Unleashing the Potential of Vision-Language Models for Long-Tailed Visual Recognition

Teli Ma[1]
telima9868@gmail.com
Shijie Geng[2]
sg1309@rutgers.edu
Mengmeng Wang[3]
mengmengwang@zju.edu.cn
Sheng Xu[4]
shengxu@buaa.edu.cn
Hongsheng Li[1,5]
hsli@ee.cuhk.edu.hk
Baochang Zhang[4]
bczhang@buaa.edu.cn
Peng Gao[1]
gaopeng@pjlab.org.cn
Yu Qiao[1,✉]
qiaoyu@pjlab.org.cn

[1] Shanghai AI Laboratory, Shanghai, China
[2] Rutgers University, New Jersey, US
[3] Zhejiang University, Zhejiang, China
[4] Beihang University, Beijing, China
[5] MMLab, The Chinese University of Hong Kong, Hong Kong

Abstract

The visual world naturally exhibits a long-tailed distribution of open classes, which poses great challenges to modern visual systems. Existing approaches either perform class re-balancing strategies or model ensembling based on image modality. In this paper, we explore strategies of leveraging large-scale pretrained vision-language models for visual long-tailed recognition inspired by the success of powerful multimodal representations that are promising to handle data deficiency and unseen concepts. We first introduce a **BALLAD** method to finetune vision-language models, transferring open-vocabulary knowledge into long-tailed domain dataset in a contrastive manner. Moreover, we propose a non-contrastive and non-parametric learning strategy named **TACKLE** to transfer conceptual knowledge from visual-linguistic model parameters into generated images to balance the training of visual representations. Extensive experiments have been conducted on three popular long-tailed recognition benchmarks to demonstrate the effectiveness of proposed methods.

1 Introduction

During past years, visual recognition tasks, such as image classification [9, 32, 39], object detection [20, 29], semantic segmentation [2, 23, 43], and instance segmentation [10, 13, 21]

have been significantly improved. The performance gains can be largely attributed to the availability of large-scale high-quality datasets [5, 18, 19]. However, the problem of data imbalance has inevitably emerged since real-world data often abide by a long-tailed distribution (e.g., Pareto distribution[26] or Zipf's law [46]). In other words, a few head classes dominate the majority of training examples, whereas many rare or fine-grained classes only have limited relevant data points.

To alleviate the issue, previous efforts either carefully create more balanced datasets (e.g., ImageNet [5], MSCOCO [19], and Kinetics-400 [17]) with human labors or develop more robust algorithms to handle data imbalance. However, since the former is notoriously laborious and expensive, many researchers have been devoted to the latter. Formally, long-tailed recognition (LTR) is a research field seeking robust models that 1) are resistant to significant imbalanced class distribution; 2) can deal with few-shot learning of tail classes. Many methods [42] have been proposed for solving LTR problems. According to the core technical contributions, they can be divided into two categories. Methods in the first line focus on class re-balancing strategies [11, 15, 24, 40] such as data re-sampling, loss re-weighting, and logit adjustment. The second category focuses on improving network modules [3, 4, 16, 30, 33, 41, 45] by classifier designing, decoupled training, and representation learning. While these methods have achieved significant progress, the performance of LTR remains unsatisfactory. When delving deeper into the utilization of the existing imbalance datasets, we have observed that almost all previous efforts are confined to a predetermined manner which designs models entirely relying on the visual modality. That is to say, they totally ignore the semantic features of the raw label text, which may be a promising solution to exert additional supervision on inadequate data sources. Therefore, this paper explores whether language modality can be effective and complementary information for this task. In the meantime, we could also broaden generalization abilities to few-shot categories and zero-shot novel instances.

Recently, contrastive vision-language (VL) models such as CLIP [27] and ALIGN [14] brought a breath of fresh air into the vision community. They learn to align vision and language representations with a contrastive loss given large-scale noisy image-text pairs collected from the web. Motivated by this, we present a simple framework based on contrastive vision-language models for LTR, termed as **BALLAD** (**BAL**anced **L**inear **AD**apter). The training procedure is broken into two phases. In Phase A, we keep finetuning both vision and language branches on a specific LTR dataset through contrastive learning. It enables our framework to fully exploit available training examples and update visual-language representations on a new domain. Then, during Phase B, we freeze the visual and linguistic networks and employ an auxiliary linear adapter for refining on re-balanced training samples. The adapter dynamically combines fixed image-text representations and balanced features via a residual connection to refine the visual representations of tail classes.

