Computer Science and Application 计算机科学与应用, 2026, 16(5), 1-12
Published Online May 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.165158

Hans汉斯

面向长尾识别的自适应加权融合策略

陈　佳

广东工业大学数学与统计学院，广东　广州

收稿日期：2026年4月1日；录用日期：2026年5月2日；发布日期：2026年5月11日

摘　要

在深度学习的早期发展阶段，研究多聚焦于如ImageNet、CIFAR等类别平衡的基准数据集。然而，在现实世界的视觉场景中，数据分布往往遵循幂律分布。这意味着极少数类别占据了绝大部分的样本量，而绝大多数类别仅拥有极少量的观察样本。这种长尾分布是自然界和人类社会的常态，广泛存在于物种识别、医疗诊断、自动驾驶物体检测以及工业缺陷检测等关键领域。然而，传统的深度学习模型在长尾数据上表现出明显的性能失衡，这源于头部偏见：受经验风险最小化驱动，模型会过度拟合样本丰富的头部类别，导致预测结果向头部偏移。尾部塌陷：对于样本稀缺的尾部类别，模型由于缺乏足够的辨识性特征，分类精度往往出现断崖式下跌。这种现象严重制约了人工智能系统在实际复杂环境中的鲁棒性和可靠性。针对长尾类别不平衡带来的严峻挑战，如何构建鲁棒的识别模型已成为计算机视觉领域的核心课题。研究基于混合专家模型架构，深入长尾数据分布下的图像分类难题，针对性地提出了创新深度视觉识别策略。文章的主要研究内容与贡献概括如下：提出了一种用于多专家长尾图像分类的测试阶段自适应集成的方法，通过在测试阶段引入Test-Time Augmentation (TTA)即测试阶段的图像增强后，并计算两个可靠性分数：1) Stability，通过增强特征间的平均余弦相似度衡量；2) Certainty，由平均概率分布的归一化熵导出。并将由这两个可靠性分数决定的不同权重分配给不同专家，使各个专家对于图像判断的偏好具有侧重性，更好地对数据集进行图像分类任务。该方法无需额外参数，开销极小，能有效抑制不可靠专家，在长尾分布下显著提升尾部类性能，同时保持整体准确率。

关键词

深度学习，长尾图像分类，动态权重

Adaptive Weighted Fusion Strategy for Long Tail Recognition

Jia Chen

School of Mathematics and Statistics, Guangdong University of Technology, Guangzhou Guangdong

Received: April 1, 2026; accepted: May 2, 2026; published: May 11, 2026

文章引用: 陈佳. 面向长尾识别的自适应加权融合策略[J]. 计算机科学与应用, 2026, 16(5): 1-12.
DOI: 10.12677/csa.2026.165158

Abstract

In the early development stage of deep learning, research focuses on the balanced benchmark data sets of categories such as ImageNet and CIFAR. However, in the real-world visual scene, the data distribution often follows the power law distribution. This means that a few categories account for the vast majority of the sample size, while the vast majority of categories only have a small number of observation samples. This long tail distribution is the normal state of nature and human society, and widely exists in key fields such as species identification, medical diagnosis, automatic driving object detection, and industrial defect detection. However, the traditional deep learning model shows an obvious performance imbalance on the long tail data, which is due to the head bias: driven by the minimization of empirical risk, the model will overfit the head categories with rich samples, leading to the deviation of prediction results to the head. Tail collapse: for tail categories with scarce samples, the classification accuracy of the model often drops precipitously due to the lack of sufficient identification features. This phenomenon seriously restricts the robustness and reliability of the artificial intelligence system in the actual complex environment. In view of the severe challenge brought by the imbalance of long tail categories, how to build a robust recognition model has become a core topic in the field of computer vision. Based on the hybrid expert model architecture, this research delves into the image classification problem under the long tail data distribution and proposes an innovative depth vision recognition strategy. The main research contents and contributions of this paper are summarized as follows: A method of adaptive integration in the test phase for multi-expert long tail image classification is proposed. After introducing Test Time Augmentation (TTA), that is, image enhancement in the test phase, two reliability scores are calculated: 1) Stability, which is measured by the average cosine similarity of features between enhancements; 2) Certainty, which is derived from the normalized entropy of the average probability distribution. And different weights determined by these two reliability scores are allocated to different experts, so that each expert has a preference for image judgment, and can better perform image classification tasks on data sets. This method requires no additional parameters and has minimal overhead. It can effectively restrain unreliable experts, significantly improve tail class performance under the long tail distribution, and maintain the overall accuracy.

