

Article

A Unified Complementary Regularization Framework for Long-Tailed Image Classification

Xingyu Shen ¹, Lei Zhang ¹, Lituan Wang ^{1,*} and Yan Wang ²

¹ Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, China; shenxy@stu.scu.edu.cn (X.S.); leizhang@scu.edu.cn (L.Z.)

² Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore 138632, Singapore; wangyan@a-star.edu.sg

* Correspondence: lituanwang@scu.edu.cn

Abstract

Class imbalance is a formidable and ongoing challenge in image classification tasks. Existing methods address this issue by emphasizing minority classes through class redistribution in the feature space or adjusting decision boundaries. Although such approaches improve the accuracy of minority classes, they often lead to unstable training and performance degradation on majority classes. To alleviate these challenges, we propose a unified redistribution framework termed as ComReg, which explicitly enforces complementary regularization on feature learning and decision boundary optimization in long-tailed image classification. Specifically, ComReg employs a multi-expert learning framework combined with prior-knowledge-guided online distillation to construct distribution-aware decision boundaries. From the feature space learning perspective, we enhance intra-class compactness and inter-class separability through decoupled-balanced contrastive learning. To further align the distributions in both spaces, we introduce a delay-weighted prototype learning strategy, which incorporates the decision boundary constructed by the head-class expert into the decoupled-balanced contrastive learning process. Extensive experiments on widely used long-tailed benchmarks, including CIFAR10-LT and CIFAR100-LT, as well as the real-world long-tailed datasets such as subsets of MedMNIST v2, demonstrate that our method achieves state-of-the-art performance.

Keywords: long-tailed image classification; supervised contrastive learning; knowledge distillation

1. Introduction

Visual recognition is a fundamental task that aims to learn the mapping between visual inputs and their corresponding labels [1–4], and it has achieved remarkable success in recent years. Most research has focused on enhancing the overall performance. This emphasis stems from the design of large-scale datasets such as ImageNet [5], where each class contains a roughly equal number of training samples. Consistent with this tendency, significant efforts have been made to advance the field through crowdsourced data collection. Unfortunately, data in real-world applications typically follow a long-tailed distribution, i.e., a small number of classes with abundant samples are referred to as head (majority) classes, and the remaining classes with few samples are referred to as tail (minority) classes. Training on class-imbalanced datasets often induces models to bias toward majority classes, resulting in poor generalization and degraded accuracy on underrepresented classes.



Academic Editor: Thomas Lindner

Received: 7 January 2026

Revised: 29 January 2026

Accepted: 4 February 2026

Published: 6 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Class re-balancing is a straightforward strategy to solve this problem, typically implemented through data resampling [6–8] or loss re-weighting [9]. These methods aim to improve classification performance by rebalancing the class contribution during training. While they yield notable improvements for minority classes, these gains often come at the expense of accuracy on majority classes.

In contrast to explicitly re-balancing class contributions, contrastive learning and multi-expert learning that focus on balanced feature learning have been widely explored in recent years to tackle the long-tailed classification. A representative example is supervised contrastive learning (SCL) [10], which constructs a feature space with high intra-class compactness and inter-class separability in a fully supervised setting by leveraging label information. Variants of SCL have been proposed in recent years to construct an effective feature space, thereby boosting the performance of long-tailed image classification by re-balancing the contribution of positive and negative samples [11–15]. However, the constructed models usually exhibit suboptimal decision boundaries, predominantly stemming from the limited batch size and inherent class imbalance (as shown in Figure 1b).

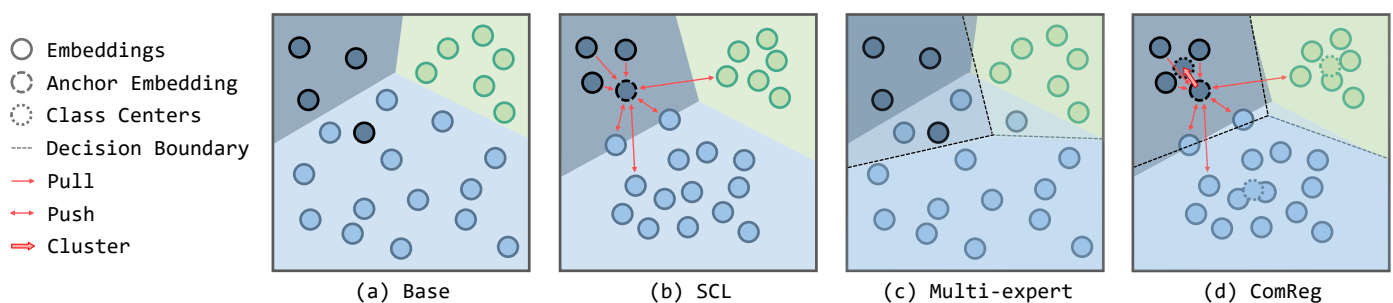


Figure 1. Illustration of feature distribution and decision boundaries with different methods. (a) The baseline without any imbalance methods, (b) Supervised Contrastive Learning method, (c) Multi-expert learning method, and (d) Our method. Dashed lines indicate decision boundaries. One-way arrow denotes intra-class compactness, two-way arrow signifies enhanced inter-class separability, and boxed one-way arrow indicates the clustering of class centers.

To construct more effective decision boundaries, multi-expert learning [16,17] leverages the strengths of multiple specialized experts to enhance the recognition of diverse classes. Existing approaches [18,19] primarily assign different classification heads to distinct domains while sharing a common feature extractor. However, they often overlook collaboration among experts, which compromises the construction of a discriminative and well-generalized feature space for long-tailed image classification (as shown in Figure 1c). Furthermore, recent studies in multi-task learning have demonstrated that training models on multiple tasks sharing the same feature extractor may lead to training instability [20], which is also observed in multi-expert long-tailed image classification.

To this end, we propose a unified framework termed ComReg, which leverages the complementary strengths of multi-expert learning in implicitly influencing decision boundary behavior and SCL in discriminative feature space construction. Unlike existing approaches that treat multi-expert learning and supervised contrastive learning as largely independent or loosely connected components, ComReg is designed to regulate how their optimization signals interact during training. Specifically, ComReg employs a multi-expert learning framework incorporating prior-knowledge-guided online distillation to encourage distribution-aware classification behavior, rather than explicitly modeling geometric decision boundaries. Inspired by the success of DeiT-LT [19], we employ two specialised experts for head and tail classes to capture class-frequency-dependent boundary information. On this basis, we introduce a triple-decoupled balanced contrastive learning (DBCL) strategy to mitigate semantic bias among positive samples under class-imbalanced

settings and to promote compact and discriminative intra-class representations. Crucially, to bridge decision boundary behavior and feature-space regularization within a unified optimization process, we incorporate class prototypes generated by the tail-class expert into contrastive learning. This design allows boundary-related prior knowledge to directly influence representation learning, resulting in a mutually constrained optimization of classifier behavior and feature geometry. To further stabilize training and mitigate the adverse effects of unreliable prototypes at early stages, we introduce a delayed-weighting mechanism, leading to the DW-DBCL (**D**elayed-**W**eighted **D**ecoupled **B**alanced **C**ontrastive **L**earning) variant. The main contribution of this work can be summarized as follows:

- We propose ComReg, an optimization-level coupling framework for long-tailed image classification, which explicitly links decision boundary behavior and feature space regularization by coordinating multi-expert learning and supervised contrastive learning within a unified training process.
- We introduce a prototype-guided contrastive learning mechanism, in which class prototypes derived from tail-class experts are integrated into contrastive supervision, together with a delayed weighting strategy to ensure stable and robust representation learning under severe class imbalance.
- We conduct extensive experiments on standard long-tailed benchmarks, including CIFAR10-LT, CIFAR100-LT, and representative subsets of MedMNIST v2, demonstrating consistent performance improvements and providing insights into the optimization behavior of the proposed framework.

2. Related Work

2.1. Class Re-Balancing

Class re-balancing strategies can be broadly divided into two categories: data re-sampling applied before training, and loss re-weighting applied during training. These methods modulate the influence of training samples through category-specific mechanisms at different stages of the model optimization process.

Data resampling techniques are generally categorized into undersampling and oversampling. Undersampling reduces the number of majority-class instances based on pre-defined criteria to achieve numerical balance with minority-class samples, promoting balanced feature learning across classes. To tackle the challenges of identifying representative majority-class instances and determining the optimal number to remove, Xie et al. [21] proposed a density-based progressive undersampling method that leverages spatial distribution characteristics. However, undersampling may compromise the model's generalization ability due to the loss of majority-class information. In contrast, oversampling addresses class imbalance by synthesizing minority-class samples. Foundational techniques like SMOTE [22] and ADASYN [23] inspired a range of oversampling-based approaches. In particular, input-mixing methods such as Mixup [24] and CutMix [25] can be viewed as variants that inherit and generalize the core of generating synthetic minority-class samples. Furthermore, with advancements of GANs [26], generative approaches for synthesizing minority-class samples have attracted increasing attention. Zhang et al. [27] proposed a GAN-based framework that incorporates Gaussian distribution estimation of minority-class samples to obtain prior knowledge in the latent space, while leveraging bagging mechanisms to alleviate overfitting.

Loss re-weighting encourages the model to focus on minority-class samples by adjusting the contribution of each class to the loss function. Early studies reweighted the CrossEntropy loss in multi-class classification to bias the model's learning process toward underrepresented classes. Building on this foundation, Focal Loss [28] introduced a modulation factor to dynamically adjust sample weights based on classification confidence,

thereby directing optimization toward hard-to-classify instances. Following a similar principle, Cui et al. [29] proposed a class-balanced loss that modifies training weights based on the effective number of samples per class, promoting more balanced optimization across long-tailed distributions. Cao et al. [30] introduced a class-dependent margin loss that improves generalization on minority classes by enlarging the margin around decision boundaries, effectively expanding the corresponding decision regions. Memon et al. [9] proposed a logit-space rectification method that adjusts the pre-softmax outputs during training to amplify gradient signals from minority-class samples. However, improper loss re-weighting may introduce optimization noise that distorts the learning dynamics, potentially degrading the quality of the resulting decision boundaries.

2.2. Multi-Expert Learning with Knowledge Distillation

With the advancement of computational capabilities in deep learning, multi-expert learning frameworks and knowledge distillation paradigms [31,32] have been increasingly adopted to tackle long-tailed recognition challenges. The multi-expert architecture enhances the model's representational capacity through synergistic integration of domain-specific expertise from multiple specialized submodels or modular components. Meanwhile, knowledge distillation refined decision boundary formation by transferring prior knowledge encoded in teacher models' soft labels, thereby enabling student models to achieve more balanced classification across majority and minority classes. Ren et al. [32] introduced a balanced self-distillation mechanism that leverages multi-augmented views of identical samples to mitigate teacher models' dependence on majority-class data, effectively redirecting the distillation focus toward learning discriminative features of minority classes.

