Article

Long-tailed Image Classification Method Based on Enhanced Contrastive Visual-language

Ying Song *, Mengxing Li, Bo Wang

Posted Date: 23 June 2023

doi: 10.20944/preprints202306.1679.v1

Keywords: long-tailed image classification; contrastive learning; data augmentation

Article

Long-tailed image classification method based on enhanced contrastive visual-language

Ying Song [1,2], MengXing Li [1,2] and Bo Wang [3]

[1] Beijing Key Laboratory of Internet Culture and Digital Dissemination, Beijing Information Science and Technology University, Beijing 100101, China
[2] Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, Beijing 100101, China
[3] Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou 450002, China
* Correspondence:orresponding author: Ying Song (songying@bistu.edu.cn).

Abstract: To solve the problem that the common long-tailed classification method does not use the semantic features of the original label text of the image, and the difference between the classification accuracy of most classes and minority classes is large, the long-tailed image classification method based on enhanced contrast visual language trains the head class and tail class samples separately, uses text image to pre-train the information, and uses enhanced momentum contrast loss function and RandAugment enhancement to improve the learning of tail class samples. On the ImageNet-LT long-tailed dataset, the enhanced contrastive visual-language based long-tailed image classification method has improved all class accuracy, tail class accuracy, middle class accuracy, and F1 values by 3.4%, 7.6%, 3.5%, and 11.2%, respectively, compared to the BALLAD method. The difference in accuracy between the head class and tail class is reduced by 1.6% compared to the BALLAD method. The results of three comparative experiments indicate that the long-tailed image classification method based on enhanced contrastive visual-language has improved the performance of tail classes and reduced the accuracy difference between majority and minority classes.

Keywords: long-tailed image classification; contrastive learning; data augmentation

1. Introduction

Image classification [1] is the earliest application of machine learning in the field of computer vision, and is the foundation of other visual tasks such as object detection and instance segmentation. Due to the rich semantic information contained in images (such as multiple targets, scenes, behaviors, etc.), the characteristics closest to human perception and expression ability, and the gradual optimization of the performance and cost of visual sensors (mainly cameras), image classification and its derived detection, segmentation and other visual algorithms are gradually being applied in fields such as healthcare, transportation, signal processing [2], etc. However, in the application process, due to the unique nature of the actual environment, some difficult to solve problems are gradually encountered.

In image classification tasks, input data is manually collected and annotated, and through human intervention, the amount of data in each category is balanced as much as possible, with no significant difference in sample size among different categories. The manually balanced data set simplifies the requirements for algorithm robustness, but with the gradual increase of the focus categories, maintaining the balance among various categories will bring Exponential growth in acquisition costs. For example, if an animal classification dataset is to be built, it is easier to collect millions of pictures from common data such as cats and dogs. However, considering the balance of the data set, it is also necessary to collect the same amount of samples for rare animals such as snow leopards. With the increase of the rarity of the category, the collection volume tends to grow Exponential growth, as shown in Figure 1.

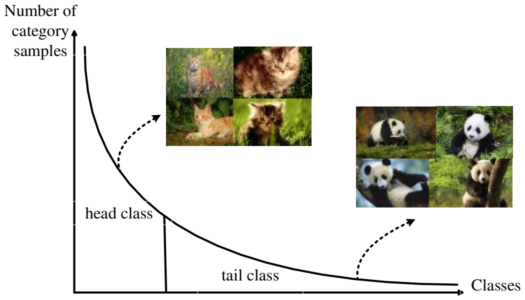


Figure 1. Schematic diagram of the long-tailed distribution of natural animal species.

In practical applications, such as facial recognition, species classification, autonomous driving, medical diagnosis, drone detection, and other fields, there is a problem of long-tailed category imbalance [3]. For example, for autonomous driving, the data on normal driving will account for the vast majority, while there is very little data on actual abnormal situations/car accident risks; For medical diagnosis, the number of people with specific diseases is also extremely uneven compared to the normal population. However, this type of imbalance problem often makes the training of deep neural networks very difficult. Classification and recognition systems that directly use long-tailed distribution data for training often tend to lean towards the head class data, making them insensitive to tail class features during prediction and affecting the correct judgment of the system [3]. In traditional methods, a series of common methods to mitigate performance degradation caused by long tail distribution data are based on category rebalancing strategy, including resampling training data and reweighting to redesign loss function [3]. These methods can effectively reduce the bias of the model to the head class in the training process, thus producing more accurate classification Decision boundary. However, because the distribution of the original data is unbalanced, and the over parameterized deep networks are easy to fit this composite distribution, they often face the risk of tail class overfitting and head class underfitting.

Given that the problem of class imbalance in long tailed distribution datasets is very widespread in practical tasks, it is crucial to train high-performance network models from a large number of images that follow the long-tailed distribution. Moreover, the difference in class distribution between training and testing data will greatly limit the practical application of neural networks. This research topic has important practical significance and is an important paradigm for promoting the implementation of deep neural networks in model implementation. How to effectively utilize long tail data to train a balanced classifier is a key issue. From a practical implementation perspective, this study will improve the speed of data collection and reduce collection costs. This article explores effective contrastive learning strategies to learn better image representations from imbalanced data, in order to better apply them to long tail image classification. We hope to provide better development ideas for the application of image classification in today's gradually developing image technology.

2. Related Work

The research content of this paper is long tail distribution image classification. Long-tail image classification methods mainly include data resampling, data reweighting, data enhancement, transfer learning, ensemble learning, etc.

2.1. Data Resampling

Data resampling solves the problem of long-tailed distribution image classification from the data level. Resampling is the most widely used method [4] [5] in processing long-tailed distribution image classification in depth learning, mainly including oversampling [6,7] , under-sampling [8-10] and mixed sampling [11,12] .

The oversampling method mainly reduces the imbalance between the head class and the tail class by increasing the number of samples of the tail class [6,7]. Inspired by this, in 2019, Gupta et al. proposed the repeated factor sampling method [13], which performs a rebalancing operation on the training data by increasing the sampling frequency of the tail image. In 2020, Peng et al. proposed the soft box sampling method [14], which utilizes class perception sampling to calculate the replication factor for each image based on the distribution of labels, and replicate the images according to the specified number of times to solve the problem of class imbalance. In 2020, Hu et al. designed an instance level class balancing scheme [15] to balance instance level samples of original images. The balanced samples are learned using meta-modules, transferring knowledge from the data rich header to the data poor tail. In 2020, Wu et al. proposed non maximum suppression resampling [16], which adaptively adjusts the threshold for non-maximum suppression based on the label frequencies of different categories, in order to retain more candidate target categories from the tail category and balance the data distribution by suppressing candidate target categories from the head category. According to the principle of the oversampling method, this method simply repeats the positive example, which will cause overemphasis on the positive example, and it is easy to over fit the positive example. In 2022, Park et al. proposed a oversampling method based on feature dictionary [6], and built a feature dictionary through a pre trained feature extractor. Synthesize samples based on feature dictionaries and enrich the diversity of minority class data by fine-tuning classifiers. In 2023, Li et al. proposed a small number of oversampling based on subspace [7]. This method believes that each type of sample is formed by common and unique features, and these features can be extracted through subspace. In order to obtain balanced data, map images belonging to minority categories are oversampling to more accurately describe minority categories. Balanced data is obtained by restoring the generated subspace product to the original space.

