A Survey on Long-tailed Visual Recognition

Lu Yang∗**, He Jiang**∗**, Qing Song**†**, Jun Guo**

arXiv:2205.13775v1 [cs.CV] 27 May 2022

Abstract The heavy reliance on data is one of the major reasons that currently limit the development of deep learning. Data quality directly dominates the effect of deep learning models, and the long-tailed distribution is one of the factors affecting data quality. The long-tailed phenomenon is prevalent due to the prevalence of power law in nature. In this case, the performance of deep learning models is often dominated by the head classes while the learning of the tail classes is severely underdeveloped. In order to learn adequately for all classes, many researchers have studied and preliminarily addressed the long-tailed problem. In this survey, we focus on the problems caused by long-tailed data distribution, sort out the representative long-tailed visual recognition datasets and summarize some mainstream long-tailed studies. Specifically, we summarize these studies into ten categories from the perspective of representation learning, and outline the highlights and limitations of each category. Besides, we have studied four quantitative metrics for evaluating the imbalance, and suggest using the Gini coefficient to evaluate the long-tailedness of a dataset. Based on the Gini coefficient, we quantitatively study 20 widely-used and large-scale visual datasets proposed in the last decade, and find that the long-tailed phenomenon is widespread and has not been fully studied. Finally, we provide several future directions for the development of long-tailed learning to provide more ideas for readers.
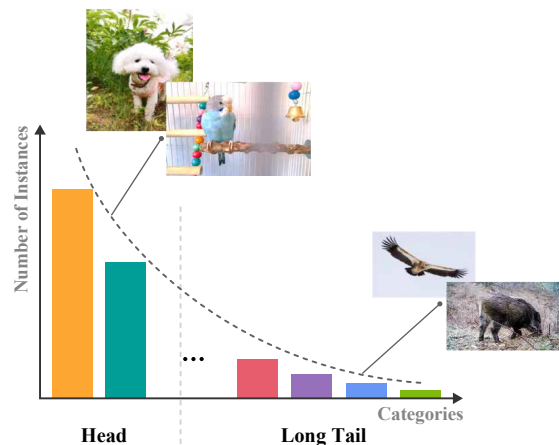
Lu Yang, Beijing University of Posts and Telecommunications (soeaver@bupt.edu.cn)

He Jiang, Beijing University of Posts and Telecommunications (JiangHe@bupt.edu.cn)

Qing Song, Beijing University of Posts and Telecommunications (priv@bupt.edu.cn)

Jun Guo, Beijing University of Posts and Telecommunications (guo-jun@bupt.edu.cn)

∗ Equal Contribution

† Corresponding author: Qing Song

Fig. 1: **Distribution of Long-tailed dataset**. In nature, there are cases where a few individuals make a large contribution and data tend to show a long-tailed distribution. For example, dog and budgie are common classes, while most other classes such as alpine vulture, tetra are uncommon classes.

1 Introduction

"From politics to public relations, from music scores to college sports, the long tail is everywhere."

– Chris Anderson, The Long Tail [5]

The advent of deep neural networks has led to remarkable breakthroughs in many fields such as computer vision [47, 79, 97, 132, 159], natural language processing [35, 86, 87], and reinforcement learning [141]. However, deep learning models learn features from large amounts of data, and thus
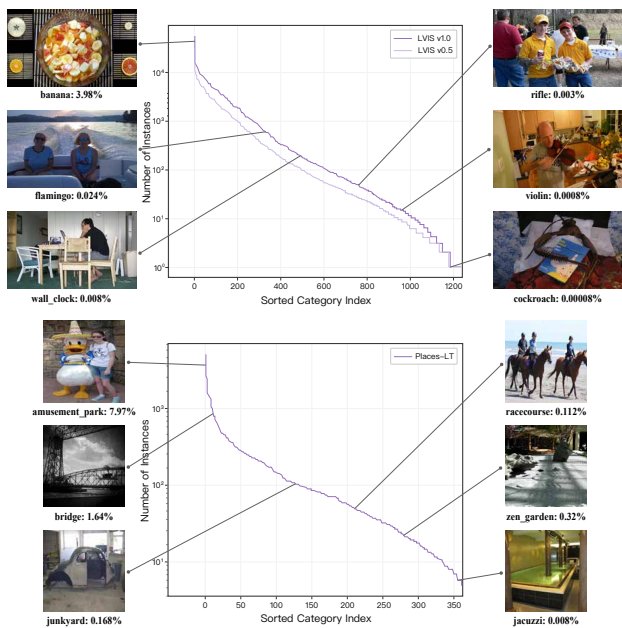
Fig. 2: **Some common long-tailed distribution datasets as well as their long-tailedness**. **Top**: examples of LVIS [53], which is a large-scale fine-grained vocabulary annotation dataset for instance segmentation. **Bottom**: examples of Places-LT [108], which is a scene-centric recognition dataset.

inevitably have a heavy dependence on it. Therefore, deep learning faces the challenges brought by the existence of problems in the data itself.

In real life, there exists a distribution of random variables that is more extensive than the positive-terrestrial distribution, *i.e.* the long-tailed distribution. It is mainly reflected in the fact that a small number of individuals usually make a large number of contributions, where a few classes occupy the majority of the dataset (*i.e.* head classes), while the majority of classes have very little data samples (*i.e.* tail classes), as shown in Fig. 1. Long-tailed distribution can be reflected in many cases. For example, in the field of economics where the long-tailed theory first emerged, head and tail are used to distinguish between red ocean markets and blue ocean markets. In business sales, "best-selling goods" have fewer classes but high sales volume, which belonging to the head classes, while the "cold goods" have a huge variety, but the sales volume of each class is low, which belonging to the tail classes. In visual recognition, there are also many subfields that involve long-tailed problems, such as instance segmentation, scene classification, *etc.* as shown in Fig. 2.

Chris Anderson [5], who first proposed the long-tailed theory, suggested that the future of business and culture lies not in the popular products but in the infinitely long-tailed demand curve, which shows the importance of the research for the tail classes. From the perspective of machine learning re-

search objectives and application implications, we should not only focus on the head classes but also give equal attention to the tail classes in data research.

We must admit that the success of deep learning is inseparable from the large-scale well-annotated datasets, such as ImageNet-1K [131], COCO [99], and Places365 [200], *etc*. These large datasets are artificially balanced, and the classes approximately obey a uniform distribution. In deep learning, we need to be able to learn well for all classes, so artificially balanced data will undoubtedly drive the development of deep learning. Therefore, we should realize that some of the progress made in the field of deep learning is partly driven by this artificial balance by force. In reality, however, this forced balancing of data is inappropriate. On the one hand, forcing class balancing within the dataset by hand is not in line with the natural conditions of data distribution. On the other hand, making the data distribution as balanced as possible by collecting more tail examples is a notoriously difficult task [41, 83, 99, 149], and the naturally existing power law can be a huge challenge in constructing a balanced dataset. Thus, the solution to the long-tailed problem is imperative.

1.1 Previous surveys and our contributions

To our knowledge, this work is not the first review to summarize the long-tailed phenomenon in visual recognition. Zhang *et al*. [191] summarized the technical guide for long-tailed visual recognition earlier this year, aiming to improve the performance of some long-tailed benchmarks through a reasonable combination of existing tricks. Although Zhang *et al*.'s work did not comprehensively introduce and analyze the long-tailed phenomenon in visual recognition, their quantitative analysis of some methods can still be regarded as an early overview of this field. In addition, Zhang *et al*. [190] conducted a survey on the topic of deep long-tailed learning in the same period of our work, which grouped the existing deep long-tailed learning studies into three categories (class re-balancing, information augmentation and module improvement), and proposed a new evaluation metric (relative accuracy). In contrast, our work analyzes the long-tailed visual recognition more deeply, divides the existing methods more finely, and quantitatively analyzes the long-tailed phenomenon of mainstream large-scale visual datasets. We recommend that readers also read the above two works [190, 191] in order to have a more comprehensive understanding of long-tailed visual recognition.

This review aims to comprehensively analyze the long-tailed problem in visual recognition, summarize the highlights and limitations of mainstream methods, and provide an outlook on future research directions. At the technical level, we not only sort out some general long-tailed problem solving methods. We also recommend to use the *Gini coefficient* [46] as a measure of the datasets' long-tailedness.

At the application level, through the general research on the long-tailed phenomenon, we find that the long-tailed problem is common in the mainstream large-scale visual datasets [3, 52, 133, 166, 194, 196, 201], which reveals that the research on the long-tailed phenomenon in many fields is not enough. In general, the contribution of this survey can be summarized as follows:

– We conduct a comprehensive review of the advanced long-tailed studies, finely summarize them into ten categories from the perspective of representation learning, and outline the highlights and limitations of each category.

– We have studied four quantitative metrics to evaluate imbalance, deeply compared their characteristics, and proposed to use Gini coefficient to evaluate the long-tailedness of a dataset.

– Beyond the existing research scope, we further study the long-tailed phenomenon of 20 widely-used and large-scale visual datasets proposed in the last decade, and reveal that this problem has not been fully studied in some fields.

– We elaborate on open problems and opportunities in this field to facilitate future research.

1.2 Organization

The rest of this paper is organized as follows. In Sec. 2, we provide the definition of the long-tailed problem and compare the similarities and differences between it and related research fields. In Sec. 3, we introduce some commonly used long-tailed datasets as well as their evaluation metrics, and use Gini coefficient to quantitatively evaluate the long-tailedness of datasets. In Sec. 4, we give an overview of approaches to solving the long-tailed problem and summarize them based on the existing studies. In Sec. 5, we report the performance of some popular studies on CIFAR-10/100-LT, ImageNet-LT, Places-LT, iNaturalist 2017 & 2018 as well as LVIS v0.5 & v1.0. In Sec. 6, we further study the long-tailed phenomenon of mainstream large-scale visual datasets proposed in the last decade. Future directions for the long-tailed problem are given in Sec. 7, and Sec. 8 concludes the whole paper.

2 Overview

To provide readers with the necessary background knowledge, in Sec. 2.1, we formulate the task, and analyze the key challenges as well as the driven factors of the long-tailed distribution. And in Sec. 2.2, we establish linkages to other relevant fields, and compare their similarities and differences.

2.1 Problem Definition

In nature or real life, there exists a distribution of random variables that is more widespread than the positive-terminus distribution, *i.e.* the long-tailed distribution. It is actually a colloquial expression for the power laws and Pareto characteristics in statistics. The protruding part in the curve is called "head", and the class corresponding to this part is called head class or frequent class. The relatively flat part on the right is called "tail", and the corresponding class is called tail class or rare class. Currently, some CNN-based models [60, 105, 128, 147] perform well on balanced datasets, but these networks tend to perform poorly on long-tailed datasets.

The long-tailed phenomenon is inherently present in large vocabulary scenarios, making model learning with long-tailed distributed data challenging in a number of ways:

From the perspective of model learning. First, since the data in the tail classes is usually insufficient to represent its true distribution, this poses a significant challenge for classifiers: a good classifier aims to provide a good decision boundary for the model, yet when a class is severely under-represented, it becomes more difficult to determine the location of the decision boundary, which can affect the performance of the model in the dataset. Besides, due to the rich training samples of the head classes, the head classes will be more adequately studied. Based on the case of tail classes severely under-learned, the positive gradient generated by the tail classes will inevitably be overwhelmed by the head classes, which makes it more difficult to learn effective feature extractors and classifiers for the tail classes.

From the perspective of transfer learning, we take training data as the source domain and inference-time data as the target domain. For the long-tailed data, the training set satisfies the long-tailed distribution, while the test set usually satisfies the uniform distribution as shown in Fig. 3. There is no guarantee of having similar data distributions between the source and target tasks due to the large gap between the head and tail item distributions [164, 189]. Using conventional methods (*e.g.*, Cross-Entropy loss, or simple fine-tuning, *etc.*) will result in the poor performance of the tail classes. This is because the traditional deep learning methods assume that the training data and the test data satisfy the independently and identically distributed condition. Therefore the quality of knowledge transfer can also be greatly affected. And the problem of *target shift* can arise because the features learned on the training set are different from the features belonging to the corresponding labels in the test set. As the number of tail classes' representative examples are insufficient, which are susceptible to noise and other factors.

In addition to the challenges posed by the long-tailed in the classification task described above, we also investigate
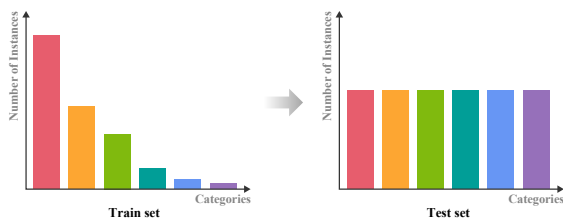
Fig. 3: **Differences in data distribution between the training and test sets**. For long-tailed dataset, the training set satisfies the long-tailed distribution, while the test set usually conforms to a balanced distribution.

the challenges raised by the long-tailed in the object detection task as well as in the instance segmentation task.

For the long-tailed object detection task, there is competition for the category of the boxes between the tail and head classes. In the long-tailed distribution, the data of the tail class is often much smaller than that of the head class, which makes it likely that the sampling of the tail class is directly lost or the tail class is classified as background in the box sampling phase. In addition, for the NMS phase, the long-tailed distribution of data may cause a large number of missed detection [28], which leads to poor detection results.

For the long-tailed instance segmentation task, the two-stage method is widely used, such as Mask R-CNN, which needs to execute the detection first, so it also faces the limitations of the long-tailed object detection task. In addition, the sparse tail samples make it difficult to go through the learning to distinguish it well from the background, resulting in the inaccurate masks of the tail classes [183].

2.2 Relevant Learning Problems

Among the research in machine learning, there are some areas with strong relevance to long-tailed visual recognition, such as imbalance learning and few-shot learning. In this section, we compare long-tailed visual recognition with these two domains (Sec. 2.2.1 and Sec. 2.2.2), respectively. And we also illustrate the similarities and differences between the three in Sec. 2.2.3.

2.2.1 Imbalanced Learning

Imbalance learning is a widespread problem in deep learning, and it does not only refer to the imbalance of training data. Kemal *et al*. [118] proposed that imbalance problems are divided into four types, namely *class imbalance*, *scale imbalance*, *spatial imbalance* and *objective imbalance*. For the long-tailed visual recognition, the current study is mainly based on the image long-tailed distribution level. For the imbalanced distribution of training data, Buda *et al*. [11]

define and investigate two types of imbalance namely *step imbalance* and *linear imbalance*, which can represent most of the real-world cases.

The long-tailed distribution has a strong correlation with the imbalance problem. Specifically, the long-tailed distribution is an extreme case of imbalance. As shown in Tab. 1, generally speaking, it is considered that the imbalance learning is usually reflected in the situation where there are fewer learning classes such as the two-classification problem, while for the long-tailed visual recognition, the number of classes is larger. When the number of classes increases to a certain level, the dataset tends to favor the long-tailed distribution. More importantly, for the long-tailed visual recognition, the tail classes are likely to lack a comprehensive data distribution due to the sparse training examples, and thus the model decision boundaries are more ambiguous, combined with the fact that the number of tail classes occupies most of the dataset, so the training of the model is more challenging, making it difficult to solve the long-tailed problem.

2.2.2 Few-Shot Learning

In many application scenarios, it is very difficult to collect labeled data, so people want to be able to learn a well-performing model with only a small amount of data. In addition, humans have the ability to learn quickly from a small number of samples, and machine learning was desired to give such a property, which gave birth to few-shot learning (FSL). Wang *et al*. [162] propose that FSL is a type of machine learning problem, specified by experience E, task T, and performance measure P, where E contains little supervised information for the target T. For the C-way K-shot problem in FSL, it simply means that we need to learn C classes with only K training images (typically no more than 20) in each class, *i.e*., we are required to learn how to distinguish these C classes in these $C \times K$ images.

The tail classes of long-tailed dataset have little supervisory information, which is similar to FSL. But the difference is that the base set of FSL is much more balanced, and head classes of long-tailed datasets are rich in supervised information. Although, generally speaking, the more data is available, the more beneficial it is for deep model learning, but this will inevitably inhibit or over-whelm the tail classes. Therefore, how to balance the relationship between the head classes and the tail classes is also an important point that needs additional consideration in long-tailed learning.

2.2.3 Differences and similarities

Long-tailed visual recognition has a strong relationship with imbalance learning and few-shot learning. The head and body classes of the long-tailed dataset can be regarded as

Table 1: **Similarities and differences between Long-tailed Recognition, Imbalance Learning, and Few-shot Learning in terms of their data and tasks**. Part of the table is extracted from [108]. We compared these three areas in terms of the training set (base set), the imbalance of the test set, the sample number of tail classes, the comparison of the class numbers, and evaluation range, respectively.

Task Setting	Imbalanced Train / Base Set	Balanced Test Set	Samples in Tail Classes	Number of Classes	Evaluation: Accuracy Over?
Imbalanced Learning	✓	x	$20 \sim 50$	less	all classes
Few-Shot Learning	x	✓	$1 \sim 20$	−	novel classes
Long-Tailed Recognition	✓	✓	$1 \sim 20$	much	all classes

the traditional imbalance problem [108]. Besides, long-tailed data has the characteristics of "long" tails, and each tail class has very little data. Therefore, long-tailed data has a few-shot problem that cannot be ignored. In general, the long-tailed phenomenon is an extreme case of the imbalance problem, and it is also a combination of data imbalance and few-shot learning. Tab. 1 summarizes their differences according to [108].

3 Long-tailed Datasets and Metrics

Over time, several researchers have proposed some mainstream long-tailed datasets which facilitate the development of long-tailed studies. In this section, we focus our analysis around the long-tailed datasets, starting with introducing some generic long-tailed datasets in Sec. 3.1, and then we constructively analyze four quantitative metrics to measure the long-tailedness of datasets in Sec. 3.2. Finally, we list the performance evaluation metrics of some long-tailed benchmarks in Sec. 3.3.

3.1 Long-tailed Benchmark

To better study the long-tailed problem, several long-tailed datasets have been proposed over the past decades. We summarize the commonly used long-tailed datasets in Tab. 2, depicts their category-instances distribution curves in Fig. 4, and give detailed review below.

• **CIFAR-10/100-LT [27].** CIFAR-10-LT and CIFAR-100-LT are the long-tailed versions of the CIFAR-10 and CIFAR-100 [84]. Both CIFAR-10 and CIFAR-100 contain 60,000 images, 50,000 for training and 10,000 for validation with class number of 10 and 100, respectively.

• **ImageNet-LT [108].** ImageNet-LT is a long-tailed version of ImageNet-1K [32], created by Liu *et al.* [108], including 115.8K images from 1,000 classes, with maximally 1,280 images and minimally 5 images per class.

• **Places-LT [108].** Places-LT is a long-tailed version of Places365 [200], which contains 184.5K images from 365 classes, with maximum of 4,980 images and minimum of 5 images per class.

