Towards Heterogeneous Long-tailed Learning: Benchmarking, Metrics, and Toolbox

Haohui Wang, **Weijie Guan**, **Jianpeng Chen**, **Zi Wang**, **Dawei Zhou**

Computer Science, Virginia Tech

`{haohuiw,skjguan,jianpengc,ziwang,zhoud}@vt.edu`

Abstract

Long-tailed data distributions pose challenges for a variety of domains like e-commerce, finance, biomedical science, and cyber security, where the performance of machine learning models is often dominated by head categories while tail categories are inadequately learned. This work aims to provide a systematic view of long-tailed learning with regard to three pivotal angles: (A1) the characterization of data long-tailedness, (A2) the data complexity of various domains, and (A3) the heterogeneity of emerging tasks. We develop HEROLT, a comprehensive long-tailed learning benchmark integrating 18 state-of-the-art algorithms, 10 evaluation metrics, and 17 real-world datasets across 6 tasks and 4 data modalities. HEROLT with novel angles and extensive experiments (315 in total) enables effective and fair evaluation of newly proposed methods compared with existing baselines on varying dataset types. Finally, we conclude by highlighting the significant applications of long-tailed learning and identifying several promising future directions. For accessibility and reproducibility, we open-source our benchmark HEROLT and corresponding results at `https://github.com/SSSKJ/HeroLT`.

1 Introduction

In the era of big data, many high-impact domains, such as e-commerce [1, 2], finance [3], biomedical science [4, 5], and cyber security [6, 7], naturally exhibit long-tailed data distributions, where a few head categories[1] are well-studied with abundant data, while massive tail categories are under-explored with scarce data. To name a few, in financial transaction networks, the majority of transactions fall into a few head classes that are considered normal, like wire transfers and credit card payments. However, a large number of tail classes correspond to various fraudulent transaction types, like synthetic identity transactions and money laundering. Although fraudulent transactions rarely occur, detecting them is essential for preventing unexpected financial loss [8, 9]. Another example is antibiotic resistance genes, which can be classified based on the antibiotic class they confer to and their transferable ability. Genes with mobility and strong human pathogenicity may not be detected frequently and are viewed as tail classes, but these resistance genes have the potential to be transmitted from the environment to bacteria in humans, thereby posing an increasing global threat to human health [10, 11].

Massive long-tailed learning studies have been conducted in recent years, proposing methods like the cost-sensitive focal loss to effectively address the data imbalance by reshaping the standard cross-entropy loss [12], and graph augmentation by interpolating tail node embeddings and generating new edges [13]. The advancements in tackling long-tailed problems drove the publication of several studies on the survey of long-tailed problems [14, 15, 16, 17, 18], which generally categorize existing

---

[1]Long-tailed problems occur in labels or input data (like degrees of nodes), collectively named as category.

[13, 30, 31, 32] **Node Classification** [1, 35, 36, 37, 38] **Recommendation** [40, 41, 42, 43, 26, 44, 45] **Image Classification** [49, 50, 51] **Multi-Label Text Classification**

A3. Task Heterogeneity [33, 34] **Entity Alignment** [39] **Classification** [46, 47, 48] **Instance Segmentation** [52] **Information Extraction**

Tabular Data **Grid Data** **Sequential Data**

A2. Data Complexity [39, 58, 59] [60, 61] **Relational Data** [41, 62, 63, 64] [65, 66, 67] [68, 69]

[53, 54, 55] [56, 57]

A1. Long-Tailedness **Data Imbalance** **Extreme # of Categories**

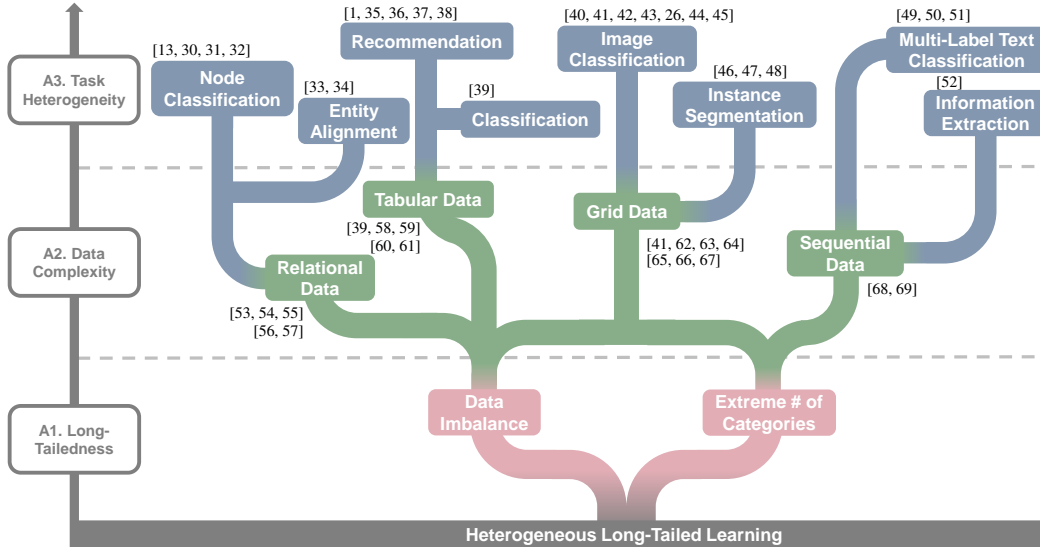**Heterogeneous Long-Tailed Learning**

Figure 1: The systematic view of heterogeneous long-tailed learning concerning three pivotal angles, including long-tailedness (colored in red), data complexity (green), and task heterogeneity (blue).

works by different types of learning algorithms (*e.g.*, re-sampling [19, 20], cost-sensitive [21, 22], transfer learning [23, 24], decoupled training [25, 26], and meta learning [27, 28, 29]).

Despite tremendous progress, some pivotal questions largely remain unresolved, *e.g.*, *how can we characterize the extent of long-tailedness of given data? how do long-tailed algorithms perform with regard to different tasks on different domains?* To fill the gap, this work aims to provide a systematic view of long-tailed learning with regard to three pivotal angles in Figure 1: **(A1) the characterization of data long-tailedness:** long-tailed data exhibits a highly skewed data distribution and an extensive number of categories; **(A2) the data complexity of various domains:** a wide range of complex domains may naturally encounter long-tailed distribution, *e.g.*, tabular data, sequential data, grid data, and relational data; and **(A3) the heterogeneity of emerging tasks:** it highlights the need to consider the applicability and limitations of existing methods on heterogeneous tasks. With three major angles in long-tailed learning, we design extensive experiments that conduct 18 state-of-the-art algorithms and 10 evaluation metrics on 17 real-world benchmark datasets across 6 tasks and 4 data modalities.