Compared with the oversampling method, the under-sampling method reduces the imbalance between the head class and the tail class by reducing the number of samples of the head class [8-10]. In 2020, the Bilateral branch network (BBN) [17] developed conventional learning branches and rebalanced branches, using a new bilateral sampling strategy to address class imbalance issues. Uniform sampling was applied to simulate the original long tail training for conventional branches, and a reverse sampler was applied to sample more tail class samples for rebalanced branches to improve tail class performance. In 2021, Lee et al. proposed a framework for classifying unbalanced data using under-sampling and Convolutional neural network [8], created a balanced training set through under-sampling, and then used Convolutional neural network for training. In 2021, Zang et al. proposed a feature and sampling adaptive strategy [18], which used model classification loss to adjust the sampling rates of different categories on the validation set, thereby sampling more tail class samples with insufficient representation. In 2022, Lehmann et al. proposed a subclass based under-sampling method [9], which selects samples from all subclasses of a class for under-sampling, and identifies subclasses by clustering the advanced features of the CNN model. In 2023, Farshidvard et al. proposed a method based on under-sampling and ensemble [10], which divides most classes into clusters, so that there are no minority class samples in the majority samples of each cluster, while controlling the size of each cluster.

Mixed sampling [11,12] is a method of combining oversampling and under-sampling to achieve sample balance. In 2020, Ding et al. proposed a KA integration method of under-sampling and oversampling [11], under-sampling the majority of classes through the kernel based adaptive synthesis method, and oversampling the minority classes at the same time, generating a set of balanced data sets to train the corresponding classifiers separately, and the final results will be voted by all these trained classifiers. By combining under-sampling with oversampling in this way, KA Ensemble is good at solving the class imbalance problem with large imbalance rate [11]. In 2022, EF Swana et al. studied the use of a Naïve Bayes classifier, support vector machine, and k-nearest neighbors together with synthetic minority oversampling technique, Tomek link, and the combination of these two resampling techniques for fault classification with simulation and experimental imbalanced data.

doi:10.20944/preprints202306.1679.v1

Resampling is the most common method to solve the problem of long-tailed distribution image classification. However, these classic methods generally have poor results. For example, in the case of Oversampling of tail categories, it may lead to overfitting of tail categories [6,7], and if there are errors or noises in the samples of tail categories, oversampling may exacerbate these problems. Under-sampling may lead to insufficient learning of head categories [8-10] and may result in the loss of valuable data in the head categories. For extremely imbalanced long tail data, under-sampling methods often lose a large amount of information due to the significant difference in data volume between the head class and tail class.

2.2. Data Reweighting

The data reweighting strategy aims to minimize the total cost of the classifier, and solves the problem of imbalanced data classification by adjusting the loss values of different categories during training and increasing the attention of minority class samples during model learning.

In 2017, Lin et al. proposed Focal Loss [19]. During training, this loss function can automatically reduce the weight of the head class, making the model focus on learning the tail class. In 2017, Hermans et al. proposed the Triplet loss function [20], and used the Gradient descent to train samples with small differences. In 2019, Class Balanced Loss (CB) [21] introduced the concept of effective sample size, which alleviated the problem of class imbalance by forcing a class balance reweighting term that is inversely proportional to the number of effective samples in the class. In 2019, Cao et al. proposed Label Distribution Aware Margin Loss (LDAM) [22], which involves the model learning the initial feature representation before reweighting. In 2020, the distribution balance loss [23] was alleviated by a new tolerant regularization method to alleviate gradient over suppression. At the same time, it also evaluates the difference between the expected sampling frequency and the actual sampling frequency for each class, and then uses the quotient of these two frequencies to recalculate the weighted loss values for different classes. In 2020, Equalization Loss [24] proposed to directly reduce the weight of loss values for tail class samples when the tail class samples are negative sample pairs for a large number of head class samples. In 2021, Equalization loss v2 [25] extended Equalization loss by introducing a new gradient reweighting mechanism that dynamically increases the weight of positive gradients and decreases the weight of negative gradients for model training on each subtask. Seesaw loss [26] rebalances the positive and negative gradients of each category using mitigation and compensation factors. LADE [27] introduces label distribution decoupling loss to disentangle the learning model from the long tail training distribution, and then adapts the model to any test class distribution when the test label frequency is available.

Balanced meta-softmax [28], optimizing sample distribution by adjusting the model on the validation set. The progressive margin Loss function [29] uses two margin items to adjust the classification margin of long-tailed learning. Sequential margins extract discriminative features and maintain category order relationships. The variational margin gradually suppresses the head class and handles class imbalance in long tail training samples. The adversarial robust long-tailed classification method [30] rebalances data through a scale invariant classifier and boundary adjustments during the inference process.

Although the data reweighting method alleviates the imbalance in gradient proportion caused by long tail distribution, for some extreme cases, such as when the sample proportion of tail categories is very small, the recognition accuracy of tail categories is still at a low level.

2.3. Data Augentation

Data augmentation aims to utilize a series of data augmentation techniques to enhance the size and quality of the dataset [31,32] for model training. In 2021, Zhang et al. proposed a data augmentation method based on neighborhood risk minimization [33], which helps correct overconfidence in the model. In the decoupling training scheme, this method has a positive impact on representation learning and a negative or negligible impact on classification learning. Based on these observations, data mixup is used in the decoupling training scheme to enhance representation learning.

In addition, the Remix method also adopts a method for long-tailed learning and introduces a rebalancing hybrid enhancement method to enhance tail classes. In 2021, Li et al. proposed Meta semantic augmentation (MetaSAug) based on meta learning [34], using a variant of implicit semantic data augmentation (ISDA) [35] to enhance tail classes. ISDA obtains semantic direction by estimating the Covariance matrix of sample features, and translates deep features along multiple semantic meaningful directions to generate diversified enhancement samples. However, due to insufficient tail class samples, it is impossible to estimate the Covariance matrix of tail class. To solve this problem, MetaSAug explores meta learning to guide the learning of each class's Covariance matrix. In this way, the Covariance matrix of tails can be estimated more accurately, thus generating rich tail class feature information. Although data augmentation methods enhance the diversity of training samples, they are prone to introducing noise and ambiguity during the training process.

2.4. Transfer Learning

Transfer learning is to transfer knowledge from the source domain to enhance model training in the target domain. The source domain is a different domain from the test sample, but it has rich supervisory information. The target domain is the domain where the test sample is located, with no labels or only a small number of labels. There are mainly four Transfer learning schemes in the deep learning processing of long-tailed distribution image classification, namely, head to tail knowledge transfer, model pre-training, knowledge distillation and self-training.

The knowledge transfer from beginning to end is to transfer the knowledge of the head class to the tail class to enhance model performance. Yin et al. proposed feature transfer learning (FTL) [36], which uses the intra class variance knowledge of the head class to enhance the characteristics of the tail class samples, so that the tail class features have higher intra class variance, so that the tail class gets better performance. The LEAP [37] method proposed by Liu et al. constructs a "feature cloud" for each class by adding tail class samples with certain interference in the feature space, seeking to transfer the knowledge of the head class feature cloud to enhance the intra class variation of the tail class feature cloud. This method effectively alleviates the problem of inter class feature variance distortion. The online feature enhancement method [38] uses class activation mapping to decouple sample features into specific class features and uncertain class features, and combines the class specific features of the tail class samples with the class unknown features of the head class samples to enhance the tail class. Then, using all enhanced and original features, the model classifier is fine-tuned using a rebalancing sampler to achieve better long tail learning performance.