• **iNaturalist 2017 & 2018 [148].** iNaturalist (iNat) is a real-world fine-grained species classification and detection dataset, covering several domains such as birds, dogs, airplanes, flowers, leaves, food, trees, cars, *etc*. iNat 2017 [148] contains 579,184 training images of 5,089 classes, and its 2018 version [1] has 437,513 training samples in 8,142 classes.

• **LVIS v0.5 & v1.0.** LVIS is proposed by Gupta *et al.* [53], which is a large-scale fine-grained vocabulary instance segmentation dataset that is based on the COCO dataset and is annotated with instances for over 1,000 classes of objects.

• **MS1M-LT [108].** MS1M-LT is a face recognition dataset, a long-tailed version of MS1M-ArcFace dataset [33, 52]. In MS1M-LT, each identity is sampled with a probability proportional to its number of images, which lead MS1M-LT to a long-tailed distribution with 887,530 images and 74,532 identities.

3.2 Long-tailedness Metrics

Accurate and objective measurement of the long-tailedness of data is an important prerequisite to solve the long-tailed visual recognition problem. Therefore, in this section, we compare four commonly used quantitative metrics in statistics, and critically analyze their advantages and disadvantages in measuring the long-tailedness.

3.2.1 Four Quantitative Metrics in Statistics

• **Imbalance Factor.** In [27], Cui *et al.* defined the imbalance factor (denoted as β) of a dataset as the number of training samples in the largest class divided by the smallest:

$$\beta = max\left\{n_1, n_2, ..., n_k\right\} / min\left\{n_1, n_2, ..., n_k\right\} \qquad (1)$$

where $n_1, n_2, ..., n_k$ represents the number of samples in different classes. Although the imbalance factor is widely-used as a measurement of the long-tailedness [27, 108, 148], it is easily affected by extreme classes and can not reflect the overall characteristics of the dataset.

• **Standard Deviation.** Standard deviation (denoted as σ) is frequently used in probability statistics as a measurement of

Table 2: **Statistics of representative long-tailed visual recognition datasets**. See Sec. 3.1 for more detailed descriptions.

Dataset	Venue	Fields	Annotation Types	Training Samples	Classes	Max Size	Min Size	Imba. Factor β
CIFAR-10-LT [27]	CVPR 2019	Object-centric	Classification	50,000 - 11,203	10	5,000	500 - 25	10 - 200
CIFAR-100-LT [27]	CVPR 2019	Object-centric	Classification	50,000 - 9,502	100	500	500 - 2	1 - 250
ImageNet-LT [108]	CVPR 2019	Object-centric	Classification	115,846	1,000	1,280	5	256
Places-LT [108]	CVPR 2019	Scene-centric	Classification	62,500	365	4,980	5	996
iNaturalist 2017 [148]	CVPR 2018	Species-centric	Classification Bounding-box	579,184	5,089	3,919	9	435
iNaturalist 2018 [1]	-	Species-centric	Classification Bounding-box	437,513	8,142	1,000	2	500
MS1M-LT [108]	CVPR 2019	Face-centric	Classification	887,530	74,532	598	1	598
LVIS v0.5 [53]	CVPR 2019	Object-centric	Bounding-box Instance-mask	56,740	1,230	26,148	1	26,148
LVIS v1.0	-	Object-centric	Bounding-box Instance-mask	99,388	1,203	50,552	1	50,552

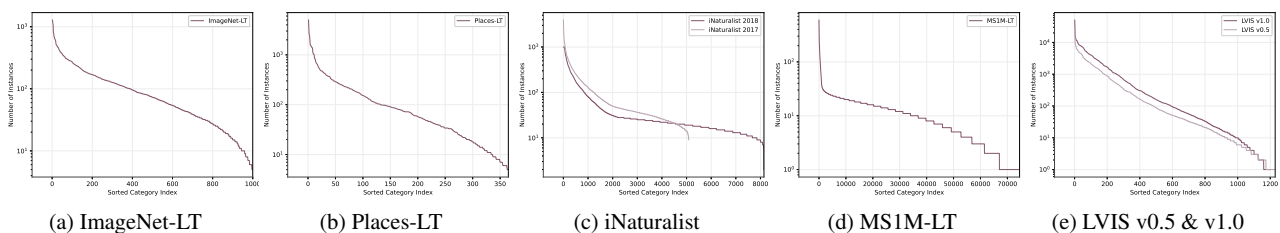| (a) ImageNet-LT | (b) Places-LT | (c) iNaturalist | (d) MS1M-LT | (e) LVIS v0.5 & v1.0 |

Fig. 4: **Distributions of common long-tailed datasets**. Figure (a)-(e) show the long-tailed distributions of ImageNet-LT, Places-LT, iNaturalist 2017 & 2018, MS1M-LT and LVIS v0.5 & v1.0, respectively.

Table 3: **Different metrics are used to quantify the long-tailedness**. The upper part of the table lists four common balanced visual datasets, and the lower part is the long-tailed datasets. The Gini coefficient is recommended in this review.

Dataset	Imba. Factor β	Std. σ	$\frac{mean}{median} \gamma$	**Gini Coef.** δ
CIFAR [84]	1.0	0.0	1.0	0.0
ImageNet-1K [32]	1.77	70	0.98	0.013
Places365 [200]	1.62	259	0.98	0.011
COCO [99]	1,325	29,360	1.76	0.564
ImageNet-LT	256	139	1.58	0.524
Places-LT	996	382	2.37	0.671
iNaturalist 2017	435	241	2.77	0.634
iNaturalist 2018	500	117	2.44	0.620
MS1M-LT	598	18	1.32	0.473
LVIS v0.5	26,148	1,516	11.7	0.825
LVIS v1.0	50,552	2,789	11.1	0.820

statistical dispersion [29, 31], and can also reflect the uncertainty of sampling in some cases [15, 30], it can be expressed as:

$$\sigma = \sqrt{\frac{1}{k} \sum_{i=1}^{k} (n_i - \mu)^2} \quad (2)$$

where k represents the number of classes; n_i represents the instance number of class i, and μ represents the average number of instances. Although standard deviation quantifies the dispersion degree between the number of classes within a dataset, it is also affected by the absolute number of samples,

so it is difficult to objectively express the long-tailedness of data. In the third column of the Tab. 3, we counted the standard deviations of some balanced datasets as well as long-tailed datasets, and it can be found that the balanced dataset, COCO (σ=29,360), has the largest standard deviation, and the long-tailed dataset, MS1M-LT (σ=18) has the smallest one, which shows that the standard deviation can not well identify the long-tailedness.

● **Mean / Median.** Median is a proper term in statistics and is widely used in economics [44], sociology [43] and medicine [36]. Compared with the mean, the median is not affected by the maximum or minimum of data, and can better represent the distribution of data to a certain extent. Therefore, the ratio of mean to median (denoted as γ) can also reflect the skew distribution of data, which can be expressed by:

$$\gamma = \frac{mean(n_1, n_2, ...n_k)}{median(n_1, n_2, ...n_k)} \quad (3)$$

When γ is closed to 1, it indicates that the dataset is uniformly distributed, and when γ is significantly greater than 1, it indicates that the dataset is of imbalance, including a large gap between the instance number in head and tail classes. As shown in Tab. 3, although γ accurately distinguishes between balance datasets and long-tailed datasets. However, like imbalance factor, it is easily affected by individual cases and cannot reflect the overall distribution. And the value range
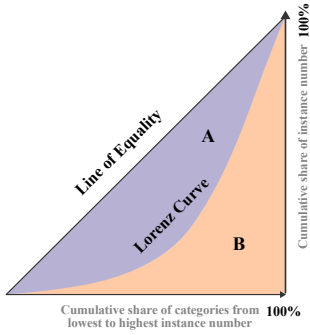
Fig. 5: **Calculation of the Gini coefficient**. The horizontal axis is the cumulative distribution of the class proportion and the vertical axis is the cumulative distribution of the instance number proportion.

of γ is an open interval (no upper limit), which is not a good characteristic for a measure.

● **Gini Coefficient.** Gini coefficient (denoted as δ) was originally proposed by the Italian economist Gini in 1912 [46] as an indicator to judge the degree of distribution equality based on Lorentz curve. It is always used to represent income inequality or wealth inequality [76, 178]. Since long-tailedness is similar to inequality between each category, Gini coefficient can serve as a long-tailed metric. As shown in Tab. 3, Gini coefficient can effectively distinguish balanced datasets and long-tailed datasets.

The calculating process of Gini coefficient consists of three steps. First, we suppose that the set of number samples of k classes dataset $n_i, (i = 1, 2, ..., k)$ is in ascending order, we calculate the normalized cumulative distribution $\{C_i\}$ by:

$$C_i = \frac{1}{k} \sum_{j=1}^{i} n_j \qquad (4)$$

Intuitively, C_i indicates the probability of the is smallest categories. By defining $C_0 = 0$, we can normalize the x-axis to the share of total categories and interpolate linearly to obtain continuous Lorentz curve $L(x), x \in [0, 1]$ (as shown in Fig. 5). The Lorentz curve $L(x)$ follows:

$$L(x) = \begin{cases} C_i, & x = \frac{i}{k} \\ C_i + (C_{i+1} - C_i)(kx - i), & \frac{i}{k} < x < \frac{i+1}{k} \end{cases} \qquad (5)$$

where $i = 1, 2, \ldots, k$. B represents the area to the lower right of the actual instance number distribution curve. Since Lorentz curve is linearly interpolated by $\{C_i\}$, we can calcu-

late the area by trapezoids:

$$B = \int_0^1 L(x)dx = \sum_{i=1}^{k} \frac{C_i + C_{i-1}}{2} \cdot \frac{1}{k} \qquad (6)$$

For balanced dataset, Lorentz curve is an identity line and A presents the area between the identity line and the Lorentz curve of an actual dataset. Thanks to the normalization of Lorentz curve, A can be simply calculated by: $A = 0.5 - B$. Finally, the Gini coefficient can be expressed as:

$$\delta = \frac{A}{A + B} \qquad (7)$$

The Gini coefficient conforms to the closed interval distribution of (0,1), so it can better quantify the degree of imbalance and make the datasets more comparable with each other. Usually, the smaller the Gini coefficient δ of a dataset is, the more imbalanced the dataset is, and vice versa.

3.2.2 Long-tailedness Analysis

Since the Gini coefficient δ is not affected by extreme samples, is not affected by the absolute number of data, and has a bounded distribution, we recommend using the Gini coefficient to measure the long-tailedness of data. Based on Gini coefficient, we quantify the long-tailedness of some commonly used balanced datasets [32, 99] and long-tailed datasets [27, 53, 108, 148], as shown in Tab. 3.

As the most widely used visual dataset, CIFAR [84] has a perfect manual balance, and the number of instance in each class is equal, so its Gini coefficient is 0. Another widely used visual dataset, ImageNet-1K [32] contains 1,000 classes, and the instances in each category are also manually balanced. The Gini coefficient of ImageNet-1K is 0.013. In contrast, the Gini coefficients of long-tailed datasets are generally above 0.5, and some can even reach 0.8. It can be seen that using Gini coefficient to measure the long-tailedness of data is reasonable and effective, and there are great differences in the long-tailedness of existing long-tailed datasets.

On the other hand, COCO [99] is the most common dataset to evaluate the performance of object detection and instance segmentation methods, and it is considered to be balanced. The Gini coefficient of COCO is 0.564, which is much larger than the balanced datasets CIFAR, ImageNet-1K and Places365 [200], and even larger than the long-tailed datasets ImageNet-LT and MS1M-LT. However, the annotation of these datasets is image-level, and it is much simpler to manually control the data distribution than the instance-level annotation datasets. As long-tailed object detection / instance segmentation datasets, LVIS v0.5 and LVIS v1.0, their Gini coefficients are 0.825 and 0.820 respectively, which are still much larger than COCO. Therefore, for different visual tasks, different standards should be used to measure the long-tailedness of data distribution.

With the above analysis, we can quantitatively analyze the long-tailedness of most visual datasets. This provides guidance for long-tailed visual recognition, that is, different solutions are adopted through the long-tailedness of data. In Sec. 6, we will also use this standard to study 20 mainstream large-scale long-tailed visual recognition datasets, so as to deeply reveal the research status and future direction of this field.

3.3 Performance Evaluation Metrics

Presently, there are several general metrics that are widely-used to measure how long-tailed methods perform on classification and detection tasks. In this section, we provide a review of the major performance evaluation metrics for long-tailed recognition.

In terms of evaluation metric for classification, the top-1 accuracy is frequently adopted in the research community. For ImageNet-LT, Places-LT, and iNaturalist 2018 datasets, the accuracy can be split into four types based on the set of classes followed Liu *et al*. [108]: *Many-shot* (classes each with over training 100 images), *Medium-shot* (classes each with 20∼100 training images), *Few-shot* (classes under 20 training images) and the *Overall* accuracy. In CIFAR-10/100-LT, datasets with different long-tailedness is be sampled according to the imbalance factor $\beta \in \{200, 100, 50, 20, 10, 1\}$, and their evaluation metric is to measure the top-1 accuracy of the datasets under different imbalance factors respectively. To make a full comparison between different methods, we report benchmarks on CIFAR-10/100-LT (Tab. 5), ImageNet-LT and Places-LT (Tab. 6), as well as iNaturalist (Tab. 7), in Sec. 5, respectively.

For object detection and instance segmentation tasks, there are several evaluation metrics that are used in LVIS v0.5 and v1.0, such as AP_r(mask AP for rare classes), AP_c(mask AP for common classes) and AP_f(mask AP for frequent classes). To make a full comparison, we keep the common evaluation metrics in detection and instance segmentation tasks, like mask AP for $\{AP, AP_{50}, AP_{75}\}$, $\{AP_r, AP_c, AP_f\}$ and bounding-box AP, on LVIS v0.5 and v1.0 benchmark in Sec. 5 (see Tab. 8 and Tab. 9).

4 Long-tailed Visual Recognition

In the past few years, a growing number of research has investigated the long-tailed distribution of data as shown in Fig. 6. In this section, we review deep learning based methods for long-tailed visual recognition from 2016 to present and introduce the related earlier work in context. Although many studies have mixed a variety of methods to solve the long-tailed problem, in order to highlight the contribution of each study, we mainly reviewed their core methods, finely summarized them into ten categories from the perspective of representation learning, and outlined the highlights and limitations of each category (Tab. 4).

In order to make it easier for readers to understand the characteristics and differences of various methods, we use color scatter diagram to express the principle of each category (Fig. 7 - Fig. 9, Fig. 11 - Fig. 17). Among them, dots of different colors represent different classes, gray dots represent unlabeled data, and the number of dots represents the instance number in that category. The circle outside the dot indicates that this data is sampled multiple times.

4.1 Data Processing

For dataset's long-tailed distribution, an intuitive idea is to make the model learn relatively balanced classes from the perspective of data. There are three ways to handle the data, namely over-sampling, under-sampling and data augmentation.

4.1.1 Over-sampling

Over-sampling is one of the most common methods in deep learning [55,70,88]. As shown in Fig. 7 (a), the over-sampling method emphasizes the tail classes and increases the instance number of the tail classes [11, 12, 134] to reduce the imbalance between the head classes and the tail classes.

Shen *et al*. [134] propose a sampling strategy *Class-Aware Sampling (CAS)* to ensure that each class has the same probability of occurrence in each batch as much as possible. We denote the *CAS* probability of the $i - th$ class as $P_a(i)$, *i.e.*, for a total of C classes, following the definition of [121], the sampling probability for each class is:

$$P_a(i) = \frac{1}{C} \tag{8}$$

Dhruv *et al*. [110] compute a replication factor for each image based on the distribution of labels and repeated the images several times based on the replication factor. Inspired by this, the work of Gupta *et al*. [53] propose *Repeat Factor Sampling (RFS)* to perform rebalancing operations on training data by increasing the sampling frequency of images containing tail instances. *Soft-balance Sampling with Hybrid Training* [121] combined the conventional sampling scheme and *CAS*, which first trains the detector using the conventional strategy, and then introducing hyper-parameters to control the degree of ordinary sampling with $P_o(i) = \frac{n_i}{N}$, where n_i represents the instance number of class i and N is the total number of instances.

Legend: Long-tailed classification · Long-tailed detection / instance segmentation · Both

2016

CAS (ECCV2016) Peng et al.

FiFDM (CVPR 2016) Ouyang et al.

2017

CRL (CVPR 2017) Dong et al.

LMT (NeurIPS 2017) Wang et al.

Range Loss (ICCV 2017) Zhang et al.

2018

L2RW (ICML 2018) Ren et al.

LSFC (CVPR2018) Cui et al.

2019

UDFR (CVPR 2019) Zhong et al.

OLTR (CVPR2019) Liu et al.

CB Loss (CVPR2019) Cui et al. 2019

MW-Net (NIPS 2019) Shu et al.

DCL (ICCV 2019) Wang et al.

LDAM (NIPS 2019) Cao et al.

FTLFR (CVPR 2019) Yin et al.

2020

LST (CVPR 2020) Hu et al.

BAGS (CVPR 2020) Li et al.

Forest R-CNN (ACM MM 2020) Wu et al.

EQL (CVPR 2020) Tan et al.

LSOD (CVPR 2020) Peng et al.

CDB loss (ACCV 2020) Sinha et al.

Domain Balancing (CVPR 2020) Cao et al.

RCBM (CVPR 2020) Jamal et al.

Decoupling (ICLR 2020) Kang et al.

BBN (CVPR 2020) Zhou et al.

ALEAP (CVPR 2020) Liu et al.

FSA (ECCV2020) Chu et al.

Deep-RTC (ECCV 2020) Wu et al.

LADE arxiv 2020 Hong et al.

SimCal (ECCV 2020) Wang et al.

De-confound-TDE (NeurIPS 2020) Tang et al.

MFM arxiv 2020 Wang et al.

LFME (ECCV 2020) Xiang et al.

BALMS (NeurIPS 2020) Ren et al.

DB loss (ECCV 2020) Wu et al.

SSP (NeurIPS 2020) Yang et al.

Remix (ECCV 2020) Chou et al.

2021

EQL v2 (CVPR 2021) Tan et al.

GistNet (arxiv 2021) Liu et al.

PML (CVPR 2021) Deng et al.

Seesaw loss (CVPR 2021) Wang et al.

DRO-LT (arxiv 2021) Samuel et al.

FASA (arxiv 2021) Zang et al.

Fed Loss (arxiv 2021) Zhou et al.

ALA Loss (arxiv 2021) Zhao et al.

CReST (CVPR 2021) Wei et al.

Drop Loss (arxiv 2021) Hsieh et al.

Breadcrumbs (arxiv 2021) Liu et al.

Logit Adjustment (ICLR 2021) Menon et al.

MiSLAS (CVPR 2021) Zhong et al.

ResLT (arxiv 2021) Cui et al.

DisAlign (CVPR 2021) Zhang et al.

Hybrid-SC/PSC (CVPR 2021) Wang et al.