Keywords

Deep Learning, Long Tail Image Classification, Dynamic Weight

1. 引言

在深度学习时代，图像识别技术取得了显著进展，得益于大规模标注数据集和卷积神经网络的强大表示能力，模型在平衡分布下的分类任务上达到了接近人类水平的性能。然而，真实世界中的视觉数据往往遵循长尾分布：少数头部类别拥有大量训练样本，而大量尾部类别仅含有极少甚至单个样本。这种极度不平衡的数据分布严重违背了传统机器学习中独立同分布和类别平衡的假设，导致训练出的模型在头部类别上表现优异，却在尾部类别上泛化能力极差，甚至出现近乎随机的预测准确率。这种现象被称为长尾识别问题，是当前计算机视觉领域从实验室基准向真实场景部署转型所面临的核心挑战之一。

混合专家模型(Mixture of Experts, MoE)作为一种稀疏激活架构，近年来被引入长尾识别领域。其核心思想是将模型分解为多个“专家”子网络，每个专家专攻数据分布的不同部分，通过门控/路由机制动态分配输入样本到合适专家，最后集成多个专家的输出。这种设计天然契合长尾分布：能促进专家多样性，让不同专家学习互补知识，避免单一模型的偏置问题。

传统 MoE 在测试阶段多采用简单的平均加权融合，这种固定策略易使头部专家主导预测，压制尾部专家作用，且无法根据样本差异动态调整，导致长尾分类中尾部精度低、预测校准差。为此，本文在自异构集成与知识挖掘 SHIKE [1]方法模型框架的基础上，提出一种基于稳定性与确信度的自适应权重融合方法，通过动态分配专家权重，强化尾部专家对对应样本的贡献，缓解头部偏置问题，提升预测鲁棒性与校准能力，在保持 MoE 高效性的同时，实现更均衡的头部和尾部性能，有效提升长尾图像分类效果。

2. 相关工作

2.1. 长尾图像识别

长尾图像识别是计算机视觉领域的一个核心挑战，主要源于真实世界数据分布的幂律特性：少数头部类别拥有大量样本，而大量尾部类别样本极少。这种不平衡导致传统深度学习模型(如基于经验风险最小化 ERM 的训练)严重偏向头部类别，尾部性能接近随机。典型基准数据集包括 ImageNet-LT [2]、CIFAR-100-LT [3]、iNaturalist 2018 [4]、Places-LT 等，这些数据集模拟真实场景，推动了从平衡基准向开放世界泛化的转型。当前长尾图像识别研究持续活跃，自 2023 年 *Deep Long-Tailed Learning*: *A Survey* [5]发表后，领域进入快速发展期。2024~2026 年间，顶级会议涌现大量工作，性能在基准上不断刷新 SOTA。在 ImageNet-LT 和 iNaturalist 2018 上，结合混合专家模型(MoE)和视觉语言大模型的方案已成为主流，尾部准确率显著提升，同时整体性能更均衡。

近年来，MoE 开始被引入长尾识别领域，早期工作探索专家分组策略；当前，随着 MoE 在大语言模型(如 Mixtral [6])和视觉模型中的成功，其在长尾视觉任务中的应用迅速升温，成为提升尾部性能、实现均衡泛化的关键范式。

2.2. 测试阶段自适应加权融合

在长尾图像识别(Long-Tailed Visual Recognition)中，混合专家模型(Mixture of Experts, MoE)通过多个专精专家(如头部/尾部类别专精)和路由机制缓解类别不平衡问题。然而，早期的 MoE 在测试阶段多采用简单平均权重(Average Weighting)或固定加权集成专家输出，这种静态策略忽略了输入样本的个体差异(如难度、分布异质性)，导致头部专家主导预测、尾部贡献被稀释，以及较高的校准误差。为解决这一问题，测试阶段动态加权融合(Test-Time Dynamic Weighted Fusion/Adaptive Weighting)应运而生：基于输入样本的实时信号(如置信度、熵、难度分数或语义相似性)动态计算每个专家的权重，实现自适应专家集成。测试阶段动态加权融合的概念最早源于集成学习(Ensemble Learning)，但在深度学习测试阶段的应用较晚兴起，早期测试阶段多为静态集成，如简单平均或多数投票，缺乏自适应性。

在混合专家模型的应用发展中，MoE 天然适合动态加权融合：其自带的路由器已提供专家激活基础，测试阶段还可进一步利用 Test-Time 信号(如 Logit Entropy、类准确率或 OOD 分数)计算权重，避免训练阶段的固定偏置。

自适应动态加权融合的发展历程如下：

最开始 MoE 长尾工作(如 RIDE、BalPoE)多用固定或简单加权集成，动态性有限。而后 Balanced Product of Experts (BalPoE [7]前身工作)提出 Product of Experts (PoE)机制，通过乘积集成校准专家输出，避免平均偏置，但仍非完全动态。

后来测试时适应(Test-Time Adaptation, TTA)兴起，如 TENT [8]、MEMO [9]等，动态调整模型参数或批次统计；同时，在 MoE 中出现初步自适应权重，如基于路由分数的加权，包括引入内部蒸馏和初步自适应，如 MEID [10] (2024)在视频长尾中的动态协作。

