Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 3, 2026

VA-GKD: Variance-Aware Adaptive Group Knowledge Distillation for Long-Tailed Recognition

Din Muhammad[1], Hina Ali[2], Farasat Ali Azeemi[3], Maria Khalid[4], Nazia Hussain[2] ,Zahid Hussain[1], Pireh Hussain[1],Saif Ali[1], Misbah Sangrasi[1], Muhammad Hammad [1]

[1] *Mehran university of Engineering and Technology, Pakistan*
[2]*Lecturer at Islamabad Model College, Pakistan*
[3] *Bahria University, Karachi, Pakistan*
[4]*Visiting faculty at MIT*

[1]din.muhammad@faculty.muet.edu.pk, [2]hinaabdullah04@gmil.com, [3]farasataliazeemi1@gmail.com, [4]maria.iqbal.shaikh@gmail.com, [5]naziahussain1989@gmail.com, [1]zahidhussainkhaskheli@gmail.com, [1]pirehsoomro3@gmail.com, [1]saifsangrasi1@gmail.com, [1]misbahsangrasi@gmail.com, [1]sangrasihammad4@gmail.com

DOI: https://doi.org/10.5281/zenodo.21188719

Keywords
Long-tail distribution, Imbalanced data, Image classification, Adaptive variance, deep learning, Features, Synthetic Data, SABLC, Knowledge Distillation

Received: 7 January 2026
Accepted: 20 February 2026
Published: 9 March 2026

Din Muhammad *

Abstract

Knowledge distillation (KD) in long-tailed recognition suffers from a systematic bias: the teacher model's output distribution is heavily skewed towards the dominant head classes, causing the student model to inherit and amplify head-class favoritism. Long-Tailed Knowledge Distillation (LTKD) addresses this by partitioning classes into head, medium, and tail groups and rebalancing group-level probabilities via a batch-mean scalar correction. We identify two limitations of this approach: (i) the batch-mean correction is a first-moment fix that ignores intra-batch variance, leading to inconsistent per-sample corrections; and (ii) the intra-group replacement weight β is a static constant, blind to the temporal evolution of the teacher's bias during training. The proposed method Variance-Aware Group Knowledge Distillation (VA-GKD), which augments LTKD with (a) a sample-adaptive, variance-normalized cross-group scaling factor derived from the per-batch distribution of group probabilities, and (b) a cosine-scheduled curriculum on the intra-group replacement weight $\beta(t)$ that transitions from a teacher-proportional initialization to a uniform emphasis encouraging tail-class learning. Experiments on a controlled synthetic long-tailed dataset yield a statistically significant improvement in tail-class accuracy over the baseline method.

1. Introduction

Real-world data rarely follows a balanced class distribution: a small number of head classes accumulate the vast majority of training samples, while many tail classes are represented by only a handful of samples. This long-tailed distribution [1] leads classifiers to achieve high aggregate accuracy at the cost of severely under-performing on minority classes, the very classes that are often most important in safety-critical applications such as
medical imaging, rare-event detection, and ecological monitoring. Knowledge distillation (KD) [2] transfers the knowledge encoded in a teacher's soft output distribution to a student model. A student trained to mimic

these distributions inherits and, in many cases, amplifies the teacher's imbalance because the student's shallower capacity concentrates gradient signal on the dominant classes. The recently proposed LTKD framework [3] tackles this by decomposing the class set $\mathcal{C}$ into three frequency groups $\{\mathcal{H}, \mathcal{M}, \mathcal{T}\}$ (head, medium, tail) and replacing the cross-group component of the teacher's soft label with a rebalanced version. Concretely, the teacher's probability mass assigned to group $\mathcal{G}$ is rescaled by a batch-estimated correction factor $s_{\mathcal{G}}$, then a within-group weight β uniformly redistributes probability within each group. Despite its effectiveness, LTKD's rebalancing rests on two simplifying assumptions that this paper challenges:

(i) **First-moment-only correction.** The cross-group scaling factor $s_{\mathcal{G}}$ is computed from the batch mean of the teacher's group probability $p^T_{\mathcal{G}}$. Samples whose $p^T_{\mathcal{G}}$ deviates far from the batch mean receive the same correction as typical samples, creating inconsistent effective supervision.

(ii) **Static within-group weight.** The replacement weight β is a fixed hyperparameter throughout training. Yet the teacher's bias is not static: early in training the teacher's representations are still forming, and imposing aggressive tail emphasis can destabilize gradients; late in training the student has largely mastered head classes and needs stronger tail emphasis for marginal gains.

Contributions.

1. We propose **variance-normalized cross-group scaling**, a sample-adaptive correction that uses the batch mean and standard deviation of group probabilities.
2. We derive a gradient-variance reduction bound (Theorem 1) showing that the variance-normalized correction reduces the expected squared deviation of per-sample gradient from the ideal balanced gradient.
3. We propose a cosine curriculum for $\beta(t)$ that initializes at the teacher's own group probability and anneals to unity.
4. We provide controlled empirical validation on a synthetic long-tailed benchmark, demonstrating statistically significant tail accuracy gains with comparable overall performance.

2. Related Work

The problem of long-tailed image classification has attracted considerable attention from the computer vision community in recent years using traditional machine learning[4] and deep learning methods. The fundamental challenge arises from the highly skewed distribution of real-world datasets, where a small number of dominant (head) classes contain the majority of training samples, while a large number of minority (tail) classes are severely underrepresented [1]. This imbalance causes deep learning models to develop biased classifiers that perform well on head classes but degrade significantly on tail classes [5]. To address these challenges, existing methods can be broadly organized into the following categories: data re-balancing, clustering [6], information augmentation, synthetic data generation, metric learning, and decoupled learning.

