 1/|\mathbb{K}^2|$.

(2-4). In this setting, LPM$_\lambda$-MS is exactly same to the LPM in imbalanced case when the number of classes is K^2. For this reason, we omit steps (2-4) and conclude Theorem 2.

For simplicity, we remove the superscript λ in Theorem 2.

Theorem 2. *Assume $p \ge 2K^2$ and the loss function $\mathcal{L}$ is convex in the first argument. Let $\boldsymbol{X}^\star$ be any minimizer of the convex program with $n_{(1,1)} = n_{(1,2)} = \cdots = n_{(K_A,K_A)} = n_A$ and $n_{(K_A+1,K_A+1)} = n_{(K_A+1,K_A+2)} = \cdots = n_{(K,K)} = w_r n_B$. Define $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ as*

$$\left[\boldsymbol{h}_{(1,1)}^\star, \boldsymbol{h}_{(1,2)}^\star, \ldots, \boldsymbol{h}_{(K,K)}^\star, \ (\boldsymbol{W}^\star)^\top\right] = \boldsymbol{P}(\boldsymbol{X}^\star)^{1/2}, \tag{10}$$

$$\boldsymbol{h}_{k,i}^\star = \boldsymbol{h}_k^\star, \ \textit{for all } i \in \mathcal{I}_k^\lambda, k \in \mathbb{K}_A^2, \ \ \boldsymbol{h}_{k,i}^\star = \boldsymbol{h}_k^\star, \ \textit{for all } i \in \mathcal{I}_k^\lambda, k \in \mathbb{K}_B^2,$$

where $\boldsymbol{P} \in \mathbb{R}^{p\times 2K^2}$ is any partial orthogonal matrix such that $\boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}_{2K^2}$. Then $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ is a global minimizer of the mixed-label balanced LPM$_\lambda$-MS.

Proof. Theorem 2 follows directly from the same arguments applied to oversampling-adjusted LPM in imbalanced case, which has already been proven in Fang et al. (2021). We omit the proof here. □

6 Experimental Results

To empirically validate the effectiveness of our analysis and proposed solutions, we conducted experiments in various imbalanced environments. We used CIFAR10/100-LT, Places-LT, ImageNet-LT and iNaturalist2018, with five repeated experiments with random seeds in CIFAR10/100-LT and three in others. The tables presenting the experimental results show the average of test accuracies. In all tables, *imb* refers to the imbalance factor, *C10/100* represents the CIFAR10/100-LT datasets, *Clf.* refers to the classifier, and BMLS$_{\text{MS}}$ denotes the method using both BMLS and MS. Unless otherwise specified, all experiments include Mixup. Best in bold. Implementation details are illustrated in Appendix D.

6.1 Empirical Validation

BMLS. As shown in Figure 3, epoch-wise imbalance not only of singleton labels but also of

Table 1: Experiments on CIFAR10/100-LT datasets with imbalance factor 200 and 100 for effectiveness of Multi-Singleton Classifier (higher imbalance factor is more imbalanced)

Sampler	Dataset	Clf.	imb200				imb100			
			many	med	few	all	many	med	few	all
BMLS	C10	FC	**90.49**	**74.12**	54.43	73.13	88.53	**77.84**	70.53	78.85
		MS	88.94	72.97	**62.77**	**74.70**	**89.14**	76.34	**74.63**	**79.67**
		diff.	-1.55	-1.15	+8.34	+1.57	+0.61	-1.50	+4.10	+0.82
BMLS	C100	FC	**65.77**	41.73	7.19	40.36	**68.98**	46.13	14.98	45.32
		MS	63.24	**44.86**	**11.19**	**41.71**	66.31	**49.80**	**21.80**	**47.62**
		diff.	-2.53	+3.13	+4.00	+1.35	-2.67	+3.67	+6.82	+2.30

Table 2: Experiments on CIFAR10/100-LT datasets with various imbalance factors (higher imbalance factor is more imbalanced) (More references in Table 7)

Method	CIFAR10-LT				CIFAR100-LT			
	imbalance factor				imbalance factor			
	200	100	50	10	200	100	50	10
Mixup (Zhang et al., 2018)	67.30	72.80	78.60	87.70	38.70	43.00	48.10	58.20
CAS+ERM (Shen & Lin, 2016)	N/A	68.40	N/A	86.90	N/A	31.90	N/A	55.00
LOM (Zhang et al., 2022)	N/A	74.20	N/A	**89.40**	N/A	41.50	N/A	59.90
random	66.77	72.94	78.64	88.05	39.06	42.88	48.31	63.03
CBS	70.17	76.63	81.15	89.24	39.61	44.24	49.99	63.90
CAS	69.90	76.43	81.42	89.24	40.28	44.65	50.07	63.57
BMLS_{MS}	**74.70**	**79.67**	**83.46**	88.51	**41.71**	**47.62**	**52.74**	**64.47**

mixed ones affects model performance. While class-balanced sampling methods such as CBS and CAS oversamples singleton label samples within each mini-batch, Mixup ruins the balance of both singleton labels and mixed ones by randomly permuting input samples and blending them each other. Empirically, we observe that enforcing balance among mixed labels through BMLS improves model performance, promoting more balanced classifier, as demonstrated on Figure 4.

Mixed-Singleton Classifier. To validate the Mixed Singleton Classifier and support the conjecture in §5.3, we compared a singleton classifier (FC) and a Mixed-Singleton classifier (MS). As shown in Table 1, MS further boosts performance, particularly for few classes. This improvement indicates that MS facilitates less minority collapse in few classes, and the effect still maintains even though the degree of imbalance increases.

6.2 Standard Imbalanced Learning Benchmarks

Results and Analysis on Small Datasets. To evaluate the performance of our method, we selected Mixup, CAS, and LOM—the latter being the most similar to our approach—as baselines. As shown in Table 2, our proposed method achieves the highest performance on CIFAR10-LT and CIFAR100-LT across all settings, except for the case with an imbalance factor of 10, where class imbalance is relatively mild. Furthermore, when classes are categorized into *many*, *medium*, and *few* based on their sample frequency, and test accuracy is measured accordingly (cf. Table 8 Appendix G), BMLS demonstrates the largest improvement for *few* classes compared to other baselines. These results indicate that BMLS mitigates minority collapse more effectively than other class-balanced samplers.

Table 3: Experiments on large datasets. (PL: Places-LT, *:use pre-trained model, IN: ImageNet-LT, iNat18: iNaturalist2018)

Method	PL	PL*	IN	iNat18
random	22.06	25.90	45.19	64.62
CBS	24.79	37.32	47.49	67.06
CAS	24.26	37.44	47.31	**67.55**
BMLS	27.33	37.39	**48.83**	66.98
BMLS_{MS}	**27.95**	**37.81**	47.54	56.60

Results and Analysis on Large Datasets. In practical experimental settings, both BMLS and MS exhibit limitations depending on the number of classes K. First, BMLS struggles when K^2 is bigger than the dataset size, as it fails to generate mixed samples uniformly across all classes in each epoch. This leads to the same issue seen in traditional class-balanced samplers, we already introduced, epoch-wise label imbalance. MS, in addition to the issues faced by BMLS, suffers from an exponential increase in the number of class vectors for mixed labels as K grows. Concurrently, the number of samples available for learning each class vector decreases significantly, raising the potential for underfitting. As shown in the results in Table 3, the effect of BMLS+MS diminishes as the number of classes increases (e.g., $K_{PL} = 365 < K_{IN} = 1000 < K_{iNat18} = 8142$). However,

despite these limitations, BMLS+MS demonstrates superior performance compared to other class-balanced samplers on Place-LT, and when only BMLS is used on ImageNet-LT, it achieves the highest performance, while improving the accuracy on few classes (cf. Tables 9 and 10 in Appendix G). Even in the most challenging case, iNaturalist2018, using only BMLS still results in competitive performance compared to other class-balanced samplers.

6.3 Ablation Study and Synergy with Fixed ETF Classifier

Ablation Study (Table 4 in Appendix E.) To empirically validate whether our proposed methods effectively address the minority collapse issue and improve model performance in imbalanced learning environments, we conducted an ablation study. As shown in Table 4, applying both BMLS and MS together resulted in the largest performance improvement. Moreover, in scenarios where the number of samples in *few* classes is extremely small (e.g., imbalance factors of 200 and 100 in CIFAR100-LT), where both MS and FC face the same issue about underfitting, MS alone actually outperforms.

Synergy with Fixed ETF Classifier (Table 5 in Appendix F.) In Yang et al. (2022), it was proven that by fixing the classifier as a K-simplex ETF, NC is satisfied regardless of class balance, and that using this fixed ETF classifier along with a specialized loss (Dot-Ridge; DR) improves model performance in imbalanced learning environments. Leveraging the advantages of the fixed ETF classifier, we hypothesized that our method could produce synergies with this approach, and we conducted experiments applying our method to this framework. As shown in Table 5, our proposed methods significantly enhance the performance of the original ETF approach through seamless integration. Notably, even when the original method was used without any modifications for our approach on CIFAR10-LT, we observed an improvement in model performance.

7 Conclusion

The research problem targeted in this study is the issue of minority collapse in imbalanced learning environments, where class imbalance negatively impacts model performance, particularly for minority classes. We analyzed the impact of Mixup on this problem and identified two key findings: first, minority collapse is influenced by the frequency balance of mixed labels, and second, when mixed labels are balanced, interpreting them as singletons enhances reducing the minority collapse. Based on these findings, we proposed BMLS and MS as solutions. BMLS balanced mixed-label frequencies more effectively, while MS leveraged class vector interpretation to further enhance classifier performance. These methods demonstrated significant effectiveness in mitigating minority collapse and improving model performance, particularly for minority class samples. Through experiments, we validated the utility and versatility of the proposed methods, showing that both BMLS and MS consistently improved performance compared to existing baselines and demonstrated their applicability across different datasets and imbalance factors.