Bag of Tricks (AAAI 2021) Zhangi et al.

RIDE (arxiv 2021) Wang et al.

RoBal (CVPR 2021) Wu et al.

MOSAICOS (arxiv 2021) Zhang, Pan et al.

MetaSAug (CVPR 2021) Li et al.

ACSL (CVPR 2021) Wang et al.

DiVE (arXiv 2021) He et al.
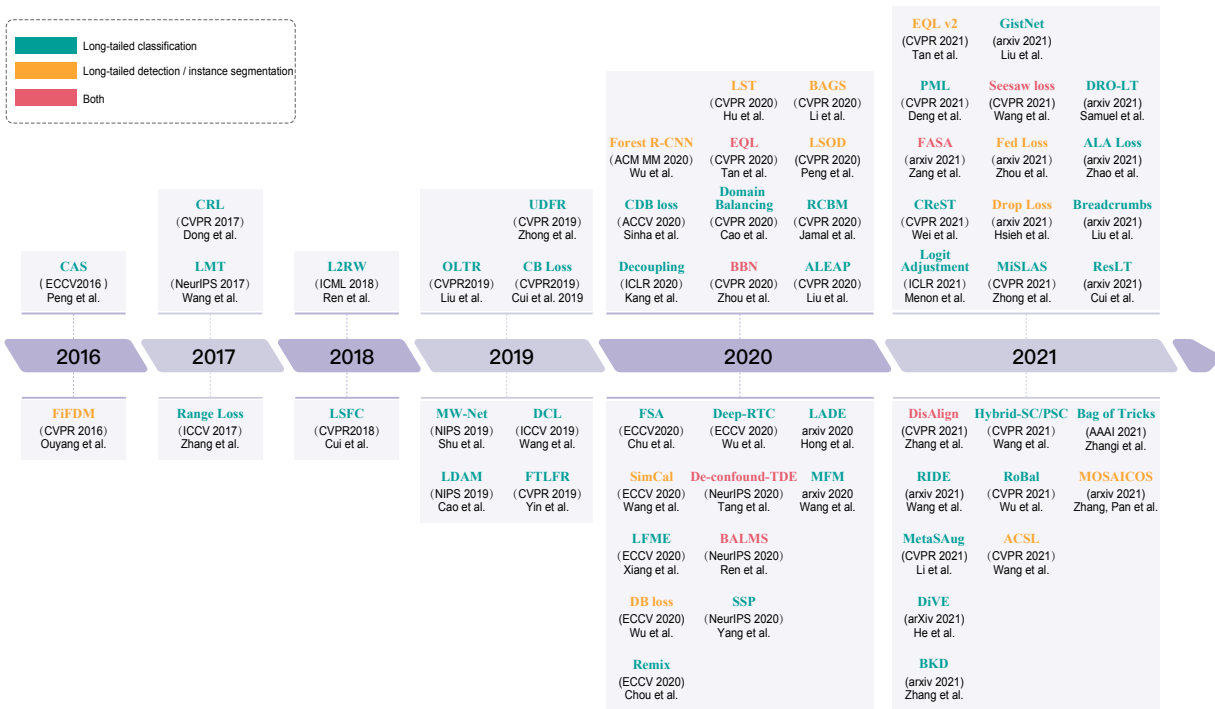
BKD (arxiv 2021) Zhang et al.

Fig. 6: **A chronological overview of recent representative work in long-tailed recognition**. Visual recognition study on long-tailed distribution started with [120] in 2016 and has become more and more abundant since then. Work marked in green represents long-tailed classification, marked in orange represents long-tailed object detection / instance segmentation, marked in red represents both. We abbreviate some studies for ease of presentation.
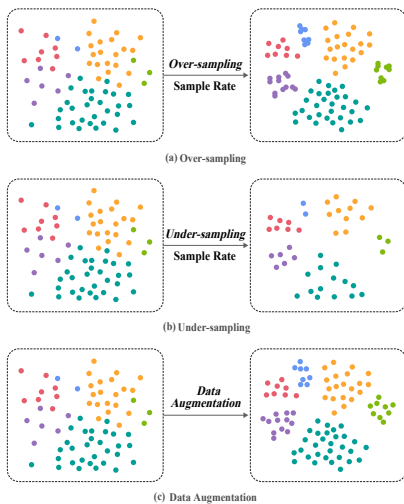
Fig. 7: **Data processing methods for long-tailed problem**. Data points with the same color represent the same class. (a) and (b) represent data over-sampling and under-sampling, respectively. *i.e.*, assigning different sampling rates for head/tail classes. (c) represents data synthesis, *i.e.*, synthesizing new data for tail classes to increase their weighting.

In addition to the above-mentioned works that perform re-sampling at the image level, some works perform instance-level re-sampling to prevent unnecessary duplication of certain non-tail class instances in the oversampled images. Hu *et al.* [65] presente *Instance-level Data Balanced Replay* strategy. At stage t, for each class, a certain number of images containing that class are randomly sampled, and among these images, only annotations belonging to that class are considered valid during the training process. *NMS Resampling* [169] adaptively adjusts the *non-maximum suppression (NMS)* threshold for different classes according to their label frequencies, so as to balance the data distribution by retaining more proposal candidates from the tail classes while suppressing those from the head classes.

4.1.2 Under-sampling

In contrast to the over-sampling method, the under-sampling method reduces the imbalance between the head classes and the tail classes by reducing the sample times of the head classes [11, 55, 58, 71], as shown in Fig. 7 (b).

Random under-sampling is performed by randomly deleting the head classes data until it has the same number of instances as the other classes [11]. *EasyEnsemble* [106] divides the frequent classes into several subsets, each with

the same number of instances as the rare classes, and combines them separately with the rare classes to train multiple classifiers and eventually combine the outputs of multiple classifiers in an ensemble fashion, thus dealing with the information loss problem in traditional random under-sampling. *BalanceCascade* trains the classifiers sequentially, where in each step, the majority class examples that are correctly classified by the current trained learners are removed from further consideration. *NearMiss* [111] is another method to alleviate the information loss problem in random under-sampling, which is essentially a prototype selection method that uses *KNN* to select the most representative samples from frequent class samples for training.

There are also some data cleaning methods, which mainly clean overlapping data to achieve the purpose of under-sampling. In *Edited Nearest Neighbours (ENN)* [168], for a sample belonging to a frequent class, if more than half of its K nearest neighbors do not belong to the frequent class, this sample is eliminated. Modified *Tomek Link* method [34] performs data cleaning by finding pairs of samples whose nearest neighbors are each other and belong to different classes, and removing the one that belongs to the frequent class.

Over / under-sampling is one of the most common and easily considered operations. However, there are some drawbacks involved. For example, in the case of over-sampling the tail classes, it may lead to over-fitting [17, 154] the tail classes [25, 27, 139, 144] and if there are errors or noise in the samples of the tail classes, then over-sampling may aggravate these problems. Under-sampling may lead to under-learning of the head classes [27, 139], and may potentially missing valuable data in the head classes. For extremely long-tailed data, the under-sampling method usually loses a lot of information because of the large difference in the amount of data between the head class and the tail class [144].

4.1.3 Data Augmentation

Data augmentation is another way of data processing to solve the problem of long-tailed distribution, as shown in Fig. 7 (c). Due to the small sample size of the tail classes, it is difficult to learn the complete features. Therefore the tail classes can be compensated by data augmentation methods such as generating and synthesizing new samples with the help of similar samples [17] or other data sources [45, 57]. There are some common ways of data augmentation such as random image flipping, scaling, rotating and cropping and so on. But these naive methods are not good enough for the tail classes where samples and features are extremely sparse. In order to reduce the over-fitting risk of the tail classes and to improve the generalization ability, some methods expand the tail classes by data synthesis, which can be divided into two approaches: image space and feature space.

For image space, some data augmentation methods help to improve the performance for tail classes. Zhang *et al.* [184] proposed that although *Empirical Risk Minimization (ERM)* allows large-scale neural networks to memorize (rather than generalize) training data, validation on samples outside the training distribution (adversarial samples) can dramatically change the prediction results, *i.e.*, when the distribution of the test set are different from the training set, the *ERM* method no longer has good interpretation and generalization performance. In contrast, data augmentation methods can improve the generalization of the model to the training data [138]. *Mixup* is a data-independent data augmentation approach that performs data augmentation by constructing a virtual training sample, expressed by the formula:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j,$$
$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j. \tag{9}$$

where (x_i, y_i) and (x_j, y_j) are two samples randomly selected from the training data and $\lambda \in [0, 1]$. Recently, Chou *et al.* [21] improved the mixup method and proposed a new data augmentation method *Remix*. It assigns the label in favor of the minority class by providing a distributed higher weight to the minority class, which makes the classifier push the decision boundary to the majority class and balance the generalization error between the majority class and the minority class. The formulation of *Remix* is:

$$\tilde{x}^{RM} = \lambda_x x_i + (1 - \lambda_x)x_j,$$
$$\tilde{y}^{RM} = \lambda_y y_i + (1 - \lambda_y)y_j. \tag{10}$$

where λ_x is sampled from the beta distribution and λ_y takes the form of:

$$\lambda_y = \begin{cases} 0, & n_i/n_j \geq \kappa \ and \ \lambda < \tau; \\ 1, & n_i/n_j \leq 1/\kappa \ and \ 1 - \lambda < \tau; \\ \lambda, & otherwise \end{cases} \tag{11}$$

where the hyper-parameters κ and τ are used to set synthetic labels by comparing the number relationship between samples and to control the degree of synthetic labels, respectively.

Some studies are based on the feature level for feature synthesis, which can enrich the features of the tail classes and create clearer decision boundaries. At the beginning of this century, some classic data synthesis work provided ideas for some current advanced research.*SMOTE* [17] is an effective method for data synthesis. For each minority class sample, a sample is randomly selected from its nearest neighbors, and then a point on the line between these two samples is randomly selected as the newly synthesized minority class sample. However, *SMOTE* method has some drawbacks, it has some blindness in the selection of nearest neighbors, and it is also prone to the problem of distribution marginalization due to the imbalance of the data. Therefore there are many works to improve it. *Borderline-SMOTE* [56] judges

the boundary samples for the minority class samples and generates new samples for the boundary samples, and the rule for judging the boundary samples is that more than half of the K-nearest neighbors of the sample are majority class samples. Another classical approach is *ADASYN* [57], which can adaptively decide how many synthetic samples to generate for each minority class based on the distribution of the samples. First the degree of imbalance as well as the number of new synthetic samples to be generated are calculated, then the distribution of each minority class sample is calculated and the distribution is used to determine the number of synthetic samples for each class.

In recent years, some studies use data synthesis and feature synthesis to solve the long-tailed problem. Oversampling the tail classes is a very common strategy. However, this can easily lead to tail classes over-fitting. Towards this question, Kim *et al*. [80] utilize an optimization phase so that the head class samples are modified into tail class samples and then added to the original dataset for the purpose of balancing the dataset. Chu *et al*. [22] decompose the class activation map into class-generic features and class-specific features. In order to make up for the missing information of the tail classes, the specific features of the tail classes are fused with the common features of the head classes, which can expand the feature space and generate augmented samples to recover the base distribution of the tail classes. Finally, the online generated augmented samples are used to fine-tune the network trained in the first stage to improve the performance of the tail classes. *FASA* [180] generates virtual features by obtaining the mean and standard deviation of the corresponding class features. And the number of generated virtual features is dynamically decided by an adaptive feature sampling scheme, thus effectively avoiding over-fitting and under-fitting triggered by feature augmentation. *Breadcrumbs* [101] proposes a new feature augmentation strategy that tracks features backwards to access the large number of feature vectors available for each training image from previous epochs, in a way that is more diverse than the features obtained by simply copying and pasting. To overcome the lack of discriminative information in existing re-sampling methods, Zhang *et al*. propose a novel data augmentation approach based on *Class Activation Maps (CAM)* [199], which is tailored for two-stage training and generates discriminative images by transferring foregrounds while keeping backgrounds unchanged.

Data augmentation aims at applying enhancement techniques in the image space or feature space to synthesize new samples, thus expanding the data in knowledge-poor tail classes [180]. We need to acknowledge that the data produced by the data synthesis approach is more economical and efficient in many cases, especially when the data is difficult to obtain, and can also complement the real-world data. However, this artificial way of creating data does not come from

real scenarios, and its impact on the model lacks theoretical guidance, and the synthesis of data close to real scenarios is a highly complex operation, such as VAE [81, 119] and GAN *etc*. [9, 48, 78]. At the same time, the simple data synthesis method cannot accurately avoid the adverse effects of noise and other undesirable factors from the original dataset.

4.2 Cost Sensitive Weighting

Cost sensitive learning can be traced back to a classical approach in statistics that considers the cost of misclassified samples, some studies refer to this as importance sampling [27, 75]. It is shown that there is a strong link between cost sensitive learning and imbalance learning [58], so this approach can be naturally used to deal with extremely unbalanced data like long-tailed data. As shown in Fig. 8(a), cost sensitive weighting assigns different weights to different classes in an explicit or implicit way, so that the influence of the tail samples can be improved. Cost sensitive weighting can also assign weights at the sample level for more fine-grained control, as shown in Fig. 8(b). This strategy can be applied in many directions besides long-tailed learning, such as the classification of foreground and background for detection task.
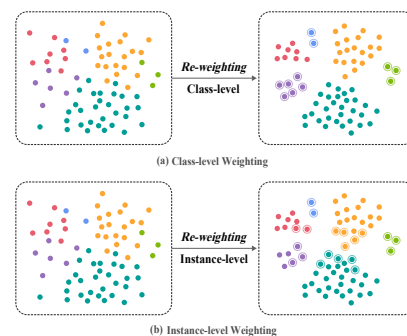

(a) Class-level Weighting

(b) Instance-level Weighting

Fig. 8: **Cost sensitive weighting methods for long-tailed problem**. We use a ring to represent increasing the learning weight of data points. (a) Represents class-level re-weighting, *i.e.*, assigning different learning weights to different classes so that the model enhances the learning of the tail classes. (b) represents instance-level re-weighting, *i.e.*, adjusting the sample weights by controlling at a more fine-grained level to make model focus more on learning hard samples.

4.2.1 Class-level Re-weighting

For cost-sensitive weighting, one of the most intuitive approaches is to re-weight classes proportionally by the inverse of their frequencies [66, 164]. But this naive method tends

to perform poorly, some works tend to use a "smoothed" version of weights that are empirically set to be inversely proportional to the square root of label frequency [110, 115].

However, it is often inefficient to set weights directly based on the instances number of the classes for class-level re-weighting methods, and it is difficult to find an appropriate weight that is valid. Therefore, some methods implicitly distinguish between head and tail classes, and turn their attention to other factors such as effective number of samples, contribution gradient, sample difficulty, *etc*.

Cui *et al*. [27] argue that there is information overlap among data, as the number of samples increases, the marginal benefit a model can extract from the data diminishes. Based on this view the concept of effective number of samples was proposed. Where *Effective Number E_n* is defined as

$$E_n = (1 - \beta^n)/(1 - \beta), where \ \beta = (V - 1)/V, \qquad (12)$$

n is sample number, V is the total volume of all possible data in the class. *Class-Balanced Loss* solves the training problem for unbalanced data by adding a class-balanced weighting term $\alpha_i \propto 1/E_{n_i}$ to the loss function for class i that is inversely proportional to the number of valid samples, where n_i is the number of samples for class i.

Tan *et al*. [144] argue that negative gradients from the head classes severely inhibit the learning of tail classes during training. For the tail classes, the gradients from negative samples are larger than those from positive samples. Therefore, *Equalization Loss (EQL)* sets a weight term w for class j:

$$w_j = 1 - E(r)T_\tau(f_j)(1 - y_j). \qquad (13)$$

For a region proposal r, $E(r)$ outputs 1 means r is a foreground region proposal and otherwise 0. f_j means frequency of class j, and $T_\tau(x)$ is a threshold function. y is the ground truth distribution with one-hot representation. It aims to ignore the gradients from frequent classes on rare classes while preserving the gradients from background samples, thus ensuring a fair training for each class. However, *DropLoss* [64] observes that most of the gradients that inhibit the tail class actually come from correct background classification rather than incorrect foreground prediction, and therefore *DropLoss* is proposed to adaptively rebalance the ratio of background prediction loss between the rare/common class and the frequent class. Tan *et al*. [143] propose *EQL v2* from the gradient perspective. It chooses the gradient statistic as an indicator to indicate whether a task is in balanced training. The ratio of accumulated positive gradients to negative gradients for each classifier is used to independently increase the weight of positive gradients and reduce the weight of negative gradients for each classifier. Li *et al*. [90] extend the idea of equalization loss to the single-stage object detector [98], independently rebalances the loss contribution of positive and negative samples of different categories according to their imbalance degrees, and effectively solve the long-tailed problem under the imbalance of positive and negative samples.

Wu *et al*. [170] propose that for the multi-label recognition problem under long-tailed distribution, the general re-sampling scheme leads to undesirable effects due to the presence of label co-occurrence. Therefore, *Distribution-Balanced Loss* is proposed to address the undesirable effects caused by label co-occurrence, while over-suppression of negative labels is overcome by regularization to mitigate the tail classes over-fitting problem. Peng *et al*. [121] address the case where multiple labels explicitly exist for an object. In order to avoid the traditional softmax function suppressing coexisting classes, concurrent softmax is proposed to avoid unnecessarily large losses due to the multi-label problem, and the gradient could focus on more valuable knowledge.

Sinha *et al*. [139] propose *Class-Wise Difficulty-Balanced Loss (CDB loss)* to assign loss weights by measuring the learning difficulty for each class. *Seesaw Loss* [153] sets a mitigation factor, and the penalty for the class with fewer instances is dynamically adjusted according to the ratio of the number of instances in the tail class to the instance number in the head classes. *Federated loss* [202] is proposed for solving the federal annotation of LVIS. It selects a subset of classes for each training image, including all positive annotations as well as a random negative subset. A binary Cross-Entropy loss is used for all classes in this subset during training, and classes outside are ignored. Wang *et al*. [158] propose to treat all object classes as tail classes regardless of the instance number of each class. In addition, *Adaptive Class Suppression Loss (ACSL)* is introduced to adaptively balance the negative gradients between different classes, which can effectively improve the discriminative power for the tail classifier. *ResLT* [25] is rebalanced from a parameter space perspective. The shared part of the model parameters is used to learn the classes' common features. The dedicated part retains the specific capacities of the head, middle, and tail classes through three branches, where the main branch learns to recognize images from all classes, and then augments images from the middle+tail and tail classes through two other residual branches to progressively augment the classification results on the tail classes, respectively. Finally, the branches are aggregated into a final result by additive shortcuts, which is an adaptive, incremental learning method.

