通过节省参数的微调方式进行长尾识别

摘要

语言图像多模态模型和其他预训练模型迅速兴起，引起了广泛关注。人们也逐渐研究微调预训练模型在长尾视觉识别任务上的应用。最近已经有很多研究将预训练模型迁移到长尾任务上，最终也都达到了较好的性能。然而，这些工作都面临着共同问题：需要很多轮训练才能让模型收敛或者需要额外的训练数据才能达到较好的性能。针对这些问题，最近有研究提出了一种名为 PEL 的长尾识别方法。该方法借助现有的节省参数微调方法，只引入少量可训练参数来微调预训练模型，以减轻尾部类的过拟合；另外，提出了一种新的分类器初始化方式：借助多模态模型产生的语义信息初始化分类器，显著加快了微调预训练模型收敛的速度。本报告的工作基于该篇研究，并在此方法的基础上进行了改进，开展了一些实验探究。

关键词：长尾视觉识别；多模态模型；预训练；微调；加速收敛

1 引言

近年来，深度学习迅速发展，在目标检测、模式识别等各种计算机视觉任务上获得了巨大的成功。其中一个重要原因是用于训练的数据集样本充足且各类别所含样本数均衡。然而，在现实世界中数据大多呈现出长尾分布，即只有少部分类拥有大量且充足的样本，而大部分类别只有极少量的样本 [20]。因此，长尾问题非常具有挑战性，受到了广泛的关注。有大量研究都在长尾视觉识别任务上取得了非常好的效果，尤其是对尾部类。这些方法大体可分为两类：类平衡方式 [1,4,16]、提升模型本身的泛化能力 [19,24,25]。这些方式都显著的提升了分类精度，但是依旧受到很严重的不平衡数据分布的影响。

以前上述两类方式大多是从头训练深度神经网络。最近，预训练模型风靡一时，也有许多研究致力于通过微调预训练模型的方式来提升长尾识别的性能，并且也取得了不错的效果，例如 BALLAD [13]、RAC [12]、VL-LTR [18] 和 LPT [5] 等方式。在 ImageNet 数据集上预训练的 ViT 模型 [6]、语言图像多模态预训练模型（CLIP） [15] 已经被很多研究用来解决长尾问题。

有研究指出，现存的这都些使用预训练模型处理长尾问题的方法存来三个问题：训练轮次较多、需要两阶段训练、需要额外的训练数据集 [17]。于是，该篇工作提出了一种名为 PEL [17] 的方式，很好地缓解了上述提到的问题，使得微调预训练模型在长尾视觉识别任务上有更好的表现。

本报告中也重点研究了 PEL 方法，并且基于该方法做出了一些改进，并且进行了实验探究。PEL 是一种在长尾识别任务上微调预训练模型的新方法。该方法整体采用一下策略：固

定住整体的预训练模型只引入少量可调节参数，即节省参数的微调方式。以此来减少参数量以减轻样本数极少的尾部类过拟合的问题。另外，PEL 中提出了两个新方法：利用类标签的语义特征初始化分类器以加速收敛；集成多个扰动输入的预测结果以增强模型的鲁棒性。而且 PEL 方式能与各种节省参数的微调方式组合使用，如 VPT [9]、LoRA [8]、Adapter [7]、AdaptFormer [2]。

大规模的预训练模型近几年迅速兴起，其在长尾任务上的应用也逐渐成为热点。而 PEL 方法是最近的研究，且很好地分析并缓解了一些预训练模型应用到长尾任务时出现的问题。PEL 方法效果卓著，在微调预训练方式中比较通用，并且该篇工作总结了现存的几种常用微调方式，也对一些超参的选择、微调方式的性能以及分类器的选择做了详细的分析。因此深刻研习该篇工作对日后的研究很有帮助，故而本篇报告决定复现该篇工作，基于该篇工作做出部分改进，并进行一些实验探究。