Key Takeaways: Through extensive experiments (see Section 3), we find (1) most works mainly focus on data imbalance while paying less attention to the extreme number of categories in the long-tailed distribution; (2) surprisingly none of the algorithms statistically outperforms others across all tasks and domains, emphasizing the importance of algorithm selection in terms of scenarios.

In general, we summarize the main contributions of HEROLT as below:

- **Comprehensive Benchmark.** We conduct a comprehensive review and examine long-tailed learning concerning three pivotal angles: (A1) the characterization of data long-tailedness, (A2) the data complexity of various domains, and (A3) the heterogeneity of emerging tasks.

- **Insights and Future Directions.** With comprehensive results, our study highlights the importance of characterizing the extent of long-tailedness and algorithm selection while identifying open problems and opportunities to facilitate future research.

- **The Open-Sourced Toolbox.** We provide a fair and accessible performance evaluation of 18 state-of-the-art methods on multiple benchmark datasets using accuracy-based and ranking-based evaluation metrics at `https://github.com/SSSKJ/HeroLT`.

2 HeroLT: Benchmarking Heterogeneous Long-Tailed Learning

2.1 Preliminaries and Problem Definition

Here we provide a general definition of long-tailed learning as follows. Given a long-tailed dataset $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of n samples from C categories, let $\mathcal{Y}$ denote the label set of category, where a
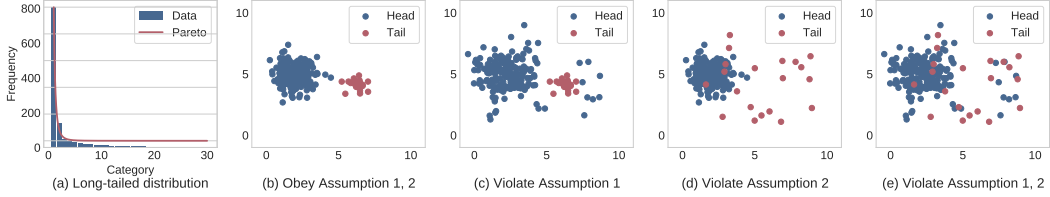
Figure 2: Illustrative figures of a synthetic long-tailed distributed data. (a) long-tailed distribution of categories. (b) Visualization of obeying of Assumption 1, 2. (c) Visualization of violating of Assumption 1. (d) Visualization of violating of Assumption 2. (e) Visualization of violating of Assumption 1, 2.

sample $\mathbf{x}_i$ may belong to one or more categories. For example, in image classification, one sample belongs to one category; in document classification, one document is often associated with multiple categories (*i.e.*, topics). For simplicity, we represent $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_C\}$, where $\mathcal{D}_c, c = 1, \ldots, C$ denotes the subset of samples belong to category c, and the size of each category $n_c = |\mathcal{D}_c|$ following a descending order. As an instantiation, a synthetic long-tailed dataset is given in Figure 2(a). Categories with massive samples refer to head categories, while categories with only a few samples refer to tail categories. Without the loss of generality, we make the following assumptions in long-tailed learning regarding the decision region of head and tail categories on the embedding space.

Assumption 1 (Smoothness Assumption for Head Category). *Given a long-tailed dataset, the distribution of the decision region of each head category is sufficiently smooth.*

Assumption 2 (Compactness Assumption for Tail Category). *Given a long-tailed dataset, the tail category samples can be represented as a compact cluster in the feature space.*

These assumptions ensure that tail categories are identifiable and meaningful. If Assumption 1 of smoothness is violated, then it is challenging to identify the head category clearly. For example in Figure 2(c), it is difficult to depict the decision region of the head category. Similarly, if Assumption 2 of compactness is violated, as seen in Figure 2(d), it is difficult to determine whether it is noise data or data from tail category. Based on the assumptions, we give the definition of long-tailed learning.

Problem 1. *Long-Tailed Learning.*
Given: a training set $\mathcal{D}$ of n samples from C distinct categories following long-tailed distribution, and the label set of category $\mathcal{Y}$. The data follows long-tailed distribution, i.e., the frequency of the categories is approximated as $\lim_{y \to \infty} e^{ty} P(Y > y) = \infty, \forall t > 0$, where Y is a random variable.
Find: a function $f : \mathcal{X} \to \mathcal{Y}$ that gives accurate label predictions on both head and tail categories.

2.2 Benchmark Angles in HEROLT

Angle 1: Long-Tailedness in Terms of Data Imbalance and Extreme Number of Categories.
While extensive public datasets and algorithms are available for benchmarking, these datasets or algorithms often exhibit varying extents of long-tailedness or focus on specific characteristics of long-tailed problems, making it challenging to select appropriate datasets and baselines to evaluate new algorithms. Two key properties of long-tailed data lie in highly skewed data distribution and an extreme number of categories. The former introduces a significant difference in sample sizes between head and tail categories, resulting in a bias towards the head categories [17], while the latter poses challenges in learning classification boundaries due to the increasing number of categories [70, 71].

To measure the long-tailedness of a dataset, we introduce three metrics in Table 1. Firstly, a commonly used metric is imbalance factor (IF) [72], and a value closer to 1 means a more balanced dataset. Secondly, Gini coefficient can be used as a metric to quantify long-tailedness [18]. It ranges from 0 to 1, where a smaller value indicates a more balanced dataset. IF quantifies data imbalance between the most majority and the most minority categories, while Gini coefficient measures overall imbalance, unaffected by extreme samples or absolute data size. However, the two metrics pay more attention to data imbalance and may not reflect the number of categories. Therefore, Pareto-LT Ratio is proposed to jointly evaluate the two aspects of the data long-tailedness. Intuitively, the higher the skewness of the data distribution, the higher Pareto-LT Ratio may be; the more number of categories, the higher

3

Table 1: Three metrics for measuring long-tailedness of datasets. $Q(p) = min\{y : Pr(\mathcal{Y} \leq y) = p, 1 \leq y \leq C\}$ is the quantile function of order $p \in (0, 1)$ for $\mathcal{Y}$.

Metric name	Computation	Description
Imbalance Factor [72]	n_1/n_C	Ratio to the size of largest and smallest categories.
Gini Coefficient [18]	$\frac{\sum_{i=1}^{C} \sum_{j=1}^{C} |n_i - n_j|}{2nC}$	Relative mean absolute differences between category sizes.
Pareto-LT Ratio	$\frac{C - Q(0.8)}{Q(0.8)}$	The number of categories of the last 20% samples to the number of categories of the rest 80% samples.

the value of Pareto-LT Ratio. In light of three long-tailedness metrics, we gain a better understanding of which datasets and baselines to consider when evaluating a newly proposed algorithm.