4.2.2 Instance-level Re-weighting

Models have difficulty in learning on hard examples as well as their features, so some studies have identified these hard samples for targeted treatment. Although hard example mining are not specifically designed for the long-tailed problem, some studies [27, 37, 69] illustrate the effectiveness of instance-level re-weighting for long-tailed learning. For long-tailed data, the extreme imbalance of the data can lead to the

tail classes learning fewer iterations and gradually becoming a kind of hard-to-score sample, so the instance-level reweighting methods can be effective in improving the performance of the tails.

In *OHEM* [135], each example is scored by its loss, non-maximum suppression (NMS) is then applied, and a mini-batch is constructed with the highest-loss examples. Positive and negative samples are taken as 3:1 to calculate the loss, and the other negative sample weights are set to 0. Although the *OHEM* algorithm increases the weights of hard samples, it ignores the samples that are easy to classify. Lin *et al*. [98] propose *Focal Loss* to solve the problem of severe imbalance in the ratio of positive and negative samples in one-stage object detection. It improves on the Cross-Entropy loss by adding a modulating factor, which distinguishes between simple and hard samples. The weight of the simple samples is reduced, while paying more focus on the hard samples. As hard samples are mostly composed of tail classes for the long-tailed data, so *Focal Loss* can effectively improve the learning of tail classes.

The essential effect of sample imbalance from the perspective of gradient distribution is explored by *Gradient Harmonizing Mechanism (GHM)* [89], which points out that in one-stage detectors, the number of simple samples is very large, so they tend to dominate the model update, and since they are already well discriminated by themselves, the parameter update caused by this part does not improve the model much and the gradients generated by the samples are small. The class imbalance can thus be attributed to an imbalance in the degree of difficulty, and the imbalance in the degree of difficulty can be attributed to an imbalance in the distribution of the gradient parametrization. Specifically, by counting the gradients of the samples and designing *GHM* based on this distribution, the gradients generated by different samples are weighted so as to change the amount of their contributions and eliminate the negative effects of outliers. Zhao *et al*. designe an *adaptive logic adjustment (ALA) loss* [195], which contains an instance-specific adjustment term that adapts to the logic of each sample and can make the model more focused on hard samples.

The re-weighting approach is an important strategy to solve the long-tailed problem by giving different learning weights to different categories or samples. However, it is not feasible to set simple learning weights, such as the inverse of the category frequency or a smoothed version, based only on the size of the category data. Therefore finding a loss weight that fits the model and the data takes effort. Some work assigns weights at the instance level for more fine-grained control. However, some researchers [139] point out that this training strategy still results in a focus on learning the head classes because the absolute number of head data is dominant so that the number of hard samples in the head classes is still more than the number of hard samples in the tail classes.

For large-scale real-world data, re-weighting tends to make the deep model difficult to optimize during training [66, 67]. Moreover, re-weighting methods are susceptible to sensitive hyper-parameters, and the optimal settings may vary widely from dataset to dataset [143].
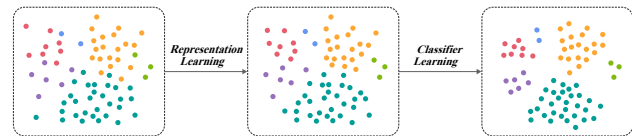
4.3 Decoupling Methods

Fig. 9: **Decoupling methods for long-tailed problem**. Because the rebalancing approach can severely damage the learned representations, some studies decouple representation learning and classifier learning. In the first stage, ordinary learning is used for representation learning. In the second stage, the network parameters of the representation learner are frozen and using a rebalancing approach to learn a good classifier.

Some studies found that although rebalancing strategies are important for long-tailed data, manipulating the data by re-sampling or re-weighting methods can harm the feature representation during the representation learning phase, while regular sampling tends to give more general representations. Therefore, uniform sampling is used to train the deep learning model to obtain the features of the data, and then class-balanced sampling is performed on the classifier to balance the head and tail classes, as shown in Fig. 9.

Kang *et al*. [77] propose that unbalanced data would not be a problem in learning high-quality feature representation, while strong long-tailed recognition could be achieved by tuning the classifier only. The learning process is decoupled into representation learning and classifier learning. The former is learned by standard instance-balanced sampling, class-balanced sampling, or a mixture of the two, the results shows that instance-balanced sampling yields better results, which shows that the rebalance methods can impair the learned representations. The latter learns the classifier by three methods, namely *Classifier Re-training* (*cRT*), *Nearest Class Mean classifier* (*NCM*), τ-*normalized classifier* (τ-*normalized*).

Inspired by [77], *BAGS* [96] also decouple the representation learner and classifier. The balanced group softmax module is introduced in the classification head of the detection framework. Grouping according to the instance number of the classes, and respectively executing softmax operation. The group-by-group training is used to separate the classes with disparate numbers of instances, thus balancing the clas-

sifiers in the detection framework and effectively reducing the control of the head classes over the tail classes.

BBN [198] designed a conventional learning branch as well as a re-balancing branch, where the former learns the generic pattern of the original distribution from the original long-tailed data, while the latter models the tail data in a back-sampling manner. Finally, the weights of the feature vectors of the two branches f_c and f_r are controlled by the adaptive trade-off parameter φ and input to the two classifiers W_c and W_r, respectively, and combined to obtain the final logits results z:

$$z = \varphi W_c^\top f_c + (1 - \varphi) W_r^\top f_r. \tag{14}$$

SimCal [157] corrects for biases in the classification head through a decoupled learning scheme. The model is first trained normally. A *bi-level sampling* scheme combining image-level and instance-level sampling is then used to collect class-balanced training instances. These samples are then used to calibrate the classification head to improve the tail classes performance. To mitigate the adverse effects of the above calibration on the head classes, *SimCal* also proposes a *Dual Head Inference* architecture that selects predictions for the tail and head classes directly from the new balanced classifier head and the original head.

DisAlign [187] believes that existing two-stage learning methods usually rely on heuristic design to adjust the initially learned classifier, which requires lengthy hyper-parameter tuning. At the same time, the bias of the decision boundary in the feature space can become a bottleneck. To this end, an adaptive method is designed to calibrate the output of the classifier, and a generalized weighting method is used to balance the class prior, so that the classifier output is matched to the reference distribution of the class that is beneficial to balance the prediction to calibrate the output of the classifier.

MiSLAS [197] proposes label-aware smoothing to deal with different degrees of class over-confidence in order to solve the mis-calibration problem of the two-stage method. Shifted batch in the decoupling framework is further proposed for the deviation of the dataset between the two stages due to different samplers normalization.

LADC [150] suppose that tail classes can be enriched by similar head classes and proposes a novel distribution calibration approach, which transfers the statistics from relevant head classes to infer the distribution of tail classes in the second stage.

Recently, decoupling representation learner and classifiers method has been shown to be effective in long-tailed data distributions, and it has become one of the mainstream research directions of long-tailed recognition. Moreover, decoupling method is relatively convenient to use in combination with data processing and cost sensitive weighting methods, which can obtain better model learning effect [96, 191]. But the two-stage learning strategy defies the expectation of end-to-end training sought in deep learning. At the same time, the resampling or re-weighting method adopted in the second stage still has the limitations mentioned above.

4.4 Other Long-tailed Visual Recognition Methods

In addition to the above methods, many studies use one or mixed multiple machine learning methods (metric learning, transfer learning, meta learning, knowledge distilling, mixture-of-experts, *etc.*) to solve the long-tailed problem, as shown in Fig. 10.
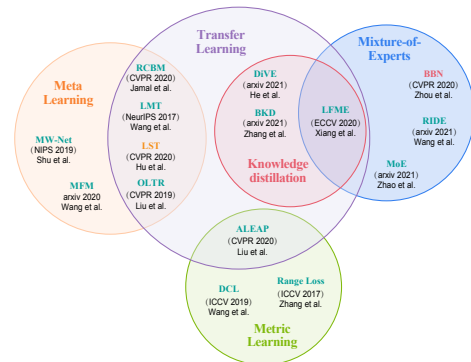

Fig. 10: **Domain relevance of some long-tailed methods**. To illustrate the domain relevance of the long-tailed approach, we list some of the mainstream studies, which draw on knowledge from multiple research domains.
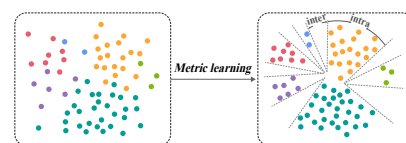
4.4.1 Metric Learning

Fig. 11: **Metric learning for long-tailed problem**. The method aims to clarify the decision boundary between classes while expanding the inter-class distance and reducing the intra-class distance.

Long-tailed data can be further clarified by performing feature similarity metrics in feature space with the help of metric learning to specify the boundaries between classes. Metric learning [117, 140], also known as *Distance Metric Learning (DML)*, which aims to learn an embedding function that can embed data to a feature space where the inter-data relationships are preserved [139]. Specifically, metric learning

is a spatial mapping method. By learning to an embedding space in which all data converted to feature vectors are measured for similarity and distances are reduced for similar samples and expanded for dis-similar samples, as shown in Fig. 11. This is helpful for classifiers to classify certain tail classes where the decision boundary is not clear enough.

Zhang *et al*. [188] propose *Range Loss*, the intra-class compactness constraint minimizes the two maximum intra-class distances for each class. The inter-class separation constraint calculates the centers of each class and makes the two classes with the smallest class center distances greater than a set margin. *Class Rectification Loss (CRL)* [38] argued that the traditional Cross-Entropy loss is not applicable to data with high imbalance and can cause model suffers from generalizing inductive decision boundaries biased towards majority classes. Therefore, *CRL* is proposed to gradually enhance the decision boundaries of minority classes. Specifically, the triplet ranking loss [104] is considered to model the relative relationship constraints within and between classes. In order to fully learn and exploit the minority classes, each minority class sample is considered as an "anchor" in the triplet structure to compute the batch loss balance regularization. *Dynamic Curriculum Learning (DCL)* [161] argues that this way of setting anchors tends to cause feature push-pull confusion.Therefore only simple samples of minority classes are set as anchors, instead of all samples of minority classes. Thus, the problem of sample feature push and pull confusion can be avoided to a certain extent when the decision boundary is blurred.

Cao *et al*. [14] believe that encouraging a large margin can be viewed as regularization, and propose to regularize the minority classes more strongly than the frequent classes. Therefore, minority classes are encouraged to obtain higher margins. Assume that γ_i is the margin of class i and n_i is the number of samples of class i. *LDAM* designs a *label-distribution-aware loss* function that finds the best trade-off between the margins of the classes:

$$\gamma_i \propto n_i^{-1/4} \tag{15}$$

and forces it to be a multi-class class-dependent margin.

Hybrid SC/PSC [155] is designed to include a *supervised contrastive learning (SCL)-based* feature learning branch and a Cross-Entropy loss based classifier learning branch. The hybrid network structure progressively adjusts the weight of the two branches in the learning process, and jointly performs feature learning and classifier learning. *Prototypical supervised contrastive loss* is designed to learn the prototype of each class for comparative learning and to force the different augmented views of each sample to be close to the prototype of their class and away from the prototype of the remaining classes.

Wang *et al*. [163] study the relationship between the margins and logits (classification scores) and empirically ob-

serve the biased margins and the biased logits are positively correlated. Based on this observation, they propose *Margin Calibration (MARC)*, a simple yet effective margin calibration function to dynamically calibrate the biased margins for unbiased logits.

Cui *et al*. [26] observe supervised contrastive loss tends to bias on high-frequency classes and thus increases the difficulty of imbalance learning. Therefore, *Parametric Contrastive Learning (PaCo)* is proposed to solve the problem of long-tailed recognition. *PaCo* introduces a set of parametric class-wise learnable centers to rebalance from an optimization perspective. When more samples are pulled together with their corresponding centers, *PaCo* can adaptively enhance the intensity of pushing the same class of samples closer, and is conducive to hard example learning. In addition, this study also found that RandAugment [24] and longer training epochs can further improve the effect of long-tailed learning, and achieved very competitive accuracy on multiple benchmarks combined with *PaCo*.

The basic purpose of metric learning is to make the sample features of similar classes closer together and those of different classes farther apart. Metric-based learning methods are usually based on loss functions to perform metrics between features, so it is necessary to consider the appropriate way for the combination of training samples as well as to choose the appropriate loss function. And for the head classes with large absolute numbers, it is still necessary to consider how to avoid the bias of the model for the head classes [155].
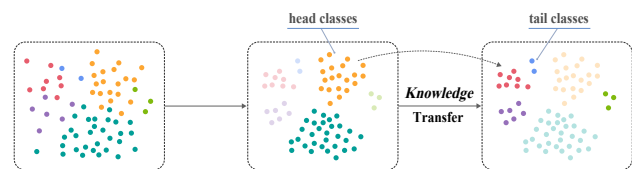
4.4.2 Transfer Learning

Fig. 12: **Transfer learning for long-tailed problem**. The method aims to make full use of the sufficient head classes and to transfer the knowledge acquired from the head classes to the tail classes.

Due to the large absolute amount of data in the head classes of the dataset, it has a richer and more complete training resource compared to the tail classes. Therefore some studies hope the head knowledge can be fully utilized to guide the learning of feature under-represented tail classes, as shown in Fig. 12.

Wang *et al*. [164] emphasis the strong correlation between meta-networks and model parameters. The basic assumption is proposed that model parameters share similar

dynamics across classes, laying the foundation for transferring model parameters between classes. Specifically, this research constructs a meta-network with a deep residual network as the basic unit to learn the head classes on the model parameter space and gradually transfer the meta-knowledge to the tail classes.

Hu *et al*. [65] divide the LVIS dataset of more than one thousand classes into sections according to the number of instances within the classes. The learning and merging of each part was performed by a divide-and-conquer strategy. A knowledge distillation strategy is used in order to retain the knowledge learned from the head classes. And a *Meta Weight Generator (MWG)* is designed to dynamically generate the weight matrix of the current stage using the fundamental knowledge learned and inherited from the head classes.

Due to the small spatial span of the tail classes and the large spatial span of the head classes, in order to compensate for the intra-class diversity of the tail classes, Liu *et al*. [102] construct to a feature cloud for each feature, transferring from the head classes to extend the distribution of the tail classes. The ideas of *CosFace* [151] and *ArcFace* [33] are borrowed to learn corner features. The overall variance of the head classes is obtained by computing the angle distributions and means between the head class features and their corresponding class centers. And the angular variance of the head classes is passed to the tail classes by constructing additional distributions, thus constructing feature clouds for each tail instance and expanding the space of the tail classes. *GistNet* [100] implements geometric structure transfer by implementing constellations of classification parameters, transferring the geometric structure of the head classes to the tail classes.

In order to make full use of the knowledge in the head classes and compensate for the lack of knowledge in the tail classes, some work uses transfer learning to transfer knowledge from the head classes to the tail classes. How to improve the performance of the tail classes without damaging the performance of the head classes is one of the issues to be considered. Meanwhile, some works usually carry out complex model and module design for knowledge transfer [164], which is not conducive to the training of the model and the convergence of the network.

4.4.3 Meta Learning

Meta Learning, also called Learning to Learn, is another important branch of research after Reinforcement Learning [74, 142]. It aims to learn a model that acquires general knowledge from different tasks and equips the model with the ability to learn in order to quickly adapt to new tasks. In long-tailed problem, meta-learning can guide the model to train, construct metamaps from head classes to tail classes, learn model parameters adaptively, assign sample weights, adjust classification network features, *etc*., as shown in Fig. 13. The
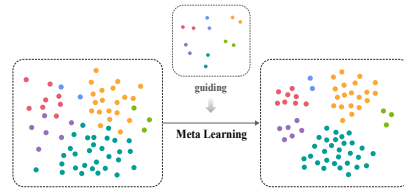

Fig. 13: **Meta Learning for long-tailed problem**. Deep models are trained adaptively by constructing a small amount of balanced meta-data, building meta-models, or setting adaptive parameters.

difficulty and uncertainty of human intervention to guide the model to learn are alleviated.

Ren *et al*. [127] propose that there is a tendency to select samples with small training losses in noisy images, and a tendency to select examples with large training losses in data imbalance problems. To address these two conflicting views, *L2RW* guides the updating of the weights of training losses by constructing a small unbiased validation set (*i.e*., meta-data). To reduce the computational cost, an online approximation method is used to adjust the weights of the training losses online in a mini batch in an approach similar to SGD. Shu *et al*. [136] argued that the possible way of learning weights implicitly in *L2RW* leads to unstable weighting behavior and non-generalizability during the training process, and therefore adopted the meta-network mechanism of learning weights explicitly, and proposed an adaptive sample weighting strategy for setting up sample loss. Specifically, a simple meta-network, *MW-Net*, is set up so that it adaptively learns the weights of the $i - th$ sample loss and uses meta-data to guide all parameters.

Jamal *et al*. [69] propose that the conditional distribution $P_s(x|y) = P_t(x|y)$ may not hold, which can result in target-shift. For this reason, this research proposes to relax the assumption that the source and target domains share the same conditional distribution $P_s(x|y)$ and $P_t(x|y)$ to enhance class-balanced learning. Specifically, this research relates the expected error in the target domain to the error in the source domain, and sets a balanced meta-data set to guide the meta-framework in estimating the conditional weights which is set for the target-shift part. Also, Ren *et al*. [126] find that in the long-tailed case, according to Bayes theorem, the regular softmax regression is affected by the label distribution shift, which will make the classifier more inclined to consider the samples as belonging to the head class. For this purpose, the work explicitly considers the label distribution shift, and re-derive the softmax function. The final *Balanced Softmax* $\hat{\phi}_j$ can be expressed as:

$$\hat{\phi}_j = \frac{n_j e^{\eta_j}}{\sum_{i=1}^{k} n_i e^{\eta_i}}, \tag{16}$$

where n_i denotes the number of class i, and η denotes model output.

Meta feature modulator (MFM) [156] is proposed to model the difference between long-tailed data and balanced meta-data from a representational learning perspective. Specifically, modulation parameters are introduced to channel-wisely scale or shift the intermediate features of the classification network. The modulation parameters and the classification network parameters are gradually optimized under the guidance of the balanced meta-data. Eventually, the model is made to have similar preferences for all classes.

MetaSAug [92] performs data augmentation with the help of *Implicit Semantic Data Augmentation (ISDA)* technique to obtain more semantically informative features. For tail classes, a reasonable covariance matrix cannot be obtained. Therefore, *MetaSAug* validates a small balanced validation set in each training iteration, minimizing the validation loss to update and learn the appropriate class-level covariance to achieve more meaningful augmentation results.

The meta-learning based approach allows the model to be more automated for adaptive learning. Some work guides the training of models for balance by employing meta-data [92], or employs meta-models to learn model parameters adaptively [136]. However, the guidance of meta-data is weak, and the model will inevitably be more inclined to learn head data. At the same time, the design of some meta-models or modules is complicated.
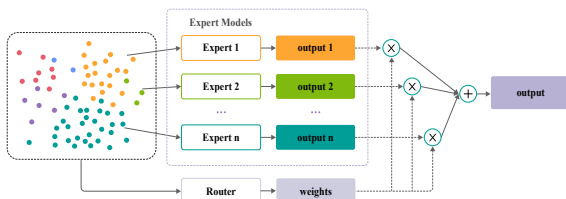
4.4.4 Mixture-of-Experts

Fig. 14: **Mixture-of-Experts for long-tailed problem**. Training multiple expert networks with a routing network. Each expert network in MoE has a data region in which it excels and on which it performs better than the other experts.