2 相关工作

预训练模型兴起之后，将对于长尾问题的研究划分成了两个大类：从头训练神经网络、微调预训练模型。以前从头训练神经网络的方式中诞生出了许多性能卓著的方法，其中有许多依旧可以迁移到微调预训练模型的方式中。本章节将对于这两种训练方式下的工作做出简要介绍。

2.1 基于从头训练网络的长尾识别

传统的长尾识别方法大多直接从头训练神经网络，这类方法大多有许多个提升模型性能的入手点，如提升网络的表征能力、平衡分类器。为了提升网络的表征能力，可以使用对比学习的方式来训练特征提取器 [3,11]，对比学习能使网络具有很好的表征能力，有利于之后的分类。对于分类器的平衡可以使用重采样的策略，或者损失重加权的策略。也有很多研究利用类的先验信息设计出更平衡的损失函数。如 [1,14,16]。另外有种经常采用的训练策略：解耦学习，即将特征学习和分类器的学习分开以将两者提升至最佳 [21,23]。在从头训练的网络上，数据增强也能起到很好的作用。

2.2 基于微调预训练模型的长尾识别

相对于从头训练的网络，使用微调预训练模型解决长尾问题的方式还不是很多。预训练模型最大的优点是，其在大规模数据集上预训练，因而具有很强的特征学习能力，因此微调预训练模型成为了一种解决数据不平衡问题的有效策略，在长尾识别领域逐渐受到关注，基于此诞生了一些方法 [5,12,13,17,18]。这些方式有的使用 ViT 预训练模型，有的使用 CLIP 模型，然后与现存的几种微调方式相结合来处理长尾问题，有也沿用了之前在从头训练的网络上诞生的方法，取得了很好的效果。

3 本文方法

PEL 方法中主要有两个创新点，加快了模型的收敛速度、提升了模型在长尾任务下的分类精度。另外，该论文的最佳性能的得出使用了先前工作提出的平衡的损失函数。本小节将

对 PEL 中的两个创新点以及使用的一个先前工作提出的损失函数做详细介绍。

3.1 语义初始化模块

PEL 方法中使用 CLIP 模型，其整体框架如图 1所示，虚线左侧是模型的整体架构，右侧是各种节省参数的微调方式。按照模型架构图所示，将类标签提取之后生成文本提示词，送入文本编码器中得到各标签的文本特征编码，然后用文本特征编码初始化分类器。

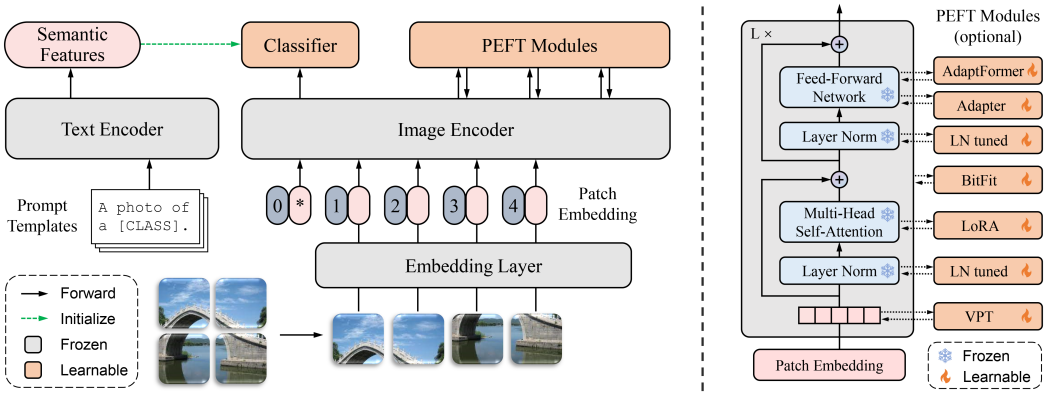

图 1. 整体架构图

该初始化方式简单又高效，具体是用 CLIP 模型提取出的文本特征来初始化分类器的权重。因为在文本数据中不会受到长尾问题的困扰，因此文本编码器能对类标签进行很好的语义编码，使其在特征空间有较好的分布。故以文本特征来初始化分类器权重相当于给分类器一个比较好的起始点，给出了一个较好的类中心，因此在训练时，会大大加快收敛。