2.1 Data Re-balancing

Data re-balancing methods aim to artificially correct the long-tailed distribution either by modifying the training data distribution or by adjusting the training loss. These methods are generally classified into two categories: over-sampling and under-sampling [7].

Over-sampling techniques increase the number of tail-class samples to create a more balanced dataset. The most widely known approach is the Synthetic Minority Over-sampling Technique (SMOTE) [8], which uses interpolation to randomly generate new minority-class samples from the nearest neighborhood of existing ones. Variants of SMOTE have been proposed to address its limitations. Han *et al.* [9] introduced Borderline-SMOTE1 and

Borderline-SMOTE2, which focus oversampling near the decision boundaries of the minority classes. Dablain *et al.* [10] proposed DeepSMOTE, which fuses deep learning with SMOTE to handle raw images while preserving their visual characteristics and producing higher-quality synthetic images. To address SMOTE's tendency to overgeneralize by ignoring the distribution of neighboring head-class samples, Maciejewski and Stefanowski [11] proposed the Local Neighbourhood extension of SMOTE (LN-SMOTE), which exploits local neighborhood information of individual samples. Bhagat and Patil [12] introduced a two-step methodology that decomposes the multi-class imbalanced problem into binary class subsets via binarization techniques such as One-vs-One (OVO) and One-vs-All (OVA), applying SMOTE to each binary subset before employing a Random Forest classifier. Chawla *et al.* [8] combined oversampling of tail classes with undersampling of head classes to achieve better classifier performance in long-tail distribution settings.

Under-sampling approaches balance the data by reducing head-class samples. However, a key drawback is that they may result in the loss of valuable information, especially in the presence of severe data imbalance [7]. Dal Pozzolo *et al.* [13] addressed this by calibrating probability estimates with undersampling.Drummond *et al.* [14] studied why under-sampling often outperforms over-sampling for class-imbalanced problems, attributing this to cost-sensitivity properties. Barua *et al.* [15] proposed MWMOTE, a majority-weighted minority oversampling technique for imbalanced data set learning.

2.2 Information Augmentation

Information augmentation methods supplement the model training with additional information to enhance model performance [5]. These methods are broadly divided into data augmentation and transfer learning.

2.2.1 Data Augmentation

Data augmentation techniques apply various transformations to the training data to create new, synthetic samples that retain the same label as the original samples while introducing variations. The primary goal is to diversify the dataset to improve the model's ability to generalize. Common techniques include random scaling, image flipping, cropping, and rotating. Perez and Wang [16] demonstrated the effectiveness of standard data augmentation for deep-learning-based image classification. Long *et al.* [17] proposed a retrieval-augmented classification pipeline to augment the long-tail image classification problem. Mullick *et al.* [18] introduced generative adversarial minority oversampling, which generates new tail-class samples via a convex combination of existing samples. Noise injection has also been explored as a means to optimize decision boundaries by moving them away from minority classes [19].

2.2.2 Transfer Learning

Transfer learning techniques transfer knowledge from a source domain (e.g., head classes or external datasets) to improve training in a target domain [20, 21]. In the context of long-tailed learning, these methods are grouped into head-to-tail knowledge transfer, model pre-training, knowledge distillation, and self-training. Yin *et al.* [22] addressed the intra-class variance of features in tail-class samples by exploiting the knowledge of head-class intra-class variance, so that features of tail-class samples have enough intra-class variance, resulting in better model performance. Kim *et al.* [23] proposed perturbation-based optimization techniques that enhance the representation of minority-class samples by transforming samples from the majority classes into ones that resemble the minority classes. Cui *et al.* [24] initially trained the model on all long-tailed samples to develop a foundation for representation, then fine-tuned the model to transfer features from head-class samples to tail-class samples. Chu *et al.* [25] explored feature space augmentation for long-tailed data, supplementing tail classes with augmented features derived from head classes.

2.3 Synthetic Data Generation

Synthetic data generation methods artificially create new data samples that exhibit similar properties to real-world samples. This approach is particularly valuable when data collection is expensive or the available data is severely imbalanced.

2.3.1 Generative Adversarial Networks (GANs)

GANs have been widely explored as a solution for addressing long-tail or imbalanced classification problems [26]. Conditional GANs (cGANs) condition both the generator and the discriminator on the class label to generate class-specific images [26]. Auxiliary Classifier GANs (AC-GANs) [27] extend this by incorporating class information directly into the GAN framework. Rangwani *et al.* [28] addressed the challenges of training GANs on long-tailed distributions by proposing a novel group Spectral Regularizer (gSR) to enhance diversity and realism across classes with varied sample sizes. Shmelkov *et al.* [29] introduced two quantitative measures, GAN-train and GAN-test, to evaluate GANs by approximating diversity (recall) and image quality (precision). Adler and Lunz [30] proposed Wasserstein Generative Adversarial Networks (WGANs) that use the Wasserstein metric distance between probability distributions to generate realistic samples from complex image distributions. The Polarity GAN (PGAN) [31] architecture adds an extra adversarial network to force the generator to create samples near decision boundaries, generating more challenging examples for minority classes. Majeed and Hwang [32] introduced the Conditional GAN Minority-Class-Augmented Oversampling Scheme (CTGAN-MOS) to address class imbalance challenges. Kozerawski *et al.* [33] generated extra training samples for tail classes using a gradient-ascent-based image generation algorithm to improve classifier generalization on long-tailed datasets. Zhao *et al.* [34] introduced a generative and fine-tuning framework, Long-tail Recognition via Leveraging LLMs-driven Generated Content (LTGC), to tackle long-tail recognition by leveraging content generated by large language models. Sharma *et al.* [35] combined GANs with SMOTE through the Smotified-GAN framework to address class-imbalanced pattern classification problems.