However, the linguistic backbone brings about heavy computational overheads during finetuning, especially when the text descriptions are complex or target data distributions are fine-grained. Moreover, current finetuning method relies on conceptual knowledge contained in the pretrained vision-language parameters, limiting the available choices of visual backbones in transfer learning. Therefore, we propose a non-parametric retrieval strategy to convert conceptual knowledge from visual-linguistic pretrained weights into images, named as **TACKLE** (**T**r**A**nsfer **C**onceptual **K**nowledge from **L**anguage to imag**E**). To be specific, visual encoders of VL models are leveraged to project images from web data into feature embeddings $\mathcal{E} \in \mathbb{R}^{N \times d}$, where N is the number of web images. Afterwards, similarities between $\mathcal{E}$ and prompts of target categories are calculated via the language backbone. Based

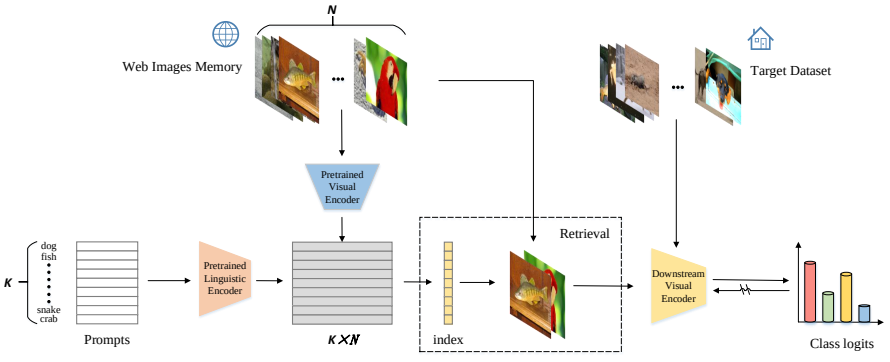


Figure 1: Overview of our TACKLE framework. We retrieve images from web-data memory according to the cosine similarity of web-images and target categories. The retrieved images are combined with target dataset to compensate for data insufficiency. Note that only parameters of downstream visual encoder are updated during the whole process.

on the similarity scores, *k-nearest* web images are retrieved as supplement to alleviate the data insufficiency. In this way, we prevent direct finetuning of both visual and linguistic encoders, while successfully transferring the conceptual knowledge into the visual domain. More importantly, it is capable of incorporating any inductive bias into the design of long-tailed visual backbone architectures. Our contributions are *four* folds:

- We point out the shortcomings of training with fixed class labels and propose to leverage language modality via contrastive vision-language backbone to facilitate long-tailed recognition.
- We develop the BALLAD framework consisting of two phases to handle head and tail classes successively. Specifically, we keep training the visual and language branches of the pretrained vision-language model simultaneously at the first stage. Then we adopt a linear adapter to tackle tail classes with vision-language parameters frozen.
- We introduce TACKLE, a novel retrieval-based strategy that utilizes abundant conceptual knowledge to retrieve incremental images from web collected datasets for solving data insufficiency. The non-parametric characteristic of TACKLE prevent complicated finetuning process of linguistic backbones and can be easily transferred to any visual backbones.
- We conduct extensive experiments to demonstrate the effectiveness of both BALLAD and TACKLE. In a fair comparison, both BALLAD and TACKLE can outperform previous approaches.

2 Method

Our goal is to explore effective pattern of utilizing linguistic hints to alleviate knowledge deficiency of long-tailed distribution. To do so, we first introduce BALLAD, a finetuning

strategy that achieves effective multi-modal representations in long-tailed domain. BALLAD includes two steps, the first step keeps finetuning the VL model in target datasets via contrastive objectives (Sec 2.1). The second step adapts and fuses the former representation with a balanced linear adapter while keeping backbone frozen to reserve open-vocabulary capabilities (Sec 2.2). Moreover, we explore a non-parametric and model-agnostic knowledge transfer method TACKLE to circumvent huge computational overheads of contrastive finetuning and prevent presuming the downstream visual backbone to be the same as the visual encoder of certain pretrained VL model (Sec 2.3).

