Text-Guided Mixup Towards Long-Tailed Image Categorization

Richard Franklin[1]
rsamfranklin@gmail.com
Jiawei Yao[1]
jwyao@uw.edu
Deyang Zhong[1]
dyzhong@uw.edu
Qi Qian[2]
qi.qian@alibaba-inc.com
Juhua Hu[1]
juhuah@uw.edu

[1] School of Engineering and Technology
University of Washington
Tacoma, WA

[2] Alibaba Group
Bellevue, WA

Abstract

In many real-world applications, the frequency distribution of class labels for training data can exhibit a long-tailed distribution, which challenges traditional approaches of training deep neural networks that require heavy amounts of balanced data. Gathering and labeling data to balance out the class label distribution can be both costly and time-consuming. Many existing solutions that enable ensemble learning, re-balancing strategies, or fine-tuning applied to deep neural networks are limited by the inert problem of few class samples across a subset of classes. Recently, vision-language models like CLIP have been observed as effective solutions to zero-shot or few-shot learning by grasping a similarity between vision and language features for image and text pairs. Considering that large pre-trained vision-language models may contain valuable side textual information for minor classes, we propose to leverage text supervision to tackle the challenge of long-tailed learning. Concretely, we propose a novel text-guided mixup technique that takes advantage of the semantic relations between classes recognized by the pre-trained text encoder to help alleviate the long-tailed problem. Our empirical study on benchmark long-tailed tasks demonstrates the effectiveness of our proposal with a theoretical guarantee.

1 Introduction

In recent years, deep learning has made state-of-the-art advancements in computer vision tasks such as image categorization, object detection, and semantic segmentation [20, 38]. Deep learning models are highly dependent on large-scale and balanced training data, but real-world data are typically class-imbalanced [2, 21, 32]. When training data is abundant for a subset of classes (i.e., head classes) but scarce for the other (i.e., tail classes), the distribution of the data is said to be long-tailed [42]. Taking image categorization as an example, deep neural networks (DNNs) aim to minimize the empirical risk on the training

data by incrementally adjusting the learnable parameters. However, given a long-tailed training data, this happens more on the head-class instances that appear more frequently, augmenting the model's performance bias towards head classes but reducing the model's generalization performance on tail classes [2, 35].

Long-tailed learning proves to be a significantly challenging task as addressed by many previous studies [23, 35, 37, 41, 44]. Intuitively, under-sampling the head classes and over-sampling the tail classes is a reasonable technique. Although class-level re-sampling or re-weighting can help balance out the data distribution and mitigate the model's performance bias on head classes, these techniques can cause the model's overfitting on tail classes and/or degenerate the performance on head classes [27]. There is evidently more success in module improvement techniques [12, 25, 42], especially those that use ensemble learning [35, 41, 44]. There are a number of additional techniques [42] that aim to mitigate the long-tailed problem such as class-level re-margining [2], data augmentation [4, 24], and transfer learning [39]. However, these methods are still limited by the scarce information found among tail classes.

Recently, vision-language models such as CLIP [26] and ALIGN [13] have demonstrated good performance in zero-shot classification and few-shot learning [3]. These models are trained on large-scale data containing image-text pairs that elicit the forming of connections between text and image embedding. By capturing the contrastive locality of image and text features, vision-language models can generalize to unseen categories well, which is a potential information source of tail classes in long-tailed learning. However, existing multi-modal works [8, 14, 22, 31] are limited by the general domain knowledge of the CLIP's pre-trained text encoder and must continue linguistic training on the downstream task.

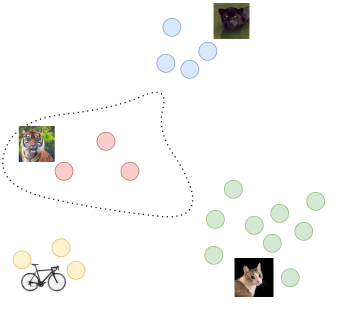


Figure 1: The decision boundary of 'tiger' stretches towards that of 'leopard' and 'cat' and away from 'bicycle' as text-guided mixup allows semantically similar classes to be mixed more frequently.