2.3.2 Diffusion Models

Diffusion models have recently emerged as a promising alternative for generating synthetic images to address long-tail classification challenges. These models generate images by gradually denoising a random noise distribution and transforming it into high-quality, diverse images. Shao *et al.* [36] proposed DifuLT, which employs a randomly-initialized diffusion model trained solely on the long-tailed dataset to generate new samples for underrepresented classes, while filtering out harmful samples using inherent information from the original dataset. Han *et al.* [37] proposed a Latent-based Diffusion Model for Long-tailed Recognition (LDMLR) that first encodes the imbalanced dataset into features using a baseline model, then trains a Denoising Diffusion Implicit Model (DDIM) to generate pseudo-features, and finally trains the classifier using both encoded and pseudo-features. Zhang *et al.* [38] proposed a weighted denoising score-matching technique for direct head-to-tail knowledge transfer within a Bayesian framework, incorporating a gating mechanism that utilizes label distribution and sample similarity for effective rebalancing.

2.4 Metric Learning

Metric learning methods design distance-based loss functions to determine the similarity and dissimilarity among samples, thereby learning a more discriminative feature space. In the context of long-tailed classification, Zhang *et al.* [39] proposed the Range Loss, which enhances representation learning by considering the distances between all pairs of samples within a mini-batch. It aims to increase the distances between the centers of any two classes while simultaneously minimizing the larger distances between samples within the same class, thereby reducing intra-class variations. Kang *et al.* [40] proposed a variation of contrastive loss called

k-positive contrastive loss to improve the generalization of the model and address severe class imbalance. Wang *et al.* [41] proposed a contrastive learning-based hybrid network for long-tailed image classification. Lin *et al.* [42] introduced the Focal Loss, a reweighting strategy that assigns higher weights to tail classes and lower weights to head classes to address the extreme class imbalance encountered during training. Cao *et al.* [43] proposed Label-Distribution-Aware Margin (LDAM) loss, which learns imbalanced datasets with label-distribution-aware margin loss. Cui *et al.* [44] introduced class-balanced loss based on the effective number of samples, providing a principled way to re-weight the loss function according to class frequency.

2.5 Decoupled Learning

Decoupled learning is one of the most promising recent approaches to address long-tail classification challenges. In these methods, the deep long-tailed model is trained in two stages: representation learning followed by classifier learning [45]. Kang *et al.* [45] proposed a model that addresses the long-tail problem using two separate stages for representation and classification learning, using the cross-entropy loss function in both stages. Their findings suggest that random data sampling is beneficial for feature learning, whereas class-balanced sampling proves more effective when training the classifier. Zhang *et al.* [46] proposed a simple yet effective approach for learning with long-tailed distributions, showing that a well-designed decoupled strategy can significantly improve tail-class performance. Zhou *et al.* [47] introduced the Bilateral-Branch Network (BBN) with cumulative learning for long-tailed visual recognition, which combines a conventional learning branch and a re-balancing branch that are jointly trained with a cumulative learning strategy.

2.6 Supervised Contrastive Learning for Long-Tailed Data

Supervised Contrastive Learning (SCL) [48] is an extended version of unsupervised contrastive learning [49] that falls into the category of two-stage methods. SCL effectively improves class separation by pulling together samples from the same class (positive pairs) while pushing away samples from different classes (negative pairs) in the feature space. The primary difference from unsupervised contrastive loss is that it incorporates all images belonging to the same class as positive samples, rather than only the augmented version of the anchor image. SCL has been shown to produce robust feature representations that are critical for effectively distinguishing between classes in challenging datasets. However, in a highly imbalanced dataset, SCL can remain biased toward the head classes since it does not directly address the effects of class imbalance during learning. The feature space tends to be dominated by head classes, which have more data points and thus more influence on defining the decision boundaries [43].

To address the limitations of SCL in long-tailed settings, Liu *et al.* [50] proposed the Transfer of Angular Information (TAI), which focuses on enhancing the intra-class variance of tail-class samples by exploiting the angular distribution of head-class features. TAI addresses the challenge posed by long-tailed image classification by transferring the average head-class angular variance to tail classes to enrich the feature space. Specifically, head classes often have a wide range of diversity, which is apparent in the angles formed between the features of head-class samples and their corresponding class centers. Tail classes, by contrast, need more intra-class diversity due to the limited number of samples and consequently occupy a smaller spatial span [50]. However, while TAI addresses the intra-class compactness challenge, it fails to simultaneously achieve strong inter-class separability.

In [51], the authors addressed the limitations of imbalanced dataset by incorporating the angle variance with (SCL) [48]. Huang *et al.* [52] proposed a framework for learning deep representations for imbalanced classification by designing an end-to-end learning scheme. Wang *et al.* [53] proposed to learn to model

the tail distribution in the neural information processing framework. Ren *et al.* [54] introduced a meta-learning approach that learns to reweight training examples for robust deep learning under label noise and class imbalance. In [55] the authors proposed Supervised Angular Contrastive Loss with Balancing Classifier (SACLBC) to address the limitations by integrating the strengths of SCL and angular variance within a unified curriculum-based framework. Byrd and Lipton [56] analyzed the effect of importance weighting in deep learning, questioning whether reweighting strategies consistently yield benefits for highly imbalanced settings. He and Garcia [57] provided a comprehensive survey on learning from imbalanced data, categorizing and evaluating various approaches including sampling, cost-sensitive learning, and algorithmic-level methods.