Recent advancements [33,34] have increasingly integrated these two strategies, aiming to further improve model performance under severe class imbalance. Building on the NCL framework [16], Tan et al. [17] proposed an improved architecture, NCL++, which incorporates dual auxiliary learning and balanced online distillation to enhance the model's discriminative capability, particularly for hard-to-classify categories. The RIDE framework, proposed by Wang et al. [18], introduces a distribution-aware diversity loss to promote divergent predictions for minority-class samples, alongside a dynamic instance-to-expert routing mechanism that adaptively assigns samples to suitable experts based on feature complexity. Rangwani et al. [19] addressed the efficiency-oriented distillation task for Vision Transformers (ViT), and proposed DeiT-LT, an enhanced framework building upon DeiT [35]. By leveraging error predictions from a ResNet teacher model, DeiT-LT introduces an inductive bias that aligns the ViT student model more closely with CNN-like representations, effectively endowing the dual-head structure with expert specialization capabilities.

2.3. Supervised Contrastive Learning

Supervised Contrastive Learning (SCL) [10] extends self-supervised contrastive learning by incorporating principles from metric learning. It refines feature representations by constructing sample pairs that encourage higher similarity among semantically related instances and lower similarity among unrelated ones, guided by a supervised contrastive loss.

To reduce the dominant influence of majority classes on the optimization direction of SCL under long-tailed distributions, Zhu et al. [12] proposed Balanced Contrastive Learning (BCL), which incorporates class-complement and class-averaging strategies to ensure the global lower bound of the loss function becomes a category-agnostic constant. Further extending the SCL paradigm, Du et al. [13] developed Probabilistic Contrastive Learning (ProCo), which models feature distributions using the von Mises–Fisher (vMF) distribution. This probabilistic framework effectively reduces the batch-size sensitivity of SCL by enforcing explicit distributional alignment in the feature space. Building upon SCL,

Xuan et al. [14] proposed a decoupled positive term to ensure that gradient contributions during early training and loss convergence remain invariant to class frequency. Despite these advancements, existing supervised contrastive paradigms often enhance minority-class performance at the expense of majority-class discriminability.

3. Methods

In this section, we provide a detailed introduction to our proposed framework from three aspects: the overall framework, the decoupled balanced contrastive learning, and the prototypes delayed weighting learning strategy. Rather than introducing new standalone components, the core objective of this section is to clarify how different supervision signals are coupled within a unified optimization process to jointly regulate decision boundary behavior and feature space geometry under long-tailed distributions.

3.1. Overall Framework

Effective training of deep models under long-tailed data distributions is often hindered by biased optimization dynamics, where decision boundary learning and feature representation learning are influenced by imbalanced supervision in different and sometimes conflicting ways. Existing approaches typically address these issues using either classifier-level interventions or feature-level regularization, with limited interaction between the two. In contrast, our framework is designed to explicitly coordinate these optimization signals during training.

To influence class-wise prediction behavior under long-tailed supervision, we adopt a multi-expert learning framework as the backbone of our method. Motivated by the DeiT-LT [19] framework, we adopt an online knowledge distillation setup with a dual-classifier student model. Specifically, the student is equipped with two classifiers: one supervised by ground-truth labels, and the other by soft targets from the teacher. This design allows the student model to adjust its classification behavior under imbalanced supervision by leveraging the prior knowledge distilled from the teacher. At the same time, it encourages the classifiers to specialize into different experts under distinct loss functions adaptively. Since the classification head is directly supervised by ground-truth labels, it better aligns with the true data distribution in the input space. In contrast, the distillation head, guided by the teacher's soft targets, tends to inherit the majority-class bias. We designate the classification head as the tail-class expert and the distillation head as the head-class expert.

While such a multi-expert architecture is effective for shaping classifier behavior, the resulting competition among experts over shared features can lead to a mismatch between decision boundary optimization and feature space organization. Here, decision boundary optimization refers to improved class-wise discrimination behavior induced by loss-level supervision, rather than an explicit geometric characterization of the boundary. To address this issue, we introduce a contrastive learning branch into the multi-expert framework, with the explicit goal of coupling feature regularization with boundary-related supervision. Building on this design, ComReg jointly coordinates boundary-related supervision signals and feature space regularization within a unified optimization loop. Crucially, contrastive learning in ComReg is not applied as an independent auxiliary objective. Instead, it is explicitly conditioned on boundary-related information provided by the tail-class expert. To this end, we incorporate class prototypes derived from the transposed weight matrix of the tail-class expert into the contrastive learning process. This enables class-frequency-aware boundary information to directly influence representation learning, particularly for tail classes that are underrepresented in mini-batch sampling.

In addition, we propose a triple decoupled balanced contrastive learning strategy to mitigate semantic bias among positive samples under class-imbalanced conditions.

Through this design, the gradients induced by contrastive learning are functionally aligned with the optimization behavior of the classification heads, resulting in mutually constrained updates to both decision boundaries and feature representations. Since class prototypes may be unreliable at early training stages due to random parameter initialization, we further introduce a prototype delayed weighting strategy to ensure stable and smooth optimization throughout training. The overview of the ComReg framework is shown in Figure 2.

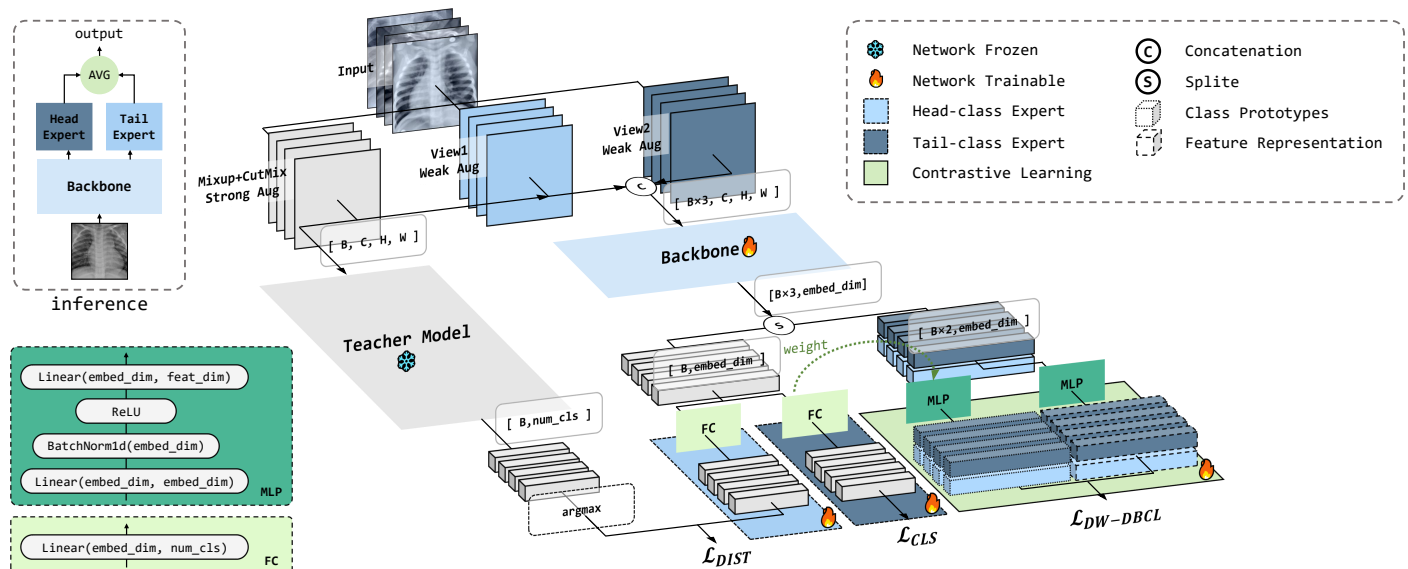


Figure 2. Overview of the ComReg framework. The left side of the diagram illustrates the fully connected layers and MLP projection used in the student model, along with the model inference process. B, C, H, W represent batch size, channels, height, and width, respectively.

3.2. Decoupled Balanced Contrastive Learning

Supervised contrastive learning (SCL) effectively models the feature space by enforcing intra-class compactness and inter-class separability. Although some methods [36] attempt to mitigate the reliance on batch size by introducing memory queues to expand the number of contrastive pairs, these approaches incur substantial memory overhead and still fail to address the semantic imbalance of sample pairs under long-tailed distributions. To tackle these issues, we introduce class prototypes into each mini-batch, ensuring that samples from all classes are represented in every training iteration. In addition, we decouple the positive pairs and assign distinct weights to each type, enabling differentiated emphasis during optimization. The formulation and analysis presented in this section are intended to provide qualitative intuition for the proposed design under simplified settings, rather than a rigorous theoretical characterization of optimization dynamics under extreme class imbalance.

Specifically, we define the input space as $\mathcal{X} = \{x_1, x_2, \dots, x_i\}$, the label space as $\mathcal{Y} = \{1, 2, \dots, K\}$, and the feature embedding space as $\mathcal{Z} = \{z_1, z_2, \dots, z_i\}$. Each sample $x_i \in \mathcal{X}$ is associated with a label $y \in \mathcal{Y}$, indicating the class to which x_i belongs. Each $z_i \in \mathcal{Z}$ denotes the representation of input $x_i \in \mathcal{X}$ in the embedding space. To ensure that contrastive learning can adequately represent each class, we treat each column in the weight matrix of the fully connected layer from the tail-class expert as the class center c_y , i.e., class prototype, for class y . These prototypes are then incorporated into the mini-batch [37]. The weight matrix from the tail-class expert is adopted, as its class prototypes are more conducive to the convergence of hard-to-classify minority class representations in

the feature space. While introducing c_y helps alleviate the imbalance in sampled batches, it does not address the semantic bias arising from intra-class imbalance.

The notations for feature embeddings used in SCL are summarized in Table 1, along with detailed explanations. In the embedding space, z_p denotes all the representations that belong to the same class y as z_i . Specifically, z_p comprises three components: z_i^+ , z_y , and c_y . We denote the positive embedding pairs as (z_i, z_p) , assuming that z_i belongs to the majority class. In this case, z_y appears in the mini-batch B with much higher frequency than z_i^+ and c_y . As a result, when modeling the feature space, the loss function primarily focuses on the similarities between (z_i, z_y) , while potentially neglecting the richer semantics encoded in z_i^+ and c_y . Conversely, if z_i belongs to a minority class, the occurrence of z_y is much lower. Nevertheless, z_i^+ , as an augmented view of z_i , still retains sufficient semantic relevance. Therefore, for majority-class anchors z_i , greater emphasis should be placed on the feature relationships between z_i^+ and c_y . In contrast, for minority-class anchors, the similarity between z_i and the limited same-class samples z_y should be prioritized.

Table 1. Notations.

Symbols	Descriptions
x_i	An input sample from the input space $\mathcal{X}$.
y	The ground-truth class label of x_i , i.e., $y \in \mathcal{Y}$.
z_i	The embedding of x_i in the feature space $\mathcal{Z}$.
z_i^+	The embedding of an augmented view of x_i .
B	A mini-batch of training samples.
$ B_y $	The number of samples in B that belong to class y .
z_y	All embeddings in B that belong to class y , excluding z_i^+ .
c_y	The prototype representation of class y .
z_p	The set of positive embeddings related to x_i , defined as $z_p = \{z_i^+, z_y, c_y\}$.

