

RetFormer: Multimodal Retrieval for Enhancing Image Recognition

Tianrui Yu, Xiubo Liang*, Hongzhi Wang
 School of Software Technology
 Zhejiang University
 Ningbo, China

tianruiyu@zju.edu.cn, xiubo@zju.edu.cn, whzzju@zju.edu.cn

Abstract

The rapid advancements in Transformers and the availability of high-quality multimodal datasets have enabled deep neural networks to achieve remarkable performance in vision and language tasks. However, catastrophic forgetting and the long-tailed distribution of real-world data make it non-trivial to apply these advances in real-world applications. We propose to explore the internal structure of the neural network for modeling and utilizing relationships between multimodal samples. Specifically, we introduce RetFormer, which incorporates a retrieval cross-fusion module connected to an external knowledge base. RetFormer establishes a robust relationship between image and text modalities by integrating information from external knowledge bases into the model's decision-making process, thus overcoming the limitations of traditional approaches on fine-tuning paradigm and datasets. Our experiments demonstrate the benefits of integrating large-scale image-text datasets into vision tasks and exemplify the importance of modeling the relationship between image and text modalities. We have evaluated our approach on long-tailed recognition and learning with noisy labels and have shown that it achieves state-of-the-art accuracies.

1. Introduction

The conjunction of large-scale Transformers with extensive pre-trained datasets has met with significant success in both vision and NLP domains. Large language models such as GPT series [29] and LLaMA series [40], along with multimodal large language models like Gemini [38], and Claude 3 [1], have demonstrated state-of-the-art performance across various downstream tasks, owing to their robust capabilities in understanding, reasoning, and generation. According to the scaling law [17], extensive multimodal and multilingual datasets such as LAION [33] and

DataComp [10] are as important as large Transformers. However, this approach relies on high-quality training samples and a large number of model parameters, while catastrophic forgetting further complicates its deployment in real-world scenarios.

In the training of large-scale Transformers, world knowledge is implicitly encoded within the vast number of model parameters, which can exacerbate certain challenges inherent in the current machine-learning paradigm. These challenges include catastrophic forgetting, difficulties in model updating, limited interpretability, and scalability issues. Furthermore, real-world data often features a long-tailed distribution along with noisy labels. These issues often occur simultaneously, complicating the estimation of the true distribution of the dataset. While previous research [44, 51] efforts aimed at addressing these challenges have made significant contributions, most of them have been limited to image-centered solutions [43], and the integration of retrieval-augmented into the field of vision recognition also remains an underexplored area. It remains a formidable challenge to devise a unified, flexible, and potent approach that can model the relationships among samples for robust representation learning without resorting to the simplistic strategy of merely increasing the number of model parameters.

As illustrated in Figure 1 (c-e), we apply k-NN search to a query image in the external knowledge base. For samples of the same class in the search results, the image modality presents concrete low-level invariant features (e.g., shape, color, texture), and the textual modality usually contains much high-level and abstractly relevant information. Even samples from different classes may share transferable knowledge with the query image across both image and text modalities. There may also be optional semantic information in the textual modality for noisy labels. In addition, textual descriptions are a priori knowledge that can be summarized by experts, which may be useful when there are not enough images to learn a general class representation for recognition.

*Corresponding author

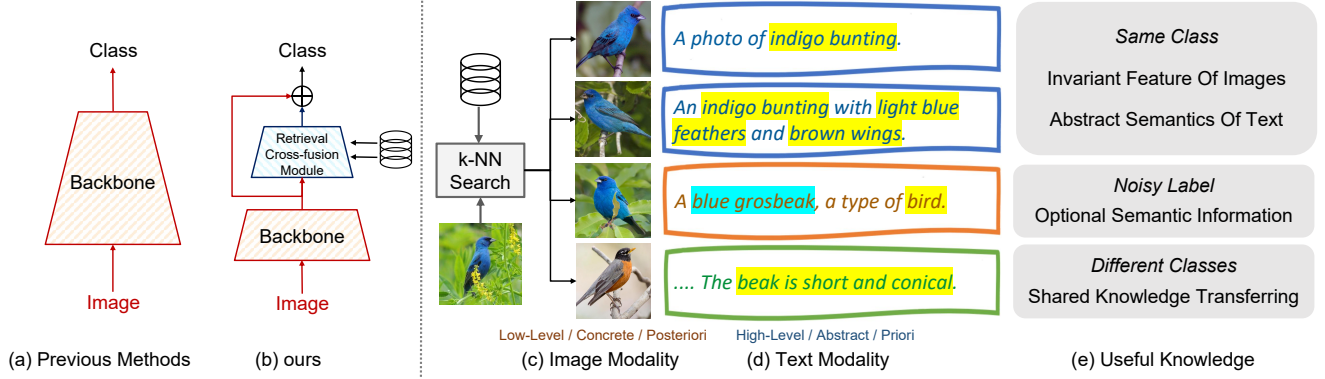


Figure 1. **Compare previous methods with ours and the knowledge contained in the different retrieved content.** (a) Previous methods focus only on image modalities and have a large number of model parameters, while (b) our method retrieves multimodal information from the knowledge base. (c) and (d) give intuitive explanations for the correlations and differences between the image and text modalities. (e) describes useful knowledge that can be retrieved from the knowledge base.