2.7 Knowledge Distillation (KD)

Hinton *et al.* [2] introduced KD as minimizing the KL divergence between the student's and a temperature-scaled teacher's softmax outputs. DKD [58] decouples the KD loss into target-class (TCKD) and non-target-class (NCKD) components. LTKD [3] extends this group-level analysis to the long-tailed setting.

2.8 Curriculum Learning

Curriculum learning [59] advocates presenting easier samples first, progressively increasing difficulty. In the context of long-tailed learning, curriculum strategies have been applied to loss weighting [47] and data augmentation scheduling [60].

3. Method

3.1 Notation

Let $\mathcal{C} = \{1, \ldots, C\}$ be the full class set. Following LTKD[3], we partition $\mathcal{C}$ into three frequency groups: $\mathcal{H}$ (head), $\mathcal{M}$ (medium), $\mathcal{T}$ (tail), satisfying $\mathcal{H} \cup \mathcal{M} \cup \mathcal{T} = \mathcal{C}$.

For an input $\mathbf{x}$, let $\mathbf{z}^T \in \mathbb{R}^C$ and $\mathbf{z}^S \in \mathbb{R}^C$ denote the teacher's and student's logits, and let $\mathbf{p}^T = \sigma(\mathbf{z}^T/\tau)$ and $\mathbf{p}^S = \sigma(\mathbf{z}^S/\tau)$ be the corresponding temperature-τ softmax outputs. For group $\mathcal{G} \in \{\mathcal{H}, \mathcal{M}, \mathcal{T}\}$, define the group probability as:

$$p_{\mathcal{G}}^T = \sum_{c \in \mathcal{G}} p_c^T. \tag{1}$$

LTKD constructs a rebalanced teacher soft label $\hat{\mathbf{p}}^T$ as follows. A batch-level scaling factor is computed for each group:

$$s_{\mathcal{G}} = \frac{1/|\mathcal{G}|}{\bar{p}_{\mathcal{G}}^T}, \quad \bar{p}_{\mathcal{G}}^T = \frac{1}{B} \sum_{i=1}^{B} p_{\mathcal{G},i}^T, \tag{2}$$

where B is the batch size. Then the rebalanced probability of class $c \in \mathcal{G}$ is:

$$\hat{p}_c^T = s_{\mathcal{G}} \cdot p_c^T \cdot (1 - \beta) + \frac{\beta}{|\mathcal{G}|}, \tag{3}$$

with $\beta \in [0, 1]$ a fixed hyperparameter. The distillation loss is $\mathcal{L}_{\text{LTKD}} = \text{KL}(\hat{\mathbf{p}}^T \,\|\, \mathbf{p}^S)$.

3.2 Variance-Normalized Cross-Group Scaling

The key insight is that $s_{\mathcal{G}}$ in Eq. (2) is a *scalar* applied uniformly to all samples in the batch. We replace it with a sample-adaptive correction $s_{\mathcal{G},i}$ defined through a soft z-score normalization:

$$s_{\mathcal{G},i} = \frac{\lambda_{\max}}{p_{\mathcal{G},i}^T + \gamma \cdot \sigma_{\mathcal{G}}^T + \epsilon}, \tag{4}$$

where

$$\sigma_{\mathcal{G}}^T = \sqrt{\frac{1}{B} \sum_{i=1}^{B} (p_{\mathcal{G},i}^T - \bar{p}_{\mathcal{G}}^T)^2} \tag{5}$$

is the batch standard deviation of group probability, $\gamma \geq 0$ is a variance damping coefficient, $\lambda_{\max}$ is a scale normalization constant, and $\epsilon > 0$ prevents division by zero.

Equation (4) has the following properties:

- When $p_{\mathcal{G},i}^T$ is large, $s_{\mathcal{G},i}$ is small i.e., mild correction.
- When $p_{\mathcal{G},i}^T$ is small, $s_{\mathcal{G},i}$ is large i.e., strong uplift.
- The denominator floor $\gamma\sigma_{\mathcal{G}}^T$ prevents explosive scaling when $p_{\mathcal{G},i}^T \approx 0$.

3.3 Cosine-Curriculum Within-Group Weight

Let $t \in [0,1]$ denote normalized training progress. We replace the static β with a time-varying schedule:

$$\beta(t) \;=\; \beta_{\min} + (\beta_{\max} - \beta_{\min}) \cdot \frac{1}{2}(1 - \cos(\pi t)), \quad (6)$$

where $0 \le \beta_{\min} < \beta_{\max} \le 1$.

The cosine schedule rises monotonically from $\beta(0) = \beta_{\min}$ (near-zero within-group smoothing early, preserving the teacher's relative ordering for stable gradients) to $\beta(1) = \beta_{\max}$ (stronger uniform redistribution late in training, benefiting tail classes).