- When all three metrics have large values, it indicates a dataset with a severe long-tailed distribution, necessitating an algorithm that addresses both data imbalance and the extreme number of categories.

- If Gini coefficient and IF are relatively small, but Pareto-LT Ratio is large, the main challenges of the dataset may lie in massive categories, as exemplified by the experiments in Section 3.4. In such cases, methods like extreme classification [73, 74] and meta learning [27, 28, 29] would be preferred.

- When Gini coefficient and IF are large, but Pareto-LT Ratio is relatively small, it suggests that the challenges of data imbalance may be more significant than the number of categories. Algorithms employing techniques like re-sampling [19, 20] or re-weighting [21, 22] may already effectively handle data imbalance, as exemplified by the experiments in Section 3.5.

- If all three metrics are small, the extent of the long-tailedness of the dataset may be small, and ordinary machine learning methods may achieve decent performance.

In addition to the above metrics, we integrate the Bayes Imbalance Impact Index (BI3 [75]) and the Complementary Cumulative Distribution Function (CCDF) into our toolbox. While BI3 and CCDF are useful for certain application domains, *i.e.*, BI3 for binary classification and CCDF for visualizing label distributions, they are not specifically designed for long-tailed learning. Therefore, we mention them here briefly without providing a detailed analysis. We give a detailed evaluation of long-tailedness metrics (AppendixB.1), and show the long-tailedness of benchmark datasets (Table 2). We analyze whether long-tailed algorithms are specifically designed to address data imbalance and an extreme number of categories (AppendixB.2), and provide a comprehensive experimental analysis (Section3).

Angle 2: Data Complexity with 17 Datasets across 4 Data Modalities. Most existing long-tailed benchmarks mainly focus on image datasets [14, 15, 16, 17, 18]. However, in real-world applications, various types of data including tabular, grid, sequential, and relational data, face long-tailed problems. To fill this gap, we consider data complexity based on data types with long-tailed distribution.

- **Tabular data** comprises samples (rows) with the same set of features (columns) and is used in practical applications (*e.g.*, medicine, finance, and e-commerce). Specifically, Glass [76] is a tabular dataset with feature attributes, where the number of samples in different categories follows a long-tailed distribution. In addition to node connections, graph data such as Amazon [55] often include node attributes, which can be regarded as tabular data.

- **Sequential data** denotes that data points in the dataset are dependent on the other points. A common example is a time series such as S&P 500 Daily Changes [39], where the input (price over date) shows a long-tailed distribution. Another example of sequential data is text composed of manuscripts, messages, and reviews, such as Wikipedia [68] containing Wikipedia articles with the long-tailed distribution of Wikipedia categories.

- **Grid data** records regularly spaced samples over an area. Images can be viewed as grid data by mapping grid cells onto pixels one for one. The labels of images often exhibit long-tailed distribution, as observed in commonly used datasets (*e.g.*, ImageNet [41], iNatural [62], and LVIS [63]), remote sensing datasets [77], and 3D point cloud datasets [78]. Furthermore, HOI categories in HICO-DET [66] and V_COCO [67], which are designed for human-object interaction detection in images, follow the long-tailed distribution. Videos can be regarded as a combination of sequential and grid data. In INSVIDEO dataset[64] for micro-video, hashtags exhibit long-tailed distribution. While the NYU-Depth dataset [65], which is composed of video sequences as recorded by depth cameras, exhibits a long-tailed per-pixel depth.

Table 2: Datasets available in HEROLT benchmark. The values in this table correspond to input data.

Dataset	Data	Data Statistics			Long-Tailedness		
		# of Categories	Size	# of Edges	IF	Gini	Pareto
Glass [76]	Tabular	6	214	-	8	0.380	0.500
Abalone [79]	Tabular	28	4177	-	689	0.172	0.333
Housing [80]	Tabular	-	1,460	-	-	-	-
EURLEX-4K [68]	Sequential	3,956	15,499	-	1,024	0.342	3.968
AMAZONCat-13K [68]	Sequential	13,330	1,186,239	-	355,211	0.327	20.000
Wiki10-31K [68]	Sequential	30,938	14,146	-	11,411	0.312	4.115
ImageNet-LT [41]	Grid	1,000	115,846	-	256	0.517	1.339
Places-LT [41]	Grid	365	62,500	-	996	0.610	2.387
iNatural 2018 [62]	Grid	8,142	437,513	-	500	0.647	1.658
CIFAR 10-LT (100) [72]	Grid	10	12,406	-	100	0.617	1.751
CIFAR 10-LT (50) [72]	Grid	10	13,996	-	50	0.593	1.751
CIFAR 10-LT (10) [72]	Grid	10	20,431	-	10	0.520	0.833
CIFAR 100-LT (100) [72]	Grid	100	10,847	-	100	0.498	1.972
CIFAR 100-LT (50) [72]	Grid	100	12,608	-	50	0.488	1.590
CIFAR 100-LT (10) [72]	Grid	100	19,573	-	10	0.447	0.836
LVIS v0.5 [63]	Grid	1,231	693,958	-	26,148	0.381	6.250
Cora-Full [53]	Relational&Tabular	70	19,793	146,635	62	0.321	0.919
Wiki [81]	Relational	17	2,405	25,597	45	0.414	1.000
Email [82]	Relational&Tabular	42	1,005	25,934	109	0.413	1.263
Amazon-Clothing [55]	Relational&Tabular	77	24,919	208,279	10	0.343	0.814
Amazon-Electronics [55]	Relational&Tabular	167	42,318	129,430	9	0.329	0.600

- **Relational data** organizes data points with defined relationships. One specific type of relational data is represented by graphs, which consist of nodes connected by edges. Therefore, in addition to long-tailed distributions in node classes, node degrees, referring to the number of edges connected to a node, may exhibit a long-tailed distribution [31]. It is frequently observed that the majority of nodes have only a few connected edges, while only a small number of nodes have a large number of connected edges, as seen in datasets like Cora [53], CiteSeer [54], and Amazon [55]. Moreover, in knowledge graphs like YAGO [56] and DBpedia [57], the distribution of entities may exhibit long-tailed distribution, with only a few entities densely connected to others.

In HEROLT, we collect 17 datasets across 4 data modalities (tabular, sequential, grid, and relational data) as discussed in Appendix B.3. Table 2 shows data statistics (*e.g.*, size, #categories) and long-tailedness of these datasets.