8 Limitations and Future Work

Scalability. As observed in the experimental results and analysis for large datasets, both BMLS and MS suffer from issues related to epoch-wise label imbalance and underfitting class vectors due to the exponential increase in the number of mixed labels, which is proportional to the number of singleton labels K. Additionally, in this study, to ensure a fair comparison, we matched the number of samples learned per epoch to those generated by a random sampler (e.g., in iNaturalist2018, we used 437,513 images, while the number of mixed labels was $K^2 = 66,292,164$ with $K = 8,142$). As explained in §4, this paper partially addresses the issue by reducing the diversity of mixed labels. However, if the number of training samples is sufficiently increased without considering the constraint, it could also serve as a technical solution.

Integration with other methods. In this study, we only extend our methods to the fixed ETF classifier inspired by neural collapse. However, both BMLS and MS are methods that can be used in conjunction with other Mixup-based methods for imbalanced learning. Through the experiments with the fixed ETF classifier extension, we demonstrated the potential for integration with other methods. We anticipate that future research will explore these integrations to more effectively mitigate minority collapse.

Reproducibility Statement

We summarize the reproducibility statement of this paper as follow.

- §4. To reproduce BMLS and MS, we define notations and provide helpful preliminaries (§3) with a theoretical support in Appendix C.3.
- §5. To prove our theorems such as Theorem 1, Proposition 1, and Theorem 2, we demonstrate the detailed proofs of them in Appendix C.
- §6. All experiments can be reproduced using our text supplementary materials (Appendix D), which provide dataset descriptions, model architectures, and hyperparameter settings, as well as our code including configuration files for each experiment. Additionally, experimental requirements, such as necessary libraries, are specified in the README files included with the code.

In addition, our codes can be accessed at *link* (T.B.A)

References

Jae Soon Baik, In Young Yoon, and Jun Won Choi. Dbn-mix: Training dual branch network using bilateral mixup augmentation for long-tailed visual recognition. *Pattern Recognition*, 147:110107, March 2024. ISSN 0031-3203. doi: 10.1016/j.patcog.2023.110107. URL http://dx.doi.org/10.1016/j.patcog.2023.110107.

Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 872–881. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/byrd19a.html.

Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *ICCV 2021*, pp. 112–121, 10 2021. doi: 10.1109/ICCV48922.2021.00018.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/621461af90cadfdaf0e8d4cc25129f91-Paper.pdf.

Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: Rebalanced mixup. In Adrien Bartoli and Andrea Fusiello (eds.), *Computer Vision – ECCV 2020 Workshops*, pp. 95–110, Cham, 2020. Springer International Publishing. ISBN 978-3-030-65414-6.

Yin Cui, Yang Song, Chen Sun, Andrew G. Howard, and Serge J. Belongie. Large scale fine-grained categorization and domain-specific transfer learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4109–4118, 2018. URL https://api.semanticscholar.org/CorpusID:43993788.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9260–9269, 2019. URL https://api.semanticscholar.org/CorpusID:58014111.

Yadolah Dodge. *Coefficient of Variation*, pp. 95–96. Springer New York, New York, NY, 2008. ISBN 978-0-387-32833-1. doi: 10.1007/978-0-387-32833-1_65. URL https://doi.org/10.1007/978-0-387-32833-1_65.

Charles Elkan. The foundations of cost-sensitive learning. *Proceedings of the Seventeenth International Conference on Artificial Intelligence: 4-10 August 2001; Seattle*, 1, 05 2001.

Cong Fang, Hangfeng He, Qi Long, and Weijie J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021. doi: 10.1073/pnas.2103091118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2103091118.

Quinn LeBlanc Fisher, Haoming Meng, and Vardan Papyan. Pushing boundaries: Mixup's influence on neural collapse. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jTSKkcbEsj.

Jintong Gao, He Zhao, Zhuo Li, and Dan dan Guo. Enhancing minority classes by mixing: An adaptative optimal transport approach for long-tailed classification. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=M7FQpIdo0X.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Anubha Kabra, Ayush Chopra, Nikaash Puri, Pinkesh Badjatiya, Sukriti Verma, Piyush Kumar Gupta, and K Balaji. Mixboost: Synthetic oversampling with boosted mixup for handling extreme imbalance. *ArXiv*, abs/2009.01571, 2020. URL https://api.semanticscholar.org/CorpusID:221470234.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1gRTCVFvB.

Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 13893–13902, 2020. ISSN 1063-6919. doi: 10.1109/CVPR42600.2020.01391. Publisher Copyright: © 2020 IEEE; 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020 ; Conference date: 14-06-2020 Through 19-06-2020.

Kisoo Kwon, Kuhwan Jeong, Sanghyun Park, Sangha Park, Hoshik Lee, Seung-Yeon Kwak, Sungmin Kim, and Kyunghyun Cho. Extramix: Extrapolatable data augmentation for regression using generative models, 2023. URL https://openreview.net/forum?id=NgEuFT-SIgI.

Pengyu Li, Xiao Li, Yutong Wang, and Qing Qu. Neural collapse in multi-label learning with pick-all-label loss. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28060–28094. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/li24ai.html.

Xuantong Liu, Jianfeng Zhang, Tianyang Hu, He Cao, Yuan Yao, and Lujia Pan. Inducing neural collapse in deep long-tailed learning. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11534–11544. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/liu23i.html.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2532–2541, 2019. URL https://api.semanticscholar.org/CorpusID:115137311.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

Haolin Pan, Yong Guo, Mianjie Yu, and Jian Chen. Enhanced long-tailed recognition with contrastive cutmix augmentation, 2024. URL https://arxiv.org/abs/2407.04911.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2015509117.

Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6877–6886, 2021. URL https://api.semanticscholar.org/CorpusID:244773284.

Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4175–4186. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/2ba61cc3a8f44143e1f2f13b2b729ab3-Paper.pdf.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Li Shen and Zhouchen Lin. Relay backpropagation for effective learning of deep convolutional neural networks. In *ECCV*, volume 9911, pp. 467–482, 10 2016. ISBN 978-3-319-46477-0. doi: 10.1007/978-3-319-46478-7_29.

Jiang-Xin Shi, Tong Wei, Yuke Xiang, and Yu-Feng Li. How re-sampling helps for long-tail learning? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=PLzCXefcpE.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6438–6447. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/verma19a.html.

Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V*, pp. 247–263, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58557-0. doi: 10.1007/978-3-030-58558-7_15. URL https://doi.org/10.1007/978-3-030-58558-7_15.

Zhengzhuo Xu, Zenghao Chai, and Chun Yuan. Towards calibrated model for long-tailed visual recognition from prior perspective. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=vqzAfN-BoA_.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=A6EmxI3_Xc.

Youngseok Yoon, Sangwoo Hong, Hyungjun Joo, Yao Qin, Haewon Jeong, and Jungwoo Lee. Mix from failure: Confusion-pairing mixup for long-tailed recognition, 2025. URL https://arxiv.org/abs/2411.07621.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Shaoyu Zhang, Chen Chen, Xiujuan Zhang, and Silong Peng. Label-occurrence-balanced mixup for long-tailed recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3224–3228, 2022. doi: 10.1109/ICASSP43922.2022.9746299.

Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:252684042.

Caidan Zhao and Yang Lei. Intra-class cutmix for unbalanced data augmentation. In *Proceedings of the 2021 13th International Conference on Machine Learning and Computing*, ICMLC '21, pp. 246–251, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450389310. doi: 10.1145/3457682.3457719. URL https://doi.org/10.1145/3457682.3457719.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16484–16493, 2021. doi: 10.1109/CVPR46437.2021.01622.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

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*, pp. 1–8, 2020.

APPENDIX

DETAILS ABOUT LARGE LANGUAGE MODELS IN PAPER WRITING

In this paper, the authors used LLMs solely for the purpose of checking mistranslations or grammar.

A ADDITIONAL RELATED WORK

A.1 MIXUP-BASED METHOD

(**Data augmentation.**) Mixup (Zhang et al., 2018) generates mixed-label samples by interpolating between input samples, extending training distribution support. Manifold Mixup (Verma et al., 2019) applies this technique to intermediate layers, regularizing the network by encouraging less confident predictions. CP-Mix, or Confusion-Pairing Mixup (Yoon et al., 2025), augments samples based on confusion pairs, addressing data deficiency by enhancing the model's ability to distinguish frequently misclassified class pairs. ExtraMix (Kwon et al., 2023) introduces a mixup technique capable of extrapolation, broadening both feature and label distributions, which minimizes label imbalance more effectively than traditional methods. CutMix (Yun et al., 2019; Zhao & Lei, 2021; Pan et al., 2024) focuses on mixed-label sample generation by cutting and pasting image patches, creating a regional dropout effect. CMO (Park et al., 2021) extends this idea by pasting minority class images onto majority class backgrounds, enriching minority class samples with context from majority class images. OTMix (Gao et al., 2023) improves upon this by using Optimal Transport to adaptively combine majority class backgrounds with minority class foregrounds, ensuring semantically reasonable mixed images. (**Architecture.**) BBN (Zhou et al., 2020), SBN, and DBN (Baik et al., 2024) utilize different architectures to enhance both representation and classifier learning. These methods incorporate bilateral mixup or decoupling strategies to optimize performance for imbalanced datasets. OTLR (Liu et al., 2019) uses dynamic meta-embedding and modulated attention to map images into a feature space that respects both closed-world classification and the novelty of the open world, improving the generalization of imbalanced datasets. (**Calibration or two-stage.**) UniMix (Xu et al., 2021) balances class distributions by introducing a novel mixing factor and sampler that favors the minority class. MiSLAS (Zhong et al., 2021) decouples representation and classifier learning, improving both calibration and performance in imbalanced data scenarios.

While many attempts have been made to address the challenges of imbalanced learning environments using Mixup, including data augmentation, architecture improvements, and calibration methods, no research has specifically focused on the balance of mixed labels in such contexts.