2.1 Contrastive VL Models Finetuning

Contrastive vision-language models such as CLIP [27] and ALIGN [14] typically follow a dual-encoder architecture with a language encoder $\mathcal{L}_{\text{enc}}$ and a visual encoder $\mathcal{V}_{\text{enc}}$. In this stage, we jointly finetune the encoders to update the multimodal representation for long-tailed recognition. Given an input image $\boldsymbol{I}$, $\mathcal{V}_{\text{enc}}$ is adopted to extract the visual feature for $\boldsymbol{I}$: $\boldsymbol{f}_v = \mathcal{V}_{\text{enc}}(\boldsymbol{I}) \in \mathbb{R}^{d_v}$. Likewise, $\mathcal{L}_{\text{enc}}$ is applied to encode an input text sequence $\boldsymbol{T}$ into its corresponding text feature: $\boldsymbol{f}_l = \mathcal{L}_{\text{enc}}(\boldsymbol{T}) \in \mathbb{R}^{d_l}$. After extracting the feature for each modality, two transformation matrices $\mathbf{W}_v \in \mathbb{R}^{d_v \times d}$ and $\mathbf{W}_l \in \mathbb{R}^{d_l \times d}$ are employed to project the original visual and text features into a shared embedding space:

$$\mathbf{v} = \frac{\mathbf{W}_v^\top \boldsymbol{f}_v}{\|\mathbf{W}_v^\top \boldsymbol{f}_v\|}, \quad \mathbf{u} = \frac{\mathbf{W}_l^\top \boldsymbol{f}_l}{\|\mathbf{W}_l^\top \boldsymbol{f}_l\|}, \tag{1}$$

where $\mathbf{v}$ and $\mathbf{u}$ are both d-dimension normalized vectors in the joint multimodal space. During pretraining, contrastive vision-language models learn to align image-text pairs inside a batch. The overall training objective consists of matching losses from two different directions, *i.e.*, $\mathscr{L}_{v\to l}$ for text retrieval and $\mathscr{L}_{l\to v}$ for image retrieval. They both maximize the scores of matched pairs while minimize that of unmatched ones, the objective function can be formulated as:

$$\begin{aligned}\mathscr{L} &= \mathscr{L}_{v\to l} + \mathscr{L}_{l\to v} \\ &= -\frac{1}{|\mathscr{T}_i^+|}\sum_{T_j\in\mathscr{T}_i^+}\log\frac{\exp\left(\mathbf{v}_i^\top \mathbf{u}_j/\tau\right)}{\sum_{T_k\in\mathscr{T}}\exp\left(\mathbf{v}_i^\top \mathbf{u}_k/\tau\right)} - \frac{1}{|\mathscr{I}_i^+|}\sum_{I_i\in\mathscr{I}_i^+}\log\frac{\exp\left(\mathbf{u}_i^\top \mathbf{v}_j/\tau\right)}{\sum_{I_k\in\mathscr{I}}\exp\left(\mathbf{u}_i^\top \mathbf{v}_k/\tau\right)},\end{aligned}$$

where $\mathscr{T}$ and $\mathscr{I}$ denote a batch of images and text descriptions respectively, and $\mathscr{T}_i^+/\mathscr{I}_i^+$ denote positive text/image subsets matched to image I_i/ text T_i. τ denotes the temperature hyperparameter.

Gururangan et al. [8] show that keeping domain-adaptive and task-adaptive model pretraining can largely improve the performances on target NLP tasks. Similarly, we find that reusing the image-text encoder weights and finetune them in a target long-tailed dataset also benefits imbalanced recognition. Such finetuning strategy is effective in boosting the performance of in-distribution targets recognition, especially in head categories with dominate number of samples. The finetuning scheme should be carefully designed to avoid disturbing the open-vocabulary zero-shot knowledge of VL models while absorbing the new knowledge of target dataset simultaneously. Moreover, to prevent catastrophic forgetting and overfitting of tail classes in finetuning large scale VL models, we introduce a balanced linear adapter module to refine the tail classes representation.