To mitigate the aforementioned issues and enhance the discriminability of features among the positive pairs (z_i, z_p) , we propose a contrastive learning loss with triple decoupling, which addresses semantic bias by assigning differentiated weights to distinct components within the positive set:

$$\mathcal{L}_{DBCL} = -\frac{1}{|B_y|} \sum_{z_p \in \{B_y \cup c_y\} \setminus \{z_i\}} \log f(z_p | z_i) \quad (1)$$

$$f(z_p | z_i) = \frac{\exp(w_p z_i \cdot z_p / \tau)}{\sum_{j \in \mathcal{Y}} \frac{1}{|B_j| + 1} \sum_{z_m \in \{B_j \cup c_j\}} \exp(z_i \cdot z_m / \tau)}$$

The function $f(z_p | z_i)$ denotes the similarity score between a positive embedding z_p and an anchor embedding z_i , computed after decoupling and class-balanced adjustment. The term w_p refers to the decoupling weight assigned to each positive sample.

3.3. Prototypes Delayed Weighting Learning Strategy

The introduction of class prototypes, along with the decoupling of the three types of positive pairs, makes the weighting of positive samples more intricate during training. To prevent the model from being misled by under-trained class prototypes in the early stages, we propose a delayed weighting strategy for class prototypes, which better guides the model toward cluster structures that align with the underlying feature space distribution.

As discussed in Section 3.2, the long-tailed class distribution leads to an uneven computation frequency of positive pairs. This imbalance makes it difficult to distinguish the semantic differences among various types of positive embeddings. Therefore, we design

the weighting scheme based on computation frequency, which can be characterized by $|B_y|$. The inclusion of class prototypes introduces further complexity to this weighting process. Specifically, the class prototype c_y , derived from the weight matrix of the fully connected layer in the tail-class expert, lacks meaningful representational capacity at the early stages of training due to random parameter initialization. As training progresses and the model gradually learns to construct a semantically structured feature space, c_y progressively gains the ability to reflect the underlying class distribution, eventually becoming a reliable embedding of the class center. To ensure that the model develops sufficient representational capacity for both class distributions and feature embeddings, we introduce a deferred weighting strategy, in which the weight w_p is defined as follows:

$$w_p^1 = \begin{cases} \alpha |B_y|, & z_p = z_i^+ \\ \frac{(1-\alpha)|B_y|}{|B_y|-1}, & z_p = z_y \in \{B_y \cup c_y\} \setminus \{z_i \cup z_i^+\} \end{cases} \quad (2)$$

$$w_p^2 = \begin{cases} \alpha |B_y|, & z_p = \{z_i^+ \cup c_y\} \\ \frac{(1-\alpha)|B_y|}{|B_y|-2}, & z_p = z_y \in \{B_y\} \setminus \{z_i \cup z_i^+\} \end{cases}$$

In the initial stage, the model has yet to form a reliable representation of the class distribution. Therefore, the weighting strategy w_p^1 emphasizes the pairwise similarity between (z_i, z_i^+) , promoting intra-instance consistency. Once the model has acquired sufficient capacity to encode class-level semantics, we introduce additional weighting on the pair (z_i, c_y) , as specified in w_p^2 , thereby gradually enhancing the role of class prototypes in the learning process.

To provide qualitative insight into the rationale behind the proposed weighting design, rather than a formal convergence or stability proof, we conduct a simplified gradient-based analysis focusing on the weighting factor. This analysis isolates the effect of frequency-based reweighting under controlled assumptions and is not intended to characterize optimization behavior under extreme imbalance. Specifically, we compute the partial derivative of the loss function $\mathcal{L}_{DBCL}$ with respect to the weighting factor w_p , while keeping the network parameters fixed, in order to isolate the effect of frequency-based reweighting:

$$\frac{\partial \mathcal{L}_{DBCL}}{\partial w_p} = -\frac{1}{\tau |B_y|} \sum_{z_p \in \{B_y \cup c_y\} \setminus \{z_i\}} (z_i \cdot z_p) \quad (3)$$

Following common practice in contrastive learning analyses, we assume that the inner products between samples within the same class are approximately homogeneous in expectation, which allows us to simplify the ratio in Equation (4). We note that this assumption is a simplification and may not strictly hold in severely imbalanced settings, where intra-class variability can be substantially higher. This assumption is introduced solely to facilitate tractable analysis and should be interpreted with caution. Therefore, we can derive the ratio of the Euclidean norms of the gradients of $\mathcal{L}_{DBCL}$ for the weight w_p at z_i^+ and z_y :

$$\frac{\left\| \frac{\partial \mathcal{L}_{DBCL}}{\partial w_p} \Big|_{z_i^+} \right\|}{\sum_{z_p \in \{B_y \cup c_y\} \setminus \{z_i \cup z_i^+\}} \left\| \frac{\partial \mathcal{L}_{DBCL}}{\partial w_p} \Big|_{z_p} \right\|} \approx \frac{|z_i \cdot z_i^+|}{(|B_y| - 1) |z_i \cdot z_y|} \quad (4)$$

We associate each component of the weight set $w_p = \{w_i^+, w_y, w_c\}$ with its corresponding positive embedding in $z_p = \{z_i^+, z_y, c_y\}$. Under the above simplifying assumptions,

by applying the frequency-based weighting strategy defined in Equation (4), we obtain the following formulation:

$$\frac{w_{z_i+}|z_i \cdot z_i^+|}{w_y(|B_y| - 1)|z_i \cdot z_y|} = \frac{\alpha|z_i \cdot z_i^+|}{(1 - \alpha)|z_i \cdot z_y|} \quad (5)$$

Under the above simplifying assumptions, Equation (5) suggests that frequency-based weighting can reduce the explicit dependence of gradient magnitude on class-specific sample counts, providing intuition for the role of the hyperparameter α . In a similar fashion, the balanced gradient formulation for the class prototype c_y can be derived, and is therefore omitted here for brevity. This observation should be interpreted as an illustrative explanation rather than a general or rigorous property of the optimization process. In practice, particularly under extreme class imbalance, the assumptions underlying this analysis may be violated, and the effectiveness of the proposed weighting strategy is therefore primarily supported by empirical observations rather than theoretical guarantees.

3.4. Loss Function

We introduce the weighting strategy on w_p term, as described in Section 3.3, into DBCL, resulting in the variant DW-DBCL. Accordingly, we define the overall training loss of the ComReg framework as follows:

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{DIST} + \xi \mathcal{L}_{CLS} + \mu \mathcal{L}_{DW-DBCL} \quad (6)$$

where λ , μ , and ξ are weighting hyperparameters. The $\mathcal{L}_{DIST}$ and $\mathcal{L}_{CLS}$ are the loss functions that constrain the distillation head and classification head, respectively.

To the configuration of loss functions for these two heads, the DeiT-LT framework employs the standard CE loss for both. Additionally, a class-frequency-based re-weighting scheme is applied to the CE loss during the later training stages to gradually specialize the model as a tail-class expert. Based on our analysis of loss re-weighting strategies in Section 2.1, together with empirical observations, we hypothesize that this specialization is closely related to the class-dependent weighting in the loss function. The effectiveness of combining Logit Adjustment (LA) with contrastive learning for feature space modeling has been demonstrated in BCL [12]. Therefore, we adopt LA loss as the class-weight-aware supervision and use standard CE loss as the weight-free baseline, assigning them to supervise the two heads respectively to validate our hypothesis.

As shown in Figure 3, when one classifier is trained with a class-weighted loss function, it tends to specialize in minority-class discrimination, while the other classifier primarily focuses on majority-class recognition. Furthermore, by comparing Figure 3a,b, we observe that the overall performance of the distillation head remains relatively stable regardless of whether class weighting is applied to its loss function, owing to the incorporation of prior knowledge from the teacher model. In contrast, when the classification head is trained under a standard CE loss without class-aware weighting, its performance on minority classes degrades significantly. It is worth noting that this design choice is empirically validated in our experiments and should not be interpreted as a theoretically optimal configuration under all imbalance regimes. Motivated by these observations, we adopt CE loss for the distillation head and LA loss for the classification head, as illustrated in Figure 3b, to leverage their respective strengths in modeling majority and minority classes.

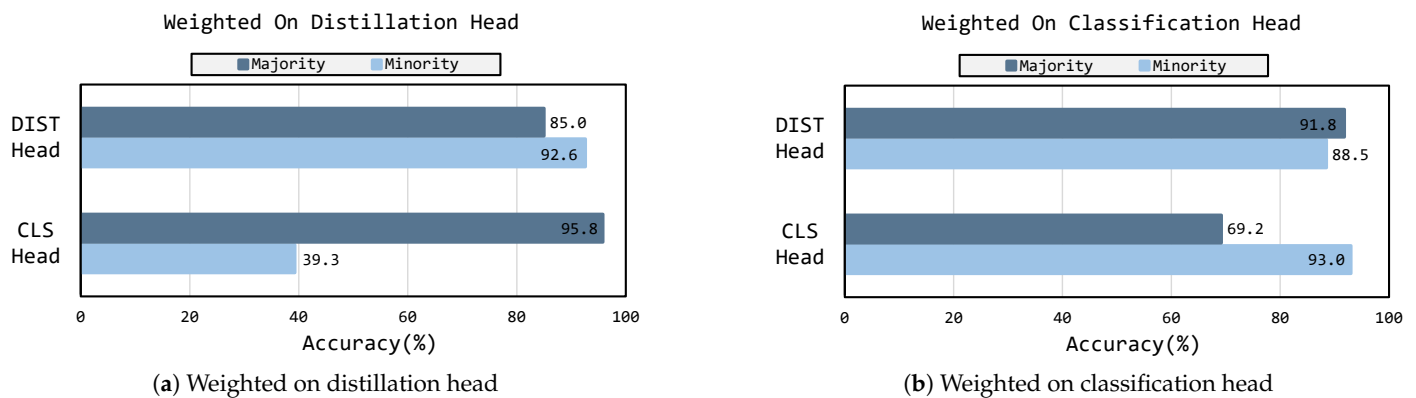


Figure 3. Comparison of expert specialization under two different loss assignment strategies on CIFAR10-LT. DIST Head and CLS Head denote the distillation head and classification head, respectively. We report the ACC (%) metric for both head-classes and minority classes.

4. Experiments

4.1. Datasets and Evaluation Metrics

We evaluate our method on the widely adopted CIFAR-LT dataset and the real-world long-tailed MedMNIST v2 benchmark [38]. Our experimental design follows standard benchmark protocols to enable fair and reproducible comparison with existing long-tailed classification methods. The imbalance factor is defined as $\rho = \frac{N_{\max}}{N_{\min}}$, where $N_{\max}$ and $N_{\min}$ denote the number of samples in the most and least frequent classes, respectively.