Angle 3: Task Heterogeneity with 18 Algorithms on 6 Tasks. While visual recognition has long been recognized as a significant aspect of long-tailed problems, real-world applications involve different tasks with unique learning objectives, presenting unique challenges for long-tailed algorithms. Despite its importance, this crucial angle has not been well explored in existing benchmarks. We aim to benchmark long-tailed algorithms based on various tasks they are designed to solve to fill the gap.

- **Object recognition** [83, 84] assigns each sample to one label. Data imbalance usually occurs in binary classification or multi-class classification, while long-tailed problems focus on multi-class classification with a large number of categories.

- **Multi-label text classification** [49, 50, 51, 85] involves assigning the most relevant subset of class labels from an extremely large label set to each document. However, the extremely large label space often leads to significant data scarcity, particularly for rare labels in the tail. Consequently, many long-tailed algorithms are specifically designed to address the long-tailed problem in this task. Additionally, tasks such as sentence-level few-shot relation classification [86] and information extraction (containing three sub-tasks named relation extraction, entity recognition, and event detection) [52] are also frequently addressed by long-tailed algorithms.

- **Image classification** [26, 40, 42, 43, 44, 45], which involves assigning a label to an entire image, is a widely studied task in long-tailed learning that received extensive research attention. Furthermore, there are some long-tailed algorithms focusing on similar tasks, including out-of-distribution detection [87, 88], image retrieval [89, 90], image generation [91, 92], visual relationship recognition [93, 94, 95, 96] and video classification [64, 97].

- **Instance segmentation** [46, 47] is a common and crucial task that gained significant attention in the development of long-tailed algorithms aimed at enhancing the performance of tail classes. It involves the identification and separation of individual objects within an image, including the

Table 3: Algorithms available in the HEROLT benchmark.

Algorithm	Venue	Long-tailedness	Task
SMOTE [19]	02JAIR	Data imbalance	Object recognition
NearMiss [104]	03ICML	Data imbalance	Object recognition
X-Transformer [49]	20KDD	Data imbalance, extreme # of categories	Multi-label text classification
XR-Transformer [50]	21NeurIPS	Data imbalance, extreme # of categories	Multi-label text classification
XR-Linear [51]	22KDD	Data imbalance, extreme # of categories	Multi-label text classification
OLTR [41]	19CVPR	Data imbalance, extreme # of categories	Image classification
BALMS [46]	20NeurIPS	Data imbalance	Image classification, Instance segmentation
TDE [42]	20NeurIPS	Data imbalance	Image classification
Decoupling [26]	20ICLR	Data imbalance	Image classification
BBN [40]	20CVPR	Data imbalance	Image classification
MiSLAS [43]	21CVPR	Data imbalance	Image classification
PaCo [105]	21ICCV	Data imbalance, extreme # of categories	Image classification
GraphSMOTE [13]	21WSDM	Data imbalance	Node classification
ImGAGN [30]	21KDD	Data imbalance	Node classification
TailGNN [31]	21KDD	Data imbalance, extreme # of categories	Node classification
LTE4G [32]	22CIKM	Data imbalance, extreme # of categories	Node classification
SmoteR [102]	13EPIA	Data imbalance	Regression
SMOGN [103]	17PKDD/ECML	Data imbalance	Regression

detection of object boundaries and the assignment of a unique label to each object. It contains several parts: object detection [98, 99] and semantic segmentation [25].

- **Node classification** [100, 13, 30, 31, 32] involves assigning labels to nodes in a graph based on node features and connections between them. The emergence of long-tailed algorithms for this task is relatively recent, but the development is flourishing. Additionally, there are some long-tailed learning studies addressing entity alignment tasks [33, 34] in knowledge graphs.
- **Regression** [101, 102, 103] involves learning from long-tailed data with continuous (potentially infinite) target values. For example, the goal in [101] is to infer a person's age from their visual appearance, where age is a continuous target that can be highly imbalanced. Unlike classification, continuous targets lack clear category boundaries, causing difficulty when directly utilizing traditional methods like re-sampling and re-weighting. Additionally, continuous labels inherently possess a meaningful distance between targets.

In HEROLT, we have a comprehensive collection of algorithms for object recognition, multi-label text classification, image classification, instance segmentation, node classification, and regression tasks (Table 3). We discuss what technologies the algorithms use and how they solve long-tailed problems in Appendix B.2.

3 Experiment Results and Analyses

We conduct extensive experiments to further answer the question: how do long-tailed learning algorithms perform with regard to different tasks on different domains? In this section, we present the performance of 18 state-of-the-art algorithms on 6 typical long-tailed learning tasks and 17 real-world datasets across 4 data modalities.

3.1 Experiment Setting

Hyperparameter Settings. For all the 18 algorithms in HEROLT, we use the same hyperparameter settings on the same task for a fair comparison. Refer to the Appendix C.1 for more information. **Evaluation Metrics.** We evaluate different long-tailed algorithms by several basic metrics, which are divided into accuracy-based metrics including accuracy (Acc) [106], precision [107], recall [107], and balanced accuracy (bAcc) [107]; ranking-based metrics such as mean average precision (MAP) [108]; regression metrics such as mean-average-error (MAE), mean-squared-error (MSE), Pearson correlation and error geometric mean (GM); and running time. The computation formula and description of the metrics are in Table 10 in Appendix C.2.

3.2 Algorithm Performance on Object Recognition

Recently, there has been very limited work considering object recognition tasks on pure tabular long-tailed data. We compare the performance of two methods in Table 4. We find: **(1)** SMOTE

Table 4: Comparing the methods on long-tailed tabular datasets. Each point is the mean and standard deviation of 10 runs. "Time" means training time plus inference time.

Dataset	Method	Acc (%)	bAcc (%)	Precision (%)	Recall (%)	mAP (%)	Time (s)
Glass	SMOTE	**97.0±1.1**	**98.5±0.5**	**95.1±1.1**	**98.5±0.5**	**99.9±0.0**	0.6
	NearMiss	76.7±0.0	88.1±0.0	92.1±0.0	88.1±0.0	98.9±0.0	**0.4**
Abalone	SMOTE	20.1±1.1	**19.1±1.1**	**11.5±0.6**	**19.1±1.1**	17.4±0.2	12.4
	NearMiss	**23.4±0.0**	10.3±0.0	7.0±0.0	10.3±0.0	**18.8±0.0**	**2.8**

Table 5: Comparison of different methods on long-tailed sequential datasets for multi-label text classification tasks. "Time" refers to inference time. The results of bAcc@k are not reported since no relevant literature discusses how to use this metric in the multi-label setting.