直到近期动态加权融合成为主流，专为长尾异质性设计。核心创新在于“难度感知”(Difficulty-Aware)和“输入特定”(Input-Specific)权重计算：使用 Test-Time 信号(如类准确率、Logit Entropy)定义样本难度，尾部/困难样本赋予尾部专家更高权重。结合 TTA 范式，实现无标签测试时动态调整专家贡献。

本文提出的是一种基于稳定性和确信度构成的权重自适应融合策略，这种机制不仅提升了尾部类别的预测影响力，还增强了模型的鲁棒性和泛化能力，通过权重分配放大专家的特长优势，弱化专家的弊端，动态融合专家输出，处理长尾异质性。

2.3. 余弦相似度与信息熵

余弦相似度作为向量空间模型的核心度量，其应用可追溯到 20 世纪 60~70 年代的信息检索领域。该方法通过计算两个向量的夹角余弦值来度量它们的方向一致性：

$$sim(u,v)=\frac{u\cdot v}{|u|\cdot|v|}\in[-1,1] \tag{1}$$

该定义的优势在于其方向不变性——只关注向量的方向而不受幅度影响，使其对特征表示的尺度变化天然鲁棒。随着深度学习的兴起，余弦相似度因其几何直觉和计算高效性，成为了特征空间度量的标准方法。

余弦相似度最早在上世纪八九十年代用于文本与信息检索，如传统 TF-IDF + 余弦相似度是文本检索的基准方法，提供了一个可解释、可证明的相似性度量框架。后来推广运用至推荐系统与协同过滤方向，用户/物品向量的相似度成为推荐的基础。相对于欧氏距离，余弦相似度在高维稀疏数据中也表现更稳定。

在深度学习方向上，孪生网络(Siamese Networks)和三元组损失(Triplet Loss)大量使用余弦相似度；多模态学习中的图文匹配(如 CLIP [11])采用余弦相似度作为对齐目标；度量学习(Metric Learning)将其作为距离函数的标准选择。

本文使用余弦相似度计算方法，并将其与欧氏距离、皮尔逊相关系数等方法进行横向对比，后两者均存在明显的适配性问题：欧氏距离 L2 对特征幅度敏感而且大幅度低质量特征可能导致虚假“接近”；皮尔逊相关系数计算成本高，需要显式中心化，无明显优势。

在神经网络特征表示中，余弦相似度的方向不变性尤为重要：由于 BatchNorm、LayerNorm 等标准化层的存在，特征的幅度信息往往不稳定，而方向信息更加可靠。这使余弦相似度成为度量特征语义一致性的自然选择。

信息熵由 Claude Shannon 在 1948 年的开创性工作中引入，是信息论的核心概念。对于概率分布 p，信息熵定义为：

$$H(p)=-\sum_{C=1}^{C}p_c\log P_c \tag{2}$$

其中 C 为类别数。该定义具有深刻的信息论意义：熵等于表示该分布所需的平均信息量(比特数)。特别地：均匀分布时，$H_{\max}=\log C$ (最大熵，完全不确定)，单点分布时，$H_{\max}=0$ (最小熵，完全确定)。

信息熵是唯一满足连续性、对称性、递归可加性和最大值位置等公理的熵定义，这保证了其在理论上的唯一性和必要性。

信息熵在深度学习中有着广泛应用，主要体现在两个方面：一是不确定性量化，其中通过 Dropout 进行多次前向推理得到多个预测，可用于估计认知不确定性；而基于单次预测概率分布计算熵，能够衡量数据不确定性。二是置信度相关任务，包括异常检测、开集识别与样本难度估计，通常正常样本、已知类别与简单样本预测熵较低，异常样本、未见类别与困难样本预测熵相对更高。

对于单一使用稳定性或确定性其中一个信号来驱动自适应权重，都有着较高的局限性。

只用余弦相似度作为特征级稳定性度量，虽然能有效评估专家对视角变化的鲁棒性，但存在过度依赖特征一致性的问题。在长尾识别场景下，尾类样本由于训练数据稀缺，特征提取的稳定性通常较低，但部分头部类专家可能会对尾类特征产生系统性偏移但一致的预测，虽然特征提取稳定，但预测完全错误。

只用信息熵作为预测级不确定性量化方法，能够衡量专家对分类的“真实确信度”，但完全忽略了特征提取的过程。在长尾识别场景下，部分头部类专家可能对尾类样本产生过度自信但错误的预测，虽然预测分布集中，但实际类别完全错误。

本文采用的是特征级增强稳定性与预测级不确定性量化的双信号融合策略，通过稳定性与确定性信号进行乘法结合，完美解决了单一信号的局限性。乘法结合意味着，只有既稳定又正确的专家才会被赋予高权重，这种双维度评估确保了融合过程的可靠性，同时提高了对尾类样本预测的鲁棒性。