In this work, we propose to leverage the frozen CLIP text encoder to obtain prompt embedding as additional supervision for long-tailed learning in vision tasks. Considering the observation that semantic relationships between class names (e.g., 'tiger' and 'cat') correlate with their localities of visual features in vision-language models, we can utilize semantically similar classes to assist the generalization among tail classes (e.g., the head class 'cat' can help assist the tail class of 'tiger' as shown in Fig. 1). However, the intra-class variance of the tail class can still be ignored. Therefore, we further propose a novel text-guided mixup strategy, named local feature mixup (LFM), to shift the label towards tail classes, so as to alleviate the long-tailed problem. The main contributions of this work are summarized as follows.

- We leverage the frozen CLIP text encoder to enhance the performance of long-tailed visual recognition tasks.
- We construct a novel mixup technique that takes advantage of the text encoder to boost the performance of tail classes with a theoretical guarantee.
- Our extensive experiments on several benchmark long-tailed data demonstrate the effectiveness of our proposal.

2 Related Work

In long-tailed visual recognition, numerous methods have been proposed to boost the performance of tail classes [42]. Module improvement methods including **ensembling** have shown recent success [15, 35, 41, 44]. In mixture of experts, TADE [41] and SHIKE [15] output an aggregation of multiple expert modules, where each expert in TADE strives to perform well in a different training distribution, and each expert in SHIKE focuses on modeling a different depth of image features. Although ensembling can boost performance, these methods are still limited by the scarce information found among instances of the tail classes.

Moreover, **class re-balancing** such as class-level re-sampling [6], re-weighting [5] (e.g. Balanced Cross Entropy [27]), and re-margining (e.g., LDAM [2]) can adjust the model's attention to classes with a lower sample rate. However, class-balanced sampling or re-weighting can lead to overfitting of the tail classes, under-represent the intra-class variance of the head classes [27, 42], and thus decrease the model's overall performance [30]. Alternatively, it can be effective to train a model with meta sampling [27], in which the optimal sample rate per class is estimated by applying a learnable parameter for each class label. Using this method can slightly avoid the overfitting of tail classes, but finding the optimal parameter or trade-off between class labels for multi-class classification is difficult.

Another instance of success is found through **pre-training** vision transformers [9, 20] in an autoencoder setup [11, 17, 28]. Once the encoder is sufficiently trained, it feeds into a classification layer that is trained using a balanced binary cross-entropy loss [2]. However, these methods still lack sufficient performance on the set of tail classes as it is an inert challenge to train deep neural networks for classes with small sample rates. Recently, pre-trained vision-language models like **Contrastive Language-Vision Pre-training** (CLIP) [26] have demonstrated strong zero-shot performance. CLIP embodies multi-modal learning through unsupervised training of image-caption pairs available on the wild web to capture the contrastive locality of image and text features. This makes CLIP more adaptable to new tasks, so that they can be leveraged to make zero-shot predictions, that is, generalize to unseen categories. Thereafter, a pre-trained vision-language model can be further fine-tuned on a downstream task in few-shot learning [1] or long-tailed learning (e.g., RAC [22], VL-LTR [31], LPT [8], TeS [34], and VPT [14]). However, most of them have been focusing more on the text encoder. For example, VL-LTR [31] requires manually retrieving text descriptions of each class from the Internet to augment the text data in preparation for linguistic training, which is resource expensive, so we instead freeze the text encoder.

3 The Proposed Method

Given a long-tailed training data $D = \{(x_i, y_i)\}$, x_i is an image associated with its target class $y_i \in \{1, \ldots, C\}$. We construct a set of text snippets T, where each T_k describes a class label for $k \in \{1, \ldots, C\}$. For example, the text snippet describing class name "dog" is a tokenized sequence generated from the string as "a photo of a dog".

We feed image and text snippets T to the image and text encoders, respectively, pre-trained by CLIP [26] as shown in Fig. 2, for which we denote as $\mathscr{F}_I$ and $\mathscr{F}_T$, respectively. Both of these encoders output feature vectors of size d. We denote the output from the text encoder as f_T and $f_T = \mathscr{F}_T(T)$ and allow f_{T_k} to denote the feature vector for class k, which does not change during the long-tailed learning. To better separate the tail-class feature embeddings from that of the head-classes following [23], we append a fully connected layer $W_I \in \mathbb{R}^{d \times d}$