A.2 Class-balanced Methods

(**Re-balance.**) Remix (Chou et al., 2020) applies a higher mixup ratio to minority classes, rebalancing the data without sampling. Re-weighting (Elkan, 2001; Byrd & Lipton, 2019; Cui et al., 2019) adjusts the loss function by tuning class weights, with methods like Balanced SoftMax (Ren et al., 2020) explicitly considering label distribution shifts during optimization. Logit Adj (Menon et al., 2021) adjusts logits based on label frequencies, promoting a larger margin between rare positive and dominant negative labels. τ-Norm (Kang et al., 2020) normalizes classifier weight norms according to class size, rebalancing decision boundaries. LDAM loss (Cao et al., 2019) improves generalization by replacing standard cross-entropy with a margin-based approach, tailored to handle imbalanced datasets. cRT (Kang et al., 2020) re-trains the classifier using class-balanced sampling, improving the model's generalization ability. LWS (Kang et al., 2020) focuses on re-scaling classifier weights to ensure a balanced learning process for imbalanced datasets. (**Re-/Over-Sampling.**) M2M (Kim et al., 2020) augments minority classes by translating samples from majority classes, enhancing generalization for minority class features. MixBoost (Kabra et al., 2020) iteratively selects and combines majority and minority class instances to create hybrid samples, improving model performance. The Meta Sampler (Ren et al., 2020), built on balanced SoftMax, adapts the sampling rate through meta-learning to alleviate over-balancing issues. CB Sampling (Kang et al., 2020) ensures that each class has an equal probability of being selected, balancing the dataset during training. Class-Aware Sampler (CAS) (Shen & Lin, 2016) is more specific method of CB Sampling, which explicitly ensures the class frequency balance on each mini-batch. Label-Occurrence Mixup (LOM) (Zhang et al., 2022) uses two CB samplers to sample input pairs, respectively. CSA (Shi et al., 2023) generates diverse training images for tail classes by maintaining a context bank from head-class images.

Various class-balanced samplers have been proposed, yet no research has specifically focused on the balance of mixed labels. Additionally, while methods such as Logit Adjustment and UniMix have concentrated on singleton-labels, they did not interpret mixed labels as singletons.

A.3 Neural Collapse in Mixup and Imbalanced Learning

NC in imbalanced learning has been studied in Fang et al. (2021). To alleviate the minority collapse, Yang et al. (2022) assumed that the classifier is fixed to the K-simplex ETF and proved that LPM with the classifier satisfies NC properties. Also, the fixed ETF classifier with Mixup has improved the model performance in imbalanced learning. Building on the theorems, Fisher et al. (2024) proved Mixup also satisfies NC properties for both same class and different class. However, Yang et al. (2022) and Fisher et al. (2024) did not consider the minority collapse from the mixed label balance in the LPM with learnable classifiers.

B Convex Optimization Program

To begin with, defining $\boldsymbol{h}_k^\lambda = \frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{h}_{k,i}^\lambda$ as the feature mean of the $\mathbb{S}_k^\lambda$ where $k \in \mathbb{K}^2$, we introduce a new decision variable $\boldsymbol{X} = [\boldsymbol{h}_{(1,1)}^\lambda, \boldsymbol{h}_{(1,2)}^\lambda, \ldots, \boldsymbol{h}_{(K,K)}^\lambda, \boldsymbol{W}^\top]^\top[\boldsymbol{h}_{(1,1)}^\lambda, \boldsymbol{h}_{(1,2)}^\lambda, \ldots, \boldsymbol{h}_{(K,K)}^\lambda, \boldsymbol{W}^\top] \in \mathbb{R}^{(K^2+K)\times(K^2+K)}$. By definition, $\boldsymbol{X}$ is positive semi-definite and satisfies

$$\frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}(k,k) = \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\|\boldsymbol{h}_k^\lambda\|^2 \overset{a}{\leq} \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}_{k,i}^\lambda\right\|^2 \leq E_H$$

and

$$\frac{1}{K}\sum_{k=K^2+1}^{K^2+K}\boldsymbol{X}(k,k) = \frac{1}{K}\sum_{k=1}^{K}\|\boldsymbol{w}_k\|^2 \leq E_W,$$

where $\overset{a}{\leq}$ follows from the Cauchy-Schwarz inequality. Thus, we consider the following semi-definite programming problem:

$$\begin{aligned}\min_{\boldsymbol{X}\in\mathbb{R}^{(K^2+K)\times(K^2+K)}} & \sum_{k\in\mathbb{K}^2}\frac{n_k}{N}\mathcal{L}(\boldsymbol{z}(k)^\lambda,\boldsymbol{y}_k^\lambda)\\ \text{s.t. } & \boldsymbol{X}\succeq 0,\\ & \frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}(k,k)\leq E_H,\ \frac{1}{K}\sum_{k=K^2+1}^{K^2+K}\boldsymbol{X}(k,k)\leq E_W,\\ & \text{for all } k\in\mathbb{K}^2,\\ & \boldsymbol{z}_k=\left[\boldsymbol{X}(k,K^2+1),\boldsymbol{X}(k,K^2+2),\ldots,\boldsymbol{X}(k,K^2+K)\right]^\top.\end{aligned} \tag{11}$$

When $\mathcal{L}$ is the cross-entropy loss with softmax function,

$$\mathcal{L}(\boldsymbol{z}^\lambda(k),\boldsymbol{y}_k^\lambda)=-\lambda\log\left(\frac{\exp(\boldsymbol{z}^\lambda(a))}{\sum_{k'=1}^K\exp(\boldsymbol{z}^\lambda(k'))}\right)-(1-\lambda)\log\left(\frac{\exp(\boldsymbol{z}^\lambda(b))}{\sum_{k'=1}^K\exp(\boldsymbol{z}^\lambda(k'))}\right),$$

where $\boldsymbol{z}^\lambda(k')$ denotes the k'-th entry of the logit $\boldsymbol{z}_i^\lambda=\boldsymbol{W}\boldsymbol{h}_{k,i}^\lambda$, and $k=(a,b)$.

C Proofs

C.1 Proofs of Lemma 1 and Proposition 1

Proof of Lemma 1. For any feasible solution $\left(\boldsymbol{W},\boldsymbol{H}^\lambda\right)$ for the original program Eq. 5, we define

$$\boldsymbol{h}_k^\lambda:=\frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{h}_{k,i},\ k\in\mathbb{K}^2,$$

and

$$\boldsymbol{X}:=\left[\boldsymbol{h}_{(1,1)}^\lambda,\boldsymbol{h}_{(1,2)}^\lambda,\ldots,\boldsymbol{h}_{(K,K)}^\lambda,\boldsymbol{W}^\top\right]^\top\left[\boldsymbol{h}_{(1,1)}^\lambda,\boldsymbol{h}_{(1,2)}^\lambda,\ldots,\boldsymbol{h}_{(K,K)}^\lambda,\boldsymbol{W}^\top\right].$$

Clearly, $\boldsymbol{X}\succeq 0$. For the other two constraints of Eq. 11, we have

$$\frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}(k,k)=\frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\|\boldsymbol{h}_k^\lambda\|^2\overset{a}{\leq}\frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}_{k,i}^\lambda\right\|^2\overset{b}{\leq}E_H$$

and

$$\frac{1}{K}\sum_{k=K^2+1}^{K^2+K}\boldsymbol{X}(k,k)=\frac{1}{K}\sum_{k=1}^{K}\|\boldsymbol{w}_k\|^2\overset{c}{\leq}E_W,$$

where $\overset{a}{\leq}$ applies Jensen's inequality and $\overset{b}{\leq}$ and $\overset{c}{\leq}$ use that $\left(\boldsymbol{W},\boldsymbol{H}^\lambda\right)$ is a feasible solution. So $\boldsymbol{X}$ is a feasible solution for the convex program Eq. 11. Letting L_0 be the global minimum of Eq. 11, for any feasible solution $\left(\boldsymbol{W},\boldsymbol{H}^\lambda\right)$, we obtain

$$\begin{aligned}\frac{1}{N}\sum_{k\in\mathbb{K}^2}\sum_{i=1}^{n_k}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{k,i}^\lambda,\boldsymbol{y}_k^\lambda)&=\sum_{k\in\mathbb{K}^2}\frac{n_k}{N}\left[\frac{1}{n_k}\sum_{k=1}^{n_k}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{k,i}^\lambda,\boldsymbol{y}_k^\lambda)\right]\\&\overset{a}{\geq}\sum_{k\in\mathbb{K}^2}\frac{n_k}{N}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}_k^\lambda,\boldsymbol{y}_k^\lambda)=\sum_{k\in\mathbb{K}^2}\frac{n_k}{N}\mathcal{L}(\boldsymbol{z}(k)^\lambda,\boldsymbol{y}_k^\lambda)\geq L_0,\end{aligned} \tag{12}$$

where in $\overset{a}{\geq}$, we use $\mathcal{L}$ is convex on the first argument, and so $\mathcal{L}(\boldsymbol{W}\boldsymbol{h}^\lambda,\boldsymbol{y}_k^\lambda)$ is convex on $\boldsymbol{h}$ given $\boldsymbol{W}$ and $k\in\mathbb{K}^2$.

For the simplicity of our expressions, we hereafter remove the superscript λ of $\boldsymbol{H}^\lambda$, $\boldsymbol{h}^\lambda$ and $\boldsymbol{z}^\lambda$.

On the other hand, considering the solution $(\boldsymbol{W}^\star,\boldsymbol{H}^\star)$ defined in Eq. 6 with $\boldsymbol{X}^\star$ being a minimizer of Eq. 11, we have $\left[\boldsymbol{h}_{(1,1)}^\star,\boldsymbol{h}_{(1,2)}^\star,\ldots,\boldsymbol{h}_{(K,K)}^\star,\boldsymbol{W}^\top\right]^\top\left[\boldsymbol{h}_{(1,1)}^\star,\boldsymbol{h}_{(1,2)}^\star,\ldots,\boldsymbol{h}_{(K,K)}^\star,\boldsymbol{W}^\top\right]=\boldsymbol{X}^\star$

($p \geq K^2 + K$ guarantees the existence of $\left[\boldsymbol{h}^\star_{(1,1)}, \boldsymbol{h}^\star_{(1,2)}, \ldots, \boldsymbol{h}^\star_{(K,K)}, (\boldsymbol{W}^\star)^\top\right]$). We can verify that $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ is a feasible solution for Eq. 5 and have

$$\frac{1}{N}\sum_{k\in\mathbb{K}^2}\sum_{i=1}^{n_k}\mathcal{L}(\boldsymbol{W}^\star\boldsymbol{h}^\star_{k,i}, \boldsymbol{y}^\lambda_k) = \sum_{k\in\mathbb{K}^2}\frac{n_k}{N}\mathcal{L}(\boldsymbol{z}(k)^\star, \boldsymbol{y}^\lambda_k) = L_0, \tag{13}$$

where $\boldsymbol{z}(k)^\star = \left[\boldsymbol{X}^\star(k, K^2+1), \boldsymbol{X}^\star(k, K^2+2), \ldots, \boldsymbol{X}^\star(k, K^2+K)\right]^\top$ for $k \in \mathbb{K}^2$.