We propose an alternative perspective based on the above insights. As shown in Figure 1 (b), instead of statically compiling world knowledge into model weights, we construct an external image-text pair knowledge base for storing world knowledge and then utilize a retrieval module to identify and obtain relevant knowledge from a predefined knowledge base, modeling the effective relationships between image and text modalities, which is then used to enhance the model’s predictive power. We overcome the limitations of traditional methods on fine-tuning-based paradigm and training datasets by seamlessly integrating information from external databases into the model’s decision-making process. This semi-parametric approach enables the model to incorporate external knowledge to improve its understanding and predictive capabilities.

To evaluate our approach, we focus on long-tailed recognition and learning with noisy labels, which are challenging and meaningful tasks. We conduct extensive experiments on five datasets: CIFAR-100-LT [4], ImageNet-LT [23], Places-LT [23], iNaturalist 2018[41], and a real-world noisy dataset WebVision [21].

In conclusion, our primary contributions are fourfold:

- (1) We analyze and highlight the interplay and distinctions between image and text modalities within the context of retrieval augmentation in vision, noting that the text modality can complement retrieval content, which aids in long-tailed recognition and learning with noisy labels.
- (2) We introduce a new image-text retrieval-augmented framework called RetFormer, which utilizes an external knowledge base to model the relationship between text and image knowledge, thereby enhancing the performance of the model without incurring significant computational overhead.
- (3) We validate RetFormer through extensive experimen-

tation on CIFAR-100-LT, ImageNet-LT, Places-LT, iNaturalist 2018 and WebVision, demonstrating its superior performance over existing state-of-the-art methods.

(4) We provide an intuitive explanation for the effectiveness of this approach from a theoretical perspective based on gradient propagation.

2. Related Work

Data Scarcity Learning. Learning with imperfect training data has proven to be a very challenging task and has been explored in a variety of data-scarce tasks. Long-tailed recognition is one of the key problems. Most of the solutions are variants of the core idea of “adjustment” (e.g., re-weighting [3, 12]). Learning with noisy labels is another important task. The main methods revolve around filtering noisy labels, including but not limited to representation calibration [51] and Dynamic Loss [15]. Some methods combine multiple techniques (e.g., Sel-CL+ [20]).

Sample Relationship. There are diverse and tight relationships between different samples, and these relationships are widely used through various types of strategies. Mixup [50] introduces prior knowledge to the model by performing simple linear mixing between training samples. Some approaches via investigating sample/class relationships to conduct transductive inference (e.g., meta embedding [53]). BatchFormer [13] applied a vanilla Transformer encoder into the batch dimension of each mini-batch to implicitly explore sample relationships during training.

Retrieval augmented in vision. Recent approaches in computer vision perform various tasks by retrieving from external memory[22, 31, 32, 46, 49, 52]. [28] store feature maps from the training set in the memory, and perform k-NN for classification. RDM [2] retrieves nearest neighbors from a memory for generative vision models. RAC [24]

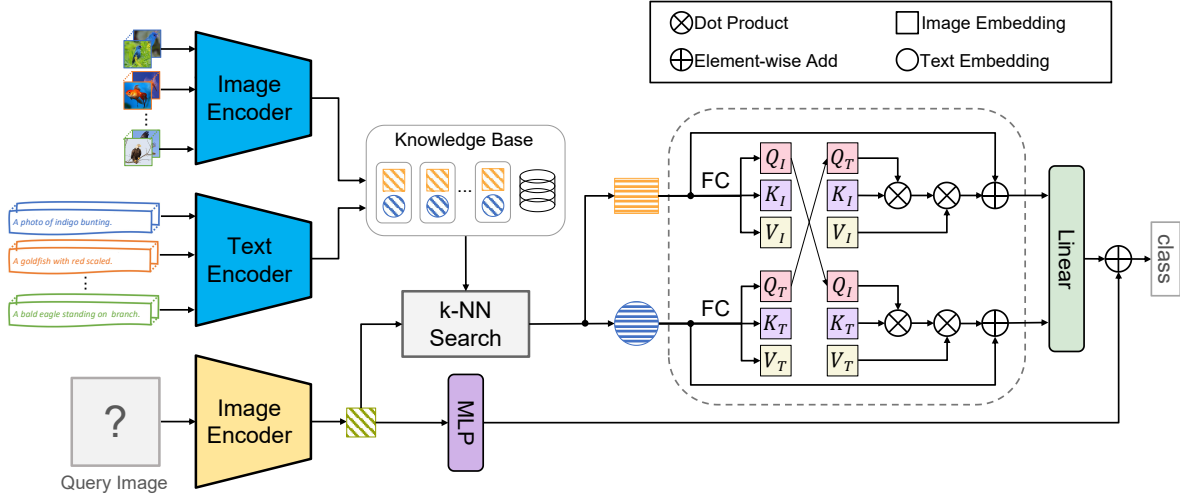


Figure 2. **Overall architecture of RetFormer** First, the image encoder maps the query image to an embedding. The image embedding will go through two branches, in the first branch the embedding goes through a simple MLP, in the second branch the image embedding first retrieves the relevant information from the knowledge base and then the contribution of each part is calculated by the retrieval cross-fusion module. The outputs of both branches are merged and trained by the same loss function.

uses only the training dataset itself as an external source of information and extracts embeddings of relevant text segments with the pre-trained CLIP model. MAM [14] considers texts to be of different importance from each other and introduces an additional dataset.