2.2 Balanced Linear Adapter

The phase of finetuning (Phase A) fully utilizes available training data and ensures the performance for classes with abundant examples. However, tail classes are short of training examples and under the few-shot settings. Directly training the whole vision-language backbone may easily overfit to them and lead to performance degradation. Inspired by parameter-efficient adapter modules [6, 12], we freeze the vision-language backbone obtained from Phase A and utilize an additional linear adapter layer to help our model refine its visual-language representation on those infrequent classes. As shown in Figure 1, the text features would remain the same as Phase A. The only difference lies in the image features. If we assume the original image feature to be $\boldsymbol{f}$, the weight matrix and bias of the linear adapter as $\mathbf{W} \in \mathbb{R}^{d\times d}$ and $\mathbf{b} \in \mathbb{R}^d$, then we can represent the refined image feature $\boldsymbol{f}^\star$ as

$$\boldsymbol{f}^\star = \lambda \cdot \mathrm{ReLU}\left(\mathbf{W}^\top \boldsymbol{f} + \mathbf{b}\right) + (1-\lambda)\cdot \boldsymbol{f}, \tag{2}$$

where λ indicates the residual factor to dynamically combine Phase-B fine-tuned image features with the original image features of Phase A.

To avoid the Phase-B training from biasing towards head classes, we also adopt class-balanced sampling strategy [16] to construct a balanced group of training samples. Suppose there are K classes that constitute a total of N training samples in the target dataset. We can represent the number of training samples for class j as n_j and thus have $N = \sum_{j=1}^{K} n_j$. If we assume these classes are already sorted in a decreasing order, then a long-tailed distribution implies $n_i \geq n_j$ if $i < j$ and $n_1 \gg n_K$. For class-balanced sampling, we define the probability of sampling each data point from class j to be $q_j = \frac{1}{K}$. In other words, to construct a balanced group of training samples, we will first uniformly choose a class out of the K candidates and then sample one data point from the selected class. Finally, we perform Phase B finetuning with $\mathscr{L}_{v\to l}$ on the balanced training data. The overall algorithm of finetuning and balanced adapting is shown in Algorithm 1 in Appendix of supplementary material.

2.3 Transfer Knowledge from Language to Image

The strategy of finetuning relies on the assumption that reusing the original structure of vision-language models and the computational overheads of back-propagating huge scale of linguistic encoders are acceptable. However, for a visual backbone network, the incorporating of linguistic sub-nets is inefficient and limits the architecture choices of visual backbones since the pretrained weights are rather significant. We hypothesize the reason that finetuning and adapter-based method BALLAD is effective in imbalanced distribution lies in abundant conceptual knowledge contained in the linguistic encoder. The linguistic hints compensate for knowledge insufficiency of tail classes. Therefore, we propose TACKLE, a recipe for leveraging conceptual knowledge to *generate* images in an annotation-free and non-parametric manner.

Given a pretrained vision-language model, we denote the visual and linguistic encoder as $\mathcal{V}$ and $\mathcal{L}$, respectively. For a theoretically infinite web-image set that is noisy and unlabeled, which is denoted as $\mathscr{D}$, we sample a sufficiently large subset of the images $\mathcal{D}$ from $\mathscr{D}$. $\mathcal{V}$ is leveraged to project the images $\mathcal{I}$, $(\mathcal{I} \in \mathcal{D})$ into the feature embeddings $f_\mathcal{I} = \mathcal{V}(\mathcal{I})_{\mathcal{I}\in\mathcal{D}}$, where $f_\mathcal{I} \in \mathbb{R}^{|\mathcal{D}|\times d}$, d is the dimension of feature embeddings. We then instantiate assembled prompts $\mathcal{P}(k)$ for each k in target dataset categories K, and then the prompts are fed into the language encoder to obtain linguistic features $f_\mathcal{P} = \mathcal{L}(\mathcal{P}(k))_{k\in K}$, $f_{\mathcal{P}\in\mathbb{R}^{|K|\times d}}$. The probability

of class k for web images can be modeled as:

$$p_k = \frac{\exp\left(f_{\mathcal{P}(k)} f_{\mathcal{I}}^{\top}\right)/\tau}{\sum_{j=1}^{K} \exp\left(f_{\mathcal{P}(j)} f_{\mathcal{I}}^{\top}\right)/\tau}, \tag{3}$$

where $P = \{p_1, p_2, ..., p_K\} \in \mathbb{R}^{K \times |\mathcal{D}|}$ represents the probability that all collected web-images are of this class for each category in target dataset. Then, top n_k images are retrieved based on the probability p_k of P for category k:

$$\mathcal{D}_k = |\mathcal{D}|_{top\{p_k, n_k\}}, k \in K. \tag{4}$$

During the whole retrieval process, the pretrained weights of $\mathcal{V}$ and $\mathcal{L}$ are frozen and no parameters updating is required. Afterwards, we concatenate all $\mathcal{D}_k$, $(k \in K)$ with target dataset $\mathcal{D}_T$ as $\{\mathcal{D}_T, \mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_K\}$ for long-tailed target visual backbone training. The whole process is visualized in Fig. 1. TACKLE makes no presumption on the architecture of target visual backbone and is capable of incorporating any inductive bias into the design of long-tailed visual backbone.

The vast linguistic knowledge encoded in VL models is the key of tackling long-tailed distribution. Therefore, the TACKLE can be perceived as one variant of knowledge distillation (KD) that transferred knowledge from linguistic encoder into downstream visual backbone. Different from conventional KD methods that distill task-specific output logits or features, TACKLE employs free web-images as intermediate modality for distilling knowledge from pretrained VL model to downstream visual backbone. We suppose the TACKLE has a great potential in leveraging linguistic instructions to tackle imbalanced data distribution in a non-parametric and model-agnostic way.

3 Experiments

3.1 Experiment Setup

Datasets. We conduct our experiments on three long-tailed benchmark datasets, namely ImageNet-LT [22], Places-LT [22], and iNaturalist-2018 (iNat) [36]. ImageNet-LT and Places-LT were first introduced in [22] for long-tailed recognition research. ImageNet-LT is a long-tailed dataset with 1,000 categories sampled from the original ImageNet [5] following the Pareto distribution with a power value of $\alpha = 6$. Places-LT is a long-tailed version of the original Places2 Database [44]. The training split of Places-LT contains with 184.5K images from 365 categories, with 4,980 images maximally per class and minimally 5 images per class. iNaturalist-2018 [36] is a real-world long-tailed dataset consisting of 437K images and 6 levels of label granularity (kingdom, genus etc.). The training and testing split of iNaturalist-2018 contains 437,513 and 24,426 samples respectively.

We also collect a repository of noisy images collected from website, mainly sourced from Conceptual 12M (CC12M) [1], Conceptual Captions 3M (CC3M) [31] and SBU Captions (SBU) [25]. The specific details of the datasets are illustrated in Appendix A.1 in supplementary material.

Implementation Details. We choose the pretrained weights of CLIP's visual and linguistic encoder to conduct BALLAD experiments. In the experiments of BALLAD, we vary among ResNet-50, ResNet-100, ViT-B/16, and ResNet-50×16, which is 16× computation

cost of ResNet-50 following the style of EfficientNet as introduced in [27] for visual encoder. The ResNet-50 is leveraged for all ablation studies by default unless specified. In TACKLE, we employ ViT-B/16 to extract features of external images memory for retrieval, and perform downstream training on ResNeXt and ViTs-like backbones respectively. All the specific configurations like *input resolution, optimizer, learning rate etc.* can be found in Appendix A.2 in supplementary material.

Evaluation Metrics. We evaluate the models for long-tailed recognition on the balanced test splits and report the commonly used top-1 classification accuracy of *all* classes. Following [16], we divide these classes into three subsets – *many-shot*, *medium-shot*, and *few-shot* categories. Specifically, *many-shot*, *medium-shot*, and *few-shot* are decided according to the amount of instances in each category, namely more than 100 images, 20-100 images, and less than 20 images, respectively.

3.2 Performance Comparison

In this section, we compare the performance of BALLAD and TACKLE with long-tailed recognition approaches that report state-of-the-art results on three benchmark datasets, *i.e.*, ImageNet-LT, Places-LT, and iNaturalist-2018. Also, zero-shot performance of CLIP is also compared with our BALLAD to demonstrate the value of our proposed finetuning method in Appendix of supplementary material.