3. 模型方法

在测试阶段，为实现对多专家预测的动态集成并提升模型在分布偏移下的稳健性，设计了一套基于特征空间稳定性与预测分布确信度双重信号的加权融合机制。整个推理流程遵循从数据增强、双视图前向传播、信号量计算、权重生成到最终融合预测的连贯路径。

3.1. 数据增强与多专家前向传播

为评估模型在数据扰动下的鲁棒性并为信号量计算提供基础，对每个原始测试批次 $x=\{x_1,x_2,\cdots,x_B\}\in R^{B\times 3\times H\times W}$ 生成其水平翻转版本，构成增强批次 x^{flip}，其中

$$x_i^{flip}=Flip\left(x_i,\dim=3\right) \tag{3}$$

原始批次与翻转批次共同作为后续双视图前向传播的输入。

将 x 与 x^{flip} 分别送入包含 E 个独立专家分类器的模型，获得两类输出：

分类 logits：第 e 个专家对原始视图的预测为 $Z_e^{orig}\in R^{B\times C}$，对翻转视图的预测为 $Z_e^{flip}\in R^{B\times C}$，其中 C 为类别数。

骨干网络特征：第 e 个专家在原始视图与翻转视图下提取的高维特征分别为 $f_e^{orig}\in R^{B\times D}$ 与 $f_e^{flip}\in R^{B\times D}$，$D$ 为特征维度。

3.2. 自适应融合

3.2.1. 预测稳定性信号

稳定性引入了余弦相似度，鲁棒好的专家在轻微扰动的情况下，特征表示应该保持一致。余弦相似度高，保持较高的稳定性。

因此，特征稳定性衡量专家在输入遭受轻微扰动(水平翻转)时，其内部特征表示的一致性。对第 e 个专家在第 i 个样本上的原始特征 f^{orig} 与翻转特征 f^{flip}，计算其余弦相似度：

$$Sim_{e,i}=\frac{f_{e,i}^{orig}f_{e,i}^{flip}}{\left\|f_{e,i}^{orig}\right\|_2\cdot\left\|f_{e,i}^{flip}\right\|_2} \tag{4}$$

对该专家在当前批次 B 个样本上的相似度取平均，得到其稳定性信号 S^{stab}：

$$S_e^{stab} = \frac{1}{B}\sum_{i=1}^{B} Sim_{e,i} \tag{5}$$

S_e^{stab} 的取值范围为[−1, 1]，值越高表明该专家的特征表示对数据扰动越不敏感，即越稳定。

3.2.2. 特征确信度信号

预测确信度衡量专家预测概率分布的集中程度，确信度信号引入的是熵，熵是信息论中衡量不确定性的标准指标。熵越低，信息分布越集中，模型越确信。在前文所提到的测试阶段数据增强。通过水平翻转加入轻微扰动，保存了原图 logits $Z_{e,i}^{orig}$ 和翻转图 logits $Z_{e,i}^{flip}$。同时，结合原图和翻转图生成平均概率，防止单个视图的偶然高置信，得到更鲁棒的概率估计。

首先对第 e 个专家在第 i 个样本上的原始与翻转 logits 分别计算 softmax 概率：

$$P_{e,i}^{orig} = \mathrm{softmax}\left(Z_{e,i}^{orig}\right) \tag{6}$$

$$P_{e,i}^{flip} = \mathrm{softmax}\left(Z_{e,i}^{flip}\right) \tag{7}$$

计算其平均概率分布：

$$\overline{P_{e,i}} = \frac{P_{e,i}^{orig} + P_{e,i}^{flip}}{2} \tag{8}$$

进而计算其归一化熵：

$$H_{e,i} = -\frac{1}{\log C}\sum_{C=1}^{C} \overline{P_{e,i}}^{(c)} \log \overline{P_{e,i}}^{(c)} \tag{9}$$

其中 $\log C$ 为最大可能熵，用于将熵值归一化到[0, 1]区间。对该专家在当前批次所有样本上的归一化熵取平均，得到其平均不确定度 $\bar{H}_e$。为将其转换为确信度信号，定义：

$$S_e^{cert} = 1 - \bar{H}_e \tag{10}$$

S_e^{cert} 的取值范围为[0, 1]，值越高表明该专家的预测越确定、越有信心。

3.2.3. 动态权重计算与融合

基于计算得到的稳定性信号 S_e^{stab} 与确信度信号 S_e^{cert}，为每个专家合成一个可靠性分数。引入可调温度参数 τ 以控制权重分布的平滑度：

$$r_e = \exp\left(\frac{S_e^{stab}}{\tau}\right) \cdot \exp\left(\frac{S_e^{cert}}{\tau}\right) \tag{11}$$

其中第一项奖励特征稳定性高的专家，第二项奖励预测确信度高的专家。对所有专家的可靠性分数进行 Softmax 归一化，得到最终的融合权重：

$$\omega_e = \frac{\exp\left(\frac{r_e}{\tau}\right)}{\sum_{k=1}^{E}\exp\left(\frac{r_k}{\tau}\right)} \tag{12}$$