Our work is different in that we consider that both image and text modalities possess exploitable knowledge. Therefore, we propose to make the neural network itself capable of learning multimodal sample relationships based on the maximum inner product search algorithm. Experiments show better robustness of our novel framework across multiple tasks.

3. Method

3.1. Preliminaries

We consider a supervised classification problem. The model has access to an imbalanced set of N training samples $S = (x_i, y_i)_{i=1}^N$, where N is the sample size, x_i denotes the i -th instance and its label $y_i \in R^C$, where C is the number of classes. Let the number of training data belonging to k -th class be n_k . Without loss of generality, we assume that the classes are sorted in decreasing order, based on the number of training data in each class, i.e., $n_1 \geq \dots \geq n_C$. Afterward, all classes can be divided into two parts: head classes (referred to as G_h) and tail classes (referred to as G_t).

As shown in Figure 1 (a), in previous approaches, the model uses a vision encoder to map the input image x_i to a d -dimensional vector $\mathbf{z}_i$. The vector $\mathbf{z}_i$ is usually converted to the class logits by a classifier h . The output of the model

can be defined as:

$$f(x_i) = h(\varepsilon_{vis}(x_i)) = h(\mathbf{z}_i), \quad (1)$$

The model parameters are trained by minimizing supervised loss function, such as cross-entropy, or LACE [3] loss when the training data is imbalanced.

3.2. RetFormer

Retrieving from an external knowledge base enables RetFormer to incorporate external knowledge to enhance its performance. We describe the retrieval augmented classification and each of RetFormer’s components in detail in this section.

3.2.1. Retrieval augmented classification

Typically, a classification model is trained in a downstream task to make predictions considering only the image x_i in the dataset. The image x_i is passed through the vision encoder to produce image representations and ultimately output class logits by the classifier.

Retrieval augmented classification aims to train more robust and accurate models by utilizing relevant information from the external knowledge base. We introduce an additional knowledge base $D = \{(I_i, T_i)\}_{i=1}^L$, consisting of images I_i and corresponding text descriptions T_i in the dataset. More specifically, in addition to x_i , the model’s prediction now depends on D . Note that D is independent of S , which means that D is a database containing additional world knowledge, so we do not assume that D contains the class labels of S .

For predicting the logits of a given image x_i , we construct a subset of D that is most relevant to x_i for improving the performance of the network. Then, two frozen encoders ε_I and ε_T act on the image and text in D to convert them into embeddings as shown in Eq. (2), where ε_I is the vision encoder and ε_T is the text encoder:

$$E_i^I = \varepsilon_I(I_i), E_i^T = \varepsilon_T(T_i), \quad (2)$$

Let $V_D = \{(E_i^I, E_i^T)\}_{i=1}^L$ be the set of feature embeddings of each pair of instances in D . We compute the cosine similarity between $\mathbf{z}_i$ and each embedding $E^I \in V_D$ to find the k-nearest neighbors. The top-k ranked embeddings and query image embedding are then used for the output logits:

$$\begin{aligned} f(x_i) &= \frac{\tau}{2}(L_I + L_R) \\ &= \frac{\tau}{2}(MLP(\mathbf{z}_i) + h(r(\mathbf{z}_i, V_{NN}(\mathbf{z}_i; V_D)))) \end{aligned} \quad (3)$$

Here, L_I and L_R are the logits based on the query image and retrieval module, respectively. $MLP(\cdot)$ denotes two linear layers with a ReLU in the middle. $r(\cdot, \cdot)$ is a retrieval module and will be discussed in Section 3.2.2. $V_{NN}(\mathbf{z}_i; V_D)$ denotes top-k ranked embeddings of $\mathbf{z}_i$ from V_D , τ serves as a balancing coefficient to control the relative contributions of the backbone and retrieval branches.

3.2.2. Retrieval cross-fusion module

The vectors in V_{NN} are multi-classes and multi-modal, so it is important to model the relationship between each vector in the V_{NN} and the query vector. We propose to compute the attentional weights between the query vector $\mathbf{z}_i$ and the retrieved V_{NN} to model these relationships. These vectors are mapped from the CLIP encoders into the same feature space to learn the contribution of each vector in the V_{NN} .

First, we concatenate the vector $\mathbf{z}_i$ with the vectors in the V_{NN} to get image embeddings $E_{NN}^I \in R^{P \times D}$ and text embeddings $E_{NN}^T \in R^{P \times D}$ (note that the text embedding corresponding to $\mathbf{z}_i$ is replaced with a $\mathbf{0}$ vector to prevent data leakage). The image embeddings E_{NN}^I are mapped to three image matrices: query matrix Q_I , key matrix K_I and value matrix V_I by three linear transformations. The text embeddings E_{NN}^T are mapped to three text matrices: query matrix Q_T , key matrix K_T and value matrix V_T by three linear transformations:

$$\begin{aligned} Q_I &= E_{NN}^I W_{Q1}, K_I = E_{NN}^I W_{K1}, V_I = E_{NN}^I W_{V1}, \\ Q_T &= E_{NN}^T W_{Q2}, K_T = E_{NN}^T W_{K2}, V_T = E_{NN}^T W_{V2}, \end{aligned} \quad (4)$$

where $W_{Q1}, W_{Q2}, W_{K1}, W_{K2}, W_{V1}$ and $W_{V2} \in R^{D \times D}$.