Some studies use Mixture-of-Experts (MoE) [68, 73] model to train multiple neural networks (*i.e.*, multiple experts), each of which is specialized to be applied to a different part of the dataset. Based on the divide-and-conquer principle, and training multiple expert networks with a routing network [112]. Each neural network in MoE (*i.e.*, each expert) will have a data region in which it excels and on which it performs better than the other experts. Thus, it precisely

solves the problem that needing to treat the head and tail classes differently under large-scale long-tailed datasets.

RIDE [160] analyzes the prediction of long-tailed classifiers in terms of bias and variance, proposing that model bias measures the prediction accuracy relative to the true value; variance measures the stability of the prediction. Routing diverse experts is proposed to reduce the model variance of the long-tailed classifier by employing multiple experts, and a distributed perceptual diversity loss is set to reduce the model bias. The accuracy of both the head classes and the tail classes can be improved. Similarly, *Mixture-of-Experts (MoE)* [195] uses multiple experts to learn diverse results, and then the routing module dynamically integrates the results of multiple experts based on each input instance and trains them jointly with the expert network in an end-to-end manner.

MoE strategies usually requires reliable expert models for better learning of long-tailed data by means of model ensemble strategies. MoE can often achieve very high accuracy, but it always implicitly or explicitly integrates multiple models [160], which brings specific computational load and is a problem that needs to be improved in the follow-up work.
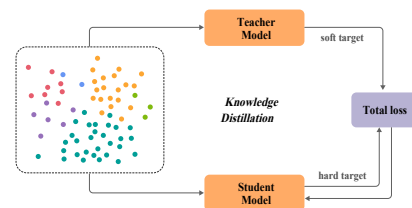
4.4.5 Knowledge Distilling

Fig. 15: **Knowledge distilling for long-tailed problem**. With the guidance of the teacher model, the student model can learn a more balanced sample.

Hinton [62] first propose the concept of knowledge distillation. By introducing a teacher network with superior inference performance, a streamlined and low-complexity student model is continuously induced for training, allowing the student model to continuously approach the predictions of the accurate network. For the long-tailed data distribution, knowledge distillation is always used to balance the predictions of head and tail classes.

In [173], *LFME* framework is proposed as a self-paced knowledge distillation method. The long-tailed dataset is split into several balanced subsets, and trained with expert models to guide the learning of the student models. A weighting scheme with automatic speed is introduced to train the training data in an easy-to-hard way, which eventually makes the student models outperform the expert models.

DiVE [61] uses the output of the teacher model as virtual examples to share knowledge among different classes by knowledge distillation, and proposes that *DiVE* can explicitly adjust the virtual example distribution to become flatter, thus improving the under-represented tail classes.

Balanced knowledge distillation (BKD) [186] first uses Cross-Entropy loss to train a common teacher model. The student model is then trained by minimizing the combination of instance-balanced classification loss and class-balanced distillation loss.

Knowledge distillation strategies also require reliable expert models for better learning of long-tailed data by means of expert-student model knowledge transfer. Moreover, for knowledge distillation strategies, existing long-tailed approaches usually no longer choose large models as teacher models, but learn better representation by teacher models [173] or balance data by adding virtual data [61], but this approach needs to choose appropriate knowledge distillation strategies with reasonable hyper-parameters.
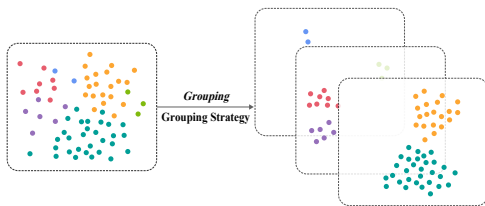
4.4.6 Grouping

Fig. 16: **Grouping method for long-tailed problem**. Group the data according to certain strategy and make the model learn separately for each group of data. The figure shows the group assignment according to the number of instances.

Some studies group vocabularies according to their relations, in addition to the most general grouping based on the number of instances of the class (*e.g.*, [96], [173], [65], *etc.*), *Forest R-CNN* [169] uses a priori knowledge of lexical, visual, and geometric relationships to construct classification trees, each of which will contribute to fine-grained classification. For example, when considering lexical relations, "school bus" and "car" have the same parent class "vehicle", while when considering geometric relations, "steering wheel" and "basketball" have the same parent class "roundness". The *Forest R-CNN* is designed to reduce the confidence scores of those classes misclassified by the fine-grained classifier, thus making the model more fault-tolerant in terms of the noise logarithm of the fine-grained classifier.

The grouping strategy requires to group the long-tailed data, and some researchers divide the categories according to the instance number, but this approach is likely to block the

knowledge interaction between groups. *Forest R-CNN* [169] divides the categories based on some semantic relations, but this requires the help of additional knowledge learning. Therefore, the automatic learning grouping strategy based on data itself is a problem worthy of research.
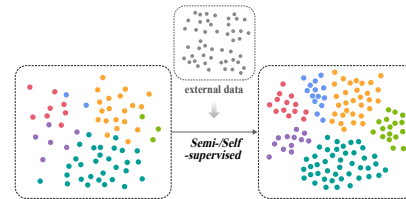
4.4.7 Semi-supervised

Fig. 17: **Semi-supervised for long-tailed problem**. Adding external data through semi-supervised strategy, thus increasing the weight of the tail classes in the overall training data.

There are also some studies [175] that use the strategy of generating pseudo labels to expand the data in semi-supervised manner [203]. Yang *et al*. [175] investigate a self-training semi-supervised learning method on long-tailed data by acquiring a certain amount of unlabeled data and generating pseudo-label for it, which is combined with labeled data to learn the final model. Wei *et al*. [165] perform data expansion of tail classes by predicting pseudo-labels for the tail classes of unlabelled data, thereby mitigating model over-fitting and alleviating data imbalance.

For object detection tasks, object-centered images may have richer data for tail classes. To make full use of these weakly labeled data and improve the tail classes performance, Ramanathan *et al*. [125] use a combination of weak supervision and full supervision. Weakly labeled images from YFCC100M [146] are used to perform enhancement for rare classes.

Also the scene-centric images have the potential to make it difficult for the detector to detect the tail objects, while the object-centric images make the detector give more attention to the target objects. Thus, Zhang *et al*. [182] use object-centric images to generate pseudo-scene-centric images to adjust the detector, thus eliminating the domain gap between the two image sources and also solving the problem of lacking bounding box labels for object-centric images. Finally, the scene centered images are combined to train and adjust the detector.

Semi-supervised learning generates pseudo-labels by learning from unlabeled data in order to expand the tail classes [175]. This approach compensates for the problem of insufficient learning of tail representation, but requires additional train-

Table 4: **Highlights of the main methods for solving the long-tailed problem as well as their limitations**. For summary purposes, some methods are not covered in this table.

Methods	Strategies	Representative Work	Highlights	Limitations
Data Processing Methods	Over-sampling	[53, 80, 100, 110, 121, 134]	Increase the number of samples for tail data.	a) Causes over-fitting of the tail class. b) Easy to amplify errors or noise present in the tail class.
	Under-sampling	[34, 106, 168]	Deletion of head data.	a) Causes under-fitting of the head class. b) It is possible to delete valuable data of head class by mistake.
	Data Augmentation	[17, 21, 22, 56, 57] [92, 180]	The tail data / feature is extended by data augmentation.	Inability to introduce new effective samples.
Cost Sensitive Weighting	Class-level Re-weighting	[27, 63, 66, 115, 143] [90, 144, 153, 170, 197]	Assigning weights to different classes and aggravating the learning of tail class.	a) It is difficult to choose appropriate weights for each class. b) Susceptible to the influence of sensitive hyper-parameters. c) There may be big differences for different data sets.
	Instance-level Re-weighting	[64, 89, 98, 135, 195]	Assign learning weights to examples based on their difficulty.	There is a high probability that the number of hard samples in the head class will exceed the tail class. So in essence, more emphasis will still be given to the head class.
Decoupling	–	[16, 77, 150, 157, 187, 197]	Decoupling representation learning and classifier learning.	a) Two-stage learning defies the end-to-end pursuit of deep learning. b) In the rebalancing phase, the same problems are faced as in other rebalancing methods.
Metric Learning	–	[38, 102, 104, 161, 188] [26, 93, 163]	Learn an embedding space in which to measure the similarity of embedded features or force a larger margin for the tail classes.	a) Select the appropriate distance function and measurement method. b) The learned embedding function still has the risk of biasing towards the head classes.
Transfer Learning	–	[65, 100, 108, 164]	Transferring the knowledge of head class to tail class.	Requiring a more complex model or module design, which can make the model difficult to train.
Meta Learning	–	[92, 126, 127, 136, 156]	Learn adaptive solutions from data or modules to make learning more automated.	a) The guidance of meta-data is weak. b) More complex model or module design is required.
Mixture-of-Experts	–	[160, 195, 198]	Multi-expert model ensemblling .	Expert model ensemble requires more computational resources,
Knowledge Distilling	–	[61, 173, 186]	Guided by the expert model, the student model is able to learn the data in a balanced manner.	a) Requires reliable expert models for better earning. b) Knowledge distillation needs to control the parameters of student model learning.
Grouping	–	[96, 169]	The data are trained in groups according to certain relationships.	A suitable grouping method needs to be found to ensure as much knowledge interaction between the groups as possible during training.
Semi-supervised	–	[125, 165, 175, 182]	Semi-supervised learning on long-tailed data by introducing other data sources.	Additional data sources are required.

ing, and it is difficult to play a role when the unlabeled data is not easy to obtain.

4.4.8 More Methods

In addition, there are several approaches that employ *Causal Inference* [145], *Adversarial Training* [171], *Distributional Robust Optimization (DRO)* [40] *etc*. to solve the long-tailed problem.

Tang *et al.* [145] propose that the momentum in the SGD optimizer is a confounder of the sample features and the classification logits, which may lead to spurious correlation between them. Therefore, causal intervention is used for de-confounding training to cut off backdoor confounding path and retain mediation path.

Wu *et al.* [171] find that long-tailed data have a negative impact on adversarial robustness and that the natural accuracy loss of the tail classes is further magnified in adversarial training. Meanwhile, they argue that suitable features as well as classifier embedding help to reduce the boundary error, and the combination of long-tailed recognition methods with the adversarial training framework helps to improve the natural accuracy. Therefore, the *RoBal* framework is designed with scale-invariant classifiers and a two-stage rebalancing method, respectively, which are thus used to improve the adversarial robustness.

DRO-LT [40] aims to improve the representation learning layer, and in order not to compromise the original data representation, a new loss based on robustness theory is proposed, which encourages the model to learn high quality representations of both head and tail classes.

Table 5: **Performance (%) summarization of some representative methods on CIFAR-10/100-LT benchmarks**. The summation of values in Epoch is meant to be a two-stage training strategy. Most methods adopt ResNet-32 as backbone. (In the 2020 and 2021 studies, the three best scores of $\beta = 100$ and $\beta = 10$ are marked in red, blue, and green, respectively.)

Year	Method	Pub.	CIFAR-10-LT (top-1) Imbalance Factor β						CIFAR-100-LT (top-1) Imbalance Factor β					
			200	100	50	20	10	1	200	100	50	20	10	1
	Softmax Loss [60]	CVPR	65.6	70.3	74.8	82.2	86.3	92.8	34.8	38.2	43.8	51.1	55.7	70.5
	Focal Loss [98]	ICCV	65.2	70.3	76.7	82.7	86.6	93.0	35.6	38.4	44.3	51.9	55.7	70.5
	L2RW [127]	ICML	66.5	74.1	78.9	82.1	85.1	89.2	33.3	40.2	44.4	51.6	53.7	64.1
2019	CB Loss [27]	CVPR	68.8	74.5	79.2	84.3	87.4	92.8	36.2	39.6	45.3	52.9	57.9	70.5
	MWNet [136]	NeurIPS	68.9	75.2	80.0	84.9	87.8	92.6	37.9	42.0	46.7	54.3	58.4	70.3
	CDB Loss [139]	ACCV	-	-	-	-	-	-	37.4	42.5	46.7	54.2	58.7	-
	RCBM-CE [69]	CVPR	70.6	76.4	80.5	86.4	88.8	92.7	39.3	43.3	48.5	55.6	59.5	71.8
	EQL [144]	CVPR	-	-	-	-	-	-	43.3	-	-	-	-	-
	BBN [198]	CVPR	-	79.8	82.1	-	88.3	-	-	42.5	47.0	-	59.1	-
2020	De-c-TDE [145]	NeurIPS	-	80.6	83.6	-	88.5	-	-	44.1	50.3	-	59.6	-
	BALMS [126]	NeurIPS	81.5	84.9	-	-	91.3	-	45.5	50.8	-	-	63.0	-
	FSA [22]	ECCV	75.5	82.0	84.4	89.2	91.2	-	41.4	48.5	52.1	59.7	65.3	-
	Remix-DRW [21]	ECCV	-	79.7	-	-	89.0	-	-	46.7	-	-	61.2	-
	LFME [173]	ECCV	-	-	-	-	-	-	37.4	42.5	46.7	54.2	58.7	-
	MBJ [103]	ArXiv	-	81.0	87.2	-	88.8	-	-	45.8	57.5	-	60.7	-
	RIDE(4 experts) [160]	ICLR	-	-	-	-	-	-	-	49.1	-	-	-	-
	LADE [63]	CVPR	-	-	-	-	-	-	-	45.4	50.5	-	61.7	-
	MetaSAug-CE [92]	CVPR	76.8	80.5	84.0	87.6	89.4	-	39.9	46.8	51.9	57.8	61.7	-
	Hybrid-SC [155]	CVPR	-	81.4	85.3	-	91.1	-	-	46.7	51.8	-	63.0	-
	Hybrid-PSC [155]	CVPR	-	78.8	83.8	-	90.0	-	-	44.9	48.9	-	62.3	-
	MiSLAS [197]	CVPR	-	82.1	85.7	-	90.0	-	-	47.0	52.3	-	63.2	-
2021	Bag of Tricks [191]	AAAI	-	80.0	83.5	-	-	-	-	47.8	51.6	-	-	-
	LDA [122]	ACM MM	-	-	-	-	-	-	-	50.6	54.6	-	61.9	-
	TSC [93]	ArXiv	-	79.7	82.9	-	88.7	-	-	43.8	47.4	-	59.0	-
	BKD [186]	ArXiv	-	81.7	83.8	-	89.2	-	-	45.0	49.6	-	61.3	-
	DRO-LT [40]	ArXiv	-	-	-	-	-	-	-	47.3	57.5	-	63.4	-
	DiVE [61]	ArXiv	-	-	-	-	-	-	-	45.3	51.1	-	62.0	-
	LADC [150]	ArXiv	81.5	84.6	87.0	-	90.8	-	46.6	50.7	54.9	-	64.6	-
	MARC [163]	ArXiv	81.1	85.3	-	-	-	-	47.4	50.8	-	-	-	-

On the whole, long-tailed learning continues to learn from other machine learning sub-fields, and more combinatorial methods will appear one after another.

4.5 Summary

We summarize the above long-tailed visual recognition approaches in this section. Highlights and limitations of these methods are shown in Tab. 4. On the whole, the characteristics of these methods are very distinct, focusing on data processing, loss function, model architecture, training methods and so on. However, there is no method that can greatly solve the long-tailed problem. In practical application, it is often a combination of multiple methods, such as using semi-supervised learning to expand the tail data, using over-sampling to improve the sampling frequency of tail data, using class-level re-weighting to balance the gradient during training, and finally using decoupling strategy to enhance the representation ability of the model. Therefore, it can be predicted that the research on solving the long-tailed problem will still be in full bloom in the future.

5 Performance Comparison

To provide readers with a straightforward statistic, we compare the performance of some mainstream long-tailed studies in this section. For classification task, we report some popular long-tailed studies via ImageNet-LT and Places-LT, CIFAR-10/100-LT, and iNaturalist 2017 & 2018 benchmarks, respectively. For object detection and instance segmentation tasks, we report some popular long-tailed studies via LVIS benchmarks. It should be noted that the experimental settings of each study are not completely consistent. We try to eliminate these effects when comparing, but we still can't be absolutely fair. Therefore, we hope that readers can only take the comparison in this section as a reference, and the specific performance comparison still needs to be analyzed based on the implementation details of the original article.

5.1 CIFAR-10/100-LT Performance Benchmarking

The performance of some representative methods on CIFAR-10/100-LT benchmark is shown in Tab. 5. In the 2020 and 2021 studies, we select two cases with imbalance factors of 100 and 10 on CIFAR-10/100-LT, respectively. The three best scores of each year are marked in red, blue, and green. In the

Table 6: **Performance (%) summarization of some representative methods on ImageNet-LT and Places-LT benchmarks**. For Places-LT dataset, we follow the experimental setup of most of the current work as a statistical benchmark, where ResNet-152 is uniformly selected as the backbone. (In the 2020 and 2021 studies, the three best scores of overall top-1 accuracy are marked in red, blue, and green, respectively.) PaCo [26]† models are trained with RandAugment [24] in 400 epochs.