Model pre-training is one of the commonly used methods for deep learning model training. Domain specific Transfer learning [39] uses all long tail samples to pretrain the model, and then fine tune the model on the training subset of class balance. Slowly transfer the learned features to the tail class to achieve a more balanced performance among all classes. In addition, the self-supervised pre training method [40] first uses self-supervised learning (such as comparative learning [41 or rotation prediction [42]) for model pre-training, and then carries out standard training on long-tailed data. This scheme is also used to process long tail data with noise labels [43]. The proposal of the visual and language pre training dataset (Conceptual 12M [44]) has promoted the development of visual language models in the field of long-tailed recognition.

Knowledge distillation is the training of student models using the output of well-trained teacher models. The Learning from multiple experts (LFME) method [45] divides the entire long tailed distribution dataset into several subsets with less imbalanced classes, and trains multiple experts with different sample subsets. Based on these experts, the LFME method utilizes adaptive knowledge distillation methods and selects difficult course examples to train a unified student model. The Routing diversity distribution aware experts (RIDE) [46] introduces a knowledge distillation method on the basis of a multi expert framework to reduce the parameters of the multi expert model by learning a student network model with fewer experts. The self-supervised distillation method [47] has invented a new self-distillation scheme to enhance decoupling training. The decoupling training scheme trains a calibration model based on supervised and self-supervised information, and then uses the calibration model to generate soft labels for all samples. Afterwards, a new student model is extracted using the generated soft labels and the original long-tailed labels, and finally a new classifier fine-tuning stage is entered. In addition, the distillation virtual instance method [48] uses a class equilibrium model as the teacher model to solve the long tail classification problem.

The purpose of self-training is to learn well performing models from a small number of labeled samples and a large number of unlabeled samples. The Class balancing self-training (CReST) method [49] studied self-training in long tail classification and found that the supervised model has high classification accuracy for tail classes. Based on this discovery, CReST proposes to select more tail class samples for online pseudo labeling in each iteration, enabling the retrained model to achieve better performance on tail classes. The MosaicOS method [50] pre trains the model using scene centered images labeled in the original detection dataset. The pre trained model is fine tuned in two stages: first, the pseudo labeled object centered image is fine-tuned, and then the original labeled scene centered image is fine-tuned, which can alleviate the negative impact of data differences and effectively improve long tail learning performance.

Due to the introduction of additional knowledge, the transfer learning method improves the performance of the tail class without sacrificing the performance of the head class, but the performance improvement is not obvious when the difference between the head class and the tail class is large. The lack of sufficient tail class samples is one of the key problems of long tail learning, and the related methods of Transfer learning deserve further exploration.

2.5. Ensemble Learning

The method based on ensemble learning solves the learning problem of long-tailed distribution image by strategically generating and combining multiple network modules (multiple experts). Long-tailed multi label visual recognition method [51] explored a bilateral branch network solution to long tail multi label classification, used sigmoid cross-entropy loss function to train each branch for multi label classification, and forced the use of logit consistency loss to improve the consistency of the two branches.

The all complete experts (ACE) method [52] divides all classes into several different subsets: one subset contains all classes, one contains intermediate and tail classes, and the other only has tail classes. ACE trains multiple experts with different class subsets, and uses distributed adaptive optimizer to adjust the Learning rate of different experts. In 2022, the ResLT [53] method proposed by Cui et al. also had an idea similar to ACE. The Test time aggregating diverse experts (TADE) [54] explores multiple expert schemes to handle long-tailed recognition problems, where the distribution of test classes can be uniform or long tail. TADE provides two solutions: one is a diversified expert learning strategy that can train experts with different class distributions based on the characteristics of long tailed distribution datasets; The second is the testing time expert aggregation strategy, which can use self-supervised methods to aggregate multiple experts to process data of various unknown test distributions. The methods based on Ensemble learning usually achieve better performance on the head and tail classes. However, such methods often result in higher computational costs due to the use of multiple experts.

3. Method

Real data often follows a long-tailed distribution, with the head class dominating the training and the tail class having only a small number of samples, which is a major challenge in the field of image classification. The existing methods either use manually balanced datasets (such as ImageNet) or develop more robust algorithms to process data, such as class rebalancing strategies and network module improvements.

Although the above methods are effective for long-tailed distribution datasets, they sacrifice the performance of the header class at different levels. To address these limitations, researchers have turned to exploring new network architecture training paradigms. Long-tailed classification models typically include two key parts: feature extractors and classifiers. For each component, there are corresponding methods, either designing better classifiers [37,55], or learning reliable representations [56,57]. In terms of the new training framework, existing work attempts to divide one stage of training into two stages. For example, the learning process of decoupling training method [58] is decoupled into representation learning and classifier training. In addition, the Ensemble learning scheme [52] [54] first learns multiple experts with different data subsets, and then combines them to deal with the long-tailed distribution image classification problem. However, these methods all use a limited set of predefined labels to train the model, ignoring the availability of semantic feature information in the original label text of the image. After research, it was found that previous work was almost limited to a predetermined approach when dealing with imbalanced datasets, which relied entirely on visual models and completely ignored the semantic features of the original label text rich in the image itself. This may be a promising solution to impose additional supervision on insufficient data sources.

The large-scale visual-language pre-training model provides a new approach for image classification. Through open vocabulary supervision, pre trained visual-language models can learn powerful multimodal representations (input information can be expressed in multiple ways). Utilize semantic similarity between visual input and text input to transform visual recognition into a visual-language matching problem. Comparative visual language models such as CLIP [59] and ALIGN [60] provide new ideas for long-tailed classification tasks. The feature extractors of these models integrate image and text modalities, focusing on learning feature matching between different modalities. They have strong robustness, but lack the ability to model complex interactions between images and text.

Due to the significant difference in classification accuracy between majority and minority classes in commonly used long tail classification algorithms, the failure to utilize the semantic features of the original image label text, and the inability of existing contrastive visual-language models to model complex interactions between images and text, this paper proposes an enhanced contrastive visual language long-tailed image classification algorithm (ECVL). The algorithm uses a two-stage training method, designs the Loss function for text and image retrieval respectively, uses enhanced momentum to compare the Loss function to measure the learning degree of samples, and applies random enhancement to the categories with insufficient learning degree to further strengthen the learning of the model for minority samples.

3.1. Overall Framework

Similar to common contrastive visual-language models, the ECVL long-tailed image classification algorithm uses a two-stage training approach to transform visual recognition into a visual-language matching problem through similarity between visual and text inputs. The first stage mainly uses the visual features of the image and the semantic features of the original label text to train for most categories. The second stage first uses class balance for a few categories, and then uses linear adapters to carry out differentiated training. Finally, use the enhancement momentum to compare the loss function to measure the memory of the model for samples. For samples with insufficient memory, use the RandAugment [63] to select random enhancement methods Enhancing breadth can further enrich feature representation.