3.2 测试集成模块

为了提高模型的性能 PEL 方法中还提出了一个测试集成的模块。ViT 模型将输入的图片划分成了几个小块，这导致连续的信息被分到了不同的块中。于是测试集成模块将输入图片进行扰动，生成其一系列扰动版本，然后再将这多个扰动版本输入模型中，得到多个预测结果，然后将预测结果都综合起来，判断最终的分类情况。具体来讲，给定一个测试数据 x，对该数据进行扰动得到其 M 个不同的扰动版本 $\alpha_i(x)$，整合这多个扰动版本的输出就可得到模型最终预测得分 z，如公式 1所示：

$$z = \log P(y|x) = \frac{1}{M}\sum_{i=1}^{M} \log P(y|\alpha_i(x)) \tag{1}$$

3.3 损失函数定义

PEL 方法中使用了先前工作提出的用于长尾任务的平衡损失：LogitAdjustment 损失 [14]，如公式所示。该损失的基本思想是对模型输出进行调整，对尾类较大的补偿，对头类较少的补偿。

$$\ell(y, f(x)) = -\log \frac{e^{f_y(x)+\tau\cdot\log \pi_y}}{\sum_{y'\in[L]} e^{f_{y'}(x)+\tau\cdot\log \pi_{y'}}} \tag{2}$$

其中 τ 是一个超参数，用来控制平衡程度；π_y 表示类别 y 的先验概率。

4 复现细节

4.1 与已有开源代码对比

与 PEL 方式对比，本报告中的研究对 PEL 方法进行了两种不同的改进，并且用一种简单且直观的微调方式对预训练 ViT 模型微调时不同模块所起的作用做了初步分析。

原始的 PEL 方法使用 AdaptFormer 微调方法使得模型达到最好的性能。该方式是在多层感知机模块又添加了一个支路，引入可调节参数进行微调。而在 ViT 模型中，每个块的归一化层对于每个样本的特征值做了不同程度的缩放，并进行归一化，因此该层能调节对不同特征的关注程度，对于新任务上的微调至关重要 [10]。因此，该报告中进行的第一个改进是将模型每个块的归一化层连同引入的 AdaptFormer 一起微调（LN+AdaptFormer），以希望预训练模型能更加适配新的任务。

第二种思考是在微调的参数量上，过多的参数对于样本数量较少的尾部类很容易过拟合，因此希望在原来的基础上进一步削减参数。原来的方法在 ViT 模型的每一层都引入了 AdaptFormer 结构进行微调，引入的参数量还是相对较多，因此我们考虑只在部分层引入。最近有研究表明，在语言图像多模态任务上微调预训练的 Transformer 模型时不同的层影响不同：对于语言任务，浅层比较敏感；对于图像任务，深层比较敏感 [10]。基于这样的考量，本报告中的第二种改进选择只在模型的后半层引入 AdaptFormer 结构，而在第一层引入可学习的提示词（VPT-S+Partial）。提示词微调的方式（VPT）能够调节注意力模块，然而浅层的提示词信息一直往深层传播也会对分类产生不好的影响，有的研究考虑控制第一层的提示词信息流入深层的比例 [22]。因此第二种改进在第一层引入可学习提示词，深层插入了可调节参数以适配浅层传递的信息，二者一定程度上可以相互促进，该方式的结构如图 2所示。