Year	Method	Pub.	Backbone	ImageNet-LT (top-1)				Places-LT (top-1)			
				>100 Many-shot	≤ 100 & >20 Medium-shot	<20 Few-shot	Overall	>100 Many-shot	≤ 100 & >20 Medium-shot	<20 Few-shot	Overall
	Softmax Loss [60]	CVPR	ResNet-10	40.9	10.7	0.4	20.9	45.9	22.4	0.36	27.2
	Focal Loss [98]	ICCV	ResNet-10	36.4	29.9	16.0	30.5	41.1	34.8	22.4	34.6
	Range Loss [188]	ICCV	ResNet-10	35.8	30.3	17.6	30.7	41.1	35.4	23.2	35.1
2019	OLTR [108]	CVPR	ResNet-10	43.2	35.1	18.5	35.6	44.7	37.0	25.3	35.9
	CDB Loss [139]	ACCV	ResNet-10	-	-	-	38.4	-	-	-	-
	EQL [144]	CVPR	ResNet-10	-	-	-	36.4	-	-	-	-
	BALMS [126]	NeurIPS	ResNet-10	50.3	39.5	25.3	41.8	41.2	39.8	31.6	38.7
	De-c-TDE [145]	NeurIPS	ResNeXt-50	62.7	48.8	31.6	51.8	-	-	-	-
	FSA [22]	ECCV	ResNet-10	47.3	31.6	14.7	35.2	42.8	37.5	22.7	36.4
2020	LFME [173]	ECCV	ResNet-10	47.1	35.0	17.5	37.2	38.4	39.1	21.7	35.2
	Joint [77]	ArXiv	ResNeXt-50	-	-	-	44.4	-	-	-	-
	NCM [77]	ArXiv	ResNeXt-50	-	-	-	47.3	-	-	-	-
	cRT [77]	ArXiv	ResNeXt-50	-	-	-	49.5	-	-	-	-
	τ-normalized [77]	ArXiv	ResNeXt-50	-	-	-	49.5	-	-	-	-
	LWS [77]	ArXiv	ResNeXt-50	-	-	-	49.9	-	-	-	-
	MBJ [103]	ArXiv	ResNeXt-50	61.6	48.4	39.0	52.1	39.5	38.2	35.5	38.1
	RIDE(4 experts) [160]	ICLR	ResNet-50	66.2	52.3	36.5	55.4	-	-	-	-
	Logit adjustment [113]	ICLR	ResNet-50	-	-	-	51.1	-	-	-	-
	LADE [63]	CVPR	ResNeXt-50	62.3	49.3	31.2	51.9	42.8	39.0	31.2	38.8
	Seesaw Loss [153]	CVPR	ResNeXt-50	67.1	45.2	21.4	50.4	-	-	-	-
	MetaSAug-CE [92]	CVPR	ResNet-50	-	-	-	47.3	-	-	-	-
	MiSLAS [197]	CVPR	ResNet-50	61.7	51.3	35.8	52.7	39.6	43.3	36.1	40.4
	DisAlign [187]	CVPR	ResNet-50	59.9	49.9	31.8	52.9	40.4	42.4	30.1	39.3
	Bag of Tricks [191]	AAAI	ResNet-10	-	-	-	43.1	-	-	-	-
	LDA [122]	ACM MM	ResNeXt-50	64.5	50.9	31.5	53.4	32.1	40.7	41.0	39.1
2021	PaCo [26]†	ICCV	ResNeXt-50	-	-	-	58.2	36.1	47.9	35.3	41.2
	TSC [93]	ArXiv	ResNet-50	63.5	49.7	30.4	52.4	-	-	-	-
	GistNet [100]	ArXiv	ResNet-10	52.8	39.8	21.7	42.2	42.5	40.8	32.1	39.6
	BKD [186]	ArXiv	ResNet-10	54.6	37.2	20.4	41.6	41.9	39.1	30.0	38.4
	DRO-LT [40]	ArXiv	ResNet-50	64.0	49.8	33.1	53.5	-	-	-	-
	ResLT [25]	ArXiv	ResNeXt-50	63.0	50.5	35.5	52.9	39.8	43.6	31.4	39.8
	DiVE [61]	ArXiv	ResNeXt-50	64.0	50.4	31.4	53.1	-	-	-	-
	Breadcrumbs [101]	ArXiv	ResNeXt-50	62.9	47.2	30.9	51.0	40.6	41.0	33.4	39.3
	ALA Loss [195]	ArXiv	ResNeXt-50	64.1	49.1	34.0	52.8	-	-	-	-
	MARC [163]	ArXiv	ResNeXt-50	60.4	50.3	36.6	52.3	39.9	39.8	32.6	38.4

2020 studies, *BALMS* [126] achieve the highest performance in three metrics. In addition, *FSA* [22] and *Remix-DRW* [21] are also competitive methods in that year. In the next year, *MARC* [163], *LADC* [150] and *MiSLAS* [197] rank in the top three of the comprehensive performance.

Through the results, we can clearly observe that the performance of the method in 2021 is slightly improved compared with that in 2020. The 2021 best method *MARC* is only 0.4 points higher (85.3% vs 84.9%) than the 2020 best method *BALMS* in 2020 at $\beta = 100$ of CIFAR-10-LT, and *MARC* has not even improved in the other three metrics. This shows that CIFAR-LT datasets have become saturated, and future research should focus on more difficult benchmarks.

5.2 ImageNet-LT & Places-LT Performance Benchmarking

Tab. 6 shows the performance of some representative methods on ImageNet-LT and Places-LT benchmarks. The three best scores of overall top-1 accuracy on 2020 and 2021 are marked in red, blue, and green, respectively.

For ImageNet-LT, the 2021 best method *PaCo* [26] achieves 58.2% overall top-1 accuracy, which is 6.1 points higher than the 2020 best method *MBJ* [103]. Although *PaCo* adopts RandAugment and longer training epochs, this improvement is still very significant. In addition, Tab. 6 records 19 studies on ImageNet-LT in 2021, which has almost doubled compared with 2020. More researchers have joined the community and greatly promoted the development of long-tailed recognition. Of course, compared with the best result of ResNeXt-50 on balanced ImageNet-1K (80.5% [167]), there are still many problems to be solved in the research of long-tailed recognition.

For Places-LT, *PaCo* is still the best method in 2021, with an overall top-1 accuracy of 41.2%, which is 2.5 points higher than the 2020 best method *BALMS* [126]. Although the methods in 2021 are better than those in 2020 on the whole, the improvement is still relatively small compared with ImageNet-LT. We consider that most of the current long-tailed recognition methods are mainly designed for object-centric data, while the scene-centric long-tailed problem needs specific solutions.

Table 7: **Performance (%) summarization of some representative methods on iNaturalist 2017 & 2018 benchmarks.** All the methods adopt ResNet-50 as backbone. (In the 2020 and 2021 studies, the three best scores of overall top-1 accuracy on iNaturalist 2018 are marked in red, blue, and green, respectively.) PaCo [26]† is trained with RandAugment [24].

Year	Method	Pub.	Epoch	iNat 2017 (top-1) Overall	iNat 2018 (top-1) >100 Many-shot	⩽ 100 & >20 Medium-shot	<20 Few-shot	Overall
	Softmax Loss [60]	CVPR	-	54.6	-	-	-	57.1
2019	CB Loss [27]	CVPR	-	58.0	-	-	-	61.1
	LDAM [14]	NeurIPS	60+30	-	-	-	-	68.0
2020	BBN [198]	CVPR	180	65.7	-	-	-	**69.6**
	RCBM-CE [69]	CVPR	-	59.3	-	-	-	67.3
	FSA [22]	ECCV	100	61.9	-	-	-	65.9
	Remix-DRW [21]	ECCV	200	-	-	-	-	**70.4**
	Joint [77]	ArXiv	90 / 200	-	72.2 / 75.7	63.0 / 66.9	57.2 / 61.7	61.7 / 65.8
	NCM [27]	ArXiv	90 / 200	-	55.5 / 61.0	57.9 / 63.5	59.3 / 63.3	58.2 / 63.1
	cRT [77]	ArXiv	90 / 200	-	69.0 / 73.2	66.0 / 68.8	63.2 / 66.1	65.2 / 68.2
	τ-normalized [77]	ArXiv	90 / 200	-	65.6 / 71.1	65.3 / 68.9	65.9 / 69.3	65.6 / 69.3
	LWS [27]	ArXiv	90 / 200	-	65.0 / 71.0	66.3 / 69.8	65.5 / 68.8	65.9 / 69.5
	MBJ [103]	ArXiv	90 / 200	-	-	-	-	**66.9 / 70.0**
2021	RIDE(4 experts) [160]	ICLR	100	-	70.9	72.4	73.1	**72.6**
	Logit adjustment [113]	ICLR	90	-	-	-	-	68.4
	Bag of Tricks [191]	AAAI	90	-	-	-	-	70.8
	LADE [63]	CVPR	200	-	-	-	-	70.0
	MetaSAug-CE [92]	CVPR	-	63.2	-	-	-	68.7
	Hybrid-SC [155]	CVPR	100	-	-	-	-	66.7
	Hybrid-PSC [155]	CVPR	100 / 200	-	-	-	-	68.1 / 70.3
	MiSLAS [197]	CVPR	200	-	73.2	72.4	70.4	**71.6**
	DisAlign [187]	CVPR	90 / 200	-	61.6 / 68.0	70.8 / 71.3	69.9 / 69.4	69.5 / 70.2
	PaCo [26]†	ICCV	400	-	-	-	-	**73.2**
	TSC [93]	ArXiv	-	-	72.6	70.6	67.8	69.7
	GistNet [100]	ArXiv	200	-	-	-	-	70.8
	BKD [186]	ArXiv	90	-	67.1	66.1	67.6	66.8
	DRO-LT [40]	ArXiv	-	-	-	-	-	69.7
	ResLT [25]	ArXiv	200	-	-	-	-	70.2
	DiVE [61]	ArXiv	90	-	70.6	70.0	67.5	69.1
	Breadcrumbs [101]	ArXiv	200	-	-	-	-	70.3
	ALA Loss [195]	ArXiv	200	-	71.3	70.8	70.4	70.7
	MARC [163]	ArXiv	200	-	-	-	-	70.4

5.3 iNaturalist 2017 & 2018 Performance Benchmarking

iNaturalist is a large species dataset, and the relevant research results of the 2017 and 2018 versions are given in Tab. 7. Since there is less research on iNaturalist 2017, we mainly analyze the studies on iNaturalist 2018. Same as Tab. 6, three best scores of overall top-1 accuracy on 2020 and 2021 are marked in red, blue, and green, respectively.

Among 2020 studies, *Remix-DRW* [21], *MBJ* [103] and *BBN* [198] rank in the best three on the overall top-1 accuracy, achieve 70.4%, 70.0% and 69.6% respectively. *PaCo* [26] still stand out in 19 studies in 2021, with 73.2% overall top-1 accuracy, an increase of 2.8 points compared with the best method in 2020. On the whole, iNaturalist 2018 is still not saturated, and its importance in the long-tailed research community is basically equal to ImageNet-LT. This domain-specific long-tailed benchmark enriches the diversity of research objectives and has great potential in industrial vision, retail, medical, *etc*.

5.4 LVIS v0.5 & v1.0 Performance Benchmarking

Tab. 8 and Tab. 9 summarize the performance of some representative work on the LVIS v0.5 & v1.0 benchmark. Due to the small number of studies, we will not discuss the studies in 2020 and 2021 separately. And as LVIS v1.0 gradually becomes the mainstream long-tailed instance segmentation benchmark, some new research does not conduct experiments on LVIS v0.5, so here we only analyze and discuss the former benchmark. The three best scores of AP^{mask}, AP_r^{mask} and AP^{bbox} are marked in red, blue, and green.

Overall, the performance of *EOD* [90], *Seesaw Loss* [153], *LDA* [122] and *EQL v2* [143] is in the leading position in LVIS v1.0 benchmark, and the gap between them is not large. Most of these studies belong to class-level re-weighting method, which can be seen as the mainstream solution to the long-tailed object detection and instance segmentation problems. These methods also use *RFS* [53] to increase the sampling frequency of tail classes, and have achieved good results. In addition, the regularly held challenge competitions [2] also significantly promoted community develop-

Table 8: **Performance (%) summarization of some representative methods on LVIS v0.5 benchmark.** All the methods adopt Mask R-CNN with ResNet-50-FPN. In the 'Epoch' column, values making additive operations represents the two-stage training strategy. (In the 2020 and 2021 studies, the three best scores of AP^{mask}, AP_r^{mask} and AP^{bbox} are marked in red, blue, and green, respectively.)

| Year | Method | Pub. | Epoch | LVIS v0.5 (mAP) | | | | | | AP^{bbox} |
| | | | | AP^{mask} | | | | | | |
				AP	AP_{50}	AP_{75}	AP_r	AP_c	AP_f	
	Softmax Loss [60]	CVPR	25	20.2	32.6	21.3	4.5	20.8	25.6	20.7
	Sigmoid Loss	-	25	20.1	32.7	21.2	7.2	19.9	25.4	20.5
	CAS [134]	ECCV	25	18.5	31.1	18.9	7.3	19.3	21.9	18.4
	Focal Loss [98]	ICCV	25	21.0	34.2	22.1	9.3	21.0	25.8	21.9
2019	RFS [53]	CVPR	25	24.4	-	-	14.5	24.3	28.4	-
	CB Loss [27]	CVPR	25	20.9	33.8	22.2	8.2	21.2	25.7	21.0
	LDAM [14]	NeurIPS	25	24.1	-	-	14.6	25.3	26.3	24.5
	EQL [144]	CVPR	25	22.8	36.0	24.4	11.3	24.7	25.1	23.3
	LST [65]	CVPR	10 + 10	23.0	36.7	24.8	-	-	-	22.6
	BAGS [96]	CVPR	12 + 12	26.2	-	-	17.9	26.9	28.7	25.7
2020	Forest R-CNN [169]	ACM MM	25	25.6	40.3	27.1	18.3	26.4	27.6	25.9
	TFA-cos [96]	ICML	-	-	-	-	-	-	-	22.7
	BALMS [126]	NeurIPS	25	27.0	-	-	19.6	28.9	27.5	27.6
	SimCal [157]	ECCV	-	23.4	-	-	16.4	22.5	27.2	-
	LWS [27]	ArXiv	25	23.8	-	-	14.4	24.4	26.8	24.5
	DisAlign [187]	CVPR	25 + 2.5	26.3	-	-	14.9	27.6	29.2	25.6
	EQL v2 [143]	CVPR	24	27.1	-	-	18.6	27.6	29.9	27.0
2021	ACSL [158]	CVPR	12 + 12	26.4	42.3	28.6	18.6	26.4	29.3	-
	Drop Loss [64]	CVPR	25	25.5	38.7	27.2	13.2	27.9	27.3	25.1
	Simp-Effe [182]	ArXiv	25	-	-	-	-	-	-	24.5

Table 9: **Performance (%) summarization of some representative methods on LVIS v1.0 benchmark.** All the methods adopt Mask R-CNN with ResNet-50-FPN. † denotes that the result is reproduced by us. s(In the 2020 and 2021 studies, the three best scores of AP^{mask}, AP_r^{mask} and AP^{bbox} are marked in red, blue, and green, respectively.)

| Year | Method | Pub. | Epoch | LVIS v1.0 (mAP) | | | | | | AP^{bbox} |
| | | | | AP^{mask} | | | | | | |
				AP	AP_{50}	AP_{75}	AP_r	AP_c	AP_f	
	Softmax Loss [60]†	CVPR	25	18.2	28.6	19.1	1.2	15.3	28.9	18.6
	Focal Loss [98]	ICCV	24	-	-	-	-	-	-	18.5
	RFS [53]†	CVPR	25	22.2	34.9	23.5	11.1	20.8	28.6	22.9
2020	EQL [143, 144]	CVPR	24	21.6	-	-	3.8	21.7	29.2	22.5
	BAGS [96]†	CVPR	25	23.7	37.7	25.1	15.4	22.8	28.3	23.4
	Drop Loss [64]	AAAI	25	22.3	34.5	23.6	12.4	22.3	26.5	22.9
	EQL v2 [143]	CVPR	24	25.5	-	-	17.7	24.3	30.2	26.1
2021	Seesaw Loss [153]	CVPR	24	25.7	-	-	19.1	25.0	29.4	26.8
	LDA [122]	ACM MM	24	25.7	-	-	-	-	-	26.6
	FASA [180]	ArXiv	24	22.6	-	-	10.2	21.6	29.2	22.6
	Fed Loss [202]†	ArXiv	25	21.4	33.4	22.6	4.5	20.5	29.7	22.2
2022	EOD [90]	ArXiv	24	-	-	-	-	-	-	27.5

ment. *EQL* [143, 144], *Seesaw Loss* and *Fed Loss* [202] have all won good rankings. However, most of these studies are the extension of image-level long-tailed recognition methods, and there is still less research on the instance-level long-tailed problem. Especially for the long-tailed instance segmentation problem, the use of mask information is very small, which is a problem that researchers need to pay attention to.

To solve long-tailed problems, recent research pays much attention to some long-tailed benchmark datasets which we collated in Sec. 3, but in other tasks, the solution of the long-tailed phenomenon is not sufficient. To investigate this, we analyze some widely-used datasets in visual recognition, and quantitatively evaluate the impact of the long-tailed phenomenon.

6 Analysis of Long-tailed Phenomenon

Although the long-tailed phenomenon is prevalent in visual recognition, the scope of current research is very limited.

6.1 Long-tailed Phenomenon in Widely-used Datasets

Long-tailed distribution is a common phenomenon. Datasets without artificial balance will basically follow this distribu-

Table 10: **Analysis of long-tailed phenomenon of 20 mainstream datasets**. Light purple indicates which has a general long-tailed distribution ($0.6 \leq \delta < 0.8$), and dark purple indicates which has a severe long-tailed distribution ($0.8 \leq \delta$). † denotes few-shots datasets.

Dataset	Venue	Fields	Anno. Types	Training Samples	Classes	Max Size	Min Size	Gini Coef. δ
ImageNet-1K [32]	CVPR 2009 IJCV 2015	Object-centric	Classification	1,281,167	1,000	1,300	732	0.013
Sports1M [7]	CVPR 2014	Human-centric	Classification	958,827	487	2,385	694	0.09
COCO [99]	ECCV 2014	Object-centric	Bounding-box Instance-mask	118,287	80	262,465	198	0.564
Market1501 [196]	ICCV 2015	Human-centric	Person-identity	12,937	752	72	1	0.329
MS1M [52]	ECCV 2016	Face-centric	Face-identity	5,822,653	85,742	602	2	0.314
DukeMTMC [130]	ECCV 2016	Human-centric	Person-identity	16,522	702	426	6	0.268
YouTube8M [3]	ArXiv 2016	Human-centric	Classification	5,786,881	3,862	788,288	123	0.839
Place365 [200]	PAMI 2017	Scene-centric	Classification	1,803,460	365	5,000	3,068	0.011
ADE20K [201]	CVPR 2017	Scene-centric	Segmentation	20,210	150	3014.9	3.84	0.801
Sth-Sth v2 [49]	ICCV 2017	Human-centric	Classification	168,913	174	3,284	91	0.352
COCO-stuff [13]	CVPR 2018	Scene-centric	Segmentation	118,287	171	10,021.1	3.81	0.653
MHP v2 [194]	ACM MM 2018	Human-centric	Human-parsing	15,403	59	840.7	0.823	0.747
OID v4 [85]	ArXiv 2018 IJCV 2020	Object-centric	Bounding-box	1,743,042	500	1,395,645	4	0.902
Object365 [133]	ICCV 2019	Object-centric	Bounding-box	608,606	365	2,120,895	28	0.845
GLD v2 [166]	CVPR 2020	Scene-centric	Classification	4,132,914	203,094	10,247	1	0.655
FSOD† [42]	CVPR 2020	Object-centric	Bounding-box	52,350	800	2,114	26	0.361
FSS-1000† [95]	CVPR 2020	Object-centric	Segmentation	20,006	1,000	21	20	0.00029
Glint360K [4]	ArXiv 2020	Face-centric	Face-identity	17,091,657	360,232	1,868	3	0.647
VSPW [114]	CVPR 2021	Scene-centric	Segmentation	197,253	124	17,191.6	13.5	0.742
LaST [137]	ArXiv 2021	Human-centric	Person-identity	71,248	5,000	140	5	0.427

tion. In order to study the long-tailedness in visual recognition, we compiled and analyzed other 20 widely-used large-scale datasets covering the fields of image classification [32,166,200], object detection [42,85,99,133], semantic segmentation [13,95,114,201], person re-identification [130, 137,196], face recognition [4,52], human parsing [194], video / action recognition [3,7,49], *etc*. as shown in Tab. 10. Based on the Gini coefficient proposed in Sec. 3, we measure the long-tailedness in these datasets and mark each of them with general long-tailed distribution in light purple and those with severe long-tailed distribution in dark purple. From the perspective of both release time and annotation type, we can observe the trend of the long-tailed phenomenon in visual recognition approximately.