3.2. Contrasting Visual-Language Pre-training Model

Compare visual language models with a dual encoder architecture, including a language encoder $\mathcal{L}_{\text{enc}}$ and a visual encoder $\mathcal{V}_{\text{enc}}$. Given an input image I, use $\mathcal{V}_{\text{enc}}$ extracts the visual features of image I using the formula shown in (1). Similarly, use $\mathcal{L}_{enc}$ encodes the input text sequence T as its corresponding text feature, as shown in the formula (2).

$$f_v = \mathcal{V}_{\text{enc}}(I) \in \mathrm{R}^{d_v} \tag{1}$$

$$f_l = \mathcal{L}_{\text{enc}}(T) \in \mathrm{R}^{d_l} \tag{2}$$

After extracting the features of each modality, use two transformation matrices $W_v \in R^{d_v \times d}$ and $W_l \in R^{d_1 \times d}$ project the original visual and textual features into a shared embedding space, where v and u are d-dimensional normalized vectors, as shown in formula (3).

$$v = \frac{W_v^\top f_v}{\|W_v^\top f_v\|}, u = \frac{W_l^\top f_l}{\|W_l^\top f_l\|} \tag{3}$$

In the pre-training stage, for text-image pairs in a batch, the training goal is to shorten the distance between the same category and different categories, $\mathcal{L}_{v\to l}$ for text retrieval, $\mathcal{L}_{l\to v}$ for image retrieval, where τ Indicates that the temperature exceeds the parameter, τ represents the number of text image pairs in a batch. $\mathcal{L}_{v\to l}$ and $\mathcal{L}_{l\to v}$ as shown in formulas (4) and (5).

$$\mathcal{L}_{v\to l} = -\frac{1}{N}\sum_i^N log \frac{exp(v_i^\top u_i/\tau)}{\sum_{j=1}^N exp\,(v_i^\top u_j/\tau)} \tag{4}$$

$$\mathcal{L}_{l\to v} = -\frac{1}{N}\ \sum_i^N log \frac{exp(u_i^\top v_i/\tau)}{\sum_{j=1}^N exp\,(u_i^\top v_j/\tau)} \tag{5}$$

By converting the category labels of an image into a text sequence of "A photo of a {Class}", the matching score between the target image and the text sequence of all categories can be obtained. The category with the highest score is selected as the final predicted category. The normalized test image features are represented as v, and the normalized text features are represented as$\{u_1, \cdots, u_K\}$. Therefore, the category probability of the test image is shown in formula (5.6). Where p_i represents the probability of class i, and K represents the total number of candidate classes. Finally, the text label with the highest probability will be selected as the prediction result.

$$p_i = \frac{exp\,(\boldsymbol{v}^\top \boldsymbol{u}_i)/\tau}{\sum_{j=1}^K exp(\boldsymbol{v}^\top \boldsymbol{u}_j)/\tau} \tag{6}$$

3.3. Balanced Linear Adapter

The performance of contrastive visual-language models on the head and tail classes is balanced, while traditional contrastive learning methods such as PaCo [61] have lower performance on the tail classes due to a lack of training samples. Inspired by the zero-shot classification ability of visual-language comparison models, improvements were made on the basis of CLIP. The training of long tail data is divided into two stages. The first stage fully utilizes existing training data and ensures the performance of most categories, while the second stage focuses on improving the learning ability of a few categories. These two stages aim at the long-tailed and balance training samples respectively, and refine the comparison Loss function.

According to the research results proposed by Gururangan et al. [62] in Phase I, model pre-training with domain adaptation and task adaptation can greatly improve the performance of the target NLP task. Similarly, this applies equally to image classification tasks. In stage one, pre-training using the contrastive visual-language backbone model on the long-tailed target dataset is also beneficial for learning most class samples, making full use of available training data. Since the input of the model in Phase I is to process image category labels into text sequences, the comparison loss function used in the pre training is formula (4). The parameters of the text encoder and image encoder are updated instantly during training. After stage one training, most classes usually achieve good results, while minority class samples require stage two balance training. The processing process of the stage model is shown in Figure 2.

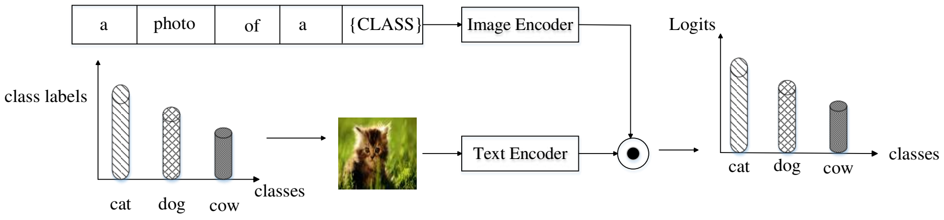


Figure 2. The model processing flow chart of Phase I.

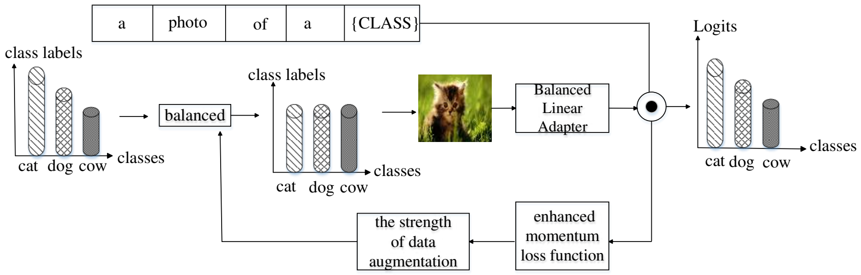


Figure 3. The model processing flow chart of Phase Ⅱ.

Due to the insufficient sample size and limited data for the tail category, direct training on the backbone in Phrase Ⅱ will result in overfitting. Therefore, in this stage, pre training is not conducted on the backbone, but instead, linear adapters and enhancements are used to optimize the visual language representation of a few category samples for momentum contrast loss. As shown in Figure 3, the processing of the semantic features of the original label text is the same as that of Stage I. Assuming the original image feature is f, the weight matrix of the linear adapter is $W \in R^{d \times d}$ The offset is R^d, and the processed image features can be expressed as formula (7).

$$f^{\star} = \lambda \cdot ReLU(W^{\top} f + b)\mathcal{L}_{DCVL} + (1 - \lambda)f \tag{7}$$

Among λ, the residual factor is used to dynamically combine the image features after fine-tuning in the second stage with the original image features in the first stage.