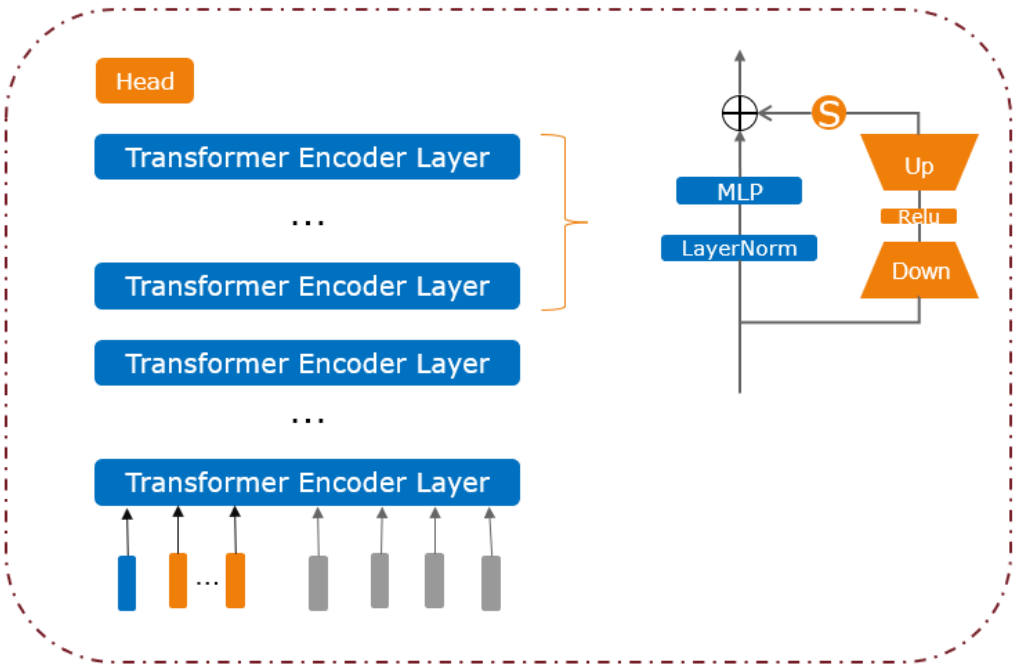

图 2. 方法示意图

另外，报告中为了探究在微调任务中注意力层、归一化层和多层感知机层的影响，采用了一种非常简单且直观的微调方式：引入几个带有权重和偏执的线性层，来分别影响注意力层、归一化层或多层感知机层，以体现在不同的任务下，哪个部分更加敏感。

4.2 实验环境搭建

实验使用的显卡为 RTX 3090，软件开发环境为 Pycharm 2023.2.2, 使用的较为关键的软件开发工具有 Python 3.8、PyTorch 2.0.0、torchvision 0.15.0，timm 0.6.12。

5 实验结果分析

5.1 两种改进方式的性能分析

在 CIFAR100-LT、ImageNet-LT、Places-LT 和 iNaturalist2018 四个数据集上进行实验验证实验的骨干网络使用 CLIP 模型，初始学习率设为 0.01，在实验过程中以余弦退火的策略衰减，权重衰减为 0.0005，动量设置为 0.9，所有输入的图像都被调整为 224×224 的分辨率再送入网络中。统计实验的结果时，使用 Top-1 分类精度来衡量多种方法下模型的性能，并且额外地统计出头部类（样本数大于 100）、中间类（样本数在 20-100 之间）和尾部类（样本数小于 20）三个部分的分类精度。在 CIFAR100-LT、ImageNet-LT、Places-LT 和 iNaturalist2018 四个数据集上的结果分别如表 1、表 2、表 3、表 4所示。

在 LN+AdaptFormer 的改进下，CIFAR100-LT、ImageNet-LT、Places-LT 三个数据集上的结果相对于原始的 PEL 有所下降，尤其是在 ImageNet-LT 和 Places-LT 两个数据集上，可以看出其尾部类的精度下降很多，体现出比原来更大的不平衡性。归一化层可能受到训练集不平衡的影响比较大。

表 1. CIFAR100-LT 数据集不同不平衡率下的结果

Method		Backbone	LearnableParam.	#Epochs	200	100	50	10
	Original PEL	ViT-B/16	0.10M	10	79.6	81.7	83.1	84.9
Ours	VPT-S+Partial	ViT-B/16	0.06M	10	76.9	79.2	79.9	82.4
	Partial	ViT-B/16	0.05M	10	76.1	78.2	79.7	82.0
	LN+AdaptFormer	ViT-B/16	0.14M	10	79.3	82.1	83.1	84.8