The outputs of retrieval cross-fusion module are repre-

sented as follows:

$$\begin{aligned} \text{Att}(Q, K, V) &= \sigma\left(\frac{QK^T}{\sqrt{d}}\right)V, \\ r(\mathbf{z}_i, V_{NN}(\mathbf{z}_i; V_D)) &= [\text{Att}(Q_T, K_I, V_I) + E_{NN}^I, \\ &\quad \text{Att}(Q_I, K_T, V_T) + E_{NN}^T], \end{aligned} \quad (5)$$

where $\sigma(\cdot)$ denotes Softmax function. $\frac{1}{\sqrt{d}}$ is the scaling factor for appropriate normalization to prevent extremely small gradients.

Note that Eq. (5) can be repeated L times, i.e. L layers. Let $E^{I,(1)}, E^{T,(1)} = r(\mathbf{z}_i, V_{NN}(\mathbf{z}_i; V_D))$ denote the output of the first layer. The output after L layers can be computed as:

$$\begin{aligned} E^{I,(L)} &= \text{Att}(Q_{L-1}^T, K_{L-1}^I, V_{L-1}^I) + E^{I,(L-1)}, \\ E^{T,(L)} &= \text{Att}(Q_{L-1}^I, K_{L-1}^T, V_{L-1}^T) + E^{T,(L-1)}, \end{aligned} \quad (6)$$

In addition, we add Position embedding and class token references to ViT's [8] setting prior to the input. Similar attention mechanisms are used for different purposes, such as feature fusion [5]. Our experiments show that this attention mechanism is a good choice for retrieval augmented, significantly outperforming other baselines.

3.2.3. Knowledge base

The knowledge base plays a crucial role in the performance of the retrieval module. In this section, we describe in detail the datasets used to construct the knowledge base.

Downstream dataset. Building a knowledge base directly using the downstream dataset is the most straightforward option. This guarantees that for each query image, there will exist at least one instance of the same class. The disadvantage of this choice is that most downstream datasets are not rich enough in textual descriptions and may contain only the labels of the images.

DataComp. DataComp [10] is a testbed for dataset experiments centered around a new candidate pool of 12.8 billion image-text pairs from Common Crawl. We use the subset of DataComp, which is the output of the Image-based and CLIP score baseline filter at the xlarge scale of DataComp and comprised of 1.4B samples.

All. To integrate world knowledge, we combined all of the above datasets; after deduplication and filtering, there are about 1.4 billion image-text pairs in the knowledge base.

Vectorize Knowledge base. To extract features and compress storage, we initialize the encoder using the parameters of the CLIP [30] pre-trained on DataComp-1B [10] and freeze the encoder parameters when vectorizing the knowledge base prior to training and validation. This choice saves significant overhead and allows us to efficiently use in-memory datasets with up to 1B images. We will calculate the complexity of RetFormer in detail in Section 3.2.4.

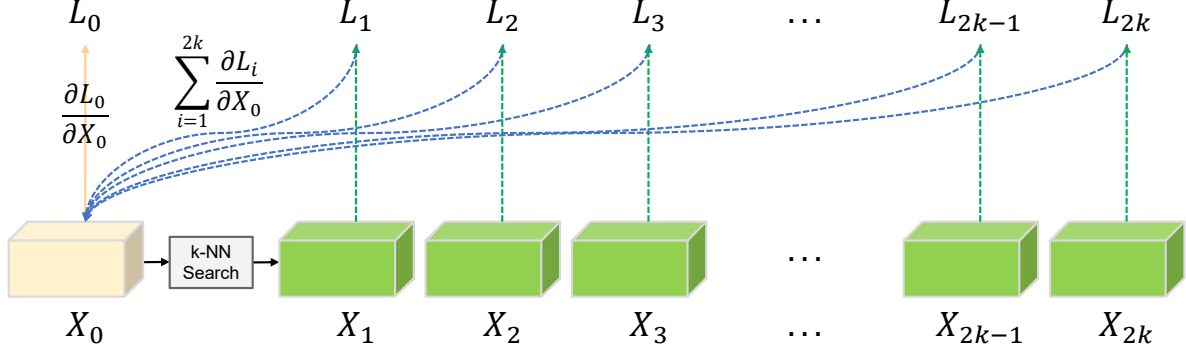


Figure 3. **Explaining RetFormer in terms of gradient propagation.** Blue dashed lines represent new gradient propagation among samples.

3.2.4. Retrieval complexity

Time complexity. We use the HNSW [27] algorithm from the Faiss [16] library to perform approximate k-NN searches in our experiments. It has sub-linear complexity, which means that for an in-memory dataset of N elements, it takes $O(\log N)$. In fact, the time consumed for querying in a 1B-element knowledge base is on the order of milliseconds. Since the encoder parameters are frozen, the results of one precomputation can be reused in subsequent training and validation, which also saves a lot of computation.

Spatial complexity. We use a fixed vision encoder in our experiments, so we pre-calculated and saved the results for all k-NN, thus saving storage space. For example, ImageNet-LT contains 115.8K images from 1000 categories. In our experiments, you need about 30GB of extra space for training.