The enhanced momentum comparison loss function is used to measure the learning of the model for samples. Assuming x_i is the training sample on the long tail dataset, x_i The comparison loss is expressed as L_i. $\{L_{i,0}, \ldots, L_{i,t}, \ldots, L_{i,T}\}$ represents the tracking loss value L_i among T Epochs. Based on this, define the moving average momentum loss, as shown in formula (8).

$$\mathcal{L}^m_{i,0} = \mathcal{L}_{i,0}, \mathcal{L}^m_{i,t} = \beta \mathcal{L}^m_{i,t-1} + (1 - \beta)\mathcal{L}_{i,t} \tag{8}$$

The β is a hyperparameter that represents the smoothness of the loss. After training T Epochs using the above moving average momentum loss, the set of momentum losses for each sample can be obtained as $\{\mathcal{L}^m_{0,t}, \ldots, \mathcal{L}^m_{i,t}, \ldots, \mathcal{L}^m_{N,t}\}$, where N is the number of training samples in the dataset. Finally, the definition of momentum loss is normalized as follows, as shown in formula (9):

$$M_{i,t} = \frac{1}{2}(\frac{\mathcal{L}^m_{i,t} - \bar{\mathcal{L}_t}^m}{max\{|\mathcal{L}^m_{i,t} - \bar{\mathcal{L}_t}^m|\}_{i=0,\ldots,N}} + 1 \tag{9}$$

Among them $\bar{\mathcal{L}_t}^m$ represents the average momentum loss of the t Epoch. The range of M_i normalized values is $[0,1]$, with an average value of 0.5, reflecting the model's level of sample memory. To promote model learning, use M_i to control the occurrence and intensity of enhancement indicators. The specific approach follows RandAugment [63], randomly selecting k types of enhancements and using probability M_i and intensity $[0, M_i]$ apply each enhancement. Assuming that the enhancement set defined by RandAugment is $A = (A_1, \ldots, A_j, \ldots, A_K)$, where K is the enhancement amount, k enhancements are applied in each step. On this basis, define a memory enhancement function, as shown in formula (10).

$$\begin{gathered} \Psi(x_i; A, M_i) = a_1(x_i) \ldots a_k(x_i), \\ a_j(x_i) = \{ \begin{array}{ll} A_j(x_i; M_i\zeta) & u \sim \mathcal{U}(0,1) \& u < M_i \\ x_i & \text{other} \end{array} \end{gathered} \tag{10}$$

Among ζ sampling from uniformly distributed $\mathcal{U}(0,1)$. $A_j(x_i; \mathrm{M}_i\zeta)$ represent x_i undergoes the j enhancement with a strength of $\mathrm{M}_i\zeta$. Apply the selected k enhancements in sequence in A. For simplicity, use $\Psi(x_i)$ to represent $\Psi(x_i; A, M_i)$. In this paper, the enhanced momentum loss function is shown in Formula (11).

$$\mathcal{L}_{DCVL} = \frac{1}{N}\sum_i^N -log \frac{exp(\frac{f(\Psi(x_i))^\top f(\Psi(x_i^+))}{\tau})}{\Sigma_{x_i' \in X'} exp\,(\frac{f(\Psi(x_i))^\top f(\Psi(x_i'))}{\tau})} \tag{11}$$

Where X' represents $X^- \cup \{x_i^+\}$, x_i and x_i^+ represents two views of a sample, $x_i' \in X^-$ is a view of other samples. Intuitively, the enhanced momentum contrast Loss function is used to measure the memory of the model for the samples, and adaptively allocate appropriate enhanced strength for the samples with insufficient memory.

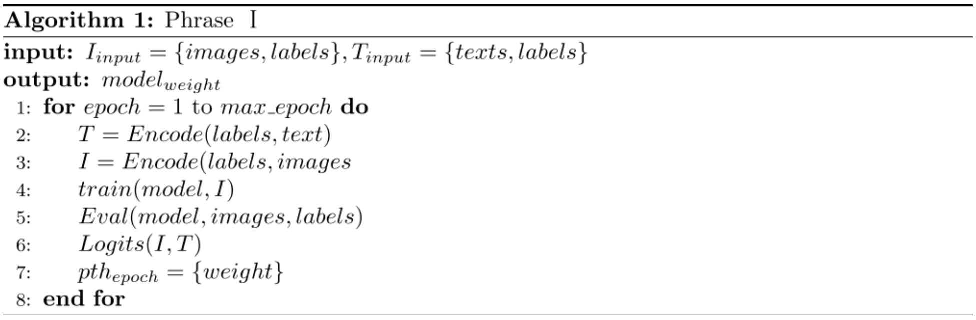
In the training process of stage 2, to avoid the model deviating from the head class, a class balance sampling strategy [8] is still used to construct a balanced training sample set. Assuming there are K classes in the target dataset to form a total of N training samples. The number of training samples for class j is expressed as n_j. Then use formula (12) to represent N.

$$N = \sum_{j=1}^{K} n_j \tag{12}$$

Assuming that classes are sorted in descending order, the long- tailed distribution means $n_i \geq n_j$ (i<j and $n_1 \gg n_K$). For class balanced sampling, the probability of sampling each data point from class j is defined as $q_j = 1/K$. In other words, to construct a balanced training sample set, first select a class from K candidate objects, and then sample a data point from the selected class. Finally, through stage two, use $\mathcal{L}_{v\to l}$ Fine tune the balanced training data.

3.4. Algorithm Description

Based on the introduction of the ECVL long tail image classification algorithm in the previous text, this section mainly introduces the training process of the long tail image classification algorithm based on enhanced contrastive visual language in two different stages: stage one and stage two, as shown in algorithm 1 and algorithm 2.



```
Algorithm 1: Phrase I
input: $I_{input} = \{images, labels\}, T_{input} = \{texts, labels\}$
output: $model_{weight}$
for $epoch = 1$ to $max\_epoch$ do
    $T = Encode(labels, text)$
    $I = Encode(labels, images$
    $train(model, I)$
    $Eval(model, images, labels)$
    $Logits(I, T)$
    $pth_{epoch} = \{weight\}$
end for
```

Algorithm 1 is the training process for model stage one, which simultaneously trains the visual and language branches of the visual language model. In each Epoch, it is preferred to input images and corresponding category text information; Afterwards, the visual features of the image and the semantic features of the original label text are extracted using formulas (1) and (2), respectively; And then use $\mathcal{L}_{v\to l}$ Perform text retrieval using $\mathcal{L}_{l\to v}$ Perform image retrieval to obtain associated image and text information; Finally, use formula (6) to predict the image category, and evaluate the prediction results using evaluation indicators after the classification is completed.

Algorithm 2 is the training process for model stage 2. The model first balances a few types of samples, and then fine tunes the linear adapter. After fine tuning, it uses the enhanced momentum Loss function described according to formula (11) to evaluate the sample learning situation. For samples with insufficient representation of learning features, RA random enhancement is used. Finally, the features learned in these two stages are dynamically fused and output.

Algorithm 2: Phrase II

```
input: $I_{input} = \{images, labels\}, T_{input} = \{texts, lables\}, model_{stage1}$
output: $weight$
$model = load(bestmodel)$
for $epoch = 1$ to $max\_epoch$ do
    if $epoch >= 2$ then
        $I = Rebalance(Momtum)$
    end if
    $Muomtum = model(I, labels, epoch)$
    $train(model, I)$
    $eval(model, images, lables)$
    $Logit(model, I, T)$
    $pth_{epoch} = \{weight\}$
end for
```

4. Experiments

The ECVL algorithm takes 229 seconds to infer 100 images on a single NVIDIA A100 40G GPU. In order to verify the performance of the proposed ECVL long-tailed image classification algorithm, experiments were carried out on three common long-tailed distribution data sets CIFAR100-LT, Places-LT and ImageNet-LT to analyze the performance of this algorithm, and ablation experiments were conducted to prove the role of enhanced momentum in comparison with Loss function and random enhancement. This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

4.1. Long-tailed Datasets

4.1.1. CIFAR100-LT

CIFAR100-LT [27] is the dataset obtained by long-tailed of the data set CIFAR100. It is created by reducing the number of training samples of each class through the Exponential function, and the test set remains unchanged.