表 2. ImageNet-LT 数据集上的结果

Method		Backbone	Learnable Param.	#Epochs	Overall	Head	Med	Tail
	Original PEL	ViT-B/16	0.62M	10	78.3	81.3	77.4	73.4
Ours	VPT-S+Partial	ViT-B/16	0.32M	10	78.4	81.5	77.2	73.7
	Partial	ViT-B/16	0.31M	10	78.4	81.3	77.3	74.0
	LN+AdaptFormer	ViT-B/16	0.65M	10	76.7	80.1	75.6	70.6

在 VPT-S+Partial 的改进下，在 ImageNet-LT 和 Places-LT 两个数据集上的性能与原方法相比略有提升，而其需要调节的参数量与原来相比减少了将近一半，该方法下尾部类精度显著提升，可能是因为参数量的减少极大地缓解了尾部类的过拟合问题。在 CIFAR100-LT 和 iNaturalist2018 两个数据集上，该方法的性能都有了不同程度的下降。

为了进一步确定 VPT-S+Partial 的有效性，表格中另外还统计了只在后半层加入 AdaptFormer（Partial）而不在第一层添加可训练提示词（VPT-S）的结果。发现在 ImageNet-LT 数据集上加不加 VPT-S 效果没有变化，但是在其他数据集上，加入了 VPT-S 之后，效果都比只用 Partial 的方式好。初步分析，这种情况与输入图像的分辨率有关，提示词对应调节注意力层，AdaptFormer 对应调节多层感知机。ImageNet-LT 数据集的图像分辨率与预训练的输入图像分辨率一致，故微调时注意力层发挥作用不大，而多层感知机层比较重要。而对于其他数据集的图像，其分辨率与预训练时的输入图像分辨率有一定差异，故在微调时注意力层会比较重要，VPT-S 对注意力进行了调整，对只用 Partial 的情况有了提升。这点与后面探究不同模块的影响时的结论也相一致。

表 3. Places-LT 数据集上的结果

Method		Backbone	Learnable Param.	#Epochs	Overall	Head	Med	Tail
	Original PEL	ViT-B/16	0.18M	10	52.2	51.7	53.1	50.9
Ours	VPT-S+Partial	ViT-B/16	0.10M	10	52.2	51.6	52.5	52.4
	Partial	ViT-B/16	0.09M	10	52.1	51.2	52.7	52.3
	LN+AdaptFormer	ViT-B/16	0.21M	10	51.6	52.5	51.6	49.8

表 4. iNaturalist2018 数据集上的结果

Method		Backbone	LearnableParam.	#Epochs	Overall	Head	Med	Tail
	Original PEL	ViT-B/16	4.75M	20	80.4	74.0	80.3	82.2
Ours	VPT-S+Partial	ViT-B/16	2.38M	20	79.6	73.2	79.5	81.4
	Partial	ViT-B/16	2.37M	20	78.4	81.3	77.3	74.0

5.2 对不同部位的实验探究

为了进一步探究在新的数据集上微调 ViT 模型时注意力层（attn）、归一化层（ln）和多层感知机（mlp）发挥的作用，我们使用线性层分别对这三层进行微调，并分别统计这三种方式下模型的分类精度。该试验在 CIFAR100-LT-IR100、ImageNet-LT 和 Places-LT 三个数据集下进行，其实验结果分别如图 3、图 4、图 5所示。最终我们得出，新数据集上图像的分辨率对于微调的部位有着较大的影响。

可以看出，在 CIFAR100-LT-IR100 数据集上，注意力层在该数据集上微调表现最好，远高于对归一化层和多层感知机的调整；而在 ImageNet-LT 和 Places-LT 两个数据集上，多层感知机的调整效果最好，远远高于对注意力层和归一化层的调整。这可能是由于输入图像的分辨率导致的。CLIP 预训练所使用的图像分辨率为 224×224，CIFAR 数据集图像的分辨率