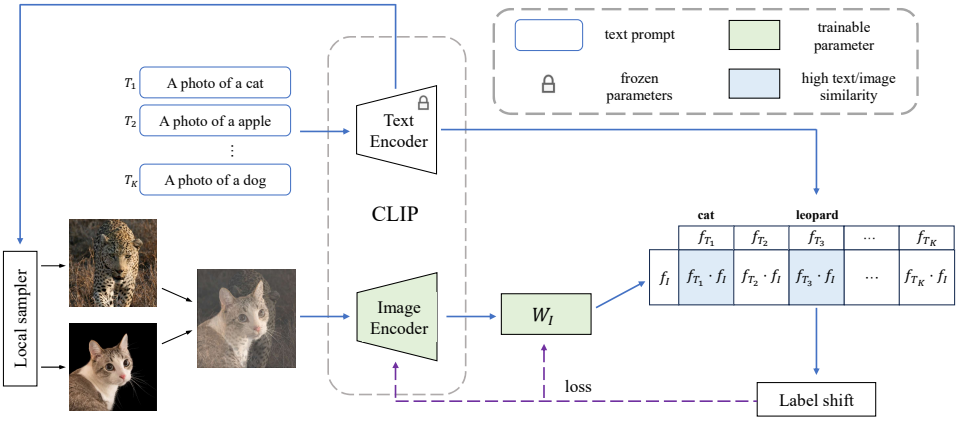


Figure 2: The proposed model architecture, in which the text encoder is fixed using the pre-trained model by CLIP [26]. The image encoder will be fine-tuned according the downstream task and $W_I \in \mathbb{R}^{d\times d}$ is appended and learnable.

that is learnable to $\mathscr{F}_I$. Thereafter, we can extract the feature vector for each image x_i as $f_I = W_I \mathscr{F}_I(x_i)$. Additionally, we normalize both f_{T_k} and f_I to be of a unit norm.

After obtaining f_I and f_T, image classification is performed as shown in Fig. 2 by computing the cosine similarity between f_I and f_{T_k} for all k, and finally, the predicted class label, $\hat{y}$, for each image is computed as $\hat{y} = \arg\max_{k\in\{1,\ldots,C\}} f_I \cdot f_{T_k}$.

Thereafter, we can adopt a decoupled training approach as suggested by [23] to learn better embeddings for tail-classes compared to joint training. In stage 1, we open $\mathscr{F}_I$ and freeze W_I for training, and in stage 2, we freeze $\mathscr{F}_I$ and open W_I. At the beginning, W_I is initialized as the identity matrix with $f_I = \mathscr{F}_I(x)$. However, by minimizing the empirical risk directly based on the training data with a long-tailed distribution, both $\mathscr{F}_I$ and W_I can still be biased to the head classes. Therefore, we propose a novel text-guided mixup technique.

Local Feature Mixup A statistical measure of class imbalance in a dataset can be defined as the imbalance factor $\gamma = n_1/n_C$, where n_k is the number of examples in class k and $n_1 \geq n_2 \geq \ldots \geq n_C$ is ordered from high to low, and typically, we have $n_1 \gg n_C$. Our main goal is to increase the few-shot accuracy (i.e., those with low n_k), while not attenuating the model's accuracy on many-shot classes (i.e., those with high n_k). We strive to boost the few-shot accuracy by making two assumptions about the data. First, we assume that classes with low n_k are underrepresented because a few examples may not fully express the complete diversity (or variance) of their associated class. For example, a cat can look different from another cat in terms of their features such as their sizes, their eye colors, and the color/pattern of their furs. When limited to observing a few examples of cats, it is difficult for DNNs to grasp the full range of features that a cat can express. Therefore, we assume that every tail class has a larger intra-class variance than that can be learned from long-tailed data.

Secondly, because both CLIP's image and text encoders map their respective inputs to d-dimensional feature vectors, we say that every class can be represented by certain feature space in $\mathbb{R}^d$. The pre-trained text encoder already has an understanding of the local relationships between words. For example, words "frog" and "toad" are close in the language model feature space, since they have similar meanings. Part of our learning objective is to closely align the outputs of our image encoder to the outputs of the pre-trained language model. That is, if we

feed an image of a frog and an image of a toad to our image encoder, their extracted feature vectors should be close in proximity as in the text feature space. Therefore, we also assume that if two classes have similar meanings (i.e., nearby in the text encoder's feature space), these two classes also share a subset of visual features and thus should also be nearby within the image encoder's feature space. In the following construction of local feature mixup, we incorporate these two critical ideas separately, that is, local sampling and label shift.