In terms of release time, a number of datasets showing long-tailed phenomenons have emerged since 2016, and the proportion of long-tailed datasets has gradually increased with each year. People are almost no longer controlling the balance of datasets artificially as CIFAR, ImageNet-1K, Places365 and COCO did.

In terms of annotation type, although we have improved the long-tailed criteria of the classification datasets, the classification task has a larger proportion of balanced datasets than the object detection and segmentation tasks in our statistics. It can be seen that for the classification task, the balance of datasets is easier to control compared to object detection and segmentation task. Such as ImageNet-1K [32], Sport1M [7], Sth-Sth v2 [49], Market1501 [196], DukeMTMC [130], MS1M [52],

and Places365 [200], many datasets in classification task demonstrate balance for which Gini coefficients are less than 0.4 and are considered as balanced datasets. Nevertheless, there are still many long-tailed datasets for classification tasks without adding artificial control over the balance, such as scene classification dataset GLD v2 [166] (δ=0.655), face recognition dataset Glint360K [4] ($\delta = 0.647$) and person re-identification dataset LaST [137](δ=0.427) which are considered as general long-tailed datasets. The large video understanding dataset YouTube8M [3] (δ=0.839) is a severe long-tailed dataset. In the field of object detection and segmentation, there are fewer balanced datasets. Except COCO [99], only few-shot learning datasets FSOD [42] and FSS-1000 [95] show to be balanced, which also reflects the difference between few-shot learning and long-tailed problems. Moreover, from the perspective of the Gini coefficient, the Gini coefficient for the object detection and segmentation task is generally higher than that for the classification tasks and is generally higher than 0.7. For example, for object detection task, OID v4 [85] (δ=0.902) and Object365 [133] (δ=0.845) both have Gini coefficients greater than 0.8, which belong to severe long-tailed datasets. And for segmentation task, human body parsing dataset MHP v2 [194] (δ=0.747) and video semantic segmentation dataset VSPW [114] (δ=0.742) are general long-tailed datasets. The scene parsing dataset ADE20K [201] (δ=0.801) is a severe long-tailed dataset. This indicates that the long-tailed phe-

nomenon is much more serious for object detection task and segmentation task, which requires more attention.

We find that the long-tailed phenomenon of the dataset is prevalent in the statistical process. Tab. 10 shows that the Gini coefficient of the dataset is gradually increasing in recent years, accompanied by a more severe long-tailed phenomenon in the dataset. We attribute this phenomenon to the fact that as people's research is more and more invested in large-scale datasets, it is increasingly difficult for researchers to control the balance of datasets, so the long-tailed phenomenon will inevitably become more and more serious, which leads to more and more long-tailed datasets in recent years.

6.2 Analysis of Performance

We select two severe long-tailed datasets in Tab. 10 (Object365 [133] and ADE20K [201]), and evaluate their performance to investigate whether the long-tailed problem is practically shown in visual recognition. In order to evaluate the impact of the long-tailed phenomenon, we split classes in descending order into three groups: the first 20% as head classes, the middle 60% as body classes, and the last 20% as tail classes. We collect some mainstream solutions without long-tailed methods on these three datasets and analyze their performance on head, body and tail classes, respectively.

Table 11: **Long-tailed performance analysis of object detection on Object365 [133]**. All the methods adopt ResNet-50-FPN as backbone and are trained by us on *Detectron2* [172].

Method	AP	AP_{tail}	AP_{body}	AP_{head}
Faster R-CNN [128]	19.8	3.7	21.5	31.0
RetinaNet [98]	18.5	3.5	19.9	29.4
FCOS [147]	20.6	4.8	22.2	31.7

For object detection task, we investigate the generic object detection dataset Object365 and evaluate the performance of Faster R-CNN [128], RetinaNet [98] and FCOS [147] on the head, body and tail classes. We adopt ResNet-50-FPN as backbone based on *Detectron2* [172], and take AP_{box} as the evaluation metric, the performance results are shown in Tab. 11. For Faster R-CNN, RetinaNet and FCOS, the head classes accuracy is $1.44\times$, $1.48\times$ and $1.43\times$ higher than the body classes, and $8.38\times$, $8.40\times$ and $6.60\times$ higher than the tail classes.

For semantic segmentation task, we investigate the performance of mainstream models on the large scene parsing dataset ADE20K [201]. We take mIoU as the evaluation metric, with the backbone of ResNet-50, the performance of SemSegFPN [82], PSPNet [193], and MaskFormer [20]

Table 12: **Long-tailed performance analysis of semantic segmentation on ADE20K [201]**. All the methods adopt ResNet-50 as backbone, the SemSegFPN and PSPNet models are taken from *mmsegmentation* [23], the MaskFormer model is taken from the officially published.

Method	mIoU	$mIoU_{tail}$	$mIoU_{body}$	$mIoU_{head}$
SemSegFPN [82]	37.48	24.03	35.33	57.40
PSPNet [193]	42.47	29.16	40.96	60.32
MaskFormer [20]	44.50	33.19	41.96	61.96

is shown in Tab. 12. For SemSegFPN, PSPNet, and Mask-Former, their head classes accuracy exceeds the body classes by $1.62\times$, $1.47\times$, and $1.47\times$, and exceeds the tail classes by $2.39\times$, $2.06\times$, and $1.87\times$.

From the performance of the above model on the head, body and tail classes, it can be found that the accuracy of model is strongly related to the instance number of classes, and their performance becomes worse as the instance number decreases, which indicates that the long-tailed problem of these datasets needs to be solved. There is no doubt that the long-tailed phenomenon is prevalent, and as people increasingly analyze large-scale datasets, people should realize the importance of solving it. Although some studies are aware of the datasets' long-tailed distribution, this problem has not been widely studied. For example, for the pixel-level semantic segmentation task, only few studies [187] have addressed the long-tailed phenomenon. Even for many other fields, none of their mainstream approaches has a targeted design for the long-tailed problem, so we believe that more effort needs to be devoted to the analysis of the long-tailed phenomenon.

7 Future Directions

As a contemporary survey for long-tailed visual recognition using deep learning, this paper has discussed the problems caused by the long-tailed distribution, summarized existing popular long-tailed datasets, provided some structural taxonomy for various methods as well as analyzed their advantages and limitations, we also find that the long-tailed phenomenon is widespread, and pointed out some valuable research areas of the long-tailed problem. Despite great progress, there are still many unsolved problems. Thus in this section, we will point out these problems and introduce some promising trends for future research. We hope that this survey not only provides a better understanding of long-tailed visual recognition for researchers but also stimulates future research activities.

Large model with Large-scale Data. Large-scale Pre-trained Language Models (PLMs) have become the new paradigm for Natural Language Processing (NLP) [10, 35, 176]. Large

model with large-scale data has demonstrated strong performances on natural language understanding and generation with zero-shot and few-shot learning. In the field of visual recognition, model parameters and data scale are limited by the characteristics of visual tasks, which is relatively small compared with NLP, but some recent studies have begun to work in this direction [39, 107, 129]. Compared with the existing methods, large models and big data do not need to explicitly model the label frequency, but learn the general representation of images through a large amount of data, so as to solve the long-tailed problem.

Long-tailed Adversarial Learning. Adversarial learning aims to deceive the model by providing deceptive input, the main research work can be simply divided into two parts: attack [51, 109, 174] and defense [152, 177]. The research of adversarial learning has greatly promoted the safety and standardization of machine learning. However, with the blowout development of the defense mode and attack mode of the model, it also gradually presents a long-tailed distribution. Therefore, it seems to be a problem worth exploring to solve the tail classed in adversarial learning through the idea of long-tailed learning.

Self-supervised Long-tailed Learning. Self-supervised learning methods regard each sample as an individual class, which can alleviate the label shifts problems and learn a relatively complete feature representation for all classes [18, 19, 59]. However, there is less work that draws on self-supervised learning with long-tailed datasets, such as [175] uses self-supervised learning to improve the performance on long-tailed datasets by considering ignoring the value of labels. SSD [94] uses the self-supervision guide feature learning method to improve the ability of the feature extractor. As work in the field of self-supervised has matured, there is great hope that self-supervision learning can surpass traditional supervised learning methods, thus making the learning of models more intelligent and automated. On this basis, the label bias problem for long-tailed datasets will hopefully be greatly improved.

Vision-Language Long-tailed Learning. Vision-Language tasks require a model to understand the visual world and to ground natural language to the visual observations [6, 8, 72, 185]. Vision-Language dataset contains two modes of annotation, and its long-tailed phenomenon is difficult to avoid [83, 146]. Recently, CLIP [124] proposes visual representation learning via natural language supervision in a similar contrastive learning setting [54], and shows amazing results on zero-shot and few-shot image classification. ViLD [50] extends CLIP to zero-shot object detection task through knowledge distillation and prompt, and goes beyond the supervised learning method in the novel class of LVIS. These studies show that Vision-Language model can learn knowledge from multiple modes and improve the representation ability of few samples, which may be the next breakthrough of the long-tailed problem.

More Task Settings. In addition to the well-known long-tailed tasks and some research directions proposed in this section, there are more long-tailed visual recognition tasks waiting to be mined. As analyzed in Sec. 6.1, data naturally satisfy the long-tailed distribution in many fields, so solving the long-tailed problem may be able to improve the performance of models in these fields. However, according to the available research results, only a few work have studied from the perspective of the long-tailed distribution in their research field, such as long-tailed distribution of object classes in UAV images [179], long-tailed distribution of driving behavior in autonomous driving [116] , and topic in the field of visual story telling [91], content-related words for video captioning tasks [192], pose inclusion in datasets for 3D human pose estimation [181], dermatological categories in dermatological diagnosis [123], *etc*. We believe that for many research fields, the existing work to analyze and solve the long-tailed distribution problem is still not enough. For future researchers, the long-tailed problem can be taken into account to solve the problem of extreme data imbalance and thus improve the performance of the task.

8 Conclusions

In this survey, we comprehensively reviewed the long-tailed visual recognition according to the datasets, methods, long-tailed phenomenon and future directions. We provided the necessary background knowledge for readers, summarize the long-tailed studies into ten categories from the perspective of representational learning, and summarized the highlights and limitations of each category. We also compiled some generalized long-tailed datasets and benchmarked the results on 8 datasets. To study the long-tailed phenomenon extensively, we also conducted a structured survey of 20 widely-used datasets and found that the long-tailed phenomenon is widespread and that many areas' mainstream studies are not aware of it. Based on the analysis of the universality of the long-tailed phenomenon, we also gave the potential innovation and future research direction. We expect this survey to provide an effective way to understand current state-of-the-arts and speed up the development of this research field.

Acknowledgements

References

1. inaturalist 2018 competition dataset. https://github.com/visipedia/inat_comp/tree/master/2018 (2018) 5, 6

2. Lvis challenge. https://www.lvisdataset.org/ (2019) 22

3. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675 (2016) 3, 24

4. An, X., Zhu, X., Xiao, Y., Wu, L., Zhang, M., Gao, Y., Qin, B., Zhang, D., Fu, Y.: Partial fc: Training 10 million identities on a single machine. arXiv preprint arXiv:2010.05222 (2020) 24

5. Anderson, C.: The long tail: Why the future of business is selling less of more. Hachette Books (2006) 1, 2

6. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: Proceedings of the European Conference on Computer Vision, pp. 382–398 (2016) 26

7. Andrej, K., George, T., Sanketh, S., Thomas, L., Rahul, S., Li, F.F.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1725–1732 (2014) 24

8. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433 (2015) 26

9. Brock, A., Jeff, D., Karen, S.: Large scale gan training for high fidelity natural image synthesis. In: International Conference on Learning Representations (2018) 11

10. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, pp. 1877–1901 (2020) 25

11. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks pp. 249–259 (2018) 4, 8, 9

12. Byrd, J., Lipton, Z.: What is the effect of importance weighting in deep learning? In: International Conference on Machine Learning, pp. 872–881. PMLR (2019) 8

13. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1209–1218 (2018) 24

14. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: Advances in Neural Information Processing Systems, pp. 1567–1578 (2019) 15, 22, 23

15. Castrup, H.: Distributions for uncertainty analysis. In: Proceedings of International Dimensional Workshop, pp. 1–12 (2001) 6

16. Chang, N., Koushik, J., Tarr, M.J., Hebert, M., Wang, Y.X.: Alpha net: Adaptation with composition in classifier space. arXiv preprint arXiv:2008.07073 (2020) 19

17. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research pp. 321–357 (2002) 10, 19

18. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020) 26

19. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) 26

20. Cheng, B., Schwing, A.G., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. arXiv preprint arXiv:2107.06278 (2021) 25

21. Chou, H.P., Chang, S.C., Pan, J.Y., Wei, W., Juan, D.C.: Remix: Rebalanced mixup. In: Proceedings of the European Conference on Computer Vision, pp. 95–110 (2020) 10, 19, 20, 21, 22

22. Chu, P., Bian, X., Liu, S., Ling, H.: Feature space augmentation for long-tailed data. In: Proceedings of the European Conference on Computer Vision, pp. 694–710 (2020) 11, 19, 20, 21, 22

23. Contributors, M.: Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation (2020) 25

24. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 702–703 (2020) 15, 21, 22

25. Cui, J., Liu, S., Tian, Z., Jia, J.: Reslt: Residual learning for long-tailed recognition. arXiv preprint arXiv:2101.10633 (2021) 10, 12, 21, 22

26. Cui, J., Zhong, Z., Liu, S., Yu, B., Jia, J.: Parametric contrastive learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 715–724 (2021) 15, 19, 21, 22

27. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9268–9277 (2019) 5, 6, 7, 10, 11, 12, 19, 20, 22, 23

28. Dave, A., Dollár, P., Ramanan, D., Kirillov, A., Girshick, R.: Evaluating large-vocabulary object detectors: The devil is in the details. arXiv preprint arXiv:2102.01066 (2021) 4

29. David, A., Hartley, O., Pearson, S.: The distribution of the ratio, in a single normal sample, of range to standard deviation. Biometrika pp. 482–493 (1954) 6

30. Davidson, L.: Uncertainty in economics. In: Uncertainty, International Money, Employment and Theory, pp. 30–37 (1999) 6

31. Delmas, R., Yan, L.: Exploring students' conceptions of the standard deviation. Statistics Education Research Journal pp. 55–82 (2005) 6

32. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009) 5, 6, 7, 24

33. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019) 5, 16

34. Devi, D., Purkayastha, B., et al.: Redundancy-driven modified tomek-link based undersampling: A solution to class imbalance. Pattern Recognition Letters pp. 3–12 (2017) 10, 19

35. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186 (2019) 1, 25

36. Dina, G., Michael, J., David, H., Julio, D., Robert, S.: Decreasing median age of covid-19 cases in the united states—changing epidemiology or changing surveillance? PLoS One (2020) 6

37. Dong, Q., Gong, S., Zhu, X.: Class rectification hard mining for imbalanced deep learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1851–1860 (2017) 12

38. Dong, Q., Gong, S., Zhu, X.: Imbalanced deep learning by minority class incremental rectification. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1367–1381 (2018) 15, 19

39. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021) 26

40. Dvir, S., Gal, C.: Distributional robustness loss for long-tail learning. arXiv preprint arXiv:2104.03066 (2021) 19, 20, 21, 22

41. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision pp. 303–338 (2010) 2

42. Fan, Q., Zhuo, W., Tang, C.K., Tai, Y.W.: Few-shot object detection with attention-rpn and multi-relation detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4013–4022 (2020) 24

43. Fogarty, A., Richard, H., John, B.: International comparison of median age at death from cystic fibrosis. Chest pp. 1656–1660 (2000) 6

44. Ghosh, M., Nangia, N., Kim, D.H.: Estimation of median income of four-person families: a bayesian time series approach. Journal of the American Statistical Association pp. 1423–1431 (1996) 6

45. Gidaris, S., Komodakis, N.: Dynamic few-shot visual learning without forgetting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4367–4375 (2018) 10

46. Gini, C.: Variabilità e mutabilità. Memorie di metodologica statistica (1912) 2, 7

47. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015) 1

48. Goodfellow, I., andd Mehdi Mirza, J.P.A., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (2014) 11

49. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Heuna, K., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5842–5850 (2017) 24

50. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Zero-shot detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021) 26

51. Gui, S., Wang, H., Yang, H., Wang, C.Y.Z., Liu., J.: Model compression with adversarial robustness: A unified optimization framework. In: Advances in Neural Information Processing Systems, pp. 1283–1294 (2019) 26

52. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: Proceedings of the European Conference on Computer Vision, pp. 87–102 (2016) 3, 5, 24

53. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5356–5364 (2019) 2, 5, 6, 7, 8, 19, 22, 23

54. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2006) 26

55. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications pp. 220–239 (2017) 8, 9

56. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: A new over-sampling method in imbalanced data sets learning. In: International Conference on Intelligent Computing, pp. 878–887. Springer (2005) 10, 19

57. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks, pp. 1322–1328 (2008) 10, 11, 19

58. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering pp. 1263–1284 (2009) 9, 11

59. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020) 26

60. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016) 3, 20, 21, 22, 23

61. He, Y.Y., Wu, J., Wei, X.S.: Distilling virtual examples for long-tailed recognition. arXiv preprint arXiv:2103.15042 (2021) 18, 19, 20, 21, 22

62. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 17

63. Hong, Y., Han, S., Choi, K., Seo, S., Kim, B., Chang, B.: Disentangling label distribution for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6626–6636 (2021) 19, 20, 21, 22

64. Hsieh, T.I., Robb, E., Chen, H.T., Huang, J.B.: Droploss for long-tail instance segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1549–1557 (2021) 12, 19, 23

65. Hu, X., Jiang, Y., Tang, K., Chen, J., Miao, C., Zhang, H.: Learning to segment the tail. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14045–14054 (2020) 9, 16, 18, 19, 23

66. Huang, C., Li, Y., Loy, C.C., Tang, X.: Learning deep representation for imbalanced classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5375–5384 (2016) 11, 13, 19

67. Huang, C., Li, Y., Loy, C.C., Tang, X.: Deep imbalanced learning for face recognition and attribute prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 2781–2794 (2019) 13

68. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation pp. 79–87 (1991) 17