3.2.5. A Gradient View of RetFormer

To better understand how RetFormer helps representation learning by exploring sample relationships, we also provide an intuitive explanation from the perspective of gradient propagation. Intuitively, without the retrieval module, all losses would only propagate gradients over the corresponding samples and categories, i.e., one-to-one. Whereas, in our approach, the retrieval module introduces additional gradient signals conditioned on retrieved neighbors, as shown in Figure 3. Specifically, given a sample x_0 with a retrieval subset $D_{NN} = \{(I_{2i-1}, T_{2i})\}_{i=1}^k$ of size k and corresponding losses $L_0, L_1, L_i, \dots, L_{2 \times k}$, we have:

$$\frac{\partial L_0}{\partial x_0} := \frac{\partial L_0}{\partial x_0} + \sum_{i=1}^{2k} \frac{\partial L_i}{\partial x_0}. \quad (7)$$

That is, the retrieval module brings in new gradient terms $\frac{\partial L_i}{\partial x_0}$. From the perspective of gradient-based optimization, L_i also optimizes the network according to samples $(I_{2i-1}, T_{2i})\}_{i=1}^k$, which is significantly different compared

to the model without the retrieval module. In other words, $(I_{2i-1}, T_{2i})\}_{i=1}^k$ can be thought of as virtual samples of y_i , where y_i is the label of x_i . We believe that the retrieval module can be considered as a data-dependent augmentation. The retrieval module implicitly extracts virtual samples and models multimodal inter-sample relationships by means of samples in their neighborhood distributions. From this perspective, the retrieval module implicitly adds virtual samples for each label y_i by modeling the relationship between samples in a k-NN search. Virtual samples are useful for tail classes because these classes lack samples. Previous approaches [54] have shown that data augmentation helps in long-tailed recognition and learning with noisy labels.

4. Experiments

4.1. Datasets and Implementation Details

We focus on two different image classification tasks: long-tailed recognition and learning with noisy labels. We now describe the datasets we used for each task.

Long-tailed Learning. We use four datasets for Long-tailed learning: CIFAR-100-LT, ImageNet-LT, iNaturalist 2018 and Places-LT. The CIFAR-100-LT is derived from the CIFAR-100 [19] with constructed imbalance ratios including 10, 50, and 100. ImageNet-LT is created by acquiring a subset of the original ImageNet dataset, so the number of images per class follows a long-tailed distribution. Places-LT contains 62.5K images from 365 classes, from a maximal 4980 to a minimum of 5 images per class. iNaturalist 2018 consists of 437.5K images distributed across 8142 species, with the number of images per species varying from as few as 2 to as many as 1000. In addition to measuring overall accuracy, we adhere to the evaluation protocol to report accuracy across three splits of classes: many-shot (more than 100 images), medium-shot (20-100 images), and few-shot (less than 20 images).

Learning with noisy labels. For learning with noisy la-

ImageNet-LT							Places-LT			
Methods	Extra Data	Backbone	Many	Med	Few	Overall	Many	Med	Few	Overall
Training from scratch										
PaCo [7]	×	ResNext-101	68.2	58.7	41.0	60.0	36.1	47.9	35.3	41.2
LiVT [48]	×	ViT-B/16	73.6	56.4	41.0	60.9	48.1	40.6	27.5	40.8
Fine-tuning pre-trained model										
TextRefiner [47]	×	ViT-B/16	81.7	62.4	39.4	66.7	52.7	32.7	22.7	38.0
BALLAD [25]	×	ViT-B/16	79.1	74.5	69.8	75.7	49.3	50.2	48.4	49.5
LIFT [35]	×	ViT-B/16	81.3	77.4	73.4	78.3	51.3	52.2	50.5	51.5
VL-LTR [39]	✓	ViT-B/16	84.5	74.6	59.3	77.2	54.2	48.5	42.0	50.1
PEL [34]	✓	ViT-B/16	81.3	77.4	73.4	78.3	51.3	52.2	50.5	51.5
RAC [24]	✓	ViT-B/16	80.9	76.0	67.5	76.7	48.7	48.3	41.8	47.2
MAM [14]	✓	ViT-B/16	80.6	77.5	74.5	78.3	52.4	52.0	48.5	51.4
Ours	✓	ViT-B/16	85.0	80.9	76.8	81.9	52.7	52.1	51.9	52.3

Table 1. Test top-1 accuracy (%) on ImageNet-LT and Places-LT. The best results are in bold for comparison.

bels, following the “mini” setting in [51], we take the first 50 classes of the Google resized image subset. We then test the trained network on the same 50 classes of the WebVision and ILSVRC12 validation sets.

Implementation details. To avoid data leakage, we initialized the vision encoder using the ViT-B/16 parameters of CLIP pre-trained on DataComp-1B instead of ImageNet-21K. Unless otherwise stated, training lasted for 25 epochs with a learning rate of 0.0005 and batch size of 256. The learning rate followed a 1-epoch warm-up schedule and then decreased at each epoch using a cosine decay schedule. We use the Adam optimizer [18] with a weight decay of 0.1. We also use label smoothing [37] and Mixup [50] during training to prevent overfitting and improve model generalization. For the attention module, we use $L = 4$. We retrieve $k = 32$ examples from memory unless otherwise stated.