The CIFAR-LT benchmark comprises CIFAR10-LT and CIFAR100-LT, following the training set partitioning protocol introduced in prior work [19], where long-tailed distributions with imbalance factors of 100, 50, and 10 are constructed in a synthetic yet widely adopted manner based solely on the original CIFAR training sets. The test sets remain unchanged from the standard CIFAR benchmarks. While such settings do not capture all real-world complexities, such as label noise or compound distributional shifts, they provide a controlled testbed for isolating the effects of class imbalance and for direct comparison with prior work.

For MedMNIST v2, to ensure a comprehensive evaluation across varied medical imaging scenarios and imbalance patterns, we base our subset selection on two key criteria: (1) diversity in imaging modalities commonly used in clinical practice, and (2) inclusion of classification tasks with varying degrees of class imbalance. Guided by these criteria, we select seven representative subsets from the MedMNIST v2 benchmark. These subsets span a range of medical imaging types, including dermoscopy, chest X-rays, breast ultrasound, blood cell microscopy, and abdominal CT scans, and cover both binary and multi-class classification tasks. This subset selection is intended to provide a controlled yet representative validation setting, rather than an exhaustive evaluation across all possible medical imaging tasks. We use the train and test sets from MedMNIST v2 for model training and performance evaluation, respectively. The class distributions of the training sets from CIFAR-LT and MedMNIST subsets are illustrated in Figure 4.

Since the test sets of CIFAR-LT follow the original CIFAR test sets without modification, we report only the accuracy (ACC), expressed as a percentage, as the evaluation metric. For the MedMNIST v2 subsets, we adopt both ACC and the area under the receiver operating characteristic curve (AUC), the latter being particularly suitable for medical image classification. We note that the reported results reflect performance under a fixed experimental configuration, and a systematic analysis of robustness to alternative dataset constructions or architectural choices is beyond the scope of this study. The AUC is computed by thresholding the predicted probabilities at 0.5.

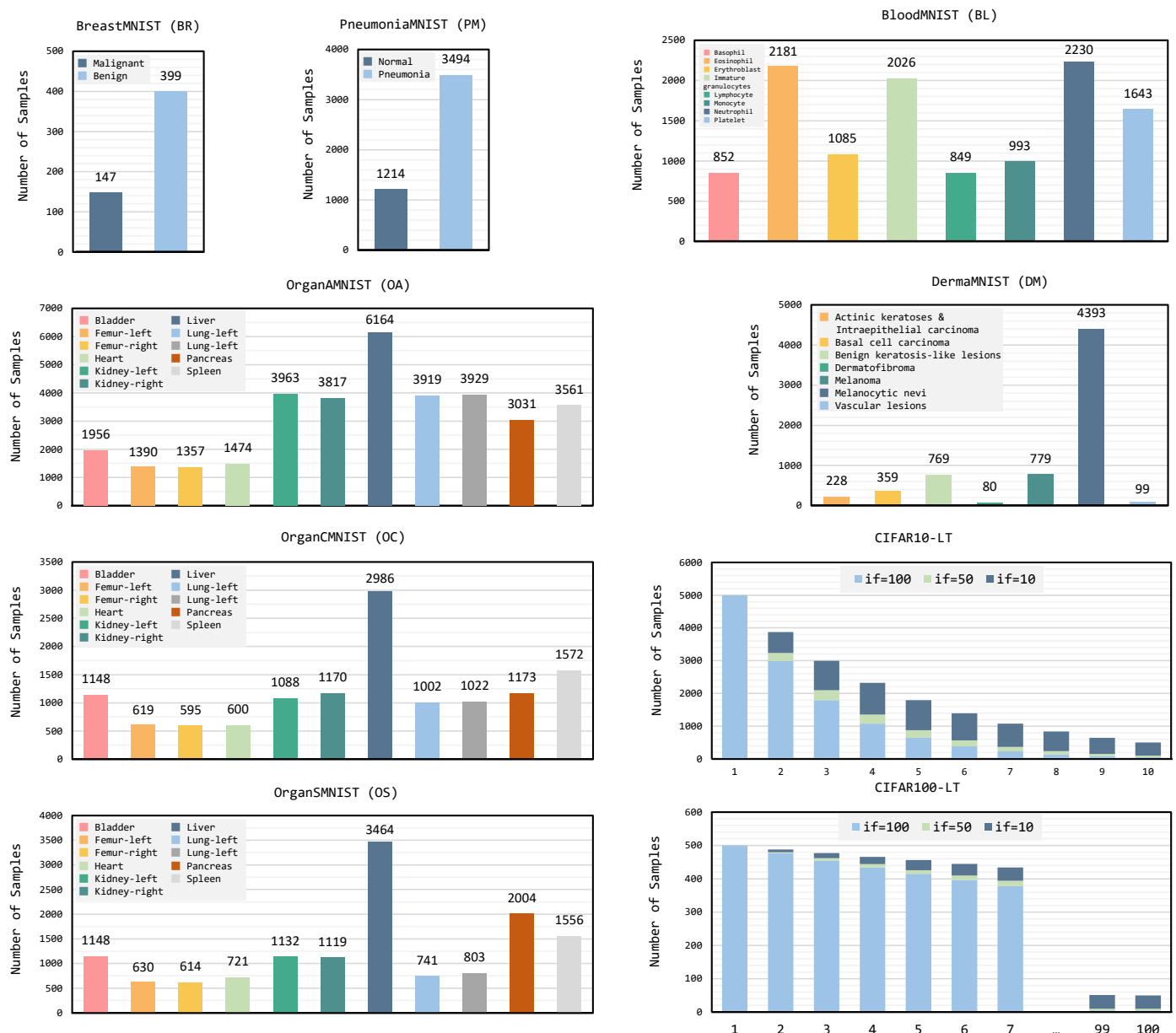


Figure 4. Class distributions in the training sets of CIFAR-LT and selected MedMNIST subsets.

4.2. Experimental Setup

In this subsection, we present the experimental configurations for both the teacher and student models on the CIFAR-LT and MedMNIST v2. All experiments, including the pretraining of the teacher model, were conducted using a single NVIDIA RTX 4090 GPU.

4.2.1. Long-Tailed Versions of CIFAR

For the CIFAR-LT benchmark, we evaluate our framework using two backbone architectures: ResNet and ViT. Specifically, following prior studies, we adopt ResNet-32 and ViT-Base as the backbones for the ResNet-based and ViT-based frameworks, respectively. ResNet-32 operates on images of size 32×32 , while ViT-Base processes images resized to 224×224 . For both architectures, the loss weights λ , μ , and ζ are all set to 1.0, and the temperature coefficient τ is set to 0.1. Under both architectures, the hyperparameter α is set to 0.3 for CIFAR10-LT and 0.4 for CIFAR100-LT.

ResNet-32 backbone. Specifically, under the ResNet-32 framework, the teacher model employs the same ResNet-32 backbone, pre-trained independently for 200 epochs

using the proposed DW-DBCL methodology. The student model is trained using SGD with a mini-batch size of 256 over 300 epochs. The learning rate is gradually increased to 0.1 over the first 5 warm-up epochs, and subsequently decayed by a factor of 0.1 at the 160th and 250th epochs. Both classifiers are trained with strongly augmented inputs (AutoAugment [39], Mixup [24], and Cutmix [25]), whereas the contrastive learning branch receives weakly augmented data (SimAugment [40]). A strongly weighted CE loss is applied to the distillation head starting from the 250th epoch. It is worth noting that for CIFAR10-LT and CIFAR100-LT, both Mixup and Cutmix augmentations are applied only during the first 250 and 160 epochs, respectively. Correspondingly, the delayed class prototype weighting strategy is activated starting from the 160th and 250th epochs for CIFAR10-LT and CIFAR100-LT, respectively.

ViT-Base backbone. In the ViT framework configuration, for teacher models addressing CIFAR-LT datasets with imbalance ratios of 100 and 50, we adopt the ResNet-32 model pre-trained using the PaCo [11] + SAM [41] strategy provided by DeiT-LT [19]. However, due to the lack of an official DeiT-LT implementation for the 10:1 imbalance setting, we employ the ResNet-32 model trained with our DW-DBCL methodology (as described previously) as the teacher model for this particular case. For student model training, while strictly adhering to DeiT-LT's experimental protocol in terms of parameter configuration, we reduce the batch size from 256 to 64 due to computational constraints. The training is conducted for 1200 epochs using the AdamW optimizer, with an initial learning rate of 5×10^{-4} scheduled by a cosine decay policy. For both CIFAR10-LT and CIFAR100-LT, strong augmentation techniques (Mixup and Cutmix) are applied during the initial 1100 epochs. Starting from the 1100th epoch, the delayed class prototype weighting strategy is activated, accompanied by an increased weighting of the CE loss at the distillation head.

4.2.2. Medical Dataset MedMNIST v2

For the MedMNIST v2 benchmark, we evaluate three commonly used backbone architectures: ResNet-18, ResNet-34, and ResNet-50. These choices are consistent with the baselines provided in the official MedMNIST v2 framework. Based on a comprehensive assessment of training time and classification accuracy, we find that ResNet-34 achieves an optimal trade-off between efficiency and performance. Therefore, we adopt ResNet-34 as the backbone architecture for both the teacher and student models. All input images are taken from the 224×224 resolution version provided by the MedMNIST v2. The teacher model is trained independently for 300 epochs under the supervision of the proposed DW-DBCL loss function. Consistent experimental configurations are maintained across all seven subsets, with student models trained for 300 epochs using a batch size of 64. The learning rate is linearly increased to 0.02 during a 5-epoch warmup phase and subsequently decayed by a factor of 0.1 at the 260th and 280th epochs. Both classifiers receive strongly augmented inputs (RandAugment [42], Mixup, and CutMix), while the contrastive learning branch operates on weakly augmented data generated via SimAugment. Notably, Mixup and CutMix are applied throughout the entire training process, and no intensified CE loss weighting is applied to the distillation head. The loss weights λ , μ , and ζ are set to 0.5, 0.5, and 1.0, respectively. The temperature coefficient τ and the hyperparameter α are fixed at 0.1 and 0.5 throughout all experiments.

4.3. Quantitative Analysis

4.3.1. Results on CIFAR-LT

Table 2 presents a comprehensive comparison between our proposed framework and previous state-of-the-art (SOTA) methods, categorized into four strategies groups: (1) loss re-weighting-based methods, e.g., CrossEntropy Loss, Focal Loss [28], LDAM-DRW [30],

and Logit Adjustment Loss [9], (2) contrastive learning-based approaches, e.g., PaCo [11], BCL [12], and ProCo [13], (3) hybrid augmentation techniques, e.g., ReMix [43] and ConCutMix [44], (4) recent proposed methods mainly focus on network architecture, e.g., Hybrid-SC [45] and Deit-LT [19]. In addition, we report baseline results obtained by applying Logit Adjustment alone for classification head regularization within our framework.