Local Sampling Existing mixup strategies often randomly sample y_i and y_j uniformly across the training data [4, 33, 40]. However, we aim to choose pairs that are semantically related supervised by the pre-trained text encoder. First, we sample an instance from class y_i uniformly across the training data as $p(y = y_i) = \frac{n_i}{\sum_k^C n_k}$. Then, we sample another instance from class y_j with probability $p_{ls}(y = y_j|y_i)$ given by Eqn. (1).

$$p_{ls}(y = y_j|y_i) = \begin{cases} \frac{exp(f_{T_i} \cdot f_{T_j}/\tau)}{\sum_{k=1}^C exp(f_{T_i} \cdot f_{T_k}/\tau)} & i \neq j \\ 0 & \text{o.w.} \end{cases} \tag{1}$$

where the hyperparameter $\tau > 0$ controls the temperature scaling on the softmax equation. A lower τ increases the likelihood that similar class pairs are chosen for mixup, but a too low temperature can lead to oversampling of nearby classes. We set $\tau = 0.05$ for most experiments. Using this strategy, we hope to extend the variance of minor class samples towards neighboring classes as our assumption is that semantically similar classes share a subset of visual features as depicted in Fig. 1.

Label Shift Then, we perform mixup by mixing images x_i and x_j sampled through our above local sampling method. With mixing factors $\lambda_x, \lambda_y \in [0, 1]$, we propose

$$\tilde{x}^{LFM} = \lambda_x x_i + (1 - \lambda_x) x_j$$
$$\tilde{y}^{LFM} = \lambda_y y_i + (1 - \lambda_y) y_j$$

where y_i, y_j are one-hot vectors and factor λ_x is chosen randomly from the beta distribution. More importantly, we generate λ_y by

$$\lambda_y = \text{clamp}\left(\lambda_x - \alpha \frac{n_i - n_j}{n_i + n_j}, 0, 1\right) \tag{2}$$

where hyperparameter $\alpha \geq 0$ adjusts the intensity of label shift and the resulting value is clamped between 0 and 1. In order to expand the margin for tail classes, we shift the decision boundary away from tail classes and towards head classes according to the difference of n_i and n_j. For example, if $n_i > n_j$ (i.e., class y_i has more samples than class y_j), we shift the target to be more in favor of class y_j, thus increasing the model's margin on the class with fewer samples. Algorithms are summarized in the supplementary and we provide a theoretical guarantee for our proposal as follows, while the overall framework is illustrated in Fig. 2.

Theorem 1 *Letting $p = n_i/(n_i + n_j)$, λ_y can be obtained by balancing the distribution between x_i and x_j*

$$\lambda_y = \arg\min_{\lambda \in [0,1]} (\lambda - \lambda_x)^2/2 + \alpha R(\lambda)$$

where $R(\lambda) = (\lambda - 1/2)^2 - (\lambda - p)^2$.

Remark The former term constrains that the obtained weight for the label should be close to the weight for the example, while the latter term is a balance regularization to incorporate the prior distribution p between two examples. By minimizing the regularization, it aims to push λ from the imbalanced initial distribution to a balanced one. When $p = 1/2$, it degenerates to the standard weight for mixup.

4 Experiments

To demonstrate the proposed LFM method, following the common practice in long-tailed learning, we use publicly available long-tailed datasets, that is, CIFAR10-LT and CIFAR100-LT [18], ImageNet-LT [21], and Places-LT [43].

Experiment Setup For CIFAR10/100-LT, we fine-tune CLIP with a single GPU, and for ImageNet-LT and Places-LT, we fine-tune CLIP with three GPUs. Each GPU is an Nvidia GeForce RTX 2080 Ti with 11GB of memory. During training, each GPU receives a batch size of 32, so for ImageNet-LT and Places-LT the effective batch size is 96. Training is performed with a fixed seed to allow for reproduceability. The hyperparameters chosen for LFM are fixed (i.e., $\alpha = 1$, $\tau = 0.05$) for all experiments on CIFAR10-LT, CIFAR100-LT, and ImageNet-LT, while they are adjusted on Places-LT as $\alpha = 1.25$, $\tau = 1.00$ in stage 1 and $\alpha = 1.50$, $\tau = 1.00$ in stage 2, due to the imbalance severity as explained in the next section. Low learning rates were picked to avoid the risk of catastrophic forgetting and losing CLIP's zero-shot performance advantage. The detailed hyperparameters used can be found in the supplementary. CLIP's default text prompt template is "a photo of a {CLASS}". For all experiments, we utilize the default text prompt template provided.