Combining Eq. 12 and Eq. 13, we conclude that L_0 is the global minimum of Eq. 5 and $(\boldsymbol{W}^\star, \boldsymbol{H}^\star)$ is a minimizer.

Suppose there is a minimizer $(\boldsymbol{W}', \boldsymbol{H}')$ that cannot be written as Eq. 6. Let

$$\boldsymbol{h}'_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{h}'_{k,i},\ k \in \mathbb{K}^2,$$

and

$$\boldsymbol{X}' = \left[\boldsymbol{h}'_{(1,1)}, \boldsymbol{h}'_{(1,2)}, \ldots, \boldsymbol{h}'_{(K,K)}, (\boldsymbol{W}')^\top\right]^\top \left[\boldsymbol{h}'_{(1,1)}, \boldsymbol{h}'_{(1,2)}, \ldots, \boldsymbol{h}'_{(K,K)}, (\boldsymbol{W}')^\top\right].$$

Eq. 12 implies that $\boldsymbol{X}'$ is a minimizer of Eq. 11. As $(\boldsymbol{W}', \boldsymbol{H}')$ cannot be written as Eq. 6 with $\boldsymbol{X}^\star = \boldsymbol{X}'$, then there is a $k' \in \mathbb{K}^2$, $i, j \in [n'_k]$ with $i \neq j$ such that $\boldsymbol{h}_{k',i} \neq \boldsymbol{h}_{k',j}$. We have

$$\begin{aligned}
&\frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}'(k,k) = \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\|\boldsymbol{h}'_k\|^2 \\
=&\frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}'_{k,i}\right\|^2 - \frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{k=1}^{n_k}\|\boldsymbol{h}'_{k,i} - \boldsymbol{h}'_k\|^2 \\
\leq&\frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}'_{k,i}\right\|^2 - \frac{1}{K}\frac{1}{n_{k'}}(\|\boldsymbol{h}'_{k',i} - \boldsymbol{h}'_{k'}\|^2 + \|\boldsymbol{h}'_{k',j} - \boldsymbol{h}'_{k'}\|^2) \\
\leq&\frac{1}{K^2}\sum_{k\in\mathbb{K}^2}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|\boldsymbol{h}'_{k,i}\right\|^2 - \frac{1}{K}\frac{1}{2n_{k'}}\|\boldsymbol{h}'_{k',i} - \boldsymbol{h}'_{k',j}\|^2 \\
<&E_H.
\end{aligned}$$

By contraposition, if all $\boldsymbol{X}^\star$ satisfy that $\frac{1}{K^2}\sum_{k=1}^{K^2}\boldsymbol{X}^\star(k,k) = E_H$, then all the solutions of Eq. 5 are in the form of Eq. 6. We complete the proof. □

Proposition 1 can be obtained by the same argument. We omit the proof here.

C.2 Proof of Theorem 1

To prove Theorem 1, we first study a limit case where we only learn the classification for partial classes. We solve the optimization program:

$$\begin{aligned}
\min_{\boldsymbol{W},\boldsymbol{H}^\lambda}\ \mathbb{E}_{\lambda\sim D_\lambda}\quad & \frac{1}{|\mathbb{K}_A^2|\cdot n_A}\sum_{k\in\mathbb{K}_A^2}\sum_{i=1}^{n_A}\mathcal{L}(\boldsymbol{W}\boldsymbol{h}^\lambda_{k,i}, \boldsymbol{y}^\lambda_k) \\
\text{s.t.}\quad & \frac{1}{K}\sum_{k=1}^{K}\|\boldsymbol{w}_k\|^2 \leq E_W, \\
& \frac{1}{K^2}\sum_{a=1}^{K}\sum_{b=1}^{K}\frac{1}{n_{(a,b)}}\sum_{i=1}^{n_{(a,b)}}\left\|\boldsymbol{h}^\lambda_{(a,b),i}\right\|^2 \leq E_H,
\end{aligned} \tag{14}$$

where $\boldsymbol{y}^\lambda_{(a,b)} = \lambda\boldsymbol{y}_a + (1-\lambda)\boldsymbol{y}_b$, $\mathbb{K}_A^2 = \{(a,b)|1 \leq a \leq K_A \wedge 1 \leq b \leq K_A\}$, $\mathbb{K}_B^2 = \{(a,b)|K_A+1 \leq a \leq K \wedge K_A+1 \leq b \leq K\}$, and

$$n_{(a,b)} = \begin{cases} n_A & \text{if } (a,b) \in \mathbb{K}_A^2 \\ n_B & \text{if } (a,b) \in \mathbb{K}_B^2 \\ 0 & \text{otherwise} \end{cases}.$$

For the simplicity of our expressions, we remove the superscript λ of $\boldsymbol{H}^\lambda$ and $\boldsymbol{h}^\lambda$.

Lemma 2 characterizes useful properties for the minimizer of Eq. 14.

Lemma 2. *Let $(\boldsymbol{W}, \boldsymbol{H})$ be a minimizer of Eq. 14. We have $\boldsymbol{h}_{k,i}^\lambda = \mathbf{0}_p$ for all $k \in \mathbb{K}_B^2$ and $i \in [n_B]$. Let L_0 be the global minimum of Eq. 14. We have*

$$L_0 = \frac{1}{|\mathbb{K}_A^2| \cdot n_A} \sum_{k \in \mathbb{K}_A^2} \sum_{i=1}^{n_A} \mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{k,i}, \boldsymbol{y}_k^\lambda).$$

Then L_0 only depends on $|\mathbb{K}_A^2|$, n_A, E_H, and E_W. Moreover, for any feasible solution $(\boldsymbol{W}', (\boldsymbol{H})')$, if there exist $k, k' \in \mathbb{K}_B^2$ such that $\|\boldsymbol{w}_k - \boldsymbol{w}_{k'}\| = \epsilon > 0$, we have

$$\frac{1}{|\mathbb{K}_A^2| \cdot n_A} \sum_{k \in \mathbb{K}_A^2} \sum_{i=1}^{n_A} \mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{k,i}, \boldsymbol{y}_k^\lambda) \geq L_0 + \epsilon',$$

where $\epsilon' > 0$ depends on ϵ, $|\mathbb{K}_A^2|$, n_A, E_H, and E_W.

Now we are ready to prove Theorem 1. The proof is based on the contradiction.

Proof of Theorem 1. Consider sequences n_A^ℓ and n_B^ℓ with $R^\ell := n_A^\ell / n_B^\ell$ for $\ell = 1, 2, \ldots$. We have $R^\ell \to \infty$. For each optimization program indexed by $\ell \in \mathbb{N}_+$, we introduce $(\boldsymbol{W}^{\ell,\star}, \boldsymbol{H}^{\ell,\star})$ as a minimizer and separate the objective function into two parts. We consider

$$\mathcal{L}^\ell\left(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell\right) = \frac{|\mathbb{K}_A^2| \cdot n_A^\ell}{|\mathbb{K}_A^2| \cdot n_A^\ell + |\mathbb{K}_B^2| \cdot n_B^\ell} \mathcal{L}_A^\ell\left(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell\right) + \frac{|\mathbb{K}_B^2| \cdot n_B^\ell}{|\mathbb{K}_A^2| \cdot n_A^\ell + |\mathbb{K}_B^2| \cdot n_B^\ell} \mathcal{L}_B^\ell\left(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell\right),$$

with

$$\mathcal{L}_A^\ell\left(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell\right) := \frac{1}{|\mathbb{K}_A^2| \cdot n_A^\ell} \sum_{k \in \mathbb{K}_A^2} \sum_{i=1}^{n_A^\ell} \mathcal{L}\left(\boldsymbol{W}^\ell \boldsymbol{h}_{k,i}^\ell, \boldsymbol{y}_k^\lambda\right)$$

and

$$\mathcal{L}_B^\ell\left(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell\right) := \frac{1}{|\mathbb{K}_B^2| \cdot n_B^\ell} \sum_{k \in \mathbb{K}_B^2} \sum_{i=1}^{n_B^\ell} \mathcal{L}\left(\boldsymbol{W}^\ell \boldsymbol{h}_{k,i}^\ell, \boldsymbol{y}_k^\lambda\right).$$

We define $\left(\boldsymbol{W}^{\ell,A}, \boldsymbol{H}^{\ell,A}\right)$ as a minimizer of the optimization program:

$$\begin{aligned} \min_{\boldsymbol{W}^\ell, \boldsymbol{H}^\ell} \quad & \mathcal{L}_A^\ell\left(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell\right) \\ \text{s.t.} \quad & \frac{1}{K} \sum_{k=1}^K \left\|\boldsymbol{w}_k^\ell\right\|^2 \leq E_W, \\ & \frac{1}{|\mathbb{K}_A^2|} \sum_{k \in \mathbb{K}_A^2} \frac{1}{n_A^\ell} \sum_{i=1}^{n_A^\ell} \left\|\boldsymbol{h}_{k,i}^\ell\right\|^2 + \frac{1}{|\mathbb{K}_B^2|} \sum_{k \in \mathbb{K}_B^2} \frac{1}{n_B^\ell} \sum_{i=1}^{n_B^\ell} \left\|\boldsymbol{h}_{k,i}^\ell\right\|^2 \leq E_H, \end{aligned} \tag{15}$$

and $\left(\boldsymbol{W}^{\ell,B}, \boldsymbol{H}^{\ell,B}\right)$ as a minimizer of the optimization program:

$$\begin{aligned} \min_{\boldsymbol{W}^\ell, \boldsymbol{H}^\ell} \quad & \mathcal{L}_B^\ell\left(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell\right) \\ \text{s.t.} \quad & \frac{1}{K} \sum_{k=1}^K \left\|\boldsymbol{w}_k^\ell\right\|^2 \leq E_W, \\ & \frac{1}{|\mathbb{K}_A^2|} \sum_{k \in \mathbb{K}_A^2} \frac{1}{n_A^\ell} \sum_{i=1}^{n_A^\ell} \left\|\boldsymbol{H}_{k,i}^\ell\right\|^2 + \frac{1}{|\mathbb{K}_B^2|} \sum_{k \in \mathbb{K}_B^2} \frac{1}{n_B^\ell} \sum_{i=1}^{n_B^\ell} \left\|\boldsymbol{h}_{k,i}^\ell\right\|^2 \leq E_H. \end{aligned} \tag{16}$$