4.1.2. Places-LT

Places-LT [64] is a dataset obtained by long-tailed transformation based on the Places [65] dataset. The Places dataset contains 10 million images classified by scene, and the label of the sample represents the meaning of the scene. It is currently the largest scene dataset in the world with the largest sample size, as shown in Figure 2. The long-tailed rate of the training set in the Places-LT dataset is 996, and the number of categories is 365. The total sample size in the training set is 62500, and the sample size in the test set is 7300. The category with the largest sample size in the training set is 4980, while the category with the smallest sample size is 5. The ratio of the maximum to minimum sample size is 996, making it the dataset with the largest long tail rate used in this article.

4.1.3. ImageNet-LT

ImageNet-LT [64] was obtained through the long-tailed ImageNet dataset, with a total of 1000 categories. The total number of samples in the dataset exceeds 186K, with 116K training samples, 20K validation samples, and 50K testing samples. In ImageNet-LT, the long-tailed rate in the training set is 256, the maximum class sample size is 1280, and the minimum class sample size is 5. This dataset simulates the distribution of long tailed data commonly found in real life. The data in the training set is divided into three parts. The header category contains categories with a sample size greater than 100, the middle category contains categories with a sample size greater than 20 but less than 100, and the tail category contains categories with a sample size smaller than 20.

4.2. Experimental Design and Validation

All experiments in this article are based on Python implementation, version 1.7.1. The server system used in the experiment is Ubuntu 20.04, CUDA version 10.1, and the AdamW optimizer and 300 Epochs are used to train the model. The experiment was trained on a NVIDIA A100 40G * 8 GPU device. The configuration details of the experimental environment are shown in Table 1.

Table 1. Experimental environment

Name	Model/parameter
CPU	Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz
GPU	NVIDIA A100 40G * 8
Memory	128G
Hard disk	1T
Operating system	Ubuntu20.04
CUDA	CUDA Version 10.1
Deep learning framework	Pytorch 1.7.1
Development language	Python 3.7

4.2.1. Experimental Results and Analysis of CIFAR100-LT

In this experiment, the backbone network used by ECVL was ResNet-50, and the experiment was conducted on the long-tailed distribution dataset CIFAR100-LT. The experimental results are shown in Table 2. The enhanced contrastive visual language long tail classification algorithm proposed in this chapter has an accuracy of 20.5% and 17.2% higher in tail categories than RIDE [46] and TADE [54], and an accuracy of 6.7% and 6.0% higher in all categories compared to RIDE [46] and TADE [54], respectively. The F1 values are 14.3% and 11.8% higher than RIDE [46] and TADE [54], respectively. ECVL improves the accuracy difference between majority and minority classes, not only improving the performance of majority classes but also improving the recognition accuracy of minority classes. It also proves that using the semantic features of the original label text as supplementary information for classification is helpful in improving the performance of the model.

Table 2. Experimental results of ECVL on CIFAR100-LT

Model	Backbone	Accuracy				F_1
		Head	Medium	Tail	All	
OLTR[64]	ResNet-32	61.8%	41.4%	17.6%	41.2%	52.3%
LDAM[22]	ResNet-32	61.5%	41.7%	20.2%	42.0%	52.9%
cRT[4]	ResNet-32	64.0%	44.8%	18.1%	43.3%	51.9%
RIDE[46]	ResNet-32	69.3%	49.3%	26.0%	49.1%	57.3%
TADE[54]	ResNet-32	65.4%	49.3%	29.3%	49.8%	58.8%
BALLAD[66]	ResNet-50	62.4%	52.3%	38.2%	51.6%	62.1%
ECVL	ResNet-50	65.0%	57.2%	46.5%	55.8%	70.6%

4.2.2. Experimental Results and Analysis of ImageNet-LT

In this experiment, the comparative experimental results are shown in Table 3. Compared with the long-tailed image classification algorithm that only uses contrastive learning, the enhanced contrastive visual language proposed in this chapter has an accuracy of 29.2% higher in tail categories than PaCo [61], 13.6% higher in all categories than PaCo [61], and a F1 value of 14.9% higher than PaCo [61]. The accuracy of CWTA in tail categories is 7.9% higher than that of BALLAD [66], 3.4% higher in all categories, and 11.2% higher in F1 values than BALLAD [66]. This not only proves that the proposed enhanced momentum contrast Loss function is more effective than only using contrast loss, but also proves that using text image pairs for pre training is helpful for improving model performance.

Table 3. Experimental results of ECVL on ImageNet-LT

Model	**Backbone**	**Accuracy**				**F$_1$**
		Head	**Medium**	**Tail**	**All**	
OLTR[64]	ResNeXt-50	43.2%	35.1%	18.5%	35.6%	47.6%
cRT[4]	ResNeXt-50	61.8%	46.2%	27.4%	49.6%	53.7%
LWS[4]	ResNeXt-152	62.2%	50.1%	35.8%	52.8%	-
	ResNeXt-50	60.2%	47.2%	30.3%	49.9%	50.6%
ResLT[53]	ResNeXt-152	63.5%	50.4%	34.2%	53.3%	-
	ResNeXt-50	63.0%	50.5%	35.5%	52.9%	55.2%
Balanced Softmax[28]	ResNeXt-101	63.3%	53.3%	40.3%	55.1%	-
	ResNet-50	66.7%	52.9%	33.0%	55.0%	-
PaCo[61]	ResNeXt-50	67.7%	53.8%	34.2%	56.2%	-
	ResNet-50	65.0%	55.7%	38.2%	57.0%	62.3%
BALLAD[66]	ResNeXt-50	67.5%	56.9%	36.7%	58.2%	-
	ResNet-50	71.0%	66.3%	59.5%	67.2%	66.0%
ECVL	ResNet-50	73.2%	69.8%	67.4%	70.6%	77.2%

4.2.3. Experimental Results and Analysis of Places-LT

In this experiment, the ECVL algorithm uses ResNet-50 as the backbone network and conducts experiments on the long-tailed distribution dataset Places-LT. The comparative experimental results are shown in Table 4. Compared with the long-tailed image classification algorithm that only uses contrastive learning, the ECVL long-tailed image classification algorithm has an accuracy of 10.1% higher on tail classes than PaCo [61], an accuracy of 6.0% higher on all classes than PaCo [61], and an F1 value of 7.3% higher than PaCo [61]; Compared with the comparative visual language model BALLAD [66], the accuracy on the tail class is improved by 1.3%. The experiment shows that the enhanced momentum contrast Loss function in ECVL is more effective than only using the contrast loss function, and it is helpful to train the model by randomly enhancing the samples with insufficient learning after processing the enhanced momentum contrast loss function.