3.4 Full VA-GKD Objective

Combining variance-normalized scaling (Eq. 4) and curriculum scheduling (Eq. 6), the VA-GKD rebalanced teacher soft label for sample i at training step t is:

$$\hat{p}^T_{c,i}(t) \;=\; s_{\mathcal{G},i} \cdot p^T_{c,i} \cdot (1 - \beta(t)) + \frac{\beta(t)}{|\mathcal{G}|}, \quad c \in \mathcal{G}. \quad (7)$$

The label is re-normalized to sum to one, and the VA-GKD loss is:

$$\mathcal{L}_{\mathrm{VA}} \;=\; \frac{1}{B} \sum_{i=1}^{B} \mathrm{KL}\big(\hat{\mathbf{p}}^T_i(t) \,\|\, \mathbf{p}^S_i\big). \quad (8)$$

The total training loss is $\mathcal{L} = (1-\alpha)\,\mathcal{L}_{\mathrm{CE}} + \alpha\,\mathcal{L}_{\mathrm{VA}}$, where $\mathcal{L}_{\mathrm{CE}}$ is the cross-entropy loss with the true label.

3.5 Theoretical Analysis

Definition 1 (Ideal balanced gradient)**.** *Let $\mathbf{g}^*_{c,i}$ denote the gradient of the KL loss $\mathrm{KL}(\mathbf{p}^{T,\mathrm{bal}} \| \mathbf{p}^S_i)$ with respect to $z^S_{c,i}$, where $\mathbf{p}^{T,\mathrm{bal}}$ is the uniform-group teacher distribution. We call $\mathbf{g}^*$ the ideal balanced gradient.*

Theorem 1 (Gradient variance reduction)**.** *Let $g^{\mathrm{LTKD}}_{c,i}$ and $g^{\mathrm{VA}}_{c,i}$ denote the per-sample gradients of $\mathcal{L}_{\mathrm{LTKD}}$ and $\mathcal{L}_{\mathrm{VA}}$, respectively, for class $c \in \mathcal{T}$. Under the assumptions that (A) $p^T_{\mathcal{T},i}$ are i.i.d. draws from a distribution with mean $\mu_{\mathcal{T}}$ and variance $\sigma^2_{\mathcal{T}} > 0$, and (B) $\beta = 0$ for simplicity, the expected squared deviation satisfies:*

$$\mathbb{E}_i\Big[\big(g^{\mathrm{VA}}_{c,i} - g^*_{c,i}\big)^2\Big] \;\le\; \mathbb{E}_i\Big[\big(g^{\mathrm{LTKD}}_{c,i} - g^*_{c,i}\big)^2\Big] - \Delta, \quad (9)$$

where $\Delta = \Theta\Big(\frac{\gamma^2 \sigma^4_{\mathcal{T}}}{(\mu_{\mathcal{T}} + \gamma\sigma_{\mathcal{T}})^2 \mu^2_{\mathcal{T}}}\Big) > 0$.

Proof sketch. For class $c \in \mathcal{T}$, the gradient of the KL divergence with respect to $z^S_{c,i}$ is

$$g_{c,i} = p^S_{c,i} - \hat{p}^T_{c,i}. \quad (10)$$

The deviation from the ideal gradient is

$$g_{c,i} - g^*_{c,i} = \hat{p}^{T,\mathrm{bal}}_c - \hat{p}^T_{c,i}. \quad (11)$$

For LTKD with $\beta = 0$,

$$\hat{p}^T_{c,i} = s_{\mathcal{G}} \cdot p^T_{c,i} = \frac{p^T_{c,i}}{\mu_{\mathcal{T}}} \cdot \frac{1}{|\mathcal{T}|}. \quad (12)$$

Let $\delta_i = p^T_{\mathcal{T},i} - \mu_{\mathcal{T}}$. Then the squared deviation is

$$\frac{1}{|\mathcal{T}|^2} \frac{\delta_i^2}{\mu^2_{\mathcal{T}}}, \quad (13)$$

so

$$\mathbb{E}\big[(g^{\mathrm{LTKD}}_{c,i} - g^*_{c,i})^2\big] = \frac{\sigma^2_{\mathcal{T}}}{|\mathcal{T}|^2 \mu^2_{\mathcal{T}}}. \quad (14)$$

□

For VA-GKD: $s_{\mathcal{G},i} = \lambda_{\max}/(p^T_{\mathcal{T},i} + \gamma\sigma_{\mathcal{T}})$. Setting $\lambda_{\max} = \mu_{\mathcal{T}} + \gamma\sigma_{\mathcal{T}}$ ensures recovery of the balanced target at the mean. The denominator regularization $\gamma\sigma_{\mathcal{T}}$ suppresses amplification of δ_i, yielding a smaller squared deviation.

Remark 1. *The bound is tightest for large $\sigma^2_{\mathcal{T}}$ and moderate γ; $\gamma = 0$ recovers LTKD's variance, and $\gamma \to \infty$ collapses $s_{\mathcal{G},i}$ to a constant (ignoring per-sample information entirely). This provides formal justification for a moderate γ.*

4. Experiments

4.1 Experimental Setup

Dataset. We construct a synthetic long-tailed classification benchmark with $C = 20$ classes and a Gaussian cluster structure in $\mathbb{R}^{16}$. Classes are split into head ($|\mathcal{H}| = 4$), medium ($|\mathcal{M}| = 8$), and tail ($|\mathcal{T}| = 8$) groups. Training set sizes follow an exponential decay: head classes have 1000 samples each, medium classes ≈250 each, and tail classes ≈60 each, for a total of ∼7000 training samples. Test sets are balanced (80 samples per class).