Note that Programs Eq. 15 and Eq. 16 and their minimizers have been studied in Lemma 2. We define:

$$L_A := \mathcal{L}_A^\ell\left(\boldsymbol{W}^{\ell,A}, \boldsymbol{H}^{\ell,A}\right) \quad \text{and} \quad L_B := \mathcal{L}_B^\ell\left(\boldsymbol{W}^{\ell,B}, \boldsymbol{H}^{\ell,B}\right).$$

Then Lemma 2 implies that L_A and L_B only depend on $|\mathbb{K}_A^2|$, K_B, E_H, and E_W, and are independent of ℓ. Moreover, since $\boldsymbol{h}_{k,i}^{\ell,A} = \mathbf{0}_p$ for all $k \in \mathbb{K}_B^2$ and $i \in [n_B]$, we have

$$\mathcal{L}_B^\ell \left(\boldsymbol{W}^{\ell,A}, \boldsymbol{H}^{\ell,A}\right) = \lambda \cdot \log(K) + (1-\lambda) \cdot \log(K) = \log(K). \tag{17}$$

Now we prove Theorem 1 by contradiction. Suppose there exists a pair (k, k') such that $\lim_{\ell\to\infty} \boldsymbol{w}_k^{\ell,\star} - \boldsymbol{w}_{k'}^{\ell,\star} \neq \mathbf{0}_p$. Then there exists $\epsilon > 0$ such that for a subsequence $\{(\boldsymbol{w}^{a_\ell,\star}, \boldsymbol{h}^{a_\ell,\star})\}_{\ell=1}^\infty$ and an index ℓ_0 when $\ell \geq \ell_0$, we have $\left\|\boldsymbol{W}_k^{a_\ell,\star} - \boldsymbol{W}_{k'}^{a_\ell,\star}\right\| \geq \epsilon$. Now we figure out a contradiction by estimating the objective function value on $(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star})$. In fact, because $(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star})$ is a minimizer of $\mathcal{L}^\ell(\boldsymbol{W}^\ell, \boldsymbol{H}^\ell)$, we have

$$\begin{aligned}
\mathcal{L}^{a_\ell}\left(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star}\right) \leq\ & \mathcal{L}^{a_\ell}\left(\boldsymbol{W}^{a_\ell,A}, \boldsymbol{H}^{a_\ell,A}\right) \\
\overset{\text{Eq. 17}}{=}\ & \frac{|\mathbb{K}_A^2| \cdot n_A^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} L_A + \frac{|\mathbb{K}_B^2| \cdot n_B^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} \log(K) \\
=\ & L_A + \frac{1}{K_R R^{a_\ell} + 1} \left(\log(K) - L_A\right) \overset{\ell\to\infty}{\to} L_A,
\end{aligned} \tag{18}$$

where we define $K_R := |\mathbb{K}_A^2|/|\mathbb{K}_B^2|$ and use $R^\ell = n_A^\ell / n_B^\ell$.

However, when $\ell > \ell_0$, because $\left\|\boldsymbol{w}_k^{a_\ell,\star} - \boldsymbol{w}_{k'}^{a_\ell,\star}\right\| \geq \epsilon > 0$, Lemma 2 implies that

$$\mathcal{L}_A^{a_\ell}\left(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star}\right) \geq L_A + \epsilon_2,$$

where $\epsilon_2 > 0$ only depends on ϵ, $|\mathbb{K}_A^2|$, K_B, E_H, and E_W, and is independent of ℓ. We obtain

$$\begin{aligned}
\mathcal{L}^{a_\ell}\left(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star}\right) =\ & \frac{|\mathbb{K}_A^2| \cdot n_A^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} \mathcal{L}_A^{a_\ell}\left(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star}\right) \\
& + \frac{|\mathbb{K}_B^2| \cdot n_B^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} \mathcal{L}_B^{a_\ell}\left(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star}\right) \\
\overset{a}{\geq}\ & \frac{|\mathbb{K}_A^2| \cdot n_A^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} \mathcal{L}_A^{a_\ell}\left(\boldsymbol{W}^{a_\ell,\star}, \boldsymbol{H}^{a_\ell,\star}\right) \\
& + \frac{|\mathbb{K}_B^2| \cdot n_B^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} \mathcal{L}_B^{a_\ell}\left(\boldsymbol{W}^{a_\ell,B}, \boldsymbol{H}^{a_\ell,B}\right) \\
=\ & \frac{|\mathbb{K}_A^2| \cdot n_A^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} (L_A + \epsilon_2) + \frac{|\mathbb{K}_B^2| \cdot n_B^{a_\ell}}{|\mathbb{K}_A^2| \cdot n_A^{a_\ell} + |\mathbb{K}_B^2| \cdot n_B^{a_\ell}} L_B \\
=\ & L_A + \epsilon_2 + \frac{1}{K_R R^{a_\ell} + 1} (L_B - L_A - \epsilon_2) \overset{\ell\to\infty}{\to} L_A + \epsilon_2,
\end{aligned} \tag{19}$$

where $\overset{a}{\geq}$ uses $\left(\boldsymbol{W}^{a_\ell,B}, \boldsymbol{H}^{a_\ell,B}\right)$ is the minimizer of Eq. 16. Thus we meet contradiction by comparing Eq. 18 with Eq. 19 and achieve Theorem 1. □

Proof of Lemma 2. The proof of Lemma 2 is the same as Lemma 5 in Fang et al. (2021) regardless of λ, as demonstrated in Eq. 17; hence, we omit the details and refer the reader to Fang et al. (2021) □

C.3 Proof of our method

To prove that only mixed labels (a, b) for the case where $a < b$ ensures Theorem 1 and Proposition 1, we demonstrate that the following statement is true.

Proposition 2. *Let $\mathbb{K}^<$ be the mixed label set where $a < b$ for all $(a, b) \in \mathbb{K}^2$ and $\boldsymbol{W}_{\mathbb{K}^<}^\lambda$ be the partial matrix of $\boldsymbol{W}^\lambda$ which has class vectors for mixed labels $(a, b) \in \mathbb{K}^<$.*

Then, $\boldsymbol{W}$ is a K-simplex ETF if $\boldsymbol{W}_{\mathbb{K}^<}^\lambda$ is a $|\mathbb{K}^<|$-simplex ETF.

For the simplicity, we remove the subscript $\mathbb{K}^<$ of $\boldsymbol{W}_{\mathbb{K}^<}^\lambda$ and $\boldsymbol{w}_{\mathbb{K}^<}^\lambda$ in the following proof.

Proof. Let $f(x;\alpha,\beta)$ be the probability density function of Beta distribution $D_\lambda(\alpha,\beta)$. For the mixup ratio λ sampled from $D_\lambda(\alpha,\alpha)$, we have

$$\begin{aligned}\boldsymbol{w}^{\lambda}_{(a,b)} =&\mathbb{E}_\lambda\left(\lambda\boldsymbol{w}_a+(1-\lambda)\boldsymbol{w}_b\right)\\ \overset{a}{=}&\frac{1}{2}\mathbb{E}_\lambda\left((\lambda\boldsymbol{w}_a+(1-\lambda)\boldsymbol{w}_b)+((1-\lambda)\boldsymbol{w}_a+\lambda\boldsymbol{w}_b)\right)\\ =&\frac{1}{2}\left(\boldsymbol{w}_a+\boldsymbol{w}_b\right),\end{aligned} \tag{20}$$

where in $\overset{a}{=}$, we use $f(\lambda;\alpha,\alpha)=f(1-\lambda;\alpha,\alpha)$.

From the definition of a simplex ETF, we get

$$\sum_{(a,b)\in\mathbb{K}^<}\boldsymbol{w}^{\lambda}_{(a,b)}=0 \tag{21}$$

Plugging the equality of Eq. 20 into Eq. 21, we have

$$\begin{aligned}\sum_{(a,b)\in\mathbb{K}^<}\boldsymbol{w}^{\lambda}_{(a,b)} =&\frac{K-1}{2}\sum_{i=1}^{K}\boldsymbol{w}_i=0\\ \therefore\ \boldsymbol{w}_i =&-\sum_{j\neq i}^{K}\boldsymbol{w}_j,\ \ \forall\, i\in[K]\end{aligned} \tag{22}$$

From the definition of $\mathbb{K}^<$, we can get $<i,j,k>$ for all $i\in[K]$, satisfying

$$\boldsymbol{w}_i=\boldsymbol{w}^{\lambda}_{\{i,j\}}-\boldsymbol{w}^{\lambda}_{\{j,k\}}+\boldsymbol{w}^{\lambda}_{\{i,k\}}, \tag{23}$$

where $\{a,b\}=(a,b)$ if $a<b$ otherwise (b,a) and $i\neq j\neq k$.