Table 2. Comparison results of the proposed method on CIFAR100-LT and CIFAR10-LT in metric of ACC (%). [†] denotes results borrowed from [45]. * represents the results reproduced using the code released by the author. Bold indicates the best result.

Method	CIFAR100-LT			CIFAR10-LT		
	$\rho = 100$	$\rho = 50$	$\rho = 10$	$\rho = 100$	$\rho = 50$	$\rho = 10$
ResNet-32 Backbone						
CE [†] [45]	38.3	43.9	55.7	70.4	74.8	86.4
Focal loss [†] [45]	38.4	44.3	55.8	70.4	76.7	86.7
LDAM – DRW [30]	42.0	46.6	58.7	77.0	81.0	88.2
Logit Adjustment [9]	50.5	54.9	64.0	84.3	87.1	90.9
PaCo [11]	52.0	56.0	64.2	-	-	-
BCL [12]	51.9	56.6	64.9	84.3	87.2	91.1
ProCo [13]	52.8	57.1	65.5	85.9	88.2	91.9
ReMix [43]	41.9	-	59.4	75.4	-	85.2
ReMix + LDAM – DRW [43]	45.0	-	59.5	79.3	-	86.8
ConCutMix + BCL [44]	53.1	57.4	64.5	86.1	88.0	91.4
Hybrid – SC [45]	46.7	49.5	63.1	81.4	85.4	91.1
Hybrid – PSC [45]	45.0	48.9	62.4	78.8	83.9	90.1
Baseline (Ours)	44.0	49.5	59.1	81.2	84.9	90.1
ComReg (Ours)	54.1	58.0	66.9	87.5	88.6	92.2
ViT Backbone						
DeiT – LT [19]	55.6	60.5	-	87.5	89.8	-
DeiT – LT *	54.9	60.1	-	87.5	89.6	-
ComReg (Ours)	55.6	61.1	68.1	87.4	89.5	93.4

As analyzed in the preceding sections, loss reweighting and hybrid augmentation strategies, two fundamental approaches for addressing long-tail recognition, exhibit limited performance improvements due to their partial focus on specific facets of the problem. Among them, LA loss is designed to optimize parameter learning during training by calibrating the output logits, and has been shown to outperform other loss re-weighting methods. However, the baseline implementation of our framework, which also adopts Logit Adjustment but applies it alongside strong augmentation strategies, i.e., Mixup and CutMix, underperforms [9], likely due to feature distribution shifts introduced by these augmentations. Among hybrid augmentation methods, ReMix extends conventional Mixup by decoupling the mixing ratios of features and labels, thereby enabling dynamic adjustment of decision boundaries to balance generalization errors between majority and minority classes. Although ReMix combined with LDAM-DRW achieves notable performance under severe imbalance conditions ($\rho = 100$), its effectiveness declines substantially in scenarios with moderate imbalance ($\rho = 50, 10$).

While both contrastive loss functions and loss re-weighting methods impose training-time constraints through loss minimization, contrastive learning offers superior performance in long-tail scenarios due to its min-max similarity optimization principle, which facilitates the learning of a more generalizable feature space. Empirical results consistently show that contrastive learning-based supervision suggest that loss re-weighting approaches.

Pan et al. [44] introduced ConCutMix, a hybrid augmentation strategy combining contrastive learning with CutMix. This method constructs semantic manifolds via contrastive relations and assigns semantically consistent labels by measuring similarities between synthetic samples and class centroids within the learned embedding space. Like other hybrid augmentation techniques, ConCutMix performs best when paired with complementary methods, particularly when integrated with BCL, where the combined approach yields a modest average accuracy improvement of approximately 0.75%.

Framework-based methodologies remain fundamentally constrained by their dependence on loss function supervision and limitations in architectural module design. DeiT-LT addresses these challenges by incorporating a dual-classifier architecture alongside out-of-distribution (OOD) image-driven knowledge distillation, thereby enhancing the generalization and robustness of ViT under long-tailed data distributions. However, its reliance on conventional delayed loss reweighting strategies primarily affects classifier outputs and may not fully capture fine-grained feature-space structure. The Hybrid-SC framework introduces a contrastive learning-based architecture with two variants: Hybrid-SC, employing supervised contrastive loss, and Hybrid-PSC, utilizing prototype-supervised contrastive loss. Although Hybrid-PSC eliminates the need for explicit negative sampling by aligning with class prototypes, its advantages diminish considerably when applied to datasets with limited categorical diversity, revealing intrinsic scalability limitations.

The proposed framework, ComReg, aims to regularize decision boundary behavior and feature space learning in a complementary manner. Under standard benchmark settings, ComReg demonstrates consistent performance improvements over representative contrastive learning methods and architectural baselines. Under the ResNet-32 configuration, the ComReg achieves average accuracy gains of 0.98% over ProCo and 1.13% over ConCutMix under the same experimental protocol, reflecting a consistent improvement trend rather than a claim of statistical significance. Notably, on CIFAR10-LT with $\rho = 100$, ComReg matches the performance of 224×224 resolution baselines despite operating on 32×32 input images. In ViT-based architectures, although ComReg shows relatively lower performance on CIFAR10-LT due to the limited number of categories and constraints in contrastive batch sampling, it shows more pronounced and consistent improvements on CIFAR100-LT, outperforming DeiT-LT-based multi-expert approaches by approximately 0.70% ($\rho = 100$) and 1.00% ($\rho = 50$) under the same experimental configuration.

Figure 5 illustrates the performance differences between ComReg and the baseline across different imbalance factors on CIFAR100-LT and CIFAR10-LT. The results indicate a consistent improvement trend across imbalance settings, with larger gains generally observed on CIFAR100-LT, where higher class cardinality increases the difficulty of long-tailed learning. These observations suggest that ComReg is particularly effective in scenarios with greater category diversity and more severe imbalance, although the reported improvements should be interpreted as indicative trends under fixed experimental conditions rather than statistically conclusive evidence.

We supplemented Table 3 with additional results on the CIFAR-LT datasets obtained by training the ComReg framework under different random initialization seeds. Unlike Table 2, where all results follow the random seed setting commonly adopted in prior work (seed = 0), we report results obtained with three different random seeds (0, 1, and 2) in this table. Overall, the results show that ComReg exhibits low variance across different random seeds under all imbalance settings, indicating that the observed performance is stable and not sensitive to random initialization.

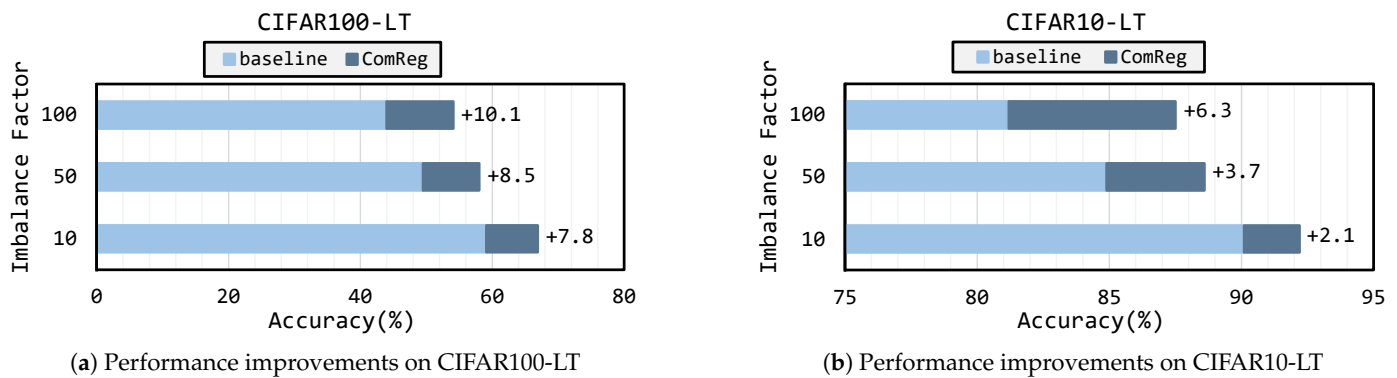


Figure 5. Comparison of classification accuracy between ComReg and the baseline across different imbalance factors ($\rho \in \{10, 50, 100\}$) on CIFAR100-LT and CIFAR10-LT. ComReg exhibits consistent accuracy improvements under all evaluated imbalance settings.

Table 3. Model performance across different random seeds (0, 1, 2) under varying imbalance ratios ($\rho = 100, 50, 10$).

Dataset	Seed	$\rho = 100$	$\rho = 50$	$\rho = 10$
CIFAR10-LT	0	87.47	88.17	92.07
	1	87.56	89.24	92.48
	2	87.34	89.15	92.37
	Mean $\pm$ std	87.46 ± 0.09	88.85 ± 0.48	92.30 ± 0.17
CIFAR100-LT	0	54.12	57.96	66.86
	1	55.76	59.30	66.82
	2	56.12	59.85	67.30
	Mean $\pm$ std	55.33 ± 0.87	59.03 ± 0.79	66.99 ± 0.21

4.3.2. Results on MedMNIST v2

We further conduct comparative evaluations on seven MedMNIST v2 subsets that exhibit realistic long-tailed distributions, benchmarking our framework against high-performing methods from recent studies. We select ResNet-18, ResNet-50, and AutoML Vision [46] as basic reference methods, following the MedMNIST v2 [38] paper, to serve as foundational baselines in our evaluation. Due to the limited prior work specifically addressing long-tailed challenges within MedMNIST v2, our primary comparisons focus on architecture-centric approaches (FPViT [47], MedViT [48], and MedMamba [49]) and the incentive-driven learning strategy Complex Mixer [50]. To further assess the effectiveness of our framework, we additionally perform ablation studies by integrating BCL [12] into the contrastive branch of ComReg. Comprehensive results across all subsets are summarized in Table 4.

Liu et al. [47] introduced the Feature Pyramid Vision Transformer (FPViT), which repurposes the foundational layers of ResNet as feature extractors by flattening multi-scale CNN feature maps into linear sequences for shallow ViT input. This design effectively leverages the hierarchical feature representations of deep CNNs, exhibiting superior extraction performance compared to pure ViT-based architectures. Along similar lines, MedViT [48] integrates the complementary strengths of CNNs and Transformers while proposing a novel augmentation strategy, Patch Momentum Changer (PMC). PMC operates in the feature space by blending feature moments across different medical images, thereby enhancing model robustness and generalization. Extending this fusion paradigm, Yue et al. [49] proposed MedMamba, which couples CNNs with Vision Mamba—an architecture that combines CNNs' localized feature extraction with State Space Models (SSMs) to capture

long-range dependencies via grouped convolutions and channel-wise mixing, specifically adapted for medical imaging tasks. In a different direction, Zheng et al. [50] introduced Complex Mixer, an incentivized learning framework that injects conditional noise into input signals to reduce modeling complexity. This self-supervised approach is particularly effective under MedMNIST v2's limited data and feature uncertainty conditions, delivering competitive performance across fully, semi-, and weakly-supervised settings.