融合预测通过对每个专家在原始与翻转视图上的 logits 进行内部平均，再按权重 ω_e 进行加权求和得到：

首先，计算每个专家在双视图上的平均 logits：

$$\bar{Z}_e = \frac{Z_e^{orig} + Z_e^{flip}}{2} \tag{13}$$

然后，进行加权融合，得到最终批次的预测：

$$Z^{fused} = \sum_{e=1}^{E} \omega_e \cdot \bar{Z}_e \tag{14}$$

4. 实验分析

本节主要在 CIFAR10-LT、CIFAR100-LT、ImageNet-LT 长尾数据集上开展一系列的实验来验证本文方法的有效性。首先介绍实验选取的数据集；其次介绍实验条件设置，包括数据预处理和模型训练参数设置；最后与近年经典的长尾图像分类算法进行比较，讨论实验结果，分析不同数据集下的分类性能表现，并进一步与基线模型做对比实验及分析，验证本方法对整体性能的贡献。

4.1. 数据集选取

为全面评估方法性能，在三个广泛使用的长尾图像分类基准数据集上进行了实验，它们分别是 CIFAR10-LT、CIFAR100-LT 和 ImageNet-LT。各数据集的统计细节如表 1 所示。

Table 1. Parameter table of different datasets
表 1. 不同数据集参数表

数据集	不平衡因子	类别数	训练样本数	头部样本数	尾部样本数
CIFAR10-LT	100	10	12,406	5000	50
	50	10	13,996	5000	100
CIFAR100-LT	100	100	10,847	500	5
	50	100	12,608	500	10
ImageNet-LT	256	1000	115,846	1280	5

4.2. 实验条件设置

4.2.1. 数据预处理

在预处理中，CIFAR10-LT 和 CIFAR100-LT 长尾数据集的训练集图像像素为 32 × 32，在数据增强方面，具体是从输入图像样本及其水平镜像变换版本中随机采样 32 × 32 像素的局部区域，每边填充 4 个像素，再从 40 × 40 中随机裁剪出 32 × 32 的区域，并做随机水平翻转，再使用 AutoAugment 数据增强策略随机选择旋转、亮度、对比度、剪切、平移等子策略组合增强，最后转为 Tensor 并进行[0, 1]归一化，使用 Cutout 随机遮挡 1 个 16 × 16 区域，最后做标准化处理。

ImageNet-LT 长尾数据集的图像大小不固定，对训练集图像和验证集图像采用不同的数据增强技术，裁剪时从[0.08, 1.0]范围随机缩放图像并随机裁剪到 224 × 224，再使用随机增强策略，而后归一化处理，最后使用 ImageNet 标准均值和标准差进行标准化。在验证时，首先保持宽高比，将图像短边调整为 256 像素，然后在中心裁剪 224 × 224 的区域，并做归一化，最后使用与训练相同的均值和标准差对图像进行归一化处理。

4.2.2. 训练参数设置

在 CIFAR-10 和 CIFAR-100 长尾数据集上，实验采用 ResNet-32 作为骨干网络，在不平衡因子为 50 和 100 上进行训练。批量大小设为 128，初始学习率为 0.05。在 ImageNet-LT 长尾数据集上，实验采用

ResNet-50 作为骨干网络，批量大小设为 256，初始学习率为 0.1。框架构建包含 3 个专家分支的多专家协同学习框架，所有专家共享骨干网络参数但拥有独立的分类器头，全部实验训练过程采用两阶段策略：总训练轮数为 200 个 epoch，前 180 个 epoch 进行特征提取器与分类器的联合训练，后 20 个 epoch 冻结共享参数仅重新训练分类器。使用 SGD 优化器，动量设为 0.9，权重衰减为 5e−4，采用带预热的余弦退火学习率调度器，预热期为 5 个 epoch。训练阶段采用随机裁剪、随机水平翻转、CIFAR10Policy 和 Cutout 等数据增强策略。测试阶段采用自适应加权融合策略。

本文的全部实验均使用 PyTorch 深度学习框架、Python 编程语言实现，硬件设备主要包括 NVIDIA RTX 3090 24 G 的 GPU 和 Intel (R) Xeon (R) Gold 5218R CPU @2.10 GHz 的 CPU。

4.3. 实验结果分析

4.3.1. CIFAR10-LT 实验结果与分析

为验证本文方法的有效性，在 CIFAR10-LT 数据集上进行了充分的实验。表 2 展示了与当前主流长尾识别方法的对比结果，其实对比方法包括经典双分支结构 BBN，类别重平衡方法 GCL、BALMS，以及采用 3 个专家的混合专家模型训练方法 ACE 和 ResLT。实验中考虑了不平衡因子为 100 和 50 的两种场景，以 ResNet-32 为骨干网络，评估了不同方法的综合性能。