Now, we show that $\boldsymbol{w}_i^\top\boldsymbol{w}_{i'}=-\frac{1}{K-1}$ is true for all $i\neq i'$

$$\begin{aligned}\boldsymbol{w}_i^\top\boldsymbol{w}_{i'}\overset{Eq.\ 23}{=}&\left(\boldsymbol{w}^{\lambda}_{\{i,j\}}-\boldsymbol{w}^{\lambda}_{\{j,k\}}+\boldsymbol{w}^{\lambda}_{\{i,k\}}\right)^\top\left(\boldsymbol{w}^{\lambda}_{\{i',j'\}}-\boldsymbol{w}^{\lambda}_{\{j',k'\}}+\boldsymbol{w}^{\lambda}_{\{i',k'\}}\right)\\ \overset{a}{=}&-\frac{1}{K-1}\end{aligned} \tag{24}$$

where in $\overset{a}{=}$, we use the property of the simplex ETF, i.e., $\left(\boldsymbol{w}^{\lambda}_{(a,b)}\right)^\top\boldsymbol{w}^{\lambda}_{(a',b')}=-\frac{1}{K-1}$ for all $(a,b)\neq(a',b')$. We complete the proof. □

D EXPERIMENTAL SETUP

Implementation Details. Our experiments follow the setups of Zhong et al. (2021) and Zhou et al. (2020) for CIFAR10-LT, ImageNet-LT, Places-LT, and iNaturalist2018 and Yang et al. (2022) for CIFAR100-LT. We employ ResNet32 for CIFAR10-LT, doubling the feature dimensions for CIFAR100-LT. For ImageNet-LT and iNaturalist2018, we use ResNet50, and for Places-LT, use ResNet152, respectively. To reproduce baseline comparisons, we adopt the same hyperparameter settings as in Zhong et al. (2021) and Zhou et al. (2020).

Datasets. Following Zhong et al. (2021); Zhou et al. (2020), we use the long-tailed variants of CIFAR10, CIFAR100, ImageNet (Russakovsky et al., 2015), Places365 (Zhou et al., 2017), and iNaturalist2018 (Cui et al., 2018).

CIFAR10-LT. 10 imbalanced classes, subsampled at exponentially decreasing rates from CIFAR10 (Zhong et al., 2021).

CIFAR100-LT. 100 imbalanced classes, constructed analogously to CIFAR10-LT.

ImageNet-LT. Derived from ImageNet for large-scale object classification. Class frequencies follow a Pareto distribution ($\alpha = 5$) with cardinalities from 5 to 1,280, totaling 115.8K images across 1,000 classes.

Places-LT. An extended version of Places, with class sizes ranging from 5 to 4,980, yielding 184.5K images from 365 classes.

iNaturalist2018. A large-scale real-world species classification dataset with extreme label imbalance, comprising 437,513 images from 8,142 categories.

Architectures. For CIFAR10-LT, we use ResNet32 (Zhong et al., 2021) with three residual blocks, producing feature dimensions of 16, 32, and 64, respectively. CIFAR100-LT doubles these dimensions. Differing from the standard ResNet architecture used for ImageNet, the ResNet32's first convolutional layer has a kernel size, stride, and padding of 3, 1, and 1, respectively. ResNet50 and 152 follow He et al. (2015).

Hyperparameters. For CIFAR10/100-LT, models are trained with mini-batch size 128 using SGD with momentum 0.9 and weight decay 2e-4 for 200 epochs. The learning rate is linearly warmed up from 0.02 and decayed by 0.1 at epochs 160 and 180. For ImageNet-LT and Places-LT, models are trained with SGD (momentum 0.9, weight decay 5e-4) and a cosine annealing scheduler. Mixup alpha is set per dataset: $\alpha = 1.0$ for CIFAR10/100-LT, $\alpha = 0.2$ for others.

E Ablation Study

Table 4: Ablation study on CIFAR10/100-LT datasets with various imbalance factors. The results are the mean of five repeated experiments with random seeds. Best in bold (CBS: Class-Balanced Sampler, CAS: Class-Aware Sampler, BMLS: Balanced Mixed Label Sampler)

Sampler	Clf.	CIFAR10-LT				CIFAR100-LT			
		imbalance factor				imbalance factor			
		200	100	50	10	200	100	50	10
random	FC	66.77	72.94	78.64	88.05	39.06	42.88	48.31	63.03
random	MS	53.11	64.08	68.56	80.56	33.42	36.87	41.66	56.71
BMLS	FC	73.13	78.85	83.07	**89.46**	40.03	45.20	51.99	**65.72**
BMLS	MS	**74.70**	**79.67**	**83.46**	88.51	**41.71**	**47.62**	**52.74**	64.47

F Synergy with Fixed ETF Classifier

In Yang et al. (2022), a scale factor is necessary for the fixed ETF classifier, due to class vectors are normalized. For this reason, we make a modified version of the fixed ETF classifier to apply our methods, named as fixed Mixed-Singleton Weighted ETF classifier (MS-WETF). The scale of class vectors is important for softmax cross-entropy loss. Thus, we remove the scale factor and add learnable parameter $s \in \mathbb{R}^K$ to control the scale of each class vectors.

$$\boldsymbol{W}_{\text{WETF}} = s \cdot \boldsymbol{W}_{\text{ETF}}$$

Then, we make $\boldsymbol{W}_{\text{WETF}}$ as Mixed-Singleton classifier

$$\boldsymbol{W}^{\lambda}_{\text{MS-WETF, (a,b)}} = [\lambda \boldsymbol{w}_{\text{WETF},a} + (1-\lambda)\boldsymbol{w}_{\text{WETF},b}]_{(a,b)\in\mathbb{K}^2}$$

Table 5: Extension to the fixed ETF classifier on CIFAR10/100-LT datasets with various imbalance factors. The results are the mean of five repeated experiments with random seeds. Best in bold (CBS: Class-Balanced Sampler, CAS: Class-Aware Sampler, BMLS: Balanced Mixed Label Sampler)

Sampler	Clf.	$\mathcal{L}$	CIFAR10-LT				CIFAR100-LT			
			imbalance factor				imbalance factor			
			200	100	50	10	200	100	50	10
random	ETF	CE†	60.06	67.00	77.20	87.00	N/A	N/A	N/A	N/A
random	ETF	DR†	71.90	76.50	81.00	87.70	40.90	45.30	50.40	N/A
random	ETF	DR	71.58	76.82	81.25	87.59	41.20	45.07	50.71	63.08
CBS	ETF	DR	69.35	75.46	81.15	88.38	38.78	42.96	48.84	62.01
CAS	ETF	DR	69.17	76.16	80.81	**88.61**	38.91	43.18	49.05	62.50
BMLS	ETF	DR	**77.77**	**80.38**	**84.30**	87.91	39.54	43.60	49.54	62.06
		diff.	+6.19	+3.56	+3.05	+0.32	-1.66	-1.47	-1.17	-1.02
BMLS	MS-WETF	CE	77.73	80.31	84.22	88.26	**42.73**	**47.10**	**52.44**	**64.10**
		diff.	+6.15	+3.49	+2.97	+0.67	+1.53	+2.03	+1.73	+1.02

G Additional Experimental Results

According to Liu et al. (2019), we also calculate top-1 test accuracy of three disjoint set: many, medium, and few classes. The classes included in each set for the respective datasets are described in Table 6. In the tables of experimental results about many, medium, and few classes, we the mean and std of top-1 test accuracies as $mean_{std}$.

Table 6: The classes in Many/Medium/Few class sets.

	CIFAR10-LT	CIFAR100-LT	Places-LT	ImageNet-LT	iNaturalist2018
Many	[0,2]	[0,35]	[0, 130]	[0, 389]	[0, 841]
Medium	[3,6]	[36,70]	[131, 287]	[390, 835]	[842, 4542]
Few	[7,9]	[71,99]	[288, 364]	[835, 999]	[4543, 8141]

Table 7: Comparison experiments of samplers on the CIFAR10/100-LT dataset with various imbalance factors. The results are the mean of five repeated experiments with random seeds. Best in bold (CBS: Class-Balanced Sampler, CAS: Class-Aware Sampler, BMLS: Balanced Mixed Label Sampler) (Gray indicates that it is not a comparison target)

Method	CIFAR10-LT				CIFAR100-LT			
	imbalance factor				imbalance factor			
	200	100	50	10	200	100	50	10
mixup								
Mixup (Zhang et al., 2018)	67.30	72.80	78.60	87.70	38.70	43.00	48.10	58.20
Remix (Chou et al., 2020)	N/A	73.00	N/A	88.50	N/A	41.40	N/A	59.50
CMO (Park et al., 2021)	N/A	N/A	N/A	N/A	N/A	43.90	48.30	59.50
SBN-mix (Baik et al., 2024)	69.87	76.33	81.04	89.84	40.30	45.07	50.39	62.37
OTMix (Gao et al., 2023)	N/A	78.30	83.40	90.20	N/A	46.40	50.70	61.60
2-stage or extra network								
BBN-mix (Zhou et al., 2020)	N/A	79.82	82.18	88.32	N/A	42.56	47.02	59.12
DBN-mix (Baik et al., 2024)	79.58	83.47	86.82	90.87	46.21	51.04	54.93	64.98
UniMix (Xu et al., 2021)	78.48	82.75	84.32	89.66	42.07	45.45	51.11	61.25
MiSLAS (Zhong et al., 2021)	N/A	82.10	85.70	90.00	N/A	47.00	52.30	63.20
CP-Mix (Yoon et al., 2025)	78.34	82.44	85.08	89.87	43.56	48.20	52.12	61.91
class-balance loss								
CB+RS (Cao et al., 2019)	N/A	70.55	N/A	86.79	N/A	33.44	N/A	55.06
CB+RW (Cui et al., 2019)	N/A	72.37	N/A	86.54	N/A	33.99	N/A	57.12
CB+Focal (Cui et al., 2019)	N/A	74.57	N/A	87.10	N/A	36.02	N/A	57.99
LDAM (Cao et al., 2019)	N/A	73.35	N/A	86.96	N/A	39.60	N/A	56.91
LDAM+DRW (Cao et al., 2019)	N/A	77.03	N/A	88.16	N/A	42.04	N/A	58.71
class-balance sampling								
CAS (Shen & Lin, 2016)	N/A	68.40	N/A	86.90	N/A	31.90	N/A	55.00
LOM (Zhang et al., 2022)	N/A	74.20	N/A	**89.40**	N/A	41.50	N/A	59.90
CAS+DRW (Shen & Lin, 2016)	N/A	73.50	N/A	87.70	N/A	41.50	N/A	57.60
LOM+DRW (Zhang et al., 2022)	N/A	78.70	N/A	89.60	N/A	46.20	N/A	61.10
reproduced results and our method								
random+Mixup	66.77	72.94	78.64	88.05	39.06	42.88	48.31	63.03
CBS+Mixup	70.17	76.63	81.15	89.24	39.61	44.24	49.99	63.90
CAS+Mixup	69.90	76.43	81.42	89.24	40.28	44.65	50.07	63.57
BMLS+MS	**74.70**	**79.67**	**83.46**	88.51	**41.71**	**47.62**	**52.74**	**64.47**

Table 8: Experimental results on Many/Medium/Few classes in the CIFAR10/100-LT datasets.