Table 4. The performance of different methods on the medical datasets. * denotes the results obtained by incorporating BCL [12] into the contrastive learning branch of ComReg. Bold indicates the best result.

Method	BL		BR		DM		PM		OA		OC		OS	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
ResNet18 [51]	0.998	96.3	0.891	83.3	0.920	75.4	0.956	86.4	0.998	95.1	0.994	92.0	0.974	77.8
ResNet50 [51]	0.997	95.0	0.866	84.2	0.912	73.1	0.962	88.4	0.998	94.7	0.993	91.1	0.975	78.5
AutoML Vision [46]	0.998	96.6	0.919	86.1	0.914	76.8	0.991	94.6	0.990	88.6	0.988	87.7	0.964	74.9
FPViT [47]	-	-	0.938	89.1	0.923	76.6	0.973	89.6	0.997	93.5	0.993	90.3	0.976	78.5
MedViT-T [48]	0.996	95.0	0.934	89.6	0.914	76.8	0.993	94.9	0.995	93.1	0.991	90.1	0.972	78.9
MedViT-S [48]	0.997	95.1	0.938	89.7	0.937	78.0	0.995	96.1	0.996	92.8	0.993	91.6	0.987	80.5
MedViT-L [48]	0.996	95.4	0.929	88.3	0.920	77.3	0.966	92.1	0.997	94.3	0.994	92.2	0.973	80.6
MedMamba-T [49]	0.999	97.8	0.825	85.3	0.917	77.9	0.965	89.9	0.999	94.6	0.997	92.7	0.982	81.9
MedMamba-S [49]	0.999	98.4	0.806	85.3	0.924	75.8	0.976	93.6	0.999	95.9	0.997	94.4	0.984	83.3
MedMamba-B [49]	0.999	98.4	0.849	89.1	0.925	75.7	0.973	92.5	0.999	96.4	0.997	94.3	0.984	83.4
MedMamba-X [49]	0.999	97.8	0.898	85.3	0.918	75.1	0.964	91.0	0.999	96.7	0.998	94.7	0.983	83.0
Complex Mixer [50]	0.998	97.9	0.939	89.9	0.940	83.3	0.997	96.6	0.997	95.1	0.993	91.7	0.977	81.0
BCL * [12] (Ours)	0.999	99.2	0.933	93.6	0.973	86.0	0.997	97.0	0.998	96.0	0.997	94.4	0.982	82.0
ComReg (Ours)	0.999	99.2	0.935	93.6	0.982	88.6	0.995	98.2	0.999	95.9	0.998	94.1	0.982	82.0

As shown in Table 4, ComReg framework also demonstrates strong applicability to real-world long-tailed medical imaging datasets such as MedMNIST v2. Notably, on the DM subset, characterized by the most severe class imbalance, ComReg achieves the largest performance improvement, surpassing the best competing method by 0.037 in AUC and 5.3% in ACC. This highlights the framework's superior capability in handling long-tailed distributions under realistic settings. Moreover, ComReg attains the highest ACC scores on the BL (99.2%), BR (93.6%), and PM (98.2%) subsets, indicating its effectiveness across both multi-class and binary classification tasks. It is also worth noting that MedMamba achieves strong results on the OA and OC subsets, while MedViT excels on OS, underscoring the robustness of SSM- and Transformer-based architectures in abdominal CT imaging scenarios.

4.4. Qualitative Analysis

To better demonstrate the intra-class compactness and inter-class separability achieved by our method in the feature space, we conduct t-SNE visualizations on the CIFAR10-LT test set. For clarity, we selectively visualize samples from both majority and minority classes. Figure 6 compares the feature distributions produced by three ResNet-32-based models: (1) the teacher model, (2) the DeiT-LT student model, and (3) the ComReg student model. Regions containing ambiguous or overlapping class clusters are marked with red boxes. As illustrated, the ComReg-trained model exhibits more compact intra-class clustering and clearer inter-class boundaries, indicating superior feature space organization. These results underscore the effectiveness of our approach, which combines decision boundary optimization from the multi-expert framework with contrastive regularization for feature space modeling. This synergy enables improved recognition of minority classes while

preserving stable performance on majority classes, constituting the central mechanism by which ComReg addresses long-tailed learning challenges.

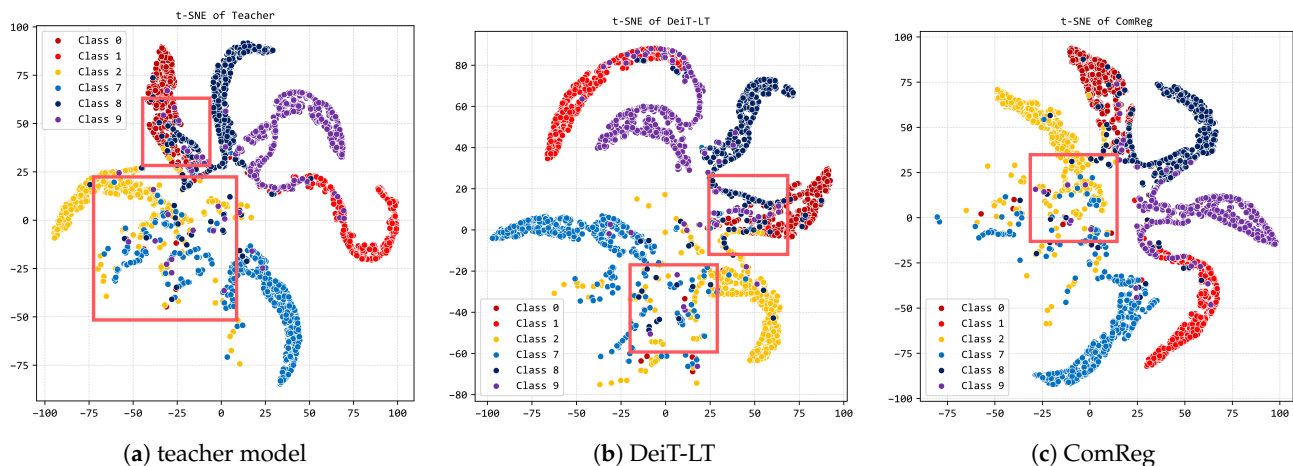


Figure 6. T-SNE visualizations of feature distributions on the CIFAR10-LT test set under ResNet-32. (a) Teacher model, (b) DeiT-LT student model, (c) ComReg student model. Only samples from selected head and tail classes (Classes 0, 1, 2, 7, 8, and 9) are visualized. Red boxes highlight regions with high inter-class overlap or ambiguous boundaries.

In addition, we provide the confusion matrices of the three models on the test set in Figure 7, to offer a more intuitive comparison in conjunction with the t-SNE visualizations in Figure 6. It can be observed that under the ComReg framework, both tail and medium-frequency classes exhibit notably improved performance, with fewer samples misclassified into head classes.

To investigate the specialization capabilities of the two expert heads in the proposed ComReg framework, we visualize their attention on minority-class samples selected from the MedMNIST v2 subsets. Specifically, we employ Smooth Grad-CAM [52] to generate activation maps for each expert head, as illustrated in Figure 8. The visualizations reveal that the tail-class expert attends to more fine-grained and localized feature regions, highlighting its specialization in recognizing minority-class patterns, in contrast to the broader focus of the head-class expert.

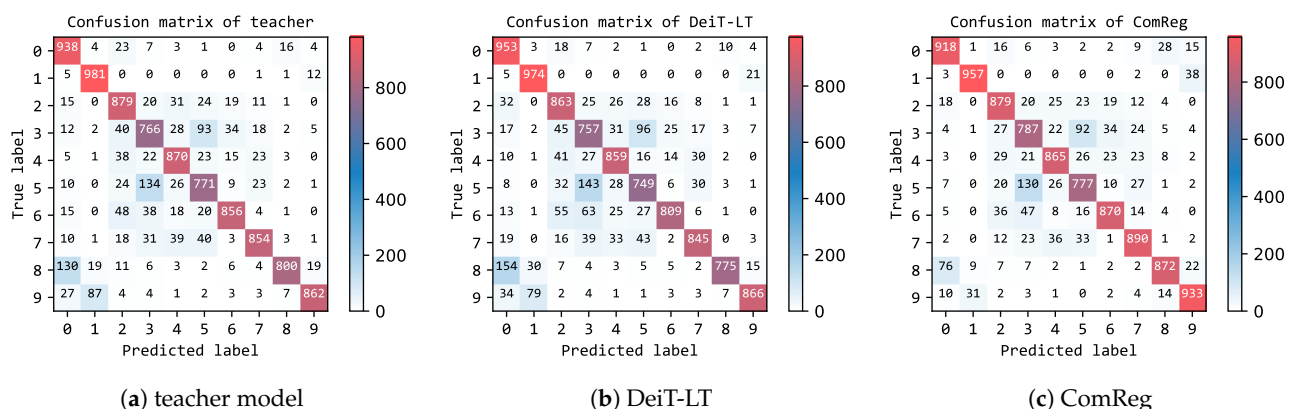


Figure 7. Confusion matrices of the CIFAR10 test set obtained with ResNet-32 as the backbone. (a) Teacher model, (b) Student model trained under the DeiT-LT framework, (c) Student model trained under the ComReg framework.

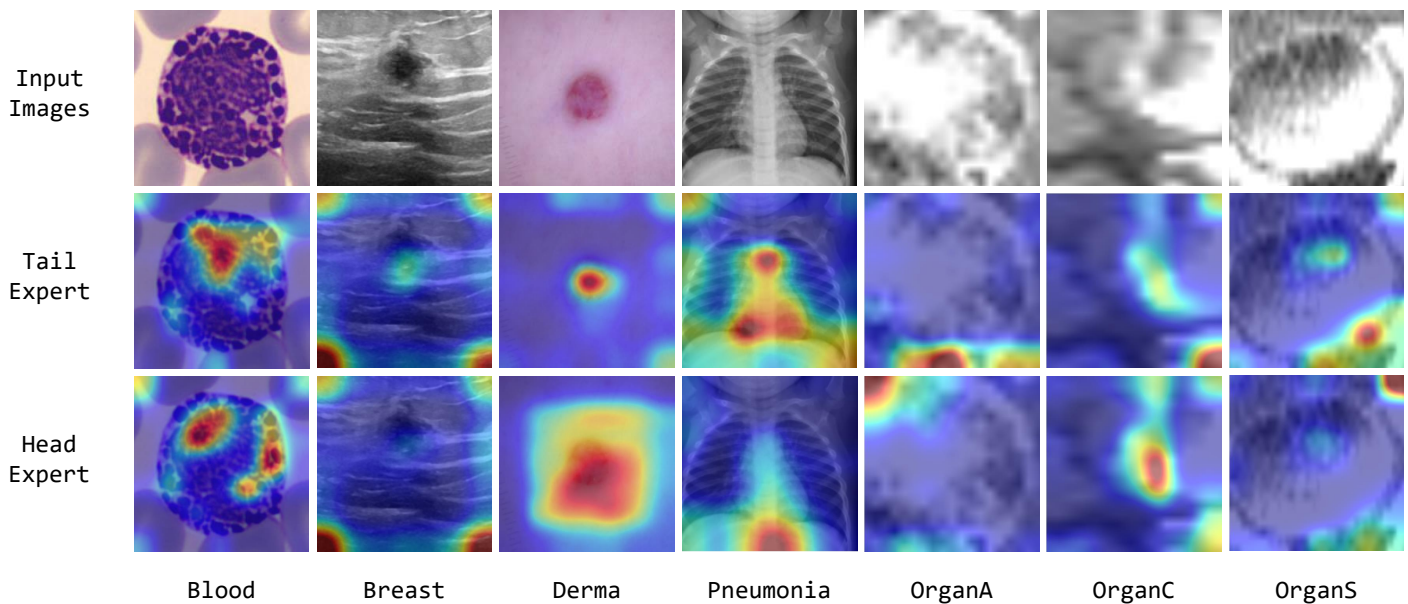


Figure 8. The Smooth Grad-CAM visualizations of the tail-class expert and head-class expert.