A model's performance is not necessarily stable across all classes, each with different sample counts, so it is important that we quantify the performance of our model in subdivisions relative to every n_k. Across all datasets, we subdivide the resulting model's accuracy into four categories, namely many-shot, medium-shot, few-shot, and overall following [16]. Many-shot classes have $n_k > 100$, medium-shot classes have $20 \leq n_k \leq 100$, and few-shot classes have $n_k < 20$. For each performance category, we report the top-1 accuracy of our model against the balanced validation set for each subdivision of our chosen datasets.

We compare our proposed method with vision-focused baseline methods and strategies that perform well in tackling the long-tailed problem. We also fine-tune the competitive image encoder (i.e., ViT-B/32) with different existing losses as baselines, i.e., Cross Entropy (CE), Balanced Cross Entropy (BalCE) [27], Focal [19], Label Distribution Aware Margin (LDAM) [2], and Margin Metric Softmax (MMS) [29]. All losses except CE were proved to be helpful for the class imbalance problem. In summary, we compare with the following baselines based on the pre-trained CLIP [26]: 1) Zero-shot: The pre-trained image and text encoders by CLIP [26] are directly used to do prediction on the balanced test data, in which ViT-B/32 is adopted; 2) CE: Fine-tuned ViT-B/32 using the cross entropy loss; 3) BalCE: Fine-tuned ViT-B/32 using the balanced loss [27]; 4) Focal: Fine-tuned ViT-B/32 using the Focal loss [19]; 5) LDAM: Fine-tuned ViT-B/32 using the loss in LDAM [2]; and 6) MMS: Fine-tuned ViT-B/32 with MMS [29].

CIFAR10/100-LT As in the literature, we can create CIFAR10-LT and CIFAR100-LT by taking a subset of the original balanced CIFAR10 and CIFAR100 datasets [18], and the imbalance factor γ is variable. We experiment with multiple imbalance factors in $\{10, 50, 100\}$.

Table 1: CLIP accuracy on CIFAR100-LT with imbalance factor 100, where the best for each is in bold.

Methods	Many	Med	Few	All
Zero-shot	63.5	60.8	61.4	62.0
CE	79.3	67.4	53.9	67.5
BalCE	74.6	69.8	57.4	67.6
Focal	80.2	65.0	54.0	66.9
LDAM	81.6	70.4	58.1	70.5
MMS	**90.3**	75.2	58.1	75.2
LFM + CE	81.2	79.6	68.6	77.3
LFM + MMS	81.0	**81.3**	**76.5**	**79.4**

First, on CIFAR100 with the imbalance factor of 100, we compare all methods based on CLIP. For our proposal, we set ViT-B/32 as the backbone and apply LFM with two different losses, i.e., cross entropy and MMS [29] that is the best in the literature. The comparison results to all baselines are summarized in Table 1. Based on the zero-short performance, we can observe that the pre-trained CLIP can help balance the performance in different categories, which demonstrates the effectiveness of pre-trained vision-language model to alleviate the class imbalance issue. Then, by fine-tuning the pre-trained image encoder, the overall accuracy can be improved. However, due to the severe imbalance, the performance of the tail classes is still lacking even when balanced losses are utilized. Our proposal can help improve the accuracy in all categories, where LFM combined with a loss well-suited for CLIP can further help improve the performance.

Table 2: Overall accuracy on CIFAR10/100-LT with varying imbalance factors (IF). The best is in bold and the 2nd best is underlined. '-' indicates that the accuracy is not available in the original paper.

Dataset	CIFAR10-LT			CIFAR100-LT		
IF	100	50	10	100	50	10
BBN [42]	79.8	82.2	88.3	42.6	47.0	59.1
LDAM [2]	77.0	-	88.2	42.0	-	58.7
LiVT [37]	86.3	-	91.3	58.2	-	69.2
RIDE [35]	-	-	-	48.0	51.7	61.8
SHIKE [15]	-	-	-	56.3	59.8	-
TADE [41]	-	-	-	49.8	53.9	63.6
GLMC [10]	87.8	90.2	94.0	57.1	62.3	72.3
MARC [36]	85.3	-	-	50.8	-	-
CLIP (ViT-B/32)						
CE	89.8	90.0	91.6	67.5	68.1	70.4
BalCE	91.3	91.6	92.4	67.6	68.8	70.8
Focal	89.8	90.0	91.6	66.9	68.6	70.4
LDAM	89.7	91.5	94.6	70.5	72.1	77.2
MMS	<u>93.3</u>	<u>94.5</u>	94.4	<u>75.2</u>	<u>77.5</u>	<u>82.0</u>
LFM + CE	**93.8**	**95.2**	<u>96.6</u>	<u>77.3</u>	<u>78.2</u>	<u>82.6</u>
LFM + MMS	90.0	91.0	**97.0**	**79.4**	**81.1**	**85.7**