	Method	Clf.	CIFAR10-LT				CIFAR100-LT			
			many	med	few	all	many	med	few	all
imb 200	random	FC	$91.17_{3.65}$	$69.99_{2.25}$	$38.09_{6.32}$	$66.77_{0.76}$	$71.16_{0.52}$	$35.22_{0.20}$	$3.85_{0.47}$	$39.06_{0.23}$
	CBS	FC	$82.63_{2.72}$	$69.29_{3.79}$	$58.89_{3.94}$	$70.17_{0.51}$	$65.92_{0.51}$	$39.44_{0.96}$	$7.15_{0.50}$	$39.61_{0.50}$
	CAS	FC	$85.65_{3.74}$	$67.67_{3.86}$	$57.14_{4.43}$	$69.90_{0.77}$	$66.32_{0.55}$	$40.54_{0.59}$	$7.62_{0.28}$	$40.28_{0.29}$
	BMLS	FC	$90.49_{0.26}$	$74.12_{1.21}$	$54.43_{2.25}$	$73.13_{0.67}$	$65.29_{0.45}$	$41.33_{0.76}$	$7.09_{0.37}$	$40.03_{0.38}$
	BMLS	MS	$88.94_{0.32}$	$72.97_{0.83}$	$62.77_{1.30}$	$74.70_{0.45}$	$63.24_{0.57}$	$44.86_{0.42}$	$11.19_{0.76}$	$41.71_{0.36}$
imb 100	random	FC	$93.39_{2.42}$	$74.05_{2.03}$	$50.99_{5.40}$	$72.94_{0.68}$	$72.09_{0.21}$	$41.10_{0.40}$	$8.77_{0.42}$	$42.88_{0.15}$
	CBS	FC	$90.89_{2.45}$	$74.31_{2.99}$	$65.46_{5.76}$	$76.63_{0.41}$	$67.07_{0.74}$	$46.29_{0.72}$	$13.43_{0.37}$	$44.24_{0.14}$
	CAS	FC	$90.54_{2.86}$	$75.54_{1.95}$	$63.51_{6.26}$	$76.43_{0.60}$	$68.28_{0.35}$	$46.47_{0.34}$	$13.12_{0.25}$	$44.65_{0.26}$
	BMLS	FC	$88.53_{1.01}$	$77.84_{0.25}$	$70.53_{1.27}$	$78.85_{0.34}$	$68.38_{0.27}$	$46.89_{0.33}$	$14.37_{0.88}$	$45.20_{0.33}$
	BMLS	MS	$89.14_{0.63}$	$76.34_{0.62}$	$74.63_{0.69}$	$79.67_{0.21}$	$66.31_{0.26}$	$49.80_{0.57}$	$21.80_{0.40}$	$47.62_{0.25}$
imb 50	random	FC	$95.25_{0.23}$	$78.52_{0.54}$	$62.19_{1.04}$	$78.64_{0.57}$	$73.72_{0.18}$	$48.62_{0.23}$	$16.40_{0.93}$	$48.31_{0.28}$
	CBS	FC	$91.57_{3.17}$	$79.62_{1.58}$	$72.78_{4.05}$	$81.15_{0.48}$	$68.80_{0.46}$	$52.97_{0.28}$	$23.06_{0.56}$	$49.99_{0.13}$
	CAS	FC	$92.78_{0.43}$	$79.28_{0.46}$	$72.89_{0.96}$	$81.42_{0.27}$	$69.16_{0.61}$	$52.71_{0.34}$	$23.18_{0.41}$	$50.07_{0.27}$
	BMLS	FC	$91.86_{0.40}$	$81.32_{0.36}$	$76.63_{1.05}$	$83.07_{0.43}$	$69.77_{0.55}$	$54.55_{0.28}$	$26.83_{0.67}$	$51.99_{0.26}$
	BMLS	MS	$89.45_{0.15}$	$79.29_{0.54}$	$83.03_{0.50}$	$83.46_{0.36}$	$67.06_{0.51}$	$55.30_{0.84}$	$31.88_{1.39}$	$52.74_{0.55}$
imb 10	random	FC	$94.79_{0.55}$	$85.38_{0.27}$	$84.86_{1.23}$	$88.05_{0.27}$	$76.06_{0.32}$	$64.10_{0.63}$	$45.56_{0.57}$	$63.03_{0.17}$
	CBS	FC	$93.95_{0.78}$	$86.04_{0.57}$	$88.81_{0.28}$	$89.24_{0.37}$	$72.42_{0.64}$	$65.76_{0.45}$	$51.08_{0.69}$	$63.90_{0.37}$
	CAS	FC	$94.14_{0.23}$	$86.34_{0.24}$	$88.21_{0.43}$	$89.24_{0.18}$	$72.59_{0.49}$	$65.20_{0.47}$	$50.40_{0.66}$	$63.57_{0.26}$
	BMLS	FC	$91.17_{0.40}$	$87.04_{0.26}$	$90.98_{0.59}$	$89.46_{0.19}$	$71.05_{0.70}$	$68.93_{0.66}$	$55.24_{0.47}$	$65.72_{0.29}$
	BMLS	MS	$91.63_{0.60}$	$84.92_{0.63}$	$90.18_{0.56}$	$88.51_{0.19}$	$71.61_{0.21}$	$65.67_{1.20}$	$54.17_{1.49}$	$64.47_{0.24}$

Table 9: Experimental results on Many/Medium/Few classes in the Places-LT datasets.

Method	Clf.	Places-LT				Places-LT (FT)			
		many	med	few	all	many	med	few	all
random	FC	$42.02_{0.76}$	$15.79_{0.54}$	$0.86_{0.12}$	$22.06_{0.50}$	$43.79_{0.29}$	$20.45_{0.27}$	$6.59_{0.26}$	$25.90_{0.06}$
CBS	FC	$38.65_{1.97}$	$22.60_{1.20}$	$5.69_{0.52}$	$24.79_{0.13}$	$41.31_{0.09}$	$39.98_{0.17}$	$25.11_{0.11}$	$37.32_{0.07}$
CAS	FC	$40.68_{0.33}$	$20.08_{0.53}$	$4.86_{0.50}$	$24.26_{0.22}$	$41.35_{0.08}$	$40.06_{0.06}$	$25.46_{0.17}$	$37.44_{0.04}$
BMLS	FC	$38.43_{0.21}$	$27.80_{0.12}$	$7.47_{0.26}$	$27.33_{0.17}$	$34.65_{0.04}$	$43.79_{0.05}$	$29.00_{0.08}$	$37.39_{0.01}$
BMLS	MS	$39.39_{0.32}$	$27.01_{0.40}$	$10.39_{0.12}$	$27.95_{0.26}$	$41.33_{0.09}$	$40.14_{0.00}$	$27.05_{0.15}$	$37.81_{0.01}$

Table 10: Experimental results on Many/Medium/Few classes in the ImageNet-LT and iNaturalist2018 datasets.

Method	Clf.	ImageNet-LT				iNaturalist2018			
		many	med	few	all	many	med	few	all
random	FC	$67.76_{0.43}$	$38.72_{0.50}$	$9.33_{0.28}$	$45.19_{0.43}$	$77.55_{0.39}$	$66.66_{0.38}$	$59.49_{0.38}$	$64.62_{0.31}$
CBS	FC	$62.46_{0.91}$	$44.55_{1.10}$	$20.00_{0.92}$	$47.49_{0.99}$	$63.25_{0.22}$	$68.36_{0.15}$	$66.63_{0.18}$	$67.06_{0.04}$
CAS	FC	$63.04_{0.31}$	$43.83_{0.34}$	$19.53_{0.40}$	$47.31_{0.33}$	$63.99_{0.63}$	$68.80_{0.02}$	$67.10_{0.08}$	$67.55_{0.09}$
BMLS	FC	$62.35_{0.69}$	$46.53_{0.43}$	$23.08_{0.54}$	$48.83_{0.55}$	$64.44_{2.52}$	$68.33_{0.37}$	$66.19_{0.87}$	$66.98_{0.19}$
BMLS	MS	$59.03_{0.89}$	$45.87_{1.01}$	$24.86_{0.92}$	$47.54_{0.94}$	$51.73_{0.83}$	$57.15_{0.14}$	$57.18_{0.28}$	$56.60_{0.18}$

Table 11: Experimental results of the ablation study on Many/Medium/Few classes in the CIFAR10/100-LT datasets. The results are the mean of five repeated experiments with random seeds.