69. Jamal, M.A., Brown, M., Yang, M.H., Wang, L., Gong, B.: Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7610–7619 (2020) 12, 16, 20, 22

70. Janowczyk, A., Madabhushi, A.: Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of Pathology Informatics (2016) 8

71. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent data analysis pp. 429–449 (2002) 9

72. Jiang, H., Misra, I., Rohrbach, M., Learned-Miller, E., Chen, X.: In defense of grid features for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020) 26

73. Jordan, M.I., Jacobs, R.A.: Hierarchical mixtures of experts and the em algorithm. Neural computation pp. 181–214 (1994) 17

74. Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey. Journal of Artificial Intelligence Research pp. 237–285 (1996) 16

75. Kahn, H., Marshall, A.W.: Methods of reducing sample size in monte carlo computations. Journal of the Operations Research Society of America pp. 263–278 (1953) 11

76. Kakwani, N.C.: Applications of lorenz curves in economic analysis. Econometrica: Journal of the Econometric Society pp. 719–727 (1977) 7

77. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: International Conference on Learning Representations (2020) 13, 19, 21, 22

78. Karras, T., Samuli, L., Timo, A.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019) 11

79. Kim, D.J., Sun, X., Choi, J., Lin, S., Kweon, I.S.: Detecting human-object interactions with action co-occurrence priors. In: Proceedings of the European Conference on Computer Vision, pp. 718–736 (2020) 1

80. Kim, J., Jeong, J., Shin, J.: M2m: Imbalanced classification via major-to-minor translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13896–13905 (2020) 11, 19

81. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 11

82. Kirillov, A., Girshick, R., He, K., Dollar, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6399–6408 (2019) 25

83. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li-Jia Li, D.A.S.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision pp. 32–73 (2017) 2, 26

84. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech Report (2009) 5, 6, 7

85. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4. International Journal of Computer Vision pp. 1–26 (2020) 24

86. Lample, G., Ott, M., Conneau, A., Denoyer, L., Ranzato, M.: Phrase-based & neural unsupervised machine translation. arXiv preprint arXiv:1804.07755 (2018) 1

87. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019) 1

88. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 34–42 (2015) 8

89. Li, B., Liu, Y., Wang, X.: Gradient harmonized single-stage detector. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8577–8584 (2019) 13, 19

90. Li, B., Yao, Y., Tan, J., Zhang, G., Yu, F., Lu, J., Luo, Y.: Equalized focal loss for dense long-tailed object detection. arXiv preprint arXiv:2201.02593 (2022) 12, 19, 22, 23

91. Li, J., Tang, S., Li, J., Xiao, J., Wu, F., Pu, S., Zhuang, Y.: Topic adaptation and prototype encoding for few-shot visual storytelling. In: Proceedings of the ACM International Conference on Multimedia, pp. 4208–4216 (2020) 26

92. Li, S., Gong, K., Liu, C.H., Wang, Y., Qiao, F., Cheng, X.: Metasaug: Meta semantic augmentation for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5212–5221 (2021) 17, 19, 20, 21, 22

93. Li, T., Cao, P., Yuan, Y., Fan, L., Yang, Y., Feris, R., Indyk, P., Katabi, D.: Targeted supervised contrastive learning for long-tailed recognition. arXiv preprint arXiv:2111.13998 (2021) 19, 20, 21, 22

94. Li, T., Wang, L., Wu, G.: Self supervision to distillation for long-tailed visual recognition. arXiv preprint arXiv:2109.04075 (2021) 26

95. Li, X., Wei, T., Chen, Y.P., Tai, Y.W., Tang, C.K.: Fss-1000: A 1000-class dataset for few-shot segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2869–2878 (2020) 24

96. Li, Y., Wang, T., Kang, B., Tang, S., Wang, C., Li, J., Feng, J.: Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10991–11000 (2020) 13, 14, 18, 19, 23

97. Li, Z., Dekel, T., Cole, F., Tucker, R., Snavely, N., Liu, C., Freeman, W.T.: Learning the depths of moving people by watching frozen people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4521–4530 (2019) 1

98. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017) 12, 13, 19, 20, 21, 23, 25

99. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision, pp. 740–755 (2014) 2, 6, 7, 24

100. Liu, B., Li, H., Kang, H., Hua, G.: Gistet: A geometric structure transfer network for long-tailed recognition. arXiv preprint arXiv:2105.00131 (2021) 16, 19, 21, 22

101. Liu, B., Li, H., Kang, H., Hua, G., Vasconcelos, N.: Breadcrumbs: Adversarial class-balanced sampling for long-tailed recognition. arXiv preprint arXiv:2105.00127 (2021) 11, 21, 22

102. Liu, J., Sun, Y., Han, C., Dou, Z., Li, W.: Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2970–2979 (2020) 16, 19

103. Liu, J., Zhang, J., Li, W., Zhang, C., Sun, Y.: Memory-based jitter: Improving visual recognition on long-tailed data with diversity in memory. arXiv preprint arXiv:2008.09809 (2020) 20, 21, 22

104. Liu, T.Y.: Learning to rank for information retrieval (2011) 15, 19

105. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Proceedings of the European Conference on Computer Vision, pp. 21–37 (2016) 3

106. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) pp. 539–550 (2008) 9, 19

107. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021) 26

108. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2537–2546 (2019) 2, 5, 6, 7, 8, 19, 21

109. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018) 26

110. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European Conference on Computer Vision, pp. 181–196 (2018) 8, 12, 19

111. Mani, I., Zhang, I.: knn approach to unbalanced data distributions: a case study involving information extraction. In: Proceedings of Workshop on Learning From Imbalanced Datasets, vol. 126. ICML United States (2003) 10

112. Masoudnia, S., Ebrahimpour, R.: Mixture of experts: a literature survey. Artificial Intelligence Review pp. 275–293 (2014) 17

113. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. In: International Conference on Learning Representations (2021) 21, 22

114. Miao, J., Wei, Y., Wu, Y., Liang, C., Li, G., Yang, Y.: Vspw: A large-scale dataset for video scene parsing in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4133–4143 (2021) 24

115. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546 (2013) 12, 19

116. Narayanan, A., Chen, Y.T., Malla, S.: Semi-supervised learning: Fusion of self-supervised, supervised learning, and multimodal cues for tactical driver behavior detection. arXiv preprint arXiv:1807.00864 (2018) 26

117. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4004–4012 (2016) 14

118. Oksuz, K., Cam, B.C., Kalkan, S., Akbas, E.: Imbalance problems in object detection: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 4

119. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems (2017) 11

120. Ouyang, W., Wang, X., Zhang, C., Yang, X.: Factors in finetuning deep model for object detection with long-tail distribution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 864–873 (2016) 9

121. Peng, J., Bu, X., Sun, M., Zhang, Z., Tan, T., Yan, J.: Large-scale object detection in the wild from imbalanced multi-labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9709–9718 (2020) 8, 12, 19

122. Peng, Z., Huang, W., Guo, Z., Zhang, X., Jiao, J., Ye, Q.: Long-tailed distribution adaptation. In: Proceedings of the ACM International Conference on Multimedia, pp. 3275–3282 (2021) 20, 21, 22, 23

123. Prabhu, V., Kannan, A., Ravuri, M., Chablani, M., Sontag, D., Amatriain, X.: Prototypical clustering networks for dermatological disease diagnosis. arXiv preprint arXiv:1811.03066 (2018) 26

124. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021) 26

125. Ramanathan, V., Wang, R., Mahajan, D.: Dlwl: Improving detection for lowshot classes with weakly labelled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9342–9352 (2020) 18, 19

126. Ren, J., Yu, C., Sheng, S., Ma, X., Zhao, H., Yi, S., Li, H.: Balanced meta-softmax for long-tailed visual recognition. In: Advances in Neural Information Processing Systems (2020) 16, 19, 20, 21, 23

127. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. In: International Conference on Machine Learning, pp. 4334–4343. PMLR (2018) 16, 19, 20

128. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28, pp. 91–99 (2015) 3, 25

129. Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Pinto, A.S., Keysers, D., Houlsby, N.: Scaling vision with sparse mixture of experts. arXiv preprint arXiv:2106.05974 (2021) 26

130. Ristani, E., Solera, F., Zou, R.S., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Proceedings of the European Conference on Computer Vision, pp. 17–35 (2016) 24

131. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision pp. 211–252 (2015) 2

132. Shaham, T.R., Dekel, T., Michaeli, T.: Singan: Learning a generative model from a single natural image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4570–4580 (2019) 1

133. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8430–8439 (2019) 3, 24, 25

134. Shen, L., Lin, Z., Huang, Q.: Relay backpropagation for effective learning of deep convolutional neural networks. In: Proceedings of the European Conference on Computer Vision, pp. 467–482 (2016) 8, 19, 23

135. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 761–769 (2016) 13, 19

136. Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., Meng, D.: Meta-weight-net: Learning an explicit mapping for sample weighting. In: Advances in Neural Information Processing Systems, vol. 32, pp. 1919–1930 (2019) 16, 17, 19, 20

137. Shu, X., Wang, X., Zang, X., Zhang, S., Chen, Y., Li, G., Tian, Q.: Large-scale spatio-temporal person re-identification: Algorithm and benchmark. arXiv preprint arXiv:2105.15076 (2021) 24

138. Simard, P.Y., LeCun, Y.A., Denker, J.S., Victorri, B.: Transformation invariance in pattern recognition—tangent distance and tangent propagation. In: Neural Networks: Tricks of the Trade, pp. 239–274. Springer (1998) 10

139. Sinha, S., Ohashi, H., Nakamura, K.: Class-wise difficulty-balanced loss for solving class-imbalance. In: Proceedings of the Asian Conference on Computer Vision (2020) 10, 12, 13, 14, 20, 21

140. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: Advances in Neural Information Processing Systems, pp. 1857–1865 (2016) 14

141. van Steenkiste, S., Greff, K., Schmidhuber, J.: A perspective on objects and systematic generalization in model-based rl. arXiv preprint arXiv:1906.01035 (2019) 1

142. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018) 16

143. Tan, J., Lu, X., Zhang, G., Yin, C., Li, Q.: Equalization loss v2: A nnew gradient balance approach for long-tailed object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1685–1694 (2021) 12, 13, 19, 22, 23

144. Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., Yan, J.: Equalization loss for long-tailed object recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11662–11671 (2020) 10, 12, 19, 20, 21, 23

145. Tang, K., Huang, J., Zhang, H.: Long-tailed classification by keeping the good and removing the bad momentum causal effect. In: Advances in Neural Information Processing Systems (2020) 19, 20, 21

146. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM pp. 64–73 (2016) 18, 26

147. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9627–9636 (2019) 3, 25

148. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8769–8778 (2018) 5, 6, 7

149. Van Horn, G., Perona, P.: The devil is in the tails: Fine-grained classification in the wild. arXiv preprint arXiv:1709.01450 (2017) 2

150. Wang, C., Gao, S., Wang, P., Gao, G., Pei, W., Pan, L., Xu, Z.: Label-aware distribution calibration for long-tailed classification. arXiv preprint arXiv:2111.04901 (2021) 14, 19, 20, 21

151. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5265–5274 (2018) 16

152. Wang, H., Xiao, C., Kossaifi, J., Yu, Z., Anandkumar, A., Wang, Z.: Augmax: Adversarial composition of random augmentations for robust training. In: Advances in Neural Information Processing Systems (2021) 26

153. Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C.C., Lin, D.: Seesaw loss for long-tailed instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9695–9704 (2021) 12, 19, 21, 22, 23

154. Wang, K.J., Makond, B., Chen, K.H., Wang, K.M.: A hybrid classifier combining smote with pso to estimate 5-year survivability of breast cancer patients. Applied Soft Computing pp. 15–24 (2014) 10

155. Wang, P., Han, K., Wei, X.S., Zhang, L., Wang, L.: Contrastive learning based hybrid networks for long-tailed image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 943–952 (2021) 15, 20, 22

156. Wang, R., Hu, K., Zhu, Y., Shu, J., Zhao, Q., Meng, D.: Meta feature modulator for long-tailed recognition. arXiv preprint arXiv:2008.03428 (2020) 17, 19

157. Wang, T., Li, Y., Kang, B., Li, J., Liew, J., Tang, S., Hoi, S., Feng, J.: The devil is in classification: A simple framework for long-tail instance segmentation. In: Proceedings of the European Conference on Computer Vision, pp. 728–744 (2020) 14, 19, 23

158. Wang, T., Zhu, Y., Zhao, C., Zeng, W., Wang, J., Tang, M.: Adaptive class suppression loss for long-tail object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3103–3112 (2021) 12, 23

159. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. In: Advances in Neural Information Processing Systems, pp. 1152–1164 (2018) 1

160. Wang, X., Lian, L., Miao, Z., Liu, Z., Yu, S.X.: Long-tailed recognition by routing diverse distribution-aware experts. In: International Conference on Learning Representations (2021) 17, 19, 20, 21, 22

161. Wang, Y., Gan, W., Yang, J., Wu, W., Yan, J.: Dynamic curriculum learning for imbalanced data classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5017–5026 (2019) 15, 19

162. Wang, Y., Yao, Q., Kwok, J., Ni, L.: Few-shot learning: A survey. arXiv preprint arXiv:1904.05046 (2019) 4

163. Wang, Y., Zhang, B., Hou, W., Wu, Z., Wang, J., Shinozaki, T.: Margin calibration for long-tailed visual recognition. arXiv preprint arXiv:2112.07225 (2021) 15, 19, 20, 21, 22

164. Wang, Y.X., Ramanan, D., Hebert, M.: Learning to model the tail. In: Advances in Neural Information Processing Systems, pp. 7029–7039 (2017) 3, 11, 15, 16, 19

165. Wei, C., Sohn, K., Mellina, C., Yuille, A., Yang, F.: Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10857–10866 (2021) 18, 19

166. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2575–2584 (2020) 3, 24

167. Wightman, R., Touvron, H., Jegou, H.: Resnet strikes back: An improved training procedure in timm. arXiv:2110.00476 (2021) 21

168. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics pp. 408–421 (1972) 10, 19

169. Wu, J., Song, L., Wang, T., Zhang, Q., Yuan, J.: Forest r-cnn: Large-vocabulary long-tailed object detection and instance segmentation. In: Proceedings of the ACM International Conference on Multimedia, pp. 1570–1578 (2020) 9, 18, 19, 23

170. Wu, T., Huang, Q., Liu, Z., Wang, Y., Lin, D.: Distribution-balanced loss for multi-label classification in long-tailed datasets. In: Proceedings of the European Conference on Computer Vision, pp. 162–178 (2020) 12, 19

171. Wu, T., Liu, Z., Huang, Q., Wang, Y., Lin, D.: Adversarial robustness under long-tailed distribution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8659–8668 (2021) 19

172. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019) 25

173. Xiang, L., Ding, G., Han, J.: Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In: Proceedings of the European Conference on Computer Vision, pp. 247–263 (2020) 17, 18, 19, 20, 21

174. Yang, L., Song, Q., Wu, Y.: Attacks on state-of-the-art face recognition using attentional adversarial attack generative network. Multimedia Tools and Applications (2021) 26

175. Yang, Y., Xu, Z.: Rethinking the value of labels for improving class-imbalanced learning. In: Advances in Neural Information Processing Systems (2020) 18, 19, 26

176. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems, pp. 5753–5763 (2019) 25

177. Yaoyao, Z., Weihong, D.: Adversarial learning with margin-based triplet embedding regularization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019) 26

178. Yitzhaki, S., Schechtman, E.: More than a dozen alternative ways of spelling gini. The Gini Methodology pp. 11–31 (2013) 7

179. Yu, W., Yang, T., Chen, C.: Towards resolving the challenge of long-tail distribution in uav images for object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3258–3267 (2021) 26

180. Zang, Y., Huang, C., Loy, C.C.: Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. arXiv preprint arXiv:2102.12867 (2021) 11, 19, 23

181. Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q., Lin, S.: Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In: Proceedings of the European Conference on Computer Vision, pp. 507–523 (2020) 26

182. Zhang, C., Pan, T.Y., Li, Y., Hu, H., Xuan, D., Changpinyo, S., Gong, B., Chao, W.L.: A simple and effective use of object-centric images for long-tailed object detection. arXiv preprint arXiv:2102.08884 (2021) 18, 19, 23

183. Zhang, G., Lu, X., Tan, J., Li, J., Zhang, Z., Li, Q., Hu, X.: Refinemask: Towards high-quality instance segmentation with fine-grained features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6861–6869 (2021) 4

184. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018) 10

185. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. arXiv preprint arXiv:2101.00529 (2021) 26

186. Zhang, S., Chen, C., Hu, X., Peng, S.: Balanced knowledge distillation for long-tailed learning. arXiv preprint arXiv:2104.10510 (2021) 18, 19, 20, 21, 22

187. Zhang, S., Li, Z., Yan, S., He, X., Sun, J.: Distribution alignment: A unified framework for long-tail visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2361–2370 (2021) 14, 19, 21, 22, 23, 25

188. Zhang, X., Fang, Z., Wen, Y., Li, Z., Qiao, Y.: Range loss for deep face recognition with long-tailed training data. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5409–5418 (2017) 15, 19, 21

189. Zhang, Y., Cheng, D.Z., Yao, T., Yi, X., Hong, L., Chi, E.H.: A model of two tales: Dual transfer learning framework for improved long-tail item recommendation. In: Proceedings of the Web Conference 2021, pp. 2220–2231 (2021) 3

190. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. arXiv preprint arXiv:2110.04596 (2021) 2

191. Zhang, Y., Wei, X.S., Zhou, B., Wu, J.: Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3447–3455 (2021) 2, 14, 20, 21, 22

192. Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., Zha, Z.J.: Object relational graph with teacher-recommended learning for video captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13278–13288 (2020) 26

193. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017) 25

194. Zhao, J., Li, J., Cheng, Y., Zhou, L., Sim, T., Yan, S., Feng, J.: Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In: Proceedings of the ACM International Conference on Multimedia, pp. 792–800 (2018) 3, 24

195. Zhao, Y., Chen, W., Tan, X., Huang, K., Xu, J., Wang, C., Zhu, J.: Improving long-tailed classification from instance level. arXiv preprint arXiv:2104.06094 (2021) 13, 17, 19, 21, 22

196. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124 (2015) 3, 24

197. Zhong, Z., Cui, J., Liu, S., Jia, J.: Improving calibration for long-tailed recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16489–16498 (2021) 14, 19, 20, 21, 22

198. Zhou, B., Cui, Q., Wei, X.S., Chen, Z.M.: Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9719–9728 (2020) 14, 19, 20, 22

199. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929 (2016) 11

200. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1452–1464 (2017) 2, 5, 6, 7, 24

201. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 633–641 (2017) 3, 24, 25

202. Zhou, X., Koltun, V., Krähenbühl, P.: Probabilistic two-stage detection. arXiv preprint arXiv:2103.07461 (2021) 12, 23

203. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European Conference on Computer Vision, pp. 289–305 (2018) 18