4.2. Results on CIFAR-100-LT

Table in the appendix of the supplementary materials shows the results for CIFAR-100-LT. The results clearly show that RetFormer outperforms other methods including PEL, LiVT, BALLAD, and various ab initio training methods. Our method is the best among all methods that use extra data, demonstrating the potential of retrieval-augmented in vision tasks. The advantage of our method is more obvious when the dataset is unbalanced, which is due to the fact that the retrieval module makes the model focus on the tail classes appropriately. In addition, we do not compare models using ViT pre-trained on the ImageNet-21K dataset due to data leakage that would introduce unfair comparisons.

4.3. Results on ImageNet-LT

Table 1 illustrates the top-1 accuracy of existing methods on ImageNet-LT. Note that the pre-training of ViT-B/16 is different between the methods. RAC and MAM use the ViT-B/16 vision encoder pre-trained on the JFT-3B [14] and WebLI [6]. BALLAD, VL-LTR, and PEL use the same ViT-B/16 pre-trained with CLIP as our method.

Our method achieved the overall highest accuracy due to the proper focus on the tailed classes. Our method outperforms other methods that also utilize extra data. RAC and MAM use retrieval-augmented as well, but they only consider the relationships between image modalities, which leads to suboptimal performance.

4.4. Results on Places-LT

Table 1 illustrates the top-1 accuracy of existing methods on Places-LT. Although our method does not achieve the highest accuracy in the Many-shot and Med-shot, it achieves the highest accuracy in the tail and overall accuracy because the retrieval module enables the model to appropriately focus on tail classes without compromising the performance of other classes. The performance of RetFormer in Places-LT is not as strong as expected. We believe there are two main reasons:

(1) Places-LT focuses on scene-oriented classification, but textual descriptions in large knowledge bases often center on objects. This mismatch can reduce the synergy between retrieved text and scene images.

(2) The retrieval might (in effect) overcorrect for imbalances, slightly hurting the performance of classes already well-covered.

We tested smaller k (8, 16) for Places-LT. Interestingly,

iNaturalist 2018						
Methods	Extra Data	Backbone	Many	Med	Few	Overall
Training from scratch						
RIDE+CR [26]	×	ResNet-50	71.0	73.8	74.3	73.5
RIDE+OTmix [11]	×	ResNet-50	71.3	72.8	73.8	73.0
PaCo [7]	×	ResNet-50	70.4	72.8	73.6	73.2
Proco [9]	×	ResNet-50	74.0	76.0	76.0	75.8
GML [36]	×	ResNet-50	-	-	-	74.5
LiVT [48]	×	ViT-B/16	78.9	76.5	74.8	76.1
Fine-tuning pre-trained model						
Decoder [42]	×	ViT-B/16	-	-	-	59.2
LIFT [35]	×	ViT-B/16	72.4	79.0	81.1	79.1
PEL [34]	✓	ViT-B/16	72.4	79.0	81.1	79.1
RAC [24]	✓	ViT-B/16	75.9	80.5	81.1	80.2
Ours	✓	ViT-B/16	77.8	81.1	82.0	81.1

Table 2. Test top-1 accuracy (%) on iNaturalist 2018. The best results are in bold for comparison.

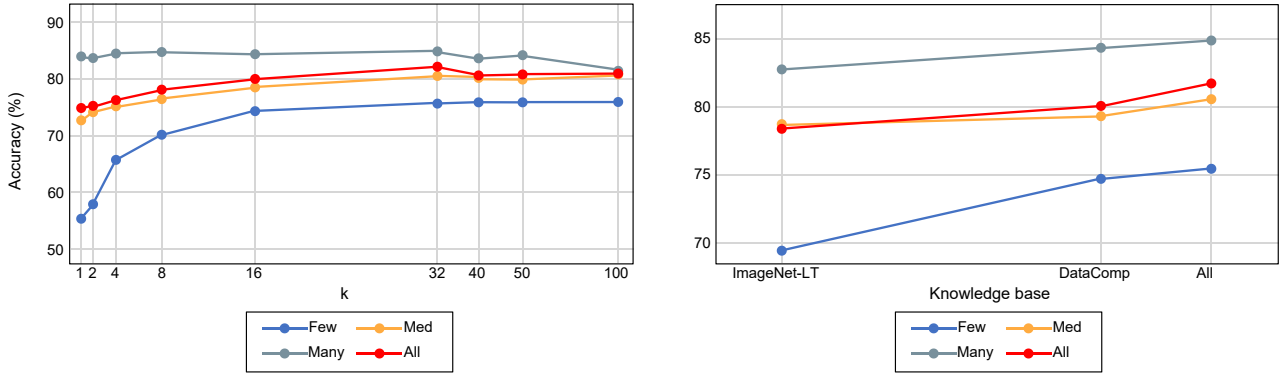


Figure 4. Ablation study on ImageNet-LT. Left: We show the effect of different k on RetFormer. Right: We show the impact of building a knowledge base with different datasets, we set $k = 32$ in this experiment.

”Many” accuracy did improve marginally (0.3–0.8%), though ”Few” classes performance dipped.

4.5. Results on iNaturalist 2018

We report the results in Table 2. Overall, our method achieves the best performance on this challenging dataset, surpassing RAC, VL-LTR, and LiVT. We acknowledge that LiVT surpasses us in Many-shot performance, but its Overall performance is far below RetFormer. Particularly, among all methods that use extra data, our method achieved the best performance in all classes.