Table 4. Experimental results of ECVL on Places-LT

Model	**Backbone**	**Accuracy**				**F$_1$**
		Head	**Medium**	**Tail**	**All**	
OLTR[64]	ResNet-152	44.7%	37.0%	25.3%	35.9%	46.4%
cRT[4]	ResNet-152	42.0%	37.6%	24.9%	36.7%	45.5%
LWS[4]	ResNet-152	40.6%	39.1%	28.6%	37.6%	46.2%
ResLT[53]	ResNet-152	39.8%	43.6%	31.4%	39.8%	51.2%
PaCo[61]	ResNet-50	37.5%	47.2%	33.9%	41.2%	52.3%
BALLAD[66]	ResNet-50	46.7%	48.0%	42.7%	46.5%	56.8%
	ResNet-101	48.0%	48.6%	46.0%	47.9%	-
	ViT-B/16	49.3%	50.2%	48.4%	49.5%	-
ECVL	ResNet-50	48.6%	48.3%	44.0%	47.2%	59.6%

4.3. *Experimental Design and Validation*

The ECVL long-tailed image classification algorithm proposed in this paper uses the visual characteristics of the image itself and the semantic characteristics of the original label text, the enhanced momentum contrastive loss function and RandAugment to complete the long tail classification, and performs well on the public long tail dataset. In order to verify the effectiveness of enhanced momentum vs. Loss function and random enhancement in the model, this section conducts ablation experimental analysis on them on different public long-tailed distribution data sets, and the experimental results are shown in Table 5 to Table 7. On CIFAR100-LT, the difference in classification accuracy between most categories and minority categories decreased by 1.8% compared with only using enhanced momentum to compare the Loss function and neither using enhanced momentum to compare the Loss function nor using random enhancement; With the enhanced momentum contrastive loss function and the random enhancement module, the classification accuracy of most categories and minority categories increased by 2.5% and 3.4% respectively than without the random enhancement module. On ImageNet-LT, compared with using only the enhanced momentum contrastive loss function module and neither the enhanced momentum comparison Loss function nor the random enhancement module, the difference between the classification accuracy of most classes and minority classes decreased by 0.7%; Compared with the loss function and the random enhancement module with enhanced momentum, the classification accuracy of most categories and minority categories increased by 0.7% and 1.2% respectively. Through analysis, it is found that although the accuracy of all categories is improved by not using the enhanced momentum contrastive loss function or the random enhancement module, there is still a large difference in the accuracy difference between the majority of categories and the minority in the final fine-tuning process; After adding the enhanced momentum contrastive loss function, the accuracy difference between the majority and minority classes has improved, but in some cases there is degradation (such as Places-LT dataset). The enhanced momentum comparison between the loss function and the random enhancement module can improve the overall accuracy and reduce the accuracy difference between the majority and minority.

Table 5. Ablation Experiment of ECVL on CIFAR100-LT

Moduel	**Accuracy**				**F$_1$**
	Head	**Medium**	**Tail**	**All**	
no momentum contrast loss + no random augment	62.4%	52.3%	38.2%	51.6%	62.1%
momentum contrast loss	62.5%	53.3%	40.1%	52.4%	65.8%
momentum contrast loss +random Augment	65.0%	57.2%	46.5%	55.8%	70.6%

Table 5. Ablation Experiment of ECVL on ImageNet-LT

Module	**Accuracy**				**F$_1$**
	Head	**Medium**	**Tail**	**All**	
no momentum contrast loss + no random augment	71.0%	66.3%	59.5%	67.2%	66.0%
momentum contrast loss	72.5%	68.7%	63.2%	69.4%	70.8%
momentum contrast loss +random Augment	73.2%	69.9%	67.4%	70.6%	77.2%

Module	**Accuracy**				**F$_1$**
	Head	**Medium**	**Tail**	**All**	
no momentum contrast loss + no random augment	46.7%	48.0%	42.7%	46.2%	56.8%

momentum contrast loss	47.0%	47.5%	43.2%	46.5%	58.3%
momentum contrast loss +random Augment	48.6%	48.3%	44.0%	47.2%	59.6%

5. Conclusions

This article first analyzes the advantages and disadvantages of existing long-tailed image classification methods, proposes a long-tailed classification algorithm based on enhanced contrastive visual-language, and then elaborates on the algorithm framework, algorithm design details, algorithm design process, and comparative experimental analysis. In addition, this article conducts comparative experiments and ablation research analysis on three long tailed datasets: CIFAR100-LT, ImageNet-LT, and Places-LT.

Compared with BALLAD method, ECVL on CIFAR100-LT reduces the difference in classification accuracy between majority and minority classes by 5.7%, and increases F_1 by 8.5%. Compared with BALLAD, ECVL on ImageNet-LT reduces the difference in classification accuracy between majority and minority classes by 1.7%, and increases F1 by 11.2%. Compared with BALLAD, the F_1 of ECVL on Places-LT has increased by 5.8%. On Places-LT, compared with using only the enhanced momentum contrast loss function module and neither the enhanced momentum contrast loss function nor the random enhancement module, the difference in classification accuracy between most classes and minority classes decreased by 1.8%. Compared with the non-random enhancement module, the accuracy rate of minority classification and F_1 of the enhanced momentum contrast loss function and random enhancement module increased by 0.7% and 1.3% respectively. The classification accuracy, difference in accuracy between majority and minority categories, F_1, and convergence of the model in different quantity categories in the experiment have demonstrated the effectiveness of the algorithm proposed in this paper.

Author Contributions: Conceptualization, Y.S., M.L. and B.W.; methodology, M.L..; formal analysis, M.L.; investigation, M.L..; writing—original draft preparation, M.L.; funding acquisition, Y.S.

Funding: This research was supported in part by the National Natural Science Foundation of China (Grant No. 61872043), State Key Laboratory of Computer Architecture (ICT, CAS) under Grant No. CARCHA202103.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tas S, Sari O, Dalveren Y, et al. Deep learning-based vehicle classification for low quality images[J]. Sensors, 2022, 22(13): 4740.
2. Berwo M A, Khan A, Fang Y, et al. Deep Learning Techniques for Vehicle Detection and Classification from Images/Videos: A Survey[J]. Sensors, 2023, 23(10): 4832.
3. Wang Z, Shen H, Xiong W, et al. Method for Diagnosing Bearing Faults in Electromechanical Equipment Based on Improved Prototypical Networks[J]. Sensors, 2023, 23(9): 4485.
4. Kang B, Xie S, Rohrbach M, et al. Decoupling representation and classifier for long-tailed recognition[J]. International Conference on Learning Representations, 2020.
5. Wang T, Li Y, Kang B, et al. The devil is in classification: A simple framework for long-tail instance segmentation[C]//Proceedings, Part XIV 16. Springer International Publishing, 2020: 728-744.
6. Park M, Song H J, Kang D O. Imbalanced Classification via Feature Dictionary-Based Minority Oversampling[J]. IEEE Access, 2022, 10: 34236-34245.Author 1, A.B. Title of Thesis. Level of Thesis, Degree-Granting University, Location of University, Date of Completion.