	Method	Clf.	CIFAR10-LT				CIFAR100-LT			
			many	med	few	all	many	med	few	all
imb 200	random	FC	$91.17_{3.65}$	$69.99_{2.25}$	$38.09_{6.32}$	$66.77_{0.76}$	$71.16_{0.52}$	$35.22_{0.20}$	$3.85_{0.47}$	$39.06_{0.23}$
	random	MS	$88.59_{0.19}$	$53.77_{1.07}$	$16.74_{1.04}$	$53.11_{0.58}$	$64.59_{0.75}$	$28.43_{0.49}$	$0.75_{0.15}$	$33.42_{0.37}$
	BMLS	FC	$90.49_{0.26}$	$74.12_{1.21}$	$54.43_{2.25}$	$73.13_{0.67}$	$65.29_{0.45}$	$41.33_{0.76}$	$7.09_{0.37}$	$40.03_{0.38}$
	BMLS	MS	$88.94_{0.32}$	$72.97_{0.83}$	$62.77_{1.30}$	$74.70_{0.45}$	$63.24_{0.57}$	$44.86_{0.42}$	$11.19_{0.76}$	$41.71_{0.36}$
imb 100	random	FC	$93.39_{2.42}$	$74.05_{2.03}$	$50.99_{5.40}$	$72.94_{0.68}$	$72.09_{0.21}$	$41.10_{0.40}$	$8.77_{0.42}$	$42.88_{0.15}$
	random	MS	$89.47_{0.46}$	$62.24_{2.21}$	$41.15_{3.22}$	$64.08_{1.59}$	$67.29_{0.31}$	$33.84_{0.61}$	$2.78_{0.32}$	$36.87_{0.24}$
	BMLS	FC	$88.53_{1.01}$	$77.84_{0.25}$	$70.53_{1.27}$	$78.85_{0.34}$	$68.38_{0.27}$	$46.89_{0.33}$	$14.37_{0.88}$	$45.20_{0.33}$
	BMLS	MS	$89.14_{0.63}$	$76.34_{0.62}$	$74.63_{0.69}$	$79.67_{0.21}$	$66.31_{0.26}$	$49.80_{0.57}$	$21.80_{0.40}$	$47.62_{0.25}$
imb 50	random	FC	$95.25_{0.23}$	$78.52_{0.54}$	$62.19_{1.04}$	$78.64_{0.57}$	$73.72_{0.18}$	$48.62_{0.23}$	$16.40_{0.93}$	$48.31_{0.28}$
	random	MS	$90.04_{0.63}$	$64.80_{1.76}$	$52.11_{1.06}$	$68.56_{0.50}$	$68.28_{0.69}$	$42.17_{0.89}$	$8.00_{0.52}$	$41.66_{0.42}$
	BMLS	FC	$91.86_{0.40}$	$81.32_{0.36}$	$76.63_{1.05}$	$83.07_{0.43}$	$69.77_{0.55}$	$54.55_{0.28}$	$26.83_{0.67}$	$51.99_{0.26}$
	BMLS	MS	$89.45_{0.15}$	$79.29_{0.54}$	$83.03_{0.50}$	$83.46_{0.36}$	$67.06_{0.51}$	$55.30_{0.84}$	$31.88_{1.39}$	$52.74_{0.55}$
imb 10	random	FC	$94.79_{0.55}$	$85.38_{0.27}$	$84.86_{1.23}$	$88.05_{0.27}$	$76.06_{0.32}$	$64.10_{0.63}$	$45.56_{0.57}$	$63.03_{0.17}$
	random	MS	$91.54_{0.43}$	$76.31_{1.02}$	$75.25_{1.44}$	$80.56_{0.76}$	$71.91_{0.35}$	$57.38_{0.90}$	$37.03_{0.72}$	$56.71_{0.48}$
	BMLS	FC	$91.17_{0.40}$	$87.04_{0.26}$	$90.98_{0.59}$	$89.46_{0.19}$	$71.05_{0.70}$	$68.93_{0.66}$	$55.24_{0.47}$	$65.72_{0.29}$
	BMLS	MS	$91.63_{0.60}$	$84.92_{0.63}$	$90.18_{0.56}$	$88.51_{0.19}$	$71.61_{0.21}$	$65.67_{1.20}$	$54.17_{1.49}$	$64.47_{0.24}$

Table 12: Experimental results of extension to the fixed ETF Classifier on Many/Medium/Few classes in the CIFAR10-LT dataset. The results are the mean of five repeated experiments with random seeds.

	Method	Clf.	$\mathcal{L}$	CIFAR10-LT			
				many	med	few	all
imb 200	random	ETF	DR	$84.13_{0.64}$	$73.89_{0.92}$	$55.94_{1.24}$	$71.58_{0.39}$
	CBS	ETF	DR	$81.05_{3.12}$	$69.26_{2.29}$	$57.77_{4.75}$	$69.35_{0.38}$
	CAS	ETF	DR	$87.67_{6.09}$	$72.17_{0.94}$	$46.67_{6.61}$	$69.17_{0.67}$
	BMLS	ETF	DR	$84.52_{0.47}$	$74.15_{0.36}$	$75.85_{0.66}$	$77.77_{0.13}$
	BMLS	MS-WETF	CE	$85.41_{0.71}$	$74.96_{0.45}$	$73.74_{0.72}$	$77.73_{0.32}$
imb 100	random	ETF	DR	$83.75_{0.92}$	$75.42_{0.30}$	$71.75_{0.95}$	$76.82_{0.20}$
	CBS	ETF	DR	$88.89_{3.19}$	$74.46_{2.41}$	$63.37_{6.15}$	$75.46_{0.37}$
	CAS	ETF	DR	$91.03_{0.54}$	$75.97_{0.44}$	$61.55_{2.15}$	$76.16_{0.56}$
	BMLS	ETF	DR	$88.85_{0.16}$	$77.51_{0.39}$	$75.74_{0.42}$	$80.38_{0.23}$
	BMLS	MS-WETF	CE	$86.71_{0.88}$	$76.28_{0.69}$	$79.27_{1.39}$	$80.31_{0.43}$
imb 50	random	ETF	DR	$85.45_{0.50}$	$78.60_{0.28}$	$80.59_{0.42}$	$81.25_{0.18}$
	CBS	ETF	DR	$91.41_{1.07}$	$79.15_{1.05}$	$73.57_{1.93}$	$81.15_{0.37}$
	CAS	ETF	DR	$91.02_{1.68}$	$79.26_{1.09}$	$72.68_{2.07}$	$80.81_{0.22}$
	BMLS	ETF	DR	$88.17_{0.24}$	$80.21_{0.19}$	$85.87_{0.20}$	$84.30_{0.07}$
	BMLS	MS-WETF	CE	$87.01_{0.89}$	$80.36_{0.67}$	$86.59_{0.29}$	$84.22_{0.43}$
imb 10	random	ETF	DR	$89.67_{0.52}$	$83.81_{0.28}$	$90.54_{0.39}$	$87.59_{0.18}$
	CBS	ETF	DR	$92.79_{0.23}$	$85.14_{0.38}$	$88.28_{0.41}$	$88.38_{0.25}$
	CAS	ETF	DR	$92.87_{0.28}$	$85.33_{0.60}$	$88.72_{0.22}$	$88.61_{0.21}$
	BMLS	ETF	DR	$88.76_{0.93}$	$85.08_{1.00}$	$90.83_{0.79}$	$87.91_{0.24}$
	BMLS	MS-WETF	CE	$91.27_{0.32}$	$85.89_{0.20}$	$88.40_{0.42}$	$88.26_{0.04}$

Table 13: Experimental results of extension to the fixed ETF Classifier on Many/Medium/Few classes in the CIFAR100-LT dataset. The results are the mean of five repeated experiments with random seeds.

	Method	Clf.	$\mathcal{L}$	CIFAR100-LT			
				many	med	few	all
imb 200	random	ETF	DR	$68.23_{0.59}$	$42.05_{0.52}$	$6.63_{0.29}$	$41.20_{0.18}$
	CBS	ETF	DR	$63.90_{1.17}$	$38.98_{0.81}$	$7.36_{0.77}$	$38.78_{0.25}$
	CAS	ETF	DR	$64.10_{0.66}$	$38.86_{0.68}$	$7.68_{0.31}$	$38.91_{0.43}$
	BMLS	ETF	DR	$63.81_{0.48}$	$39.09_{0.69}$	$9.94_{0.54}$	$39.54_{0.45}$
	BMLS	MS-WETF	CE	$65.58_{0.70}$	$45.26_{0.54}$	$11.32_{0.52}$	$42.73_{0.41}$
imb 100	random	ETF	DR	$69.85_{0.40}$	$47.22_{0.35}$	$11.72_{0.81}$	$45.07_{0.25}$
	CBS	ETF	DR	$65.43_{0.88}$	$44.78_{0.94}$	$12.88_{0.91}$	$42.96_{0.25}$
	CAS	ETF	DR	$66.04_{0.40}$	$44.73_{0.34}$	$12.93_{0.35}$	$43.18_{0.18}$
	BMLS	ETF	DR	$65.59_{0.18}$	$44.49_{0.45}$	$15.21_{0.49}$	$43.60_{0.22}$
	BMLS	MS-WETF	CE	$63.44_{0.32}$	$51.15_{0.87}$	$21.92_{0.72}$	$47.10_{0.47}$
imb 50	random	ETF	DR	$70.56_{0.39}$	$53.52_{0.65}$	$22.69_{0.70}$	$50.71_{0.24}$
	CBS	ETF	DR	$67.73_{0.54}$	$51.15_{0.13}$	$22.59_{0.50}$	$48.84_{0.16}$
	CAS	ETF	DR	$67.87_{0.55}$	$51.58_{0.62}$	$22.63_{0.78}$	$49.05_{0.36}$
	BMLS	ETF	DR	$66.21_{0.58}$	$51.02_{0.49}$	$27.06_{0.59}$	$49.54_{0.39}$
	BMLS	MS-WETF	CE	$67.02_{0.90}$	$54.66_{0.62}$	$31.66_{0.41}$	$52.44_{0.40}$
imb 10	random	ETF	DR	$72.76_{0.29}$	$64.48_{0.50}$	$49.39_{0.36}$	$63.08_{0.21}$
	CBS	ETF	DR	$70.89_{0.43}$	$63.73_{0.42}$	$48.90_{0.49}$	$62.01_{0.19}$
	CAS	ETF	DR	$71.13_{0.45}$	$63.89_{0.34}$	$50.12_{0.45}$	$62.50_{0.27}$
	BMLS	ETF	DR	$68.95_{1.20}$	$64.83_{0.71}$	$50.18_{1.26}$	$62.06_{0.22}$
	BMLS	MS-WETF	CE	$68.81_{0.40}$	$64.95_{0.46}$	$57.24_{0.28}$	$64.10_{0.25}$