4.6. Results on WebVision

Table 4 shows the results on WebVision. It can be seen that our method achieves the best results on the top-1 accuracies on both the WebVision validation set and the ImageNet ILSVRC 12 validation set compared to other state-of-the-

art methods. RCAL+ combines semi-supervised learning algorithms, resulting in slightly higher top-1 accuracy on ImageNet than our approach.

4.7. Ablation Study

In this section, we provide an in-depth analysis of RetFormer.

Modality ablation. Table 5 demonstrates the results of the ablation study. The results show that our retrieval module can effectively improve the accuracy of the tail class. This is supported by the performance of CLIP full FT. On the other hand, retrieving the image alone gives a small performance improvement, while retrieving both text and image improves the overall accuracy. The performance of Ours w/ FE is explained in Section 3.2.5.

Effect of k . The impact of k on model performance cannot be ignored. This hyperparameter controls the number of

L	Many	Med	Few	Overall
L = 1	84.4	80.3	75.6	81.2
L = 2	84.8	80.4	76.1	81.5
L = 4	85.0	80.9	76.8	81.9
L = 8	84.9	80.9	77.1	82.0

Table 3. Ablation of L on ImageNet-LT.

Train Validate Methods	mini-WebVision	
	WebVision Top1 (%)	ILSVRC12 Top1 (%)
RoLT+ [44]	77.64	74.6
NGC [45]	79.16	74.44
RCAL+ [51]	79.56	76.32
Sel-CL+ [20]	79.96	76.84
Dynamic Loss [15]	80.12	74.76
Ours	81.7	77.3

Table 4. Top-1 test accuracy on WebVision and ImageNet validation sets.

samples retrieved from the knowledge base. As illustrated in Figure 4 (left), we see that the performance gradually increases up to $k = 32$ and stabilizes thereafter. The performance of Many-shot decreases instead as k increases, which may be due to the fact that the sample size of Many-shot is sufficient for the model to find the decision boundary, and the noisy labels introduced by too large k reduces the performance instead. The performance on Few-shot is gradually stabilized as k increases, which is consistent with our theoretical analysis in Section 3.2.5.

Impact of Knowledge base. We further analyze the impact of selecting different datasets to build a knowledge base on model performance. Figure 4 (right) shows the performance of building a knowledge base using different datasets. In this experiment, we set $k = 32$. We see that the performance of the model gradually increases as the size of the knowledge base increases. Among them, the rise of the Few-shot accuracy is greater, which is in line with our theoretical analysis.

Index ablations. We examine the effect of the index type and recall on RetFormer’s performance and speed in Table 6. To quantify error induced by an approximate index, we include the accuracy on the index content itself in addition to the validation accuracy. We observe that the drop in accuracy due to use of an approximate (HNSW) instead of exact k-NN is also minor, but comes with a significant (3×) speedup on large index’s. In summary, the choice of HNSW becomes critical to ensure lookup time does not bottleneck training.

Qualitative examples. We present some of the qualita-

Methods	Many	Med	Few	Overall
CLIP zero-shot	69.2	67.6	67.7	68.3
CLIP full FT	84.3	73.1	52.9	74.6
CLIP classifier FT	77.3	73.3	64.2	73.6
Ours w/o text	81.3	74.8	65.9	76.0
Ours w/o image	83.1	76.9	68.4	78.1
Ours w/ FE	79.3	79.0	74.6	78.6
Ours	85.0	80.9	76.8	81.9

Table 5. Comparing top-1 accuracy (%) with baselines on ImageNet-LT. FT represents fine-tuning. FE represents freezing the image encoder.

Index Type	Recall	Many	Med	Few	Overall	QT (ms)
Exact	1.00	85.0	80.9	76.8	81.9	23.34
HNSW	0.97	85.1	80.7	76.5	81.8	7.69
HNSW	0.91	84.8	79.9	74.9	81.1	5.29
HNSW	0.82	84.2	79.0	73.1	80.2	3.47
HNSW	0.65	83.1	77.7	69.7	78.7	2.41

Table 6. Index ablations on ImageNet-LT. Recall represents the recall rate of the approximate k-NN algorithm under different settings. QT indicates Query Time.

tive examples in the appendix of the supplementary materials. We observed that our method assigns higher attention weights to both relevant images and text in the k-NN list, indicating that RetFormer can capture effective relationships from two different modalities. We found that even in the absence of any relevant images, our method still benefits from related shared knowledge.

Ablation of L on ImageNet-LT. Table 3 shows the ablation of L on ImageNet-LT. L=1/2 yields 0.4–0.7% lower accuracy, especially for the tail classes. L=4/8 results are close; going from 4 to 8 layers yields diminishing but still slightly positive gains (+0.11%).

5. Conclusion

We introduce RetFormer, a new multimodal retrieval-augmented vision-language framework for long-tailed recognition and learning with noisy labels. We emphasize that the image and text modality of retrieved instances have implicit intrinsic relationships, and propose to equip deep neural networks with the ability to model these sample relationships. We propose a simple, but effective, retrieval cross-fusion module that learns multimodal sample relationships and computes their contributions. Extensive experiments on a variety of long-tailed recognition and learning with noisy labels benchmarks validate that our approach works better than well-designed vision-based methods.

Acknowledgments

This work was partly supported by Ningbo Youth Science and Technology Innovation Leading Talent Project (2024QL044) and Ningbo Key R&D Program (2025Z047).