7. Li T, Wang Y, Liu L, et al. Subspace-based minority oversampling for imbalance classification[J]. Information Sciences, 2023, 621: 371-388.
8. Lee Y S, Bang CC. Framework for the Classification of Imbalanced Structured Data Using Under-sampling and Convolutional Neural Network[J]. Information Systems Frontiers, 2021: 1-15.
9. Lehmann D, Ebner M. Subclass-based Undersampling for Class-imbalanced Image Classification[C]//VISIGRAPP (5: VISAPP). 2022: 493-500.
10. Farshidvard A, Hooshmand F, MirHassani S A. A novel two-phase clustering-based under-sampling method for imbalanced classification problems[J]. Expert Systems with Applications, 2023, 213: 119003.
11. Ding H, Wei B, Gu Z, et al. KA-Ensemble: towards imbalanced image classification ensembling undersampling and over-sampling[J]. Multimedia Tools and Applications, 2020, 79: 14871-14888.
12. Swana E F, Doorsamy W, Bokoro P. Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset[J]. Sensors, 2022, 22(9): 3246.
13. Gupta A, Dollar P, Girshick R. Lvis: A dataset for large vocabulary instance segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5356-5364.
14. Peng J, Bu X, Sun M, et al. Large-scale object detection in the wild from imbalanced multi-labels[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9709-9718.
15. Hu X, Jiang Y, Tang K, et al. Learning to segment the tail[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 14045-14054.
16. Wu J, Song L, Wang T, et al. Forest r-cnn: Large-vocabulary long-tailed object detection and instance segmentation[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 1570-1578.
17. Zhou B, Cui Q, Wei X S, et al. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9719-9728.
18. Zang Y, Huang C, Loy C C. Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 3457-3466.
19. Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
20. Hermans A, Beyer L, Leibe B. In Defense of the Triplet Loss for Person Re-Identification[J]. 2017.
21. Cui Y, Jia M, Lin T Y, et al. Class-balanced loss based on effective number of samples[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 9268-9277.
22. Cao K, Wei C, Gaidon A, et al. Learning imbalanced datasets with label-distribution-aware margin loss[J]. Advances in neural information processing systems, 2019, 32.
23. Wu T, Huang Q, Liu Z, et al. Distribution-balanced loss for multi-label classification in long-tailed datasets[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. Springer International Publishing, 2020: 162-178.
24. Tan J, Wang C, Li B, et al. Equalization loss for long-tailed object recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11662-11671.
25. Tan J, Lu X, Zhang G, et al. Equalization loss v2: A new gradient balance approach for long-tailed object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 1685-1694.
26. Wang J, Zhang W, Zang Y, et al. Seesaw loss for long-tailed instance segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 9695-9704.
27. Hong Y, Han S, Choi K, et al. Disentangling label distribution for long-tailed visual recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 6626-6636.
28. Ren J, Yu C, Ma X, et al. Balanced meta-softmax for long-tailed visual recognition[J]. Advances in neural information processing systems, 2020, 33: 4175-4186.
29. Deng Z, Liu H, Wang Y, et al. Pml: Progressive margin loss for long-tailed age classification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 10503-10512.
30. Wu T, Liu Z, Huang Q, et al. Adversarial robustness under long-tailed distribution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 8659-8668.
31. Xiao L, Xu J, Zhao D, et al. Adversarial and Random Transformations for Robust Domain Adaptation and Generalization[J]. Sensors, 2023, 23(11): 5273.

32. Park S, Kim J, Jeong H Y, et al. C2RL: Convolutional-Contrastive Learning for Reinforcement Learning Based on Self-Pretraining for Strong Augmentation[J]. Sensors, 2023, 23(10): 4946.
33. Zhong Z, Cui J, Liu S, et al. Improving calibration for long-tailed recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 16489-16498.
34. Li S, Gong K, Liu C H, et al. Metasaug: Meta semantic augmentation for long-tailed visual recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 5212-5221.
35. Wang Y, Pan X, Song S, et al. Implicit semantic data augmentation for deep networks[J]. Advances in Neural Information Processing Systems, 2019, 32.
36. Yin X, Yu X, Sohn K, et al. Feature transfer learning for face recognition with under-represented data[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5704-5713.
37. Liu J, Sun Y, Han C, et al. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 2970-2979.
38. Chu P, Bian X, Liu S, et al. Feature space augmentation for long-tailed data[C]//Proceedings, Part XXIX 16. Springer International Publishing, 2020: 694-710.
39. Cui Y, Song Y, Sun C, et al. Large scale fine-grained categorization and domain-specific transfer learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4109-4118.
40. Yang Y, Xu Z. Rethinking the value of labels for improving class-imbalanced learning[J]. Advances in neural information processing systems, 2020, 33: 19290-19301.
41. He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9729-9738.
42. Li T, Wang L, Wu G. Self-supervision to distillation for long-tailed visual recognition[C]// Proceedings of the IEEE/CVF international conference on computer vision. 2021: 630-639.
43. Wei H, Tao L, Xie R, et al. Open-Sampling: Exploring Out-of-Distribution data for Re-balancing Long-tailed datasets[C]//International Conference on Machine Learning. PMLR, 2022: 23615-23630.
44. Changpinyo S, Sharma P, Ding N, et al. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3558-3568.
45. Xiang L, Ding G, Han J. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. Springer International Publishing, 2020: 247-263.
46. Wang X, Lian L, Miao Z, et al. Long-tailed Recognition by Routing Diverse Distribution-Aware Experts[C]//International Conference on Learning Representations. 2021.
47. Li T, Wang L, Wu G. Self supervision to distillation for long-tailed visual recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 630-639.
48. He Y Y, Wu J, Wei X S. Distilling virtual examples for long-tailed recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 235-244.
49. Wei C, Sohn K, Mellina C, et al. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 10857-10866.
50. Zhang C, Pan T Y, Li Y, et al. MosaicOS: a simple and effective use of object-centric images for long-tailed object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 417-427.
51. Guo H, Wang S. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 15089-15098.
52. Cai J, Wang Y, Hwang J N. Ace: Ally complementary experts for solving long-tailed recognition in one-shot[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 112-121.
53. Cui J, Liu S, Tian Z, et al. Reslt: Residual learning for long-tailed recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
54. Zhang Y, Hooi B, Hong L, et al. Test-Agnostic Long-Tailed Recognition by Test-Time Aggregating Diverse Experts with Self-Supervision[J]. 2021.

55. Tang K, Huang J, Zhang H. Long-tailed classification by keeping the good and removing the bad momentum causal effect[J]. Advances in Neural Information Processing Systems, 2020, 33: 1513-1524.
56. Zhou B, Lapedriza A, Khosla A, et al. Places: A 10 million image database for scene recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(6): 1452-1464.
57. Zhu L, Yang Y. Inflated episodic memory with region self-attention for long-tailed visual recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 4344-4353.
58. Kang B, Li Y, Xie S, et al. Exploring balanced feature spaces for representation learning[C]//International Conference on Learning Representations. 2021.
59. Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.
60. Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]//International Conference on Machine Learning. PMLR, 2021: 4904-4916.
61. Cui J, Zhong Z, Liu S, et al. Parametric contrastive learning[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 715-724.
62. Gururangan S, A Marasović, Swayamdipta S, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks[J]. 2020.
63. Lingchen TC, Khonsari A, Lashkari A, et al. UniformAugment: A Search-free Probabilistic Data Augmentation Approach[J]. 2020.
64. Liu Z, Miao Z, Zhan X, et al. Large-scale long-tailed recognition in an open world[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2537-2546.
65. Zhou B, Lapedriza A, Khosla A, et al. Places: A 10 million image database for scene recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(6): 1452-1464.
66. Ma T, Geng S, Wang M, et al. A Simple Long-Tailed Recognition Baseline via Vision-Language Model[J]. 2021.