4.5. Discussions

To gain deeper insights into our proposed method, we conduct a series of ablation studies to examine the individual contributions of each component, assess the sensitivity to key hyperparameters, and analyze specific design choices in greater detail. These supplementary experiments not only validate the robustness of our framework but also provide additional understanding of its performance and behavior across diverse settings.

4.5.1. Ablation Studies

We conduct ablation studies on CIFAR10-LT and CIFAR100-LT with an imbalance ratio of 100, using ResNet-32 as the backbone network. The results are summarized in Table 5. Specifically, we decompose the proposed framework into four key modules: (1) the baseline classifier head (corresponding to the baseline in Table 2), (2) the distillation head, (3) the decoupled contrastive learning branch, and (4) the delayed-weighting mechanism in the contrastive branch. It is worth noting that all experimental settings here are consistent with those described in Section 4.2.1.

As shown in Table 5, the components within the ComReg framework operate in a mutually constraining and synergistic fashion. Notably, the incorporation of prior knowledge through online knowledge distillation contributes the most substantial performance improvements, enhancing the baseline by 8.3% on CIFAR100-LT and 4.5% on CIFAR10-LT. Moreover, augmenting the framework with the contrastive learning branch—prior to the application of the delayed weighting strategy—further boosts performance by 1.4% and 1.1%, respectively. These results highlight the complementary strengths of the two components, culminating in a unified regularization effect that combines the decision boundary optimization introduced by the multi-expert framework, guided by online knowledge distillation, with the feature space modeling supervision enforced through contrastive learning.

Furthermore, we compute the cosine similarity between the outputs of the two expert heads on the CIFAR10-LT dataset to characterize their structural divergence, and illustrate its evolution throughout training in Figure 9. A cosine similarity approaching -1 indicates maximal dissimilarity between the two experts. It can be clearly observed that, across all imbalance ratios of CIFAR10-LT, introducing the contrastive learning branch leads to more pronounced divergence between expert heads at the early stages of training. This suggests that the contrastive learning branch facilitates the early specialization of expert heads,

enabling them to learn distinct semantic representations more effectively. Meanwhile, it helps to stabilize the shift in similarity caused by the delayed strong weighting strategy applied to the distillation head.

Table 5. Ablation study of the ComReg framework. In the table, LA denotes Logit Adjustment Loss, CE denotes Cross-Entropy Loss. ✓ indicates the inclusion of the corresponding module, while ✗ denotes its exclusion. Bold indicates the best result.

Classification Head	Distillation Head	Contrastive Learning Branch		CIFAR100-LT	CIFAR10-LT
LA	CE	Decoupled	Delayed Weighted		
✓	✗	✗	✗	44.0	81.2
✓	✗	✓	✗	45.2	82.1
✓	✗	✓	✓	46.4	82.6
✓	✓	✗	✗	52.3	85.7
✓	✓	✓	✗	53.7	86.8
✓	✓	✓	✓	54.1	87.5

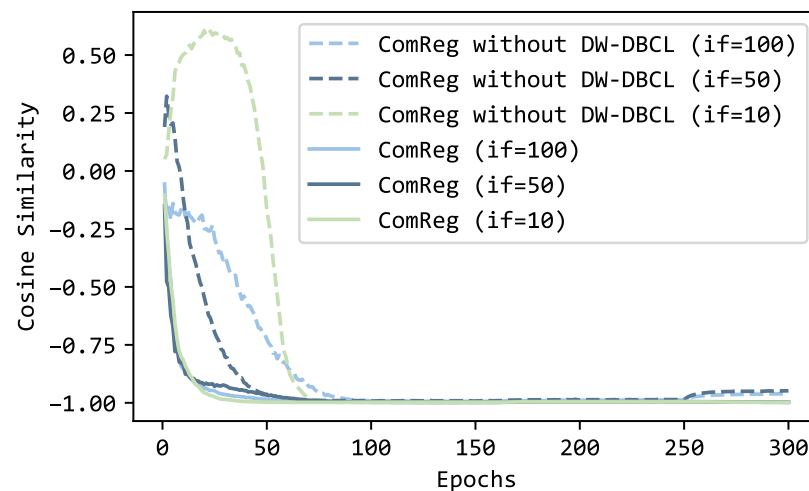


Figure 9. Cosine similarity between expert heads over training epochs. ComReg without DW-DBCL refers to ComReg without the contrastive learning branch, while ComReg denotes the full framework. The strong delayed weighting on the distillation head is applied at epoch 250, as discussed in Section 4.2.1.

4.5.2. Effect of Delayed Weighting Schedule

In Table 6, we present detailed experiments on the delayed weighting schedule. These experiments aim to validate both the effectiveness of our delayed weighting strategy and the suitability of the selected delay period. For CIFAR10-LT, which has a smaller number of classes, the model can acquire sufficient class-discriminative information at an earlier stage of training. As a result, applying class prototype weighting at epoch 160 results in improved overall performance. In contrast, CIFAR100-LT, which has a significantly larger number of classes, requires a longer training period to adequately capture the full class distribution and learn more accurate prototype representations. Therefore, applying the weighting at epoch 250 yields the best performance. Moreover, since the class prototypes are derived from the transposed weights of the tail-class expert's fully connected layer, failure to apply the weighting in a timely manner can cause the model to overfit to minority classes, which degrades overall performance.

Table 6. Effect of different delayed weighting periods on the overall model accuracy and accuracies of different class groups (head, medium, and tail classes). Accuracy (ACC, %) is used as the evaluation metric, and the best results are highlighted in bold.

Dataset	Delayed-Weighted Epoch	Overall ACC	Head ACC	Mid ACC	Tail ACC
CIFAR100-LT	0	52.81	61.11	54.69	39.71
	160	53.80	63.47	54.22	40.82
	250	54.12	63.42	55.33	40.60
	300	54.04	63.44	54.75	41.04
CIFAR10-LT	0	86.69	91.37	81.98	88.30
	160	87.47	91.80	82.48	89.80
	250	87.46	91.27	82.65	90.07
	300	86.43	91.43	82.48	90.03

4.5.3. Hyperparameter Analysis

We conduct a sensitivity analysis of the hyperparameter α in the DW-DBCL framework using the CIFAR-LT datasets. Consistent with previous experiments, we adopt the most challenging imbalance setting with $\rho = 100$ for evaluation. It is worth noting that, following common practice in long-tailed classification, we maintain a fixed experimental protocol across all methods to ensure fair comparison and statistical tractability. Within DW-DBCL, α is the primary additional hyperparameter introduced by our formulation, while other parameters, such as the temperature in contrastive learning, follow standard and widely adopted settings. Expanding the analysis to a large set of auxiliary hyperparameters would substantially increase the experimental complexity and hinder consistent evaluation; therefore, we focus our sensitivity study on α . As shown in Figure 10, α has a significant impact on the overall performance. Setting α too close to 0 or 1 leads to a noticeable degradation in performance, as also discussed in [14]. Specifically, when $\alpha = 0$, Equation (1) degenerates into clustering only within the same class; when $\alpha = 1$, the loss effectively becomes a purely self-supervised objective. Consequently, setting α around 0.5 achieves the best balance between instance-level consistency and class-level discrimination, yielding the strongest generalization performance.

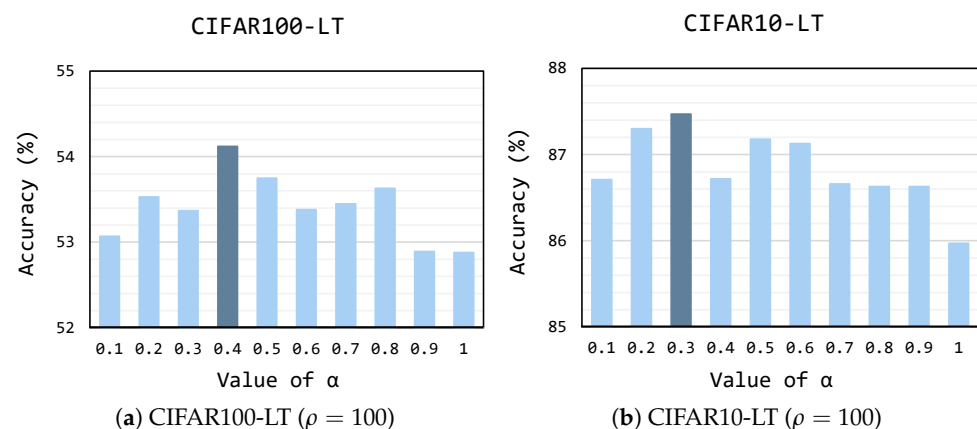


Figure 10. Effect of the hyperparameter α on CIFAR-LT with an imbalance ratio of 100. The y-axis represents accuracy (%), while the x-axis indicates the value of α .

4.5.4. Computational Cost and Efficiency Analysis

Table 7 reports a detailed comparison of model parameters as well as training and inference time for DeiT-LT and ComReg under both ResNet-32 and ViT backbones on CIFAR100-LT and CIFAR10-LT. The results provide a more concrete assessment of the computational overhead introduced by ComReg.

Table 7. Comparison of model parameter count, training time and inference time between DeiT-LT and ComReg under different backbones (ResNet-32 and ViT) on CIFAR100-LT and CIFAR10-LT ($\rho = 100$) datasets.