Table 2. Experimental results on CIFAR10-LT dataset
表 2. CIFAR10-LT 实验结果表

数据集	CIFAR10-LT	
骨干网络	ResNet-32	
不平衡因子	100	50
SHIKE	84.34	86.81
BBN	79.82	82.18
ACE (3E) [12]	81.20	84.30
ResLT (3E) [13]	81.44	-
GCL [14]	82.68	85.46
BALMS	84.61	-
Our (3E)	85.37	87.60

注：3E 表示 3 个专家组成的网络结构。

实验结果表明，本文所提方法在 CIFAR10-LT 数据集分类性能出众，整体优于现有主流长尾识别模型。在相同 200 轮训练设置下，比起经典双分支 BBN 算法，在不平衡因子为 50 和 100 时，分别实现了 5.42%和 5.55%的精度提升；与类别重平衡方法相比，在 50 和 100 的不平衡因子下分别提升 2.14%和 0.76%~2.69%。并与同样混合专家模型训练方法 ACE (3E)和 ResLT (3E)当中最高准确率对比，本文方法的提升分别达到 3.93%与 3.3%。从对比数据可以看出，本文所提方法相较于多种主流基线模型均取得了更具竞争力的分类精度。尤其在与相同专家结构的混合专家模型对比中，本文方法仍然实现了明显的性能提升，充分验证了所提策略在长尾不平衡数据场景下的有效性与先进性。

4.3.2. CIFAR100-LT 实验结果与分析

为验证本文方法的有效性，在 CIFAR100-LT 数据集上进行了充分的实验。表 3 展示了与当前主流长尾识别方法的对比结果。对比的方法除了 Focal Loss、BBN 等常规方法外，还加入了同样在长尾领域中

聚焦于测试阶段改进的方法 SADE、Logit Adjustment 和 RIDE (3E)。同时，也考虑了不平衡因子为 100 和 50 的两种场景，以 ResNet-32 为骨干网络，评估了不同方法的综合性能。

Table 3. Experimental results on CIFAR100-LT dataset
表 3. CIFAR100-LT 实验结果表

数据集	CIFAR100-LT	
骨干网络	ResNet-32	
不平衡因子	100	50
Focal Loss [15]	38.40	44.3
BBN [16]	42.6	47.0
SADE [17]	49.8	53.9
Logit Adjustment [18]	55.5	51.8
RIDE (3E) [19]	48.0	-
SHIKE	55.52	59.53
Our (3E)	56.55	60.20

注：3E 表示 3 个专家组成的网络结构。

实验结果表明，本文所提方法在分类性能上优于现有主流长尾识别模型。在相同 200 轮训练设置下，与经典双分支结构 BBN 算法相比，所提方法在不平衡因子为 50 和 100 时，分别实现了 13.2%和 13.95%的精度提升；与 Focal Loss 相比，在相同不平衡因子下分别提升 15.9%和 18.15%。与同样在长尾领域中聚焦于测试阶段改进方法对比，本文方法的提升分别达到 1.05%~8.55%与 6.3%~8.4%。从对比数据可以看出，所提策略在测试阶段的优化效果更为显著，综合性能更具优势，充分验证了本文面向测试阶段设计的改进方案在长尾图像分类任务中的有效性与优越性。

4.3.3. ImageNet-LT 实验结果与分析

为验证本文方法的有效性，在 ImageNet-LT 数据集上进行了充分的实验。实验中根据每个类别训练样本的数量，将数据集划分为三个子数据集：头部类别(超过 100 张训练图像)、中部类别(20~100 张训练图像)和尾部类别(少于 20 张训练图像)，并在表 4 中分别报告了在三个子数据集在 ImageNet-LT 数据集上的准确率，旨在验证尾部类数据的提升效果。实验以 ResNet-50 为骨干网络，评估了不同方法在头部类别、中部类别和尾部类别的综合性能。

从实验结果可知，本文所提方法在 ImageNet-LT 数据集分类性能上整体准确率能达到 54.17%，头部类别、中部类别和尾部类别分别达到 76.04%，50.32%和 31.05%。与损失重加权类方法 CB、Balanced 方法相比，尾部类别性能和整体准确率分别提升了{14.25%，3.45%}和{20.97%，4.07%}；与解耦方法 cRT 相比，本法在尾部类别性能和整体准确率提升了 4.95%和 6.87%；与多专家集成方法 DiVE、KCL 相比，在不牺牲尾部类性能的情况下，中部类别有小幅度的提升，头部类别准确率得到显著提升，整体准确率分别提升了 4.77%和 2.67%。

在分析测试阶段发现，自适应加权融合方法之所以能显著提升头部类与整体性能，核心原因在于它对模型校准度实现了系统性优化。该方法通过特征稳定性与预测置信度双指标加权融合策略，本质上是对模型输出进行“置信度校准”。对于 ImageNet-LT 数据集中的头部类，样本充足使模型学习到的特征表达更完整、预测更稳定且置信度高，改进方法能有效放大这些高质量专家的贡献；对于中尾部类，样本稀缺导致特征学习不充分、预测稳定性与置信度较低，双指标权重分配策略虽能筛选出相对更优的专