是 32×32 与其差异巨大，故而在微调时，需要很大的程度调整注意力。Places-LT 数据集图像的分辨率为 256×256，与预训练图像的分辨率差异较小，多层感知机的调整更为重要，因为分辨率也有差异，故而注意力层的调整下的结果也与多层感知机下的调整有较小的差异。对于 ImageNet-LT 数据集，其图像的分辨与预训练时图像的分辨率一致，从结果也能看出，调整多层感知机得到的效果最好，对注意力的微调取得的结果是最差的，远低于微调多层感知机得到效果。

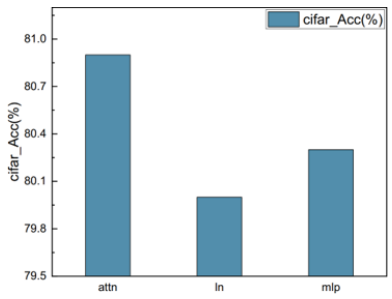

图 3. CIFAR100-LT-IR100

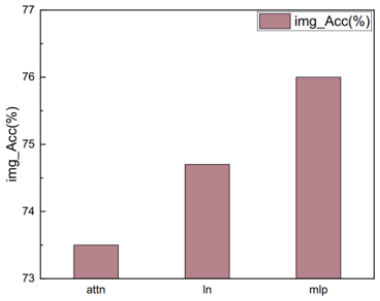

图 4. ImageNet-LT

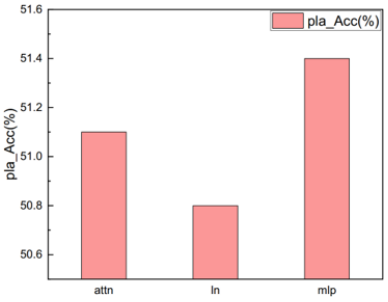

图 5. Places-LT

6 总结与展望

目前通过报告中的实验可以初步确定，减少参数量能更有利于尾部类的学习，并且新数据集上图像的分辨率与微调模型的选择有较为强烈的关系。目前实验的方法对于原始方法没有显著的提升，甚至有所下降。未来工作可能会进一步探究几种微调方式的性质，在目前已经削减了计算代价的基础上进一步提升其性能。

参考文献

[1] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, pages 1567–1578, 2019.

[2] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

[3] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 715–724, 2021.

[4] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 9268–9277, 2019.

[5] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. Lpt: Long-tailed prompt tuning for image classification. *arXiv preprint arXiv:2210.01033*, 2022.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[9] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[10] Zaid Khan and Yun Fu. Contrastive alignment of vision to language through parameter-efficient transfer learning. *arXiv preprint arXiv:2303.11866*, 2023.

[11] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested collaborative learning for long-tailed visual recognition. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 6949–6958, 2022.

[12] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6959–6969, 2022.

[13] Teli Ma, Shijie Geng, Mengmeng Wang, Jing Shao, Jiasen Lu, Hongsheng Li, Peng Gao, and Yu Qiao. A simple long-tailed recognition baseline via vision-language model. *arXiv preprint arXiv:2111.14745*, 2021.

[14] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[16] Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Proc. Conf. on Neural Information Processing Systems*, volume 33, pages 4175–4186, 2020.

[17] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Xin-Yan Han, Jie-Jing Shao, and Yu-Feng Li. Parameter-efficient long-tailed recognition. *arXiv preprint arXiv:2309.10019*, 2023.

[18] Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *European Conference on Computer Vision*, pages 73–91. Springer, 2022.

[19] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020.

[20] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Proc. Conf. on Neural Information Processing Systems*, pages 7029–7039, 2017.

[21] Tong Wei and Kai Gan. Towards realistic long-tailed semi-supervised learning: Consistency is all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3469–3478, 2023.

[22] Seungryong Yoo, Eunji Kim, Dahuin Jung, Jungbeom Lee, and Sungroh Yoon. Improving visual prompt tuning for self-supervised vision transformers. *arXiv preprint arXiv:2306.05067*, 2023.

[23] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498, 2021.

[24] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020.

[25] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 6908–6917, June 2022.