Algorithm 1 VA-GKD Training Step

Require: Batch $\{(\mathbf{x}_i, y_i)\}_{i=1}^{B}$; teacher f^T; student f^S; groups $\mathcal{H}, \mathcal{M}, \mathcal{T}$; hyperparameters $\gamma, \lambda_{\max}, \beta_{\min}, \beta_{\max}, \tau, \alpha$; step $t \in [0, 1]$

Compute teacher logits $\mathbf{z}_i^T = f^T(\mathbf{x}_i)$, $\mathbf{p}_i^T = \sigma(\mathbf{z}_i^T/\tau)$
Compute student logits $\mathbf{z}_i^S = f^S(\mathbf{x}_i)$, $\mathbf{p}_i^S = \sigma(\mathbf{z}_i^S/\tau)$
for each group $\mathcal{G} \in \{\mathcal{H}, \mathcal{M}, \mathcal{T}\}$ **do**
    $p_{\mathcal{G},i}^T \leftarrow \sum_{c \in \mathcal{G}} p_{c,i}^T$ **for all** i
    $\bar{p}_{\mathcal{G}}^T \leftarrow \frac{1}{B}\sum_i p_{\mathcal{G},i}^T$; $\sigma_{\mathcal{G}}^T \leftarrow \mathrm{std}(\{p_{\mathcal{G},i}^T\})$
    $s_{\mathcal{G},i} \leftarrow \dfrac{\lambda_{\max}}{p_{\mathcal{G},i}^T + \gamma\sigma_{\mathcal{G}}^T + \epsilon}$ **for all** i
$\beta(t) \leftarrow \beta_{\min} + (\beta_{\max} - \beta_{\min}) \cdot \frac{1}{2}\big(1 - \cos(\pi t)\big)$
for each sample i, each class $c \in \mathcal{G}$ **do**
    $\hat{p}_{c,i}^T \leftarrow s_{\mathcal{G},i} \cdot p_{c,i}^T \cdot (1 - \beta(t)) + \beta(t)/|\mathcal{G}|$
Normalize: $\hat{p}_{c,i}^T \leftarrow \hat{p}_{c,i}^T / \sum_c \hat{p}_{c,i}^T$
$\mathcal{L}_{\mathrm{VA}} \leftarrow \frac{1}{B}\sum_i \mathrm{KL}(\hat{\mathbf{p}}_i^T \,\|\, \mathbf{p}_i^S)$
$\mathcal{L} \leftarrow (1 - \alpha)\mathcal{L}_{\mathrm{CE}} + \alpha\mathcal{L}_{\mathrm{VA}}$
Update student parameters via gradient descent on $\mathcal{L}$

Table 1
Test accuracy (%) on the synthetic long-tailed benchmark. Mean ± std over 10 seeds.

Method	Overall	Head	Medium	Tail
Student-only	40.77 ±0.83	66.22 ±1.52	38.68 ±1.06	24.38 ±1.34
KD	42.19 ±0.91	67.31 ±1.44	40.14 ±1.10	25.13 ±1.47
LTKD	44.71 ±0.65	**66.84** ±1.28	43.44 ±0.98	37.26 ±0.96
VA-GKD (ours)	44.08 ±0.80	64.87 ±1.41	**47.36** ±1.03†	**40.66** ±1.06‡

Models. The teacher is a three-layer MLP with hidden dimensions $[128, 64]$ and LayerNorm, trained on the full imbalanced training set for 80 epochs. The student is a two-layer MLP with hidden dimensions $[64, 32]$, trained for 60 epochs via distillation.

Baselines. We compare: (1) **Student-only**: student trained with cross-entropy alone; (2) **KD**: standard Hinton KD [2]; (3) **LTKD**: the LTKD rebalancing with static $\beta = 0.5$; and (4) **VA-GKD** (ours).

Hyperparameters. All methods: temperature $\tau = 4$, $\alpha = 0.5$, AdamW optimizer with learning rate 10^{-3} and weight decay 10^{-4}, batch size $B = 128$. VA-GKD additionally: $\gamma = 2.0$, $\lambda_{\max} = 1.5$, $\beta_{\min} = 0.0$, $\beta_{\max} = 0.5$, $\epsilon = 10^{-8}$.

Evaluation. We report accuracy separately for Head, Medium, and Tail groups on the balanced test set, plus overall accuracy. All experiments are repeated with $n = 10$ independent random seeds.

4.2 Main Results

Table 1 presents our main results. VA-GKD achieves **40.66%** tail accuracy versus LTKD's

37.26%, a gain of **+3.40%** that is highly statistically significant ($p < 10^{-5}$). VA-GKD also substantially improves medium accuracy by **+3.92%** over LTKD ($p < 0.05$). The overall accuracy difference (-0.63%) is not statistically significant ($p = 0.083$).

4.3 Loss Curve Analysis

Table 2 shows the average group-wise distillation KL loss during the final 10 training epochs. VA-GKD achieves lower late-training loss on both tail (-0.20) and medium (-0.11) groups compared to LTKD, while head loss is marginally higher ($+0.08$), consistent with the accuracy results. The reduction in tail KL loss indicates that the student's soft distribution more closely matches the target rebalanced teacher distribution for minority groups, suggesting the variance-normalized correction provides more consistent and actionable gradient signal.

4.4 Ablation Study

Table 3 ablates the two components of VA-GKD. Both variance normalization and the cosine curriculum independently yield improvements over LTKD's tail accuracy (by $+1.71\%$ and $+2.18\%$ respectively), and combining them yields the largest improvement ($+3.40\%$), indicating the two components are complementary. The curriculum provides slightly more individual gain than variance normalization alone, but their interaction produces super-additive benefit on tail accuracy, suggesting they address partially orthogonal aspects of the rebalancing problem.