家，但提升幅度受限。整体而言，该方法更偏向于提升模型的校准度，使高置信预测更准确、低置信预测更谨慎，这种校准效果在样本充足的头部类上体现得最明显，从而带动整体性能的显著提升。

Table 4. Experimental results on ImageNet-LT dataset
表 4. ImageNet-LT 实验结果表

数据集	ImageNet-LT			
骨干网络	ResNet-50			
类别	头部类	中部类	尾部类	整体准确率
CE	64.0	33.8	5.8	41.6
CB [20]	39.6	32.7	16.8	33.2
Balanced [21]	61.1	47.5	27.6	50.1
cRT [22]	58.8	44.4	26.1	47.3
DiVE [23]	64.1	50.4	30.7	49.4
KCL [24]	61.8	49.4	30.9	51.5
SHIKE	68.7%	51.1	31.7	52.4
Our (3E)	76.04	50.32	31.05	54.17

注：3E 表示 3 个专家组成的网络结构。

从方法论维度看，损失加权类虽能提升尾部性能，但会损害头部类性能；解耦训练方法 cRT 取得一定平衡但整体性能有限；多专家集成方法展现出更优的整体性能。本文方法在 DiVE、KCL 的基础上，对头部类别和整体性能做了进一步提升，该实验结果也证明了所提方法在大规模数据集上具有良好的分类性能和泛化能力。

4.3.4. 与基线模型的对比实验及分析

为了分析不同超参数取值对模型性能的影响，本文选取基线模型 SHIKE，在 CIFAR10-LT 和 CIFAR100-LT 长尾数据集上，设置不平衡因子 $= \{10, 20, 50, 100, 200, 250\}$进行实验，实验的数据预处理和训练参数设置和 SHIKE 一致，SHIKE 实验结果均由对源论文的复现得出。表 5 和表 6 展示了模型 top-1 正确率的实验结果数据。

Table 5. Top-1 accuracy of the model on CIFAR10-LT
表 5. CIFAR10-LT 的模型 Top-1 正确率实验结果表

数不平衡因子值	10	20	50	100	200	250
SHIKE	90.91	89.55	86.55	84.53	81.42	80.06
Ours	91.84	90.64	87.17	85.37	81.80	80.39
提升(%)	0.93	1.09	0.62	0.84	0.38	0.33

Table 6. Top-1 accuracy of the model on CIFAR100-LT
表 6. CIFAR100-LT 的模型 Top-1 正确率实验结果表

数不平衡因子值	10	20	50	100	200	250
SHIKE	68.82	64.78	58.87	55.97	50.47	48.43
Ours	69.19	65.93	60.62	56.58	51.03	49.91
提升(%)	0.37	1.15	1.75	0.61	0.56	1.48

实验结果表明，本章所提的算法在 CIFAR10-LT 与 CIFAR100-LT 数据集上均能有效提升基线模型性能，验证了改进策略的有效性。值得注意的是，算法在不平衡因子为 20、50、100 的场景中性能提升更为显著，这三个不平衡因子是现实世界中最为常见的，这表明数据分布越贴近现实场景，算法的改进效果就越突出，对解决现实工程中的不平衡问题也能起到更显著的提升作用。

5. 结论

本文提出了一种基于稳定性和确信度引导的测试阶段多专家自适应加权融合框架，通过结合测试时增强策略的鲁棒性提升能力以及多专家集成的互补性优势，有效突破了传统单模型推理在长尾分布下的性能瓶颈，实现了对不同样本量类别均友好的高性能识别。框架设计了双重质量评估指标：稳定性指标通过原始图像与水平翻转图像在特征空间的余弦相似度来量化专家对输入变化的鲁棒性，确信度指标基于预测概率熵来评估专家决策的置信水平，两者经温度参数平滑后自适应融合，避免了对单一专家或简单平均策略的依赖。大量实验表明，相较于传统的平均融合策略，自适应加权融合在性能上得到显著提升，在 CIFAR-10-LT、CIFAR-100-LT 和 ImageNet-LT 多个长尾基准数据集上都保持了竞争性的结果，同时还具有较低的计算开销，为求解长尾视觉识别问题提供了一种创新且高效的推理阶段解决方案。

参考文献

[1] Jin, Y., Li, M., Lu, Y., Cheung, Y. and Wang, H. (2023) Long-Tailed Visual Recognition via Self-Heterogeneous Integration with Knowledge Excavation. 2023 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 23695-23704. https://doi.org/10.1109/cvpr52729.2023.02269

[2] Deng, J., Dong, W., Socher, R., Li, L., Li, K. and Li, F.F. (2009) ImageNet: A Large-Scale Hierarchical Image Database. 2009 *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, 20-25 June 2009, 248-255. https://doi.org/10.1109/cvpr.2009.5206848

[3] Krizhevsky, A. and Hinton, G. (2009) Learning Multiple Layers of Features from Tiny Images. Department of Computer Science, University of Toronto.