Dataset	Method	Acc (%)			Precision (%)			Recall (%)			MAP (%)			Time (s)
		@1	@3	@5	@1	@3	@5	@1	@3	@5	@1	@3	@5	
Eurlex-4K	XR-Transformer	**88.2**	**75.8**	62.8	**88.2**	**75.8**	62.8	**17.9**	**45.2**	61.1	**88.2**	**82.0**	**75.6**	66.9
	X-Transformer	87.0	75.2	**62.9**	87.0	75.2	**62.9**	17.7	44.8	**61.2**	87.0	81.1	75.1	433.9
	XR-Linear	82.1	69.6	58.2	82.1	69.6	58.2	16.6	41.4	56.6	82.1	75.9	69.9	**0.2**
AmazonCat-13K	XR-Transformer	**96.7**	83.6	67.9	**96.7**	83.6	67.9	**27.7**	63.3	79.0	**96.7**	90.5	83.1	78.3
	X-Transformer	**96.7**	**83.9**	**68.6**	**96.7**	**83.9**	**68.6**	27.6	**63.4**	**79.7**	**96.7**	**90.6**	**83.4**	428.6
	XR-Linear	93.0	78.9	64.3	93.0	78.9	64.3	26.3	59.7	75.2	93.0	86.1	78.9	**0.3**
Wiki10-31K	XR-Transformer	88.0	**79.5**	**69.7**	88.0	**79.5**	**69.7**	**5.3**	**14.0**	**20.1**	88.0	**83.9**	**79.2**	117.3
	X-Transformer	**88.5**	78.5	69.1	**88.5**	78.5	69.1	**5.3**	13.8	19.8	**88.5**	83.6	78.7	433.2
	XR-Linear	84.6	73.0	64.3	84.6	73.0	64.3	5.0	12.7	18.4	84.6	78.7	73.7	**1.1**

(using an upsampling technique) shows superior performance to NearMiss (using a downsampling technique). Specifically, SMOTE is 85.43% highly balanced accuracy than NearMiss on the Abalone dataset. **(2)** While Acc treats all samples equally, bAcc considers data imbalance by averaging across classes. Although NearMiss achieves higher accuracy on the imbalanced Abalone dataset, it tends to exhibit a bias toward majority classes and does not sufficiently improve the minority classes, resulting in a lower bAcc score. Precision provides an insight into how much we can trust the model when it identifies a sample as Positive. Precision of NearMiss is lower as it may wrongly classify minority samples into other classes. Recall assesses the model's ability to identify all the Positive samples. NearMiss fails to characterize minority classes, it typically scores lower on Recall. Similarly, MAP is calculated by averaging across all classes and exhibits a similar trend with the other metrics.

3.3 Algorithm Performance on Multi-Label Text Classification

In Table 5, we provide a performance comparison of three methods for multi-label text classification tasks on sequential datasets. We find: **(1)** Among these methods, XR-Transformer and X-Transformer demonstrate comparably superior performance on multiple metrics. On Eurlex-4K, XR-Transformer achieves a 1.37% improvement in ACC@1 compared to the second-best method. On AmazonCat-13K, X-Transformer exhibits a 1.03% improvement in Acc@5 than the suboptimal method. **(2)** Conversely, XR-Linear consistently exhibits the poorest performance across all three datasets, which verifies the effectiveness of recursive learning of transformer encoders than linear methods. However, XR-Linear is more than 100x faster than XR-Transformer, making it suitable for scenarios with large dataset sizes and strict requirements on time-consuming. **(3)** Notably, all methods exhibit a limitation in effectively recognizing certain classes, as indicated by the observed low recalls across all datasets.

3.4 Algorithm Performance on Image Classification and Instance Segmentation

Among the considered algorithms, OLTR and BALMS utilize meta-learning; BBN and MisLAS employ mixup techniques; TDE uses causal inference; BALMS, Decoupling, and MisLAS decouple the learning process; and PaCo utilizes contrastive learning. We have the following observations from the results in Table 6: **(1)** No single technique (*e.g.*, meta-learning, decoupled training, mixup, or contrastive learning) can consistently perform the best across all tasks and datasets, and there are limited algorithms that consider both classification and segmentation tasks. Therefore, in contrast to taxonomy based on techniques in methods, the three novel angles we propose (data long-tailedness, data complexity, and task heterogeneity) may be more suitable for benchmarking. **(2)** PaCo and MisLAS consistently perform well in accuracy on three natural long-tailed datasets. In particular, Paco exhibits a remarkable overall accuracy of 58.3% on ImageNet-LT dataset, surpassing the

Table 6: Comparison of different methods on natural long-tailed grid datasets. "Time" refers to inference time. Recall equals to bAcc for multi-class classification.

Dataset	Task	Method	Acc (%)				Precision (%)	Recall/bAcc (%)	MAP (%)	Time (s)
			Many	Medium	Few	Overall				
ImageNet-LT	Image classification	OLTR	37.9	36.1	30.8	36.1	52.4	36.1	20.5	69.8
		BALMS	50.1	39.6	25.3	41.6	41.2	41.6	20.6	29.0
		TDE	60.5	47.2	30.4	50.1	50.1	51.0	28.7	30.9
		Decoupling	64.0	33.8	5.8	41.6	41.6	49.9	21.1	21.8
		BBN	59.4	45.4	16.3	46.6	47.5	46.6	24.9	43.0
		MiSLAS	60.9	46.8	32.5	50.0	39.0	38.0	21.0	**18.9**
		PaCo	**67.8**	**56.5**	**37.8**	**58.3**	**58.3**	**58.3**	**37.5**	58.9
Places-LT	Image classification	OLTR	**44.0**	40.6	28.5	39.3	39.5	39.3	17.9	30.6
		BALMS	41.0	39.9	30.2	38.3	38.1	38.3	17.1	29.0
		TDE	30.5	29.3	19.5	27.8	27.8	29.1	9.8	18.6
		Decoupling	40.6	39.1	28.6	37.6	37.6	38.2	16.7	28.9
		BBN	34.8	32.0	5.8	27.5	30.7	27.5	9.4	**15.4**
		MiSLAS	42.4	41.8	34.7	40.5	41.0	40.5	19.2	16.2
		PaCo	34.8	**48.1**	**38.4**	**41.4**	41.2	**41.4**	**20.1**	56.1
iNatural 2018	Image classification	OLTR	62.5	52.2	42.2	48.8	50.8	48.8	33.3	33.9
		BALMS	37.1	31.9	7.9	28.7	32.5	28.7	10.4	52.9
		TDE	63.1	62.1	54.8	59.3	59.3	65.6	43.7	24.8
		Decoupling	69.0	65.8	63.1	65.1	65.1	71.0	50.6	23.9
		BBN	61.8	**73.5**	67.7	69.7	72.8	69.7	55.9	29.3
		MiSLAS	**72.3**	72.3	**70.7**	**71.6**	**74.6**	**71.6**	**58.0**	**19.6**
		PaCo	66.7	68.0	69.4	68.4	71.0	68.4	53.9	53.8
LVIS v0.5	Instance segmentation	BALMS	**62.9**	**34.7**	**16.1**	**60.0**	**37.1**	**46.8**	**37.1**	1436.1

Table 7: Comparing the methods on semi-synthetic long-tailed datasets with three imbalance factors (100, 50, 10). "Time" refers to inference time. Recall equals to bAcc for multi-class classification.