ImageNet-LT. Table 1 shows the long-tailed recognition results on ImageNet-LT. We present BALLAD variants (ash grey color) with ResNet-50, ResNet-101, ResNet-50×16, and ViT-B/16 as the visual backbone. We can see that BALLAD is superior to previous methods with a relative large margin. For example, when comparing performance of ResNet-50 backbone, BALLAD achieves 67.2 top-1 accuracy on overall evaluation, surpassing previous state-of-the-art PaCo [4] by 7.0% even when PaCo is also initialized of CLIP pretrained weights. When gradually increasing the size of visual backbone, we find the performance of BALLAD also enjoys an improvement. It is worth noting that BALLAD with ResNet-50×16 achieves an accuracy of 76.5%.

As for TACKLE, it enjoys the merits of making no presumption of visual backbone, therefore we compare ResNeXt-50 [39] and -101 performance with other methods (aquamarine color in Table 1). Without any re-balancing strategy, TACKLE outperforms previous well-designed SOTA method by 2.4% and 2.5% on ResNeXt-50 and ResNeXt-101 respectively.

Places-LT. We further evaluate BALLAD and TACKLE on Places-LT dataset and report the results in Table 2. It is a commonly used scheme of previous approaches to pretrain their backbones on ImageNet-1k [5] full dataset first to enrich the visual representation before finetuning on Places-LT (♠ in Table 2). Under this scheme, our TACKLE acheives 42.6% accuracy, surpassing counterparts by 1.4% without any designed inductive bias on long-tailed distribution. Meanwhile, BALLAD can directly perform training on Places-LT thanks to the additional language supervision of contrastive vision-language models. As shown in ash grey color rows in Table 2, BALLAD beats the state-of-the-art model PaCo with ResNet-152 by +5.3%, achieving better performance with smaller visual backbone.

iNaturalist-2018. We also evaluate on iNaturalist-2018, a naturally long-tailed distribution dataset to demonstrate the value of BALLAD and TACKLE on real-world scenes. As is illustrated in Table 3, BALLAD boosts the recognition accuracy by 1.0% (74.2% *vs* 73.2%), 0.4% (74.2% *vs* 73.8%) compared with randomly and CLIP initialized PaCo respectively.

Method	ImageNet-LT	
	Backbone	overall
τ-normalized [16]	RN50	46.7
LWS [16]	RN50	47.7
Blanced Softmax [28]	RN50	55.0
RIDE [3]	RN50	55.4
PaCo [4]	RN50	57.0
τ-normalized [16]	RN50*	51.3
LWS [16]	RN50*	52.1
PaCo [4]	RN50*	60.2
BALLAD	RN50*	67.2
τ-normalized [16]	RX50	49.4
LWS [16]	RX50	49.9
ResLT [5]	RX50	52.9
Blanced Softmax [28]	RX50	56.2
RIDE [3]	RX50	56.8
PaCo [4]	RX50	58.2
TACKLE	RX50	60.6
TADE [31]	RX50	58.8
τ-normalized [16]	RX101	49.6
LWS [16]	RX101	50.1
ResLT [5]	RX101	55.1
Blanced Softmax [28]	RX101	58.0
PaCo [4]	RX101	60.0
TACKLE	RX101	62.5
BALLAD	RN101*	70.5
BALLAD	V-B/16*	75.7
BALLAD	RN50×16*	76.5

Table 1: Long-tailed recognition accuracy on ImageNet-LT for different methods and backbones. $*$ means initializing visual encoder with pretrained weights of CLIP.

Method	Places-LT	
	Backbone	overall
OLTR [23]	RN152♠	35.9
cRT [16]	RN152♠	36.7
τ-normalized [16]	RN152♠	37.9
LWS [16]	RN152♠	37.6
Blanced Softmax [28]	RN152♠	38.6
ResLT [5]	RN152♠	39.8
PaCo [4]	RN152♠	41.2
TACKLE	RN152♠	42.6
BALLAD	RN50*	46.5
BALLAD	RN101*	47.9
BALLAD	V-B/16*	49.5
BALLAD	RN50×16*	49.3

Table 2: Long-tailed recognition accuracy on Places-LT for different methods.

Method	iNaturalist-2018	
	Backbone	Accuracy(%)
OLTR [23]	RN50	63.9
LWS [16]	RN50	65.9
cRT [16]	RN50	67.6
τ-normalized [16]	RN50	69.3
LADE [11]	RN50	69.3
RIDE (2 experts) [3]	RN50	71.4
ResLT [5]	RN50	72.3
RIDE (4 experts) [3]	RN50	72.6
TADE [31]	RN50	72.9
PaCo [4]	RN50	73.2
TACKLE	RN50	74.4
PaCo [4]	RN50*	73.8
BALLAD	RN50*	74.2

Table 3: Long-tailed recognition accuracy on iNaturalist-2018 for different methods. $*$ means initializing visual encoder with pretrained weights of CLIP.