[4] Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., *et al*. (2018) The iNaturalist Species Classification and Detection Dataset. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8769-8788. https://doi.org/10.1109/cvpr.2018.00914

[5] Zhang, Y., Kang, B., Hooi, B., Yan, S. and Feng, J. (2023) Deep Long-Tailed Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 10795-10816. https://doi.org/10.1109/tpami.2023.3268118

[6] Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., *et al*. (2024) Mixtral of Experts. arXiv:2401.04088.

[7] Aimar, E.S., Jonnarth, A., Felsberg, M. and Kuhlmann, M. (2023) Balanced Product of Calibrated Experts for Long-Tailed Recognition. 2023 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 19967-19977. https://doi.org/10.1109/cvpr52729.2023.01912

[8] Wang, D., Shelhamer, E., Liu, S., Olshausen, B. and Darrell, T. (2021) Tent: Fully Test-Time Adaptation by Entropy Minimization. 2021 *International Conference on Learning Representations* (*ICLR*), Vienna, 3-7 May 2021.

[9] Zhang, M., Levine, S. and Finn, C. (2022) MEMO: Test Time Robustness via Adaptation and Augmentation. 2022 *Advances in Neural Information Processing Systems* (*NeurIPS*), New Orleans, 28 November-3 December 2022, 1-14.

[10] Li, X. and Xu, H. (2023) MEID: Mixture-of-Experts with Internal Distillation for Long-Tailed Video Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 1451-1459. https://doi.org/10.1609/aaai.v37i2.25230

[11] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., *et al*. (2021) Learning Transferable Visual Models from Natural Language Supervision. 2021 *International Conference on Machine Learning* (*ICML*), Online, 18-24 July 2021, 1-16.

[12] Cai, J., Wang, Y. and Hwang, J. (2021) ACE: Ally Complementary Experts for Solving Long-Tailed Recognition in One-Shot. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 112-121. https://doi.org/10.1109/iccv48922.2021.00018

[13] Cui, J., Liu, S., Tian, Z., Zhong, Z. and Jia, J. (2022) ResLT: Residual Learning for Long-Tailed Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 3695-3706. https://doi.org/10.1109/tpami.2022.3174892

[14] Li, M., Cheung, Y. and Lu, Y. (2022) Long-Tailed Visual Recognition via Gaussian Clouded Logit Adjustment. *Proceedings of the IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022,

6929-6938.

[15] Lin, T.Y., Goyal, P., Girshick, R., *et al*. (2017) Focal Loss for Dense Object Detection. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2980-2988. https://doi.org/10.1109/iccv.2017.324

[16] Zhou, B., Cui, Q., Wei, X. and Chen, Z. (2020) BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 9719-9728. https://doi.org/10.1109/cvpr42600.2020.00974

[17] Zhang, Y., Hooi, B., Hong, L. and Feng, J. (2022) Self-Supervised Aggregation of Diverse Experts for Test-Agnostic Long-Tailed Recognition. 2022 *Advances in Neural Information Processing Systems* (*NeurIPS*), New Orleans, 28 November-9 December 2022, 34077-34090.

[18] Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A. and Kumar, S. (2021) Long-Tail Learning via Logit Adjustment. 2021 *International Conference on Learning Representations* (*ICLR*), Vienna, 3-7 May 2021.

[19] Wang, X., Lian, L., Miao, Z., Liu, Z. and Yu, S.X. (2021) Long-Tailed Recognition by Routing Diverse Distribution-Aware Experts. 2021 *International Conference on Learning Representations* (*ICLR*), Vienna, 3-7 May 2021.

[20] Cui, Y., Jia, M., Lin, T., Song, Y. and Belongie, S. (2019) Class-Balanced Loss Based on Effective Number of Samples. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 9260-9269. https://doi.org/10.1109/cvpr.2019.00949

[21] Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S. and Li, H. (2020) Balanced Meta-Softmax for Long-Tailed Visual Recognition. 2020 *Advances in Neural Information Processing Systems* (*NeurIPS*), Online, 6-12 December 2020, 4175-4186.

[22] Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J. and Kalantidis, Y. (2020) Decoupling Representation and Classifier for Long-Tailed Recognition. 2020 *International Conference on Learning Representations* (*ICLR*), Addis Ababa, 26-30 April 2020.

[23] He, Y.Y., Wu, J. and Wei, X.S. (2021) Distilling Virtual Examples for Long-Tailed Recognition. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 235-244. https://doi.org/10.1109/iccv48922.2021.00030

[24] Kang, B., Li, Y., Xie, S., Yuan, Z. and Feng, J. (2020) Exploring Balanced Feature Spaces for Representation Learning. 2020 *International Conference on Learning Representations* (*ICLR*), Addis Ababa, 23-30 April 2020.