Dataset	Method	Acc (%)			Precision (%)			Recall/bAcc (%)			MAP (%)			Time (s)
		100	50	10	100	50	10	100	50	10	100	50	10	
CIFAR-10-LT	OLTR	28.1	29.1	33.7	21.9	20.9	33.7	28.1	29.1	33.7	15.5	15.1	18.7	10.1
	BALMS	84.2	86.7	91.5	84.3	86.7	91.5	84.2	86.7	91.5	72.8	76.9	84.8	**1.7**
	TDE	80.9	82.9	88.3	80.9	82.9	88.3	81.4	83	88.2	68.0	70.7	79.2	1.8
	Decoupling	53.5	61.2	73.4	53.5	61.2	73.4	55.4	62.3	73.5	34.0	42.0	57.3	1.8
	BBN	80.7	83.4	88.7	81.0	84.0	88.8	80.7	83.4	88.7	67.5	71.8	80.0	5.2
	MiSLAS	82.5	85.7	90.0	83.4	86.1	90.1	82.5	85.7	90.1	70.5	75.2	82.3	8.2
	PaCo	**85.9**	**88.3**	**91.7**	**86.0**	**88.3**	**91.7**	**85.9**	**88.3**	**91.7**	**75.5**	**79.4**	**85.1**	4.6
CIFAR-100-LT	OLTR	7.9	9.2	12.5	5.4	6.5	12.0	7.9	9.2	12.5	1.9	2.2	3.2	12.9
	BALMS	51.0	**56.2**	55.2	50.0	**56.0**	54.8	51.0	**56.2**	55.2	29.5	**34.9**	33.6	2.2
	TDE	43.5	48.9	58.9	43.5	48.9	58.9	42.8	49.3	59.7	21.0	26.0	36.8	**1.7**
	Decoupling	34.0	36.4	51.5	33.9	36.4	51.5	32.2	35.7	51.4	14.0	15.8	29.2	1.8
	BBN	40.9	46.7	59.7	43.3	48.0	59.9	40.9	46.7	59.7	19.9	24.5	38.2	4.3
	MiSLAS	47.0	52.3	63.3	46.5	52.5	63.4	47.0	52.3	63.2	25.4	30.5	42.3	7.2
	PaCo	**51.2**	55.4	**66.0**	**51.2**	55.9	**66.2**	**51.2**	55.4	**66.0**	**34.9**	34.2	**46.0**	2.5

suboptimal method by 16.37%, MisLAS exhibits a remarkable overall accuracy of 71.6% on iNatural 2018 dataset. The superiority of PaCo and MisLAS is further evident in the tail category, where their few-shot accuracy surpasses other methods by 27.15% and 14.90% on Places-LT. **(3)** The few-shot accuracy is often lower than the many-shot accuracy for all methods. For example, on ImageNet-LT, Decoupling exhibits particularly poor performance in terms of few-shot accuracy, with a decrease of 90.94% compared to many-shot accuracy. But on iNatural 2018, PaCo achieves a few-shot accuracy of 69.4%, surpassing many-shot accuracy of 66.7% and medium-shot accuracy of 68.0%.

CIFAR-10-LT and CIFAR-100-LT are semi-synthetic datasets, where the number of samples in each class is determined by a controllable imbalance factor (typically 10, 50, 100). Similarly, we have the observations as shown in Table 7: **(1)** PaCo and BALMS show the top-two best accuracy performance on the synthetic CIFAR-10-LT and CIFAR-100-LT with varying degrees of IF. In conjunction with experimental results on natural datasets, the overall performances of decoupled learning (BALMS and MiSLAS) and contrastive learning (Paco) are generally superior. In addition, decoupled learning has the potential to be applied to multiple tasks and therefore has the potential to deal with data long-tailedness and task heterogeneity. but the stable performances in the few-shot categories still need to be considered. **(2)** As we control IF of the synthetic datasets to be larger, the long-tailed

Table 8: Comparing the methods on long-tailed relational datasets. Each point is the mean and standard deviation of 10 runs. "Time" means training time plus inference time.

Dataset	Method	Acc (%)	bAcc (%)	Precision (%)	Recall (%)	MAP (%)	Time (s)
Cora-Full	GraphSmote	60.5±0.8	51.9±0.6	60.2±0.8	51.9±0.6	54.9±0.4	718.8
	ImGAGN	4.2±0.8	1.5±0.1	0.2±0.1	1.5±0.1	2.7±0.1	**69.6**
	Tail-GNN	**63.8±0.3**	**54.6±0.6**	**63.8±0.4**	**54.6±0.6**	**66.8±0.6**	906.2
	LTE4G	60.8±0.5	**54.6±0.5**	61.5±0.6	**54.6±0.5**	51.1±0.9	281.6
Wiki	GraphSmote	**66.0±0.9**	**51.1±2.0**	**66.3±1.1**	**51.1±2.0**	64.4±1.9	52.3
	ImGAGN	45.4±6.7	24.5±4.1	44.8±5.0	24.5±4.1	64.1±6.2	**14.3**
	Tail-GNN	63.6±0.7	47.7±1.1	64.0±1.4	47.7±1.1	**65.0±2.1**	22.5
	LTE4G	58.2±18.5	48.9±12.5	60.2±20.1	48.9±12.5	59.3±16.5	53.1
Email	GraphSmote	58.3±1.2	34.2±1.4	54.4±2.0	34.2±1.4	44.6±2.4	126.2
	ImGAGN	42.7±1.9	23.0±1.4	38.4±2.5	23.0±1.4	35.7±2.2	19.7
	Tail-GNN	56.5±1.7	**34.5±1.6**	**55.6±0.4**	**34.5±1.6**	**58.0±3.0**	**8.6**
	LTE4G	**58.7±2.0**	34.3±2.1	54.1±2.5	34.3±2.1	47.0±2.1	72.1
Amazon_Clothing	GraphSmote	66.3±0.4	63.2±0.4	64.9±0.3	63.2±0.4	57.6±0.9	919.1
	ImGAGN	30.3±1.1	12.8±0.7	23.3±1.2	12.8±0.7	32.0±0.7	**89.6**
	Tail-GNN	**69.2±0.6**	**65.9±0.5**	**67.7±0.5**	**65.9±0.5**	**68.7±0.1**	768.7
	LTE4G	65.6±0.6	64.2±0.6	64.8±0.5	64.2±0.6	54.3±2.7	349.7
Amazon_Eletronics	GraphSmote	**56.2±0.5**	51.7±0.5	55.6±0.4	51.7±0.5	**36.2±1.9**	3406.2
	ImGAGN	17.1±1.1	7.4±0.6	12.3±0.7	7.4±0.6	16.8±0.7	**218.5**
	Tail-GNN	OOM	OOM	OOM	OOM	OOM	OOM
	LTE4G	55.8±0.4	**53.0±0.5**	**56.1±0.3**	**53.0±0.5**	32.7±2.2	1112.8