TACKLE achieves 74.4% with randomly initialized ResNet-50 backbone, surpassing the PaCo by 1.2% under the same configuration.

3.3 BALLAD Ablations

In this section, we first conduct extensive ablation studies to validate the design choices of BALLAD from aspects of finetuning, adapting and re-balancing.

Finetune the Vision-Language Model. To empirically discover how to finetune vision-language models contrastively in BALLAD, we probe the finetuning process by freezing the pretrained image and text encoder respectively. When both encoders are frozen, the model directly perform zero-shot predictions. From Table 4, we can easily find the following pattern – as more components are finetuned in CLIP, more accuracy improvement is obtained for *many-shot* categories whereas more accuracy drop happens in *few-shot* division. We hypothesize it is because the *many-shot* classes dominate the visual feature space during finetuning. Therefore, for finetuning phase (phase A), it is necessary to adapt CLIP on specific long-tailed dataset as much as possible, and we choose to finetune both the vision and language branches of CLIP.

Vision	Language	many	medium	few	overall
-	-	59.4	57.5	57.6	58.2
✓	-	70.4	**65.4**	**58.0**	**66.3**
-	✓	70.6	**65.4**	55.9	66.1
✓	✓	**71.3**	**65.4**	54.1	66.1

Table 4: Different methods of finetuning CLIP on ImageNet-LT.

Adapting	Decouple	overall acc
-	-	58.2
✓	-	66.0
✓	✓	**67.2**

Table 5: Influence of adapting and decoupling.

Backbones	many	medium	few	overall
DeiT-S [34]	49.7	22.7	6.3	30.8
CTN-L [7]	70.9	39.7	12.6	48.0
DeiT-S* [34]	73.2	59.3	**52.3**	63.7
CTN-L* [7]	**78.5**	**62.8**	50.2	**67.1**

Table 6: TACKLE performance of ViTs-like backbones on ImageNet-LT. * means training with additional retrieved images collected by TACKLE.

Backbones	CLIP	many	medium	few	overall
DeiT-S [34]	-	73.2	59.3	52.3	63.7
CTN-L [7]	-	78.5	62.8	50.2	67.1
DeiT-S [34]	✓	74.3	62.8	58.1	66.6
CTN-L [7]	✓	**78.8**	**65.5**	**55.9**	**69.3**

Table 7: Ensembling TACKLE and CLIP results on ImageNet-LT.

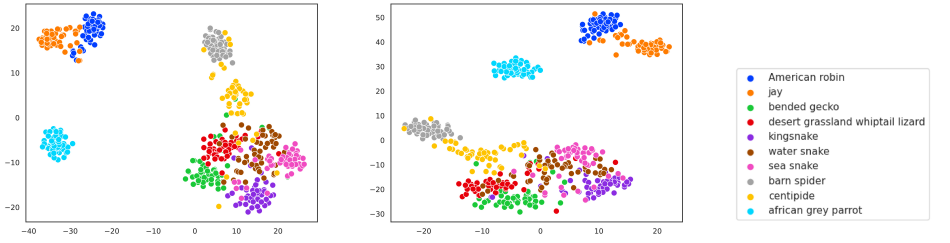
Decouple Finetuning and Adapting. As demonstrated in Sec. 2.2, we decouple the training of BALLAD into finetuning (Phase A) and adapting (Phase B). An alternative scheme is to jointly train the CLIP and linear adapter rather than decoupling the training processes. According to Table 5 the decoupled training of CLIP and linear adapter can largely boost the accuracy from 66.0% to 67.2%. We visualize the joint and decoupling training schemes using t-SNE [35] and present the results in the Fig. 2. Compared with joint training, decoupled training better separates the tail-class feature embeddings from head-classes. This demonstrates that the proposed decoupled training of vision-language model and adapter is effective to handle long-tailed distribution.