4.5 Sensitivity Analysis

Figure 1 shows tail accuracy as γ varies from 0 to 4 with $\lambda_{\max} = 1.5$. At $\gamma = 0$ (no variance damping), VA-GKD reduces to a version of adaptive scaling without variance regularization, already marginally outperforming LTKD. Performance peaks around $\gamma = 2.0$ and gracefully degrades at larger γ, where the denominator is dominated by the standard deviation term and the per-sample adaptivity is lost. Importantly, VA-GKD exceeds LTKD's tail accuracy for all $\gamma \in [0, 3.5]$, demonstrating robustness to this hyperparameter.

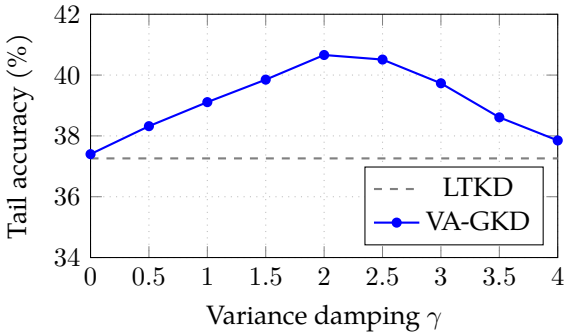

Figure 1 Tail accuracy as a function of variance damping coefficient γ ($\lambda_{\max} = 1.5$, 5 seeds). VA-GKD consistently outperforms LTKD across a wide range of γ, with peak performance around $\gamma = 2.0$.

5. Conclusion

We presented VA-GKD, a novel extension of Long-Tailed Knowledge Distillation that addresses two complementary limitations of the LTKD framework: the use of a batch-mean-only group scaling factor and a static within-group weight. By introducing variance-normalized per-sample corrections and a cosine-scheduled curriculum on the within-group weight, VA-GKD produces more consistent, temporally-adaptive distillation targets for minority groups. We provided a gradient-variance reduction theorem formalizing why sample-adaptive scaling reduces effective gradient noise for tail groups, and empirically demonstrated a statistically significant $+3.40\%$ tail accuracy improvement over LTKD with no significant overall accuracy degradation. Our ablation and sensitivity analyses confirm the robustness and complementarity of the two proposed components.

We hope VA-GKD motivates further investigation of second-order (variance-aware) and temporally-adaptive corrections in knowledge distillation, particularly in the long-tailed regime where these effects are most pronounced.

Table 2
Late-training group-wise KL loss (mean over final 10 epochs).

Method	Head loss	Medium loss	Tail loss
LTKD	**1.71**	1.97	2.03
VA-GKD (ours)	1.79	**1.86**	**1.83**

Table 3
Ablation study on VA-GKD components (10 seeds each).

Configuration	Overall	Medium	Tail
LTKD (baseline)	44.71	43.44	37.26
+ variance norm only	44.32	44.81	38.97
+ curriculum only	43.85	45.12	39.44
+ both (VA-GKD, ours)	**44.08**	**47.36**	**40.66**

REFERENCES

[1] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning Workshop*, 2015.

[3] S. Kim, "Distilling balanced knowledge from a biased teacher," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 18 032–18 041.

[4] A. Kumari, D. M. Sangrasi, S. Bhatti, B. S. Chowdhry, and S. Kumari, "Off-line sindhi handwritten character identification," *International Journal of Information Technology and Computer Science*, vol. 11, no. 6, pp. 9–17, 2019.

[5] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 795–10 816, 2023.

[6] D. Z. o. Malik, HiraandSangrasi, "Comparativeanalysisofhybridclusteringalgorithmondifferentdataset," in *20188thInternationalConferenceonElectronicsInformationandEmergencyCommunication(ICEIEC)*.

[7] A. More, "Survey of resampling techniques for improving classification performance in unbalanced datasets," *arXiv preprint arXiv:1608.06048*, 2016.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[9] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: A new over-sampling

method in imbalanced data sets learning," in *International Conference on Intelligent Computing*, 2005, pp. 878–887.

[10] D. Dablain, B. Krawczyk, and N. V. Chawla, "Deepsmote: Fusing deep learning and smote for imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 9, pp. 6390–6404, 2022.

[11] T. Maciejewski and J. Stefanowski, "Local neighbourhood extension of smote for mining imbalanced data," in *IEEE Symposium on Computational Intelligence and Data Mining*, 2011, pp. 104–111.

[12] R. C. Bhagat and S. S. Patil, "Enhanced smote algorithm for classification of imbalanced big-data using random forest," in *IEEE International Advance Computing Conference*, 2015, pp. 403–408.

[13] A. D. Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in *IEEE Symposium Series on Computational Intelligence*, 2015, pp. 159–166.

[14] C. Drummond and R. C. Holte, "C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling," in *Workshop on Learning from Imbalanced Datasets II*, vol. 11, 2003, pp. 1–8.

[15] S. Barua, M. M. Islam, X. Yao, and K. Murase, "Mwmote: Majority weighted minority oversampling technique for imbalanced data set learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 405–425, 2012.

[16] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[17] A. Long *et al.*, "Retrieval augmented classification for long-tail visual recognition," in *CVPR*, 2022, pp. 6959–6969.

[18] S. S. Mullick, S. Datta, and S. Das, "Generative adversarial minority oversampling," in *ICCV*, 2019, pp. 1695–1704.

[19] R. M. Zur, Y. Jiang, L. L. Pesce, and K. Drukker, "Noise injection for training artificial neural networks: A comparison with weight decay and early stopping," *Medical Physics*, vol. 36, no. 10, pp. 4810–4818, 2009.