Table 9: Comparing the methods on long-tailed tabular regression dataset. "Time" means training time plus inference time.

Dataset	Method	MAE				MSE				Pearson (%)				GM				Time
		Many	Med.	Few	All	Many	Med.	Few	All	Many	Med.	Few	All	Many	Med.	Few	All	(s)
Housing	SmoteR	**0.12**	**0.12**	**0.20**	**0.12**	**0.02**	**0.02**	**0.07**	**0.02**	50.2	**98.2**	**96.9**	**97.3**	**0.08**	**0.08**	**0.14**	**0.08**	56.6
	SMOGN	0.40	0.35	0.43	0.37	0.17	0.13	0.24	0.15	**53.2**	97.3	91.3	95.4	0.38	0.33	0.36	0.35	**35.0**

phenomenon is more severe (showing higher values of Gini coefficient, IF, and Pareto-LT Ratio), and the performance of nearly all methods exhibits a decline. **(3)** CIFAR-100-LT dataset appears more affected than CIFAR-10-LT under the different settings of IF, possibly because it has a more severe long-tailed phenomenon with a larger number of categories.

3.5 Algorithm Performance on Node Classification

From the results of long-tailed algorithms for node classification on relational datasets in Table 8, we have: **(1)** No method consistently outperforms all others across all datasets. The performance of different runs on the same dataset may have large variances, *e.g.*, LTE4G on Wiki. **(2)** GraphSmote exhibits promising performance on Wiki dataset. Wiki presents high IF and Gini coefficient yet a low Pareto-LT Ratio, hinting that the main challenge may stem from data imbalance as discussed in *Angle 1*. Therefore, employing the simple augmentation method may yield great results. Amazon_Clothing exhibits relatively low IFs and Gini coefficients but high Pareto-LT Ratios, which may indicate the need for increased focus on addressing the challenge caused by the number of categories (but it does not mean it necessarily contains more categories). The insight may elucidate why Tail-GNN and LTE4G, which can characterize the extensive number of categories, exhibit more significant performance. Although these metrics can give some understanding of a dataset, none can comprehensively and accurately depict all of the characteristics, and the analysis may exhibit limitations in specific datasets. **(3)** Tail-GNN exhibits superior performance on Cora-full and Amazon_clothing. Especially on Cora-full, Tail-GNN achieves a 4.93% higher accuracy than the second-best method. However, the scalability of Tail-GNN shows limitations. It faces out-of-memory problems on Amazon_Eletronics with 42,318 nodes and 129,430 edges. **(4)** The performance of ImGAGN is relatively weak since it considers only one class as the tail class by default. This limitation becomes apparent in datasets with a large number of classes. Nonetheless, ImGAGN shows a performance improvement by adjusting the number of classes considered as tail. In addition, ImGAGN is less time-consuming and is 5x faster than LTE4G on the largest Amazon_Eletronics dataset.

9

3.6 Algorithm Performance on Regression

There has been limited work considering regression tasks on long-tailed data. In Table 9, we present a comparison of the performance of two methods. It is observed that SmoteR outperforms SMOGN in terms of MAE, MSE, Pearson correlation, and GM metrics across many-shot, medium-shot, and few-shot regions, as well as overall performance. However, SmoteR requires a longer time compared to SMOGN. Additionally, SMOGN appears to overfit to the many-shot regions during training.

4 Related Work

Recently, several papers that review long-tailed problems have been published. One of the earliest works [16] aims to improve the performance of long-tailed algorithms through a reasonable combination of existing tricks. Yang et al. [18] conduct a survey on long-tailed visual recognition and shows a taxonomy of existing methods. Fu et al. [15] divide methods into three categories, namely, training, fine-tuning, and inference stage. Fang et al. [14] group methods based on balanced data, balanced feature representation, balanced loss, and balanced prediction. Zhang et al. [17] categorize them into class re-balancing, information augmentation, and module improvement. Although these papers summarize studies on long-tailed visual recognition, they pay less attention to the extent of long-tailedness and fail to consider data complexity and task heterogeneity.

Next, we highlight the similarities and differences of long-tailed learning with related areas, e.g., imbalanced learning [20] and few-shot learning [109, 110]. Imbalanced Learning focuses on learning from imbalanced data, where minority classes contain significantly fewer training samples than majority classes. In imbalance learning, the class numbers may be small, while the number of minority samples is not necessarily small. But in long-tailed learning, the number of classes is larger and samples in tail classes are often very scarce. Few-shot Learning aims to train well-performing models from limited supervised samples. Long-tailed datasets, like few-shot datasets, have limited labeled samples in tail classes but with an imbalanced distribution. In contrast, few-shot datasets tend to be more balanced.

5 Conclusions and Future Work

In this paper, we introduce HEROLT, the most comprehensive heterogeneous long-tailed learning benchmark with 18 state-of-the-art algorithms and 10 evaluation metrics on 17 real-world benchmark datasets across 6 tasks and 4 data modalities. Based on the analyses of three pivotal angles, we gain valuable insights into the characterization of data long-tailedness, the data complexity of various domains, and the heterogeneity of emerging tasks. Our benchmark and evaluations are released at `https://github.com/SSSKJ/HeroLT`.