3.4 TACKLE Experiments

ViTs-like Backbone Results. An important merit of TACKLE design is making no presumption of backbone choices while leveraging the linguistic and conceptual knowledge of pretrained vision-language models. Therefore, we can flexibly design the visual backbones to explore more effective and efficient information aggregation mechanism for feature extraction. Table 6 varies the visual encoder in TACKLE with ViTs-like backbones. The training settings are the same with the implementation details of TACKLE. The additional images obtained from TACKLE-guided retrieval obviously replenish data deficiency by dramatically improving the *few-shot* performance under the same configuration, maximally by 46.0% and 37.6% on the DeiT-S [34] and CONTAINER-LIGHT [7] backbones, respectively.

Ensemble TACKLE with VL Model. Model ensemble strategy is also widely utilized in tackling long-tailed distribution [38, 41]. We validate that ensembling our TACKLE-trained downstream visual encoder with VL model like CLIP largely boosts the accuracy as results shown in Table 7. Logits independently generated by TACKLE classifier and cosine similarity matrix of CLIP are ensembled together proportionally to fuse both the in and out-of-distribution knowledge. Notably, significant boosts of *few-shot* classes imply the two kinds of knowledge is complementary, despite the TACKLE is trained under the guidance of VL models via retrieved images.

Conceptual Knowledge Transfer. We visualize some web images retrieved by TACKLE



(a) Decoupled Training (b) Joint Training (c) Legend

Figure 2: Comparisons of training vision-language model and linear adapter decoupled and jointly.

as shown in Appendix C.6 in supplementary material. The conceptual language encoders of VL models provide linguistic hints to help conceive visual objects, which is the key of success in BALLAD. However, we show that TACKLE can transfer the conceptual knowledge without finetuning as the retrieved images complement target datasets from various perspectives (e.g., cartoon, mock-up, and design diagram). This proves that diverse images can facilitate visual encoders to understand real-world concepts more comprehensively.

4 Conclusion

This paper aims to unleash the potential of pretrained Vision-Language models for long-tailed visual recognition. Specifically, we first propose a contrastive finetuning framework named BALLAD, which decouples the whole process into finetuning and adapting. At first stage, the pretrained visual and linguistic encoders are finetuned to take long-tailed distribution knowledge into account. Then, we adapt the finetuned VL model with a balanced linear adapter to re-balance the new knowledge. The adapting can be regarded as reserving out-of-distribution knowledge of pretrained VL models as the visual and linguistic backbones are frozen. Moreover, we propose a non-parametric strategy TACKLE to leverage VL model by retrieving $k-nearest$ samples from external memory under the guidance of linguistic hints. The scheme of retrieving enjoys merits of transferring conceptual knowledge from pretrained VL model into external images and making no presumption of downstream visual encoder. Extensive experiments on both artificial and real-world long-tailed datasets demonstrate the effectiveness of the proposed BALLAD and TACKLE approaches.

5 Acknowledgement

Yu Qiao is the corresponding author. This work was supported in part by the National Natural Science Foundation of China (Grant No. 62206272) and Shanghai Committee of Science and Technology (Grant No. 21DZ1100100).

References

[1] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[3] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *arXiv preprint arXiv:2101.10633*, 2021.

[4] 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.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

[7] Peng Gao, Jiasen Lu, Hongsheng Li, Roozbeh Mottaghi, and Aniruddha Kembhavi. Container: Context aggregation network. In *Advances in Neural Information Processing Systems*, 2021.

[8] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[11] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6626–6636, 2021.

[12] 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.

[13] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4233–4241, 2018.

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021.

[15] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019.

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

[17] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[21] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[22] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019.

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

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

[25] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[26] Vilfredo Pareto. *Cours d'économie politique*. Librairie Droz, 1964.

[27] 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*, 2021.

[28] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. *arXiv preprint arXiv:2007.10740*, 2020.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[30] Dvir Samuel and Gal Chechik. Distributional robustness loss for long-tail learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9495–9504, 2021.

[31] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[33] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Advances in Neural Information Processing Systems*, 2020.

[34] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 32–42, October 2021.

[35] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[36] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[37] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2020.

[38] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=D9I3drBz4UC.

[39] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[40] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2361–2370, 2021.

[41] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. *arXiv preprint arXiv:2107.09249*, 2021.

[42] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021.

[43] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

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

[45] 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.

[46] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press, Inc., 1949.