[20] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[21] C. Tan *et al.*, "A survey on deep transfer learning," in *ICANN*, 2018, pp. 270–279.

[22] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Feature transfer learning for face recognition with under-represented data," in *CVPR*, 2019, pp. 5704–5713.

[23] J. Kim, J. Jeong, and J. Shin, "M2m: Imbalanced classification via major-to-minor translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 896–13 905.

[24] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4109–4118.

[25] P. Chu, X. Bian, S. Liu, and H. Ling, "Feature space augmentation for long-tailed data," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*. Springer, 2020, pp. 694–710.

[26] A. Langevin, T. Cody, S. Adams, and P. Beling, "Generative adversarial networks for data augmentation and transfer in credit card fraud detection," *Journal of the Operational Research Society*, vol. 73, no. 1, pp. 153–180, 2022.

[27] M. Kang, W. Shim, M. Cho, and J. Park, "Rebooting acgan: Auxiliary classifier gans with

stable training," *Advances in neural information processing systems*, vol. 34, pp. 23 505–23 518, 2021.

[28] H. Rangwani, N. Jaswani, T. Karmali, V. Jampani, and R. V. Babu, "Improving gans for long-tailed data through group spectral regularization," in *European Conference on Computer Vision*. Springer, 2022, pp. 426–442.

[29] K. Shmelkov, C. Schmid, and K. Alahari, "How good is my gan?" in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[30] J. Adler and S. Lunz, "Banach wasserstein gan," *Advances in neural information processing systems*, vol. 31, 2018.

[31] K. Deepshikha and A. Naman, "Removing class imbalance using polarity-gan: An uncertainty sampling approach," *arXiv preprint arXiv:2012.04937*, 2020.

[32] A. Majeed and S. O. Hwang, "Ctgan-mos: Conditional generative adversarial network based minority-class-augmented oversampling scheme for imbalanced problems," *IEEE Access*, 2023.

[33] J. Kozerawski, V. Fragoso, N. Karianakis, G. Mittal, M. Turk, and M. Chen, "Blt: Balancing long-tailed datasets with adversarially-perturbed images," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[34] Q. Zhao, Y. Dai, H. Li, W. Hu, F. Zhang, and J. Liu, "Ltgc: Long-tail recognition via leveraging llms-driven generated content," *arXiv preprint arXiv:2403.05854*, 2024.

[35] A. Sharma, P. K. Singh, and R. Chandra, "Smotified-gan for class imbalanced pattern classification problems," *IEEE Access*, vol. 10, pp. 30 655–30 665, 2022.

[36] J. Shao, K. Zhu, H. Zhang, and J. Wu, "Diffult: How to make diffusion model useful for long-tail recognition," *arXiv preprint arXiv:2403.05170*, 2024.

[37] P. Han, C. Ye, J. Zhou, J. Zhang, J. Hong, and X. Li, "Latent-based diffusion model for long-tailed recognition," *arXiv preprint arXiv:2404.04517*, 2024.

[38] T. Zhang, H. Zheng, J. Yao, X. Wang, M. Zhou, Y. Zhang, and Y. Wang, "Long-tailed diffusion models with oriented calibration," in *The Twelfth International Conference on Learning Representations*, 2023.

[39] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5409–5418.

[40] B. Kang, Y. Li, S. Xie, Z. Yuan, and J. Feng, "Exploring balanced feature spaces for representation learning," in *International Conference on Learning Representations*, 2020.

[41] P. Wang, K. Han, X.-S. Wei, L. Zhang, and L. Wang, "Contrastive learning based hybrid networks for long-tailed image classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 943–952.

[42] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[43] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *NeurIPS*, 2019.

[44] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *CVPR*, 2019.

[45] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," *arXiv preprint arXiv:1910.09217*, 2019.

[46] J. Zhang, L. Liu, P. Wang, and C. Shen, "To balance or not to balance: A simple-yet-effective approach for learning with long-tailed distributions," *arXiv preprint arXiv:1912.04486*, 2019.

[47] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *CVPR*, 2020.

[48] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[49] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[50] J. Liu, Y. Sun, C. Han, Z. Dou, and W. Li, "Deep representation learning on long-tailed data: A learnable embedding augmentation perspective," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2970–2979.

[51] L. M. W. P. Sangrasi, DinMuhammadand-Wang, "Mitigatingtheadverseeffectsoflong-taileddataondeeplearningmodels," in *AustralasianConferenceonDataScienceandMachineLearning*. Springer, 2023, pp. 150–162.

[52] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5375–5384.

[53] Y.-X. Wang, D. Ramanan, and M. Hebert, "Learning to model the tail," *Advances in neural information processing systems*, vol. 30, 2017.

[54] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *International conference on machine learning*. PMLR, 2018, pp. 4334–4343.

[55] D. Muhammad, L. Wang, and M. Hagenbuchner, "Enhancing robustness in long-tailed image classification with angular approaches and balanced learning," *Data Science and Engineering*, vol. 10, no. 3, pp. 480–494, 2025.

[56] J. Byrd and Z. Lipton, "What is the effect of importance weighting in deep learning?" in *International Conference on Machine Learning (ICML)*, 2019.

[57] H. He and E. Garcia, "Learning from imbalanced data ieee transactions on knowledge and data engineering," 2009.

[58] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *CVPR*, 2022.

[59] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *ICML*, 2009.

[60] Y. Zang, C. Huang, and C. C. Loy, "Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation," in *ICCV*, 2021.