References

- [1] Anthropic. Introducing the next generation of claude. Accessed 4 March 2021. <https://www.anthropic.com/news/claude-3-family>, 2024. 1
- [2] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Semi-parametric neural image synthesis. *arXiv preprint arXiv:2204.11824*, 2022. 2
- [3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 2, 3
- [4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 2
- [5] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 4
- [6] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 6
- [7] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Ji-aya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 715–724, 2021. 6, 7
- [8] 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. 4
- [9] Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 7
- [10] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 4
- [11] Jintong Gao, He Zhao, Zhuo Li, and Dandan Guo. Enhancing minority classes by mixing: An adaptive optimal transport approach for long-tailed classification. *Advances in Neural Information Processing Systems*, 36:60329–60348, 2023. 7
- [12] Xiaoxuan He, Siming Fu, Xinpeng Ding, Yuchen Cao, and Hualiang Wang. Uniformly distributed category prototype-guided vision-language framework for long-tail recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5027–5037, 2023. 2
- [13] Zhi Hou, Baosheng Yu, and Dacheng Tao. Batchformer: Learning to explore sample relationships for robust representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7256–7266, 2022. 2
- [14] Ahmet Iscen, Alireza Fathi, Ahmet Schmid, Cordelia, Alireza Fathi, and Cordelia Schmid. Improving image recognition by retrieving from web-scale image-text data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19295–19304, 2023. 3, 6
- [15] Shengwang Jiang, Jianan Li, Jizhou Zhang, Ying Wang, and Tingfa Xu. Dynamic loss for robust learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 8
- [16] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 5
- [17] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1
- [18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5
- [20] Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 316–325, 2022. 2, 8
- [21] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017. 2
- [22] Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. Learning customized visual models with retrieval-augmented knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15148–15158, 2023. 2
- [23] 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. 2
- [24] 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. 2, 6, 7

- [25] 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. 6
- [26] Yanbiao Ma, Licheng Jiao, Fang Liu, Shuyuan Yang, Xu Liu, and Lingling Li. Curvature-balanced feature manifold learning for long-tailed classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15824–15835, 2023. 7
- [27] Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018. 5
- [28] Kengo Nakata, Youyang Ng, Daisuke Miyashita, Asuka Maki, Yu-Chieh Lin, and Jun Deguchi. Revisiting a knn-based image classification system with high-capacity storage. In *European Conference on Computer Vision*, pages 457–474. Springer, 2022. 2
- [29] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 1
- [30] 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. 4
- [31] Jiahua Rao, Zifei Shan, Longpo Liu, Yao Zhou, and Yuedong Yang. Retrieval-based knowledge augmented vision language pre-training. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5399–5409, 2023. 2
- [32] Sara Sarto, Marcella Cornia, Lorenzo Baraldi, Alessandro Nicolosi, and Rita Cucchiara. Towards retrieval-augmented architectures for image captioning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(8):1–22, 2024. 2
- [33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1
- [34] 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. 6, 7
- [35] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Jie-Jing Shao, Xin-Yan Han, and Yu-Feng Li. Long-tail learning with foundation model: Heavy fine-tuning hurts. In *Forty-first International Conference on Machine Learning*, 2024. 6, 7
- [36] Min-Kook Suh and Seung-Woo Seo. Long-tailed recognition by mutual information maximization between latent features and ground-truth labels. In *International conference on machine learning*, pages 32770–32782. PMLR, 2023. 7
- [37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 6
- [38] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1
- [39] Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. VI-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *European Conference on Computer Vision*, pages 73–91. Springer, 2022. 6
- [40] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shriti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1
- [41] 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. 2
- [42] Yidong Wang, Zhuohao Yu, Jindong Wang, Qiang Heng, Hao Chen, Wei Ye, Rui Xie, Xing Xie, and Shikun Zhang. Exploring vision-language models for imbalanced learning. *International Journal of Computer Vision*, 132(1):224–237, 2024. 7
- [43] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. *Advances in neural information processing systems*, 30, 2017. 1
- [44] Tong Wei, Jiang-Xin Shi, Wei-Wei Tu, and Yu-Feng Li. Robust long-tailed learning under label noise. *arXiv preprint arXiv:2108.11569*, 2021. 1, 8
- [45] Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, and Yu-Feng Li. Ngc: A unified framework for learning with open-world noisy data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 62–71, 2021. 8
- [46] Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19265–19274, 2023. 2
- [47] Jingjing Xie, Yuxin Zhang, Jun Peng, Zhaohong Huang, and Liujian Cao. Textrefiner: Internal visual feature as efficient refiner for vision-language models prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8718–8726, 2025. 6
- [48] 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. 6, 7

- [49] Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 547–556, 2023. [2](#)
- [50] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. [2](#), [6](#)
- [51] Manyi Zhang, Xuyang Zhao, Jun Yao, Chun Yuan, and Weiran Huang. When noisy labels meet long tail dilemmas: A representation calibration method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15890–15900, 2023. [1](#), [2](#), [6](#), [8](#)
- [52] Zhongliang Zhou, Jielu Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2749–2754, 2024. [2](#)
- [53] Linchao Zhu and Yi Yang. Inflated episodic memory with region self-attention for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4344–4353, 2020. [2](#)
- [54] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1004–1013, 2018. [5](#)