Then, we compare the fine-tuned CLIP models (ViT-B/32 is adopted) including our proposal with multiple existing state-of-art long-tailed learning methods in Table 2 under

different imbalance factors. We can observe that by fine-tuning the pre-trained CLIP image encoder, the performance can be significantly improved in all scenarios. Moreover, the state-of-the-art imbalance loss MMS [29] is very helpful, while our proposal can further significantly improve the performance in most cases. This further demonstrates the proposal of alleviating the class imbalance problem using pre-trained vision-language model and the effectiveness of LFM. It should be noted that methods using the backbone of ResNet50 are performing worse in general, and thus ResNet50 is not adopted in the following experiments.

ImageNet-LT and Places-LT We construct ImageNet-LT [21] by forming a subset of the ImageNet 2014 dataset [7]. The resulting imbalance ratio of ImageNet-LT is 256. As shown in Table 3, we can observe that compared to existing methods, our method begets better performance especially on few-shot accuracy (i.e., for tail classes) by both rebalancing and leveraging semantic similarities of classes. Observing that the LFM + MMS performance for minor classes falls behind LFM + CE, we hypothesize that MMS's sole focus on exercising semantic similarities and ignorance of class sample frequencies may overfit the many classes. A technique that only focuses on one may be problematic for tasks where semantic similarities happens to exist more frequently among the many classes.

Table 3: Performance comparison on ImageNet-LT. The best is in bold and the 2nd best is underlined. '-' means not available in the original paper.

Methods	Many	Med	Few	All
CE [5]	64.0	33.8	5.8	41.6
LDAM [2]	60.4	46.9	30.7	49.8
LiVT [37]	<u>76.4</u>	59.7	42.7	63.8
RIDE [35]	68.3	53.5	35.9	56.8
SHIKE [15]	-	-	-	59.7
TADE [41]	66.5	57.0	43.5	58.8
GLMC [10]	70.1	55.9	45.5	57.2
MARC [36]	60.4	50.3	36.6	52.3
CLIP (ViT-B/16)				
Zero-shot	69.2	66.8	<u>65.8</u>	67.6
LFM + CE	69.8	**71.8**	**68.7**	<u>70.6</u>
LFM + MMS	**79.7**	<u>71.4</u>	51.3	**71.7**

In addition, we conduct experiments on Places-LT [21] using LFM with CE and MMS. Places-LT is a long-tailed subset of the original dataset Places2 [43]. It is a dataset for scene classification containing 365 classes, and it suffers from extreme imbalance ($\gamma = 996$). To account for its imbalance severity, we adjust local feature mixup hyperparameters to be highly in favor of the minority classes. We increase the value of τ, so that the probability distribution constructed by local sampling is more balanced. Additionally, we increase the value of α, so that the label is shifted to the tail classes, more heavily as shown in the supplementary.

Table 4 summarizes the results. The benefit from the pre-trained model by CLIP can be observed from the zero-shot performance on tail classes, which further demonstrates the advantage of the text supervision from CLIP. However, fine-tuning using our proposal is necessary to improve the performance. It should also be noted that due to the severe imbalance factor of this data, our proposal with CE is expected to be less effective compared to that with MMS [29]. LFM with MMS shows significantly better performance compared to

Table 4: Performance comparison on Places-LT. The best is in bold and the 2nd best is underlined. '-' indicates that the accuracy is not available in the original paper.

Methods	Many	Med	Few	All
CE [6]	<u>45.7</u>	27.3	8.2	30.2
Focal [19]	41.1	34.8	22.4	34.6
LiVT [37]	**50.7**	42.4	27.9	<u>42.6</u>
SHIKE [15]	43.6	39.2	44.8	41.9
TADE [41]	43.1	42.4	33.2	40.9
MARC [36]	39.9	39.8	32.6	38.4
CLIP (ViT-B/16)				
Zero-shot	36.8	35.8	45.1	38.1
LFM + CE	41.3	<u>43.5</u>	<u>46.2</u>	42.3
LFM + MMS	45.2	**48.5**	**46.6**	**46.9**