Framework	Model Parameter	CIFAR100-LT		CIFAR10-LT	
		Training	Inference	Training	Inference
ResNet-32 backbone					
DeiT-LT	0.60 M	1 h 20 min 31 s	9 s	20 min 55 s	3 s
ComReg	1.26 M	1 h 35 min 32 s	10 s	23 min 23 s	3 s
ViT backbone					
DeiT-LT	85.82 M	8 h 17 min 13 s	8s	8 h 32 min 11 s	3 s
ComReg	88.18 M	18 h 20 min 53 s	9s	22 h 55 min 41 s	3 s

Under the ResNet-32 backbone, the training-time difference between DeiT-LT and ComReg is relatively small across both datasets. Specifically, ComReg incurs a modest increase in training time compared to DeiT-LT, which can be attributed to the introduction of the additional contrastive learning branch during training. Given the simplicity and efficiency of the ResNet-32 architecture, this overhead remains limited, amounting to only a minor increase relative to the overall training duration. Importantly, the inference time for ComReg remains almost identical to that of DeiT-LT, indicating that the proposed framework does not introduce additional deployment-time cost under lightweight backbones. In contrast, a larger training-time gap is observed under the ViT backbone. This difference, however, is not caused by an inherent inefficiency of the ComReg framework itself, but rather by practical training constraints. Specifically, DeiT-LT is trained with a batch size of 256, whereas ComReg requires a smaller batch size of 64 to accommodate the memory demands of the contrastive learning branch. The reduced batch size leads to a higher number of iterations per epoch, which in turn results in longer overall training time. Despite this increase, the inference-time cost remains comparable between the two frameworks, with only negligible differences observed across datasets.

Overall, these results indicate that the additional computational cost introduced by ComReg is primarily confined to the training phase and is influenced by implementation-specific factors such as batch size selection, rather than fundamental architectural inefficiency. At inference time, ComReg relies solely on a single student model and exhibits computational efficiency comparable to standard single-backbone classifiers. This characteristic makes ComReg suitable for practical scenarios where training is performed offline on dedicated hardware, while inference is required to be efficient and resource-constrained, as is common in real-world and medical imaging applications.

5. Conclusions

In this work, we address long-tailed image classification by proposing ComReg, a unified framework that coordinates multi-expert learning and supervised contrastive learning to jointly regulate class-wise prediction behavior and feature space regularization. By incorporating prior-knowledge-guided online distillation, a triple-decoupled balanced contrastive learning strategy, and delayed prototype weighting, ComReg demonstrates consistent performance improvements across both majority and minority classes under standard benchmark settings. From a practical perspective, the additional computational cost introduced by ComReg is confined to the training phase, while inference relies solely on a compact student model without extra overhead, making the framework suitable for offline training and efficient deployment. Nevertheless, we acknowledge that the current evaluation does not include statistical significance analysis or sensitivity studies, and that

the multi-expert design increases training complexity. Improving computational efficiency, analyzing cost–performance trade-offs, and conducting more comprehensive statistical evaluations remain important directions for future work.

Author Contributions: Conceptualization, X.S.; methodology, X.S.; software, X.S.; validation, X.S. and L.W.; formal analysis, X.S.; investigation, L.Z. and Y.W.; resources, X.S. and Y.W.; data curation, X.S.; writing—original draft preparation, X.S.; writing—review and editing, L.Z. and L.W.; visualization, X.S. and L.W.; supervision, L.Z. and L.W.; project administration, L.Z.; funding acquisition, L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 62376174, Grant U24A20341, and Grant 62025601.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in this study are openly available from CIFAR at <https://www.cs.toronto.edu/~kriz/cifar.html> (accessed on 8 October 2024), MedMNIST at <https://zenodo.org/records/10519652> (accessed on 8 October 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cunningham, P.; Cord, M.; Delany, S.J. Supervised learning. In *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 21–49.
2. Hastie, T.; Tibshirani, R.; Friedman, J.; Hastie, T.; Tibshirani, R.; Friedman, J. Overview of supervised learning. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009; pp. 9–41.
3. Nasteski, V. An overview of the supervised machine learning methods. *Horizons. B* **2017**, *4*, 56. [[CrossRef](#)]
4. Castelli, M.; Vanneschi, L.; Largo, A.R. Supervised learning: Classification. In *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*; Ranganathan, S., Nakai, K., Schonbach, C., Eds.; Elsevier: Amsterdam, The Netherlands, 2018; Volume 1, pp. 342–349.
5. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009*; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
6. Dong, Q.; Gong, S.; Zhu, X. Class rectification hard mining for imbalanced deep learning. In *Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017*; pp. 1851–1860.
7. Park, S.; Hong, Y.; Heo, B.; Yun, S.; Choi, J.Y. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–23 June 2022*; pp. 6887–6896.
8. Shi, J.X.; Wei, T.; Xiang, Y.; Li, Y.F. How re-sampling helps for long-tail learning? *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 75669–75687.
9. Menon, A.K.; Jayasumana, S.; Rawat, A.S.; Jain, H.; Veit, A.; Kumar, S. Long-tail learning via logit adjustment. *arXiv* **2020**, arXiv:2007.07314.
10. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
11. Cui, J.; Zhong, Z.; Liu, S.; Yu, B.; Jia, J. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021*; pp. 715–724.
12. Zhu, J.; Wang, Z.; Chen, J.; Chen, Y.P.P.; Jiang, Y.G. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–23 June 2022*; pp. 6908–6917.
13. Du, C.; Wang, Y.; Song, S.; Huang, G. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5890–5904. [[CrossRef](#)] [[PubMed](#)]
14. Xuan, S.; Zhang, S. Decoupled Contrastive Learning for Long-Tailed Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024*; Volume 38, pp. 6396–6403.
15. Zeng, L.; Feng, Z.; Chen, J.; Wang, S. Long-tailed visual classification based on supervised contrastive learning with multi-view fusion. *Knowl.-Based Syst.* **2024**, *301*, 112301. [[CrossRef](#)]

16. Li, J.; Tan, Z.; Wan, J.; Lei, Z.; Guo, G. Nested collaborative learning for long-tailed visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–23 June 2022; pp. 6949–6958.
17. Tan, Z.; Li, J.; Du, J.; Wan, J.; Lei, Z.; Guo, G. NCL++: Nested Collaborative Learning for long-tailed visual recognition. *Pattern Recognit.* **2024**, *147*, 110064. [\[CrossRef\]](#)
18. Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; Yu, S.X. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv* **2020**, arXiv:2010.01809.
19. Rangwani, H.; Mondal, P.; Mishra, M.; Asokan, A.R.; Babu, R.V. DeiT-LT: Distillation Strikes Back for Vision Transformer Training on Long-Tailed Datasets. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 19–21 June 2024; pp. 23396–23406.
20. Chen, J.; Er, M.J. Mitigating gradient conflicts via expert squads in multi-task learning. *Neurocomputing* **2025**, *614*, 128832. [\[CrossRef\]](#)
21. Xie, X.; Liu, H.; Zeng, S.; Lin, L.; Li, W. A novel progressively undersampling method based on the density peaks sequence for imbalanced data. *Knowl.-Based Syst.* **2021**, *213*, 106689. [\[CrossRef\]](#)
22. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [\[CrossRef\]](#)
23. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 1–8 June 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1322–1328.
24. Zhang, H. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
25. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6023–6032.
26. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [\[CrossRef\]](#)
27. Zhang, Y.; Liu, Y.; Wang, Y.; Yang, J. An ensemble oversampling method for imbalanced classification with prior knowledge via generative adversarial network. *Chemom. Intell. Lab. Syst.* **2023**, *235*, 104775. [\[CrossRef\]](#)
28. Lin, T. Focal Loss for Dense Object Detection. *arXiv* **2017**, arXiv:1708.02002.
29. Cui, Y.; Jia, M.; Lin, T.Y.; Song, Y.; Belongie, S. Class-balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9268–9277.
30. Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
31. Zhong, K.; Zhang, L.; Wang, L.; Shu, X.; Wang, Z. Class-Center-Based Self-Knowledge Distillation: A Simple Method to Reduce Intra-Class Variance. *Appl. Sci.* **2024**, *14*, 7022. [\[CrossRef\]](#)
32. Ren, N.; Li, X.; Wu, Y.; Fu, Y. Balanced self-distillation for long-tailed recognition. *Knowl.-Based Syst.* **2024**, *290*, 111504. [\[CrossRef\]](#)
33. Xiang, L.; Ding, G.; Han, J. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part V 16; Springer: Cham, Switzerland, 2020; pp. 247–263.
34. Iscen, A.; Araujo, A.; Gong, B.; Schmid, C. Class-balanced distillation for long-tailed visual recognition. *arXiv* **2021**, arXiv:2104.05279.
35. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 10347–10357.
36. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 9729–9738.
37. An, X.; Zhu, X.; Gao, Y.; Xiao, Y.; Zhao, Y.; Feng, Z.; Wu, L.; Qin, B.; Zhang, M.; Zhang, D.; et al. Partial fc: Training 10 million identities on a single machine. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 1445–1449.
38. Yang, J.; Shi, R.; Wei, D.; Liu, Z.; Zhao, L.; Ke, B.; Pfister, H.; Ni, B. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Sci. Data* **2023**, *10*, 41. [\[CrossRef\]](#)
39. Cubuk, E.D.; Zoph, B.; Mane, D.; Vasudevan, V.; Le, Q.V. Autoaugment: Learning augmentation strategies from data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 113–123.
40. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 12–18 July 2020; pp. 1597–1607.
41. Foret, P.; Kleiner, A.; Mobahi, H.; Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *arXiv* **2020**, arXiv:2010.01412.

42. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Virtual, 14–19 June 2020; pp. 702–703.
43. Chou, H.P.; Chang, S.C.; Pan, J.Y.; Wei, W.; Juan, D.C. Remix: Rebalanced mixup. In *Proceedings of the Computer Vision–ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020*; Proceedings, Part VI 16; Springer: Cham, Switzerland, 2020; pp. 95–110.
44. Pan, H.; Guo, Y.; Yu, M.; Chen, J. Enhanced Long-Tailed Recognition With Contrastive CutMix Augmentation. *IEEE Trans. Image Process.* **2024**, *33*, 4215–4230. [[CrossRef](#)]
45. Wang, P.; Han, K.; Wei, X.S.; Zhang, L.; Wang, L. Contrastive learning based hybrid networks for long-tailed image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 943–952.
46. Bisong, E. Google automl: Cloud vision. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*; Apress: Berkeley, CA, USA, 2019; pp. 581–598.
47. Liu, J.; Li, Y.; Cao, G.; Liu, Y.; Cao, W. Feature pyramid vision transformer for MedMNIST classification decathlon. In *Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–8.
48. Manzari, O.N.; Ahmadabadi, H.; Kashiani, H.; Shokouhi, S.B.; Ayatollahi, A. MedViT: A robust vision transformer for generalized medical image classification. *Comput. Biol. Med.* **2023**, *157*, 106791. [[CrossRef](#)]
49. Yue, Y.; Li, Z. Medmamba: Vision mamba for medical image classification. *arXiv* **2024**, arXiv:2403.03849. [[CrossRef](#)]
50. Zheng, Z.; Jia, X. Complex mixer for medmnist classification decathlon. *arXiv* **2023**, arXiv:2304.10054. [[CrossRef](#)]
51. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
52. Omeiza, D.; Speakman, S.; Cintas, C.; Weldermariam, K. Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. *arXiv* **2019**, arXiv:1908.01224.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.