On top of them, we suggest intellectual challenges and promising research directions in long-tailed learning: **(C1)** Theoretic Challenge: Current work lacks sufficient theoretical tools for analyzing long-tailed models like their generalization performance. **(C2)** Algorithmic Challenge: Existing research typically focuses on one task in one domain, while there is a trend to consider multiple forms of input data (*e.g.*, text and images) by multi-modal learning [111, 112, 113], or to solve multiple learning tasks (*e.g.*, segmentation and classification) by multi-task learning [114, 115]. **(C3)** Application Challenge: In open environments, many datasets exhibit long-tailed distributions. However, long-tailed problems in domains like antibiotic resistance genes [10, 11] receive insufficient attention.

References

[1] Yin Zhang, Derek Zhiyuan Cheng, Tiansheng Yao, Xinyang Yi, Lichan Hong, and Ed H Chi. A model of two tales: Dual transfer learning framework for improved long-tail item recommendation. In *Proceedings of the web conference*, pages 2220–2231, 2021.

[2] Longfeng Wu, Bowen Lei, Dongkuan Xu, and Dawei Zhou. Towards reliable rare category analysis on graphs via individual calibration. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2629–2638. ACM, 2023.

[3] Daixin Wang, Jianbin Lin, Peng Cui, Quanhui Jia, Zhen Wang, Yanming Fang, Quan Yu, Jun Zhou, Shuang Yang, and Yuan Qi. A semi-supervised graph attentive network for financial fraud detection. In *IEEE International Conference on Data Mining*, pages 598–607. IEEE, 2019.

[4] Lie Ju, Xin Wang, Lin Wang, Tongliang Liu, Xin Zhao, Tom Drummond, Dwarikanath Mahapatra, and Zongyuan Ge. Relational subsets knowledge distillation for long-tailed retinal diseases recognition. In *Medical Image Computing and Computer Assisted Intervention*, pages 3–12. Springer, 2021.

[5] Dawei Zhou, Si Zhang, Mehmet Yigit Yildirim, Scott Alcorn, Hanghang Tong, Hasan Davulcu, and Jingrui He. A local algorithm for structure-preserving graph cut. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 655–664. ACM, 2017.

[6] Marshall A Kuypers, Thomas Maillart, and Elisabeth Paté-Cornell. An empirical analysis of cyber security incidents at a large organization. *Department of Management Science and Engineering, Stanford University, School of Information*, 30, 2016.

[7] Jiali Wang, Martin Neil, and Norman Fenton. A bayesian network approach for cybersecurity risk assessment implementing and extending the fair model. *Computers & Security*, 89:101659, 2020.

[8] Tommie W Singleton and Aaron J Singleton. *Fraud auditing and forensic accounting*, volume 11. John Wiley and Sons, 2010.

[9] Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery*, 29(3):626–688, 2015.

[10] Gustavo Arango-Argoty, Emily Garner, Amy Pruden, Lenwood S Heath, Peter Vikesland, and Liqing Zhang. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6:1–15, 2018.

[11] Yu Li, Zeling Xu, Wenkai Han, Huiluo Cao, Ramzan Umarov, Aixin Yan, Ming Fan, Huan Chen, Carlos M Duarte, Lihua Li, et al. HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes. *Microbiome*, 9:1–12, 2021.

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

[13] Tianxiang Zhao, Xiang Zhang, and Suhang Wang. Graphsmote: Imbalanced node classification on graphs with graph neural networks. In *Proceedings of the ACM international conference on web search and data mining*, pages 833–841, 2021.

[14] Chaowei Fang, Dingwen Zhang, Wen Zheng, Xue Li, Le Yang, Lechao Cheng, and Junwei Han. Revisiting long-tailed image classification: Survey and benchmarks with new evaluation metrics. *arXiv preprint arXiv:2302.01507*, 2023.

[15] Yu Fu, Liuyu Xiang, Yumna Zahid, Guiguang Ding, Tao Mei, Qiang Shen, and Jungong Han. Long-tailed visual recognition with deep models: A methodological survey and evaluation. *Neurocomputing*, 509:290–309, 2022.

[16] Yongshun Zhang, Xiu-Shen Wei, Boyan Zhou, and Jianxin Wu. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3447–3455, 2021.

[17] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[18] Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *International Journal of Computer Vision*, 130(7):1837–1872, 2022.

[19] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[20] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009.

[21] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

[22] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9695–9704, 2021.

[23] Bo Liu, Haoxiang Li, Hao Kang, Gang Hua, and Nuno Vasconcelos. GistNet: a geometric structure transfer network for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8209–8218, 2021.

[24] Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 630–639, 2021.

[25] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2361–2370, 2021.

[26] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

[27] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. 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*, pages 7610–7619, 2020.

[28] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[29] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32, 2019.

[30] Liang Qu, Huaisheng Zhu, Ruiqi Zheng, Yuhui Shi, and Hongzhi Yin. Imgagn: Imbalanced network embedding via generative adversarial graph networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1390–1398, 2021.

[31] Zemin Liu, Trung-Kien Nguyen, and Yuan Fang. Tail-gnn: Tail-node graph neural networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1109–1119, 2021.

[32] Sukwon Yun, Kibum Kim, Kanghoon Yoon, and Chanyoung Park. Lte4g: Long-tail experts for graph neural networks. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 2434–2443, 2022.

[33] Weixin Zeng, Xiang Zhao, Wei Wang, Jiuyang Tang, and Zhen Tan. Degree-aware alignment for entities in tail. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 811–820, 2020.

[34] Ermei Cao, Difeng Wang, Jiacheng Huang, and Wei Hu. Open knowledge enrichment for long-tail entities. In *Proceedings of The Web Conference*, pages 384–394, 2020.

[35] Xichuan Niu, Bofang Li, Chenliang Li, Rong Xiao, Haochuan Sun, Hongbo Deng, and Zhen-zhong Chen. A dual heterogeneous graph attention network to improve long-tail performance for shop search in e-commerce. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 3405–3415, New York, NY, USA, 2020.

[36] Zhihong Chen, Rong Xiao, Chenliang Li, Gangfeng Ye, Haochuan Sun, and Hongbo Deng. Esam: Discriminative domain adaptation with non-displayed items to improve long-tail performance. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 579–588, 2020.

[37] Jianwen Yin, Chenghao Liu, Weiqing Wang, Jianling Sun, and Steven CH Hoi. Learning transferrable parameters for long-tailed sequential user behavior modeling. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 359–367, 2020.

[38] Yoon-Joo Park and Alexander Tuzhilin. The long tail of recommender systems and how to leverage it. In *Proceedings of the ACM conference on Recommender systems*, pages 11–18, 2008.

[39] Todd Huster, Jeremy Cohen, Zinan Lin, Kevin Chan, Charles Kamhoua, Nandi O Leslie, Cho-Yu Jason