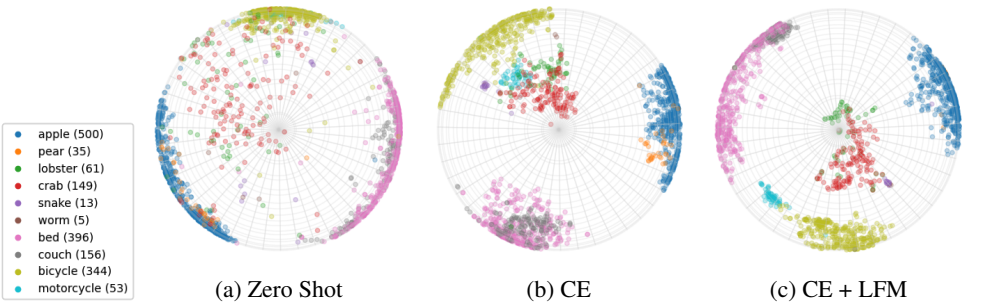
state-of-the-arts, especially on medium-shot and few-shot classes, and demonstrates strong performance on many-shot classes as well. This further demonstrates the effectiveness of our proposed method on the long-tailed problem.

Effect of Mixup Techniques To demonstrate the proposed LFM, we also compare it with the standard Mixup [40] and Remix [4] on CIFAR100-LT. Specifically, we fine-tune CLIP with the same hyperparameters and decoupled stages, using different mixup techniques. Each model is trained using cross entropy loss with the ViT-B/16 backbone. Remix is a mixup method that addresses the class-imbalance issue, and it makes a trade-off between many-shot and few-shot performances. For example, compared to the standard mixup, Remix can help improve the performance on few-shot classes but sacrifice the performance on many-shot classes. However, Remix ignores the semantic relationship between each class pair that CLIP can be used for. Our proposal shows significantly better performance in Table 5.

Table 5: Comparison of different mixup techniques on CIFAR100-LT with imbalance factor of 100.

Methods	Many	Med	Few	All
Mixup [40]	80.4	71.5	55.1	69.5
Remix [4]	79.6	71.5	55.7	69.4
LFM	**83.4**	**83.3**	**72.8**	**80.1**

Visualization To demonstrate our assumption on the local semantic relationship, we illustrate the geometric effect of fine-tuning CLIP with our proposal in Fig. 3. We demonstrate the effect by revealing the contrastive locality of image feature vector outputs, where the input is comprised of a set of randomly sampled images from 10 chosen classes {apple, pear, ..., motorcycle} in CIFAR100. Our illustration contains the following 5 pairs of semantically related categories from CIFAR100: (apple, pair), (lobster, crab), (snake, worm), (bed, couch), and (bicycle, motorcycle). The legend contains the name of the class and the sample count in parenthesis. We choose these pairs to show that their semantic relations are aligned with their visual relations in terms of contrastive locality, as perceived by the image encoding layers.



(a) Zero Shot (b) CE (c) CE + LFM

Figure 3: An illustration on image distribution of sampled classes using feature vectors extracted from different image encoders.

With the ViT-B/32 vision encoder and fully connected layer, we obtain a 512-dimensional feature vector for each image. To reduce the high-dimensional feature vectors to three dimensions for human readability, we convert them using t-SNE trained for 1000 iterations and seed set to 1. At zero shot, we can observe that semantically related classes are located nearby (e.g., 'apple' vs. 'pear' and 'bed' vs. 'couch'), although some are poorly clustered (e.g., 'lobster' vs. 'crab'). This confirms our assumption that pre-trained vision-language model can align semantically related classes together. However, the separation between non-related classes are not clear in the pre-trained model. By fine-tuning the image encoder with cross entropy loss, the separation between non-related classes becomes clear, thanks to the help of head-class training data. However, we can observe that tail-class instances are largely overlapping with semantically related head-class instances (e.g., 'lobster' vs. 'crab'). Fortunately, by incorporating our proposal of LFM, tail-class instances can be pushed a bit away from their semantically related head-class instances without sacrificing the clear boundaries between non-related classes, which further demonstrates our proposal.

5 Conclusion

Considering CLIP's ability to generalize to unseen categories, we leverage a fixed text encoder to enhance the performance of image categorization over long-tailed training distributions. We enable the accuracy boost with the construction of a novel mixup technique that takes advantage of the semantic relationships between classes by probabilistic sampling based on their locality in the text encoder's feature space and slightly shift the label towards tail classes. Our extensive experiments on several benchmark long-tailed training data demonstrate the effectiveness of our proposal in alleviating the class imbalance issue with an efficient strategy that incorporates a fixed text encoder. Local feature mixup can be easily applied to not only vision-language backbones but also non multi-modal methods (i.e. vision-only architectures), which will be studied in our future work. However, both LFM and vision-language image classification are limited by the domain knowledge of the text encoder. Without further training, pre-trained CLIP performs poorly on domain-specific tasks as suggested by [8, 14, 31] due to its generic knowledge. Our method relies on the ability of the text encoder to capture pairwise semantic similarities among the class names present in the dataset which proves to be performant for the common domain such as CIFAR and ImageNet but not for biological names present in iNaturalist [32], which will be our future work.

6 Acknowledgement

Yao and Hu's research is supported in part by NSF (IIS-2104270) and Advata Gift Funding. Zhong's research is supported in part by the Carwein-Andrews Graduate Fellowship and Advata Gift Funding. All opinions, findings, conclusions and recommendations in this paper are those of the author and do not necessarily reflect the views of the funding agencies.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[2] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[3] Jiaoyan Chen, Yuxia Geng, Zhuo Chen, Ian Horrocks, Jeff Z. Pan, and Huajun Chen. Knowledge-aware zero-shot learning: Survey and perspective. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4366–4373. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/597. Survey Track.

[4] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: rebalanced mixup. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 95–110. Springer, 2020.

[5] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[6] Damien Dablain, Bartosz Krawczyk, and Nitesh V. Chawla. Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2022. doi: 10.1109/TNNLS.2021.3136503.

[7] 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, 2009. doi: 10.1109/CVPR.2009.5206848.

[8] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. Lpt: Long-tailed prompt tuning for image classification, 2023.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold,

Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[10] Fei Du, Peng Yang, Qi Jia, Fengtao Nan, Xiaoting Chen, and Yun Yang. Global and local mixture consistency cumulative learning for long-tailed visual recognitions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15814–15823, 2023.

[11] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[12] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5375–5384, 2016. doi: 10.1109/CVPR.2016.580.

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[14] 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.

[15] Yan Jin, Mengke Li, Yang Lu, Yiu-ming Cheung, and Hanzi Wang. Long-tailed visual recognition via self-heterogeneous integration with knowledge excavation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23695–23704, 2023.

[16] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *CoRR*, abs/1910.09217, 2019.

[17] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[18] Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[21] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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

[23] 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. *CoRR*, abs/2111.14745, 2021.

[24] Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. In *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, pages 117–122, 2018. doi: 10.1109/IIPHDW.2018.8388338.

[25] Wanli Ouyang, Xiaogang Wang, Cong Zhang, and Xiaokang Yang. Factors in finetuning deep model for object detection with long-tail distribution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 864–873, 2016. doi: 10.1109/CVPR.2016.100.

[26] 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.

[27] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020.

[28] Dvir Samuel, Yuval Atzmon, and Gal Chechik. From generalized zero-shot learning to long-tail with class descriptors. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 286–295, 2021.

[29] Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing clip to out-of-distributions. In *International Conference on Machine Learning*, pages 31716–31731. PMLR, 2023.

[30] Saptarshi Sinha, Hiroki Ohashi, and Katsuyuki Nakamura. Class-wise difficulty-balanced loss for solving class-imbalance. In *Proceedings of the Asian conference on computer vision*, 2020.

[31] 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.

[32] 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.

[33] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR, 2019.

[34] Junyang Wang, Yuanhong Xu, Juhua Hu, Ming Yan, Jitao Sang, and Qi Qian. Improved visual fine-tuning with natural language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11899–11909, October 2023.

[35] 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.

[36] Yidong Wang, Bowen Zhang, Wenxin Hou, Zhen Wu, Jindong Wang, and Takahiro Shinozaki. Margin calibration for long-tailed visual recognition. In *Asian Conference on Machine Learning*, pages 1101–1116. PMLR, 2023.

[37] Zhengzhuo Xu, Ruikang Liu, Shuo Yang, Zenghao Chai, and Chun Yuan. Learning imbalanced data with vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15793–15803, 2023.

[38] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[39] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Long Beach, CA, June 2019.

[40] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[41] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. *Advances in Neural Information Processing Systems*, 35:34077–34090, 2022.

[42] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023.

[43] 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*, 2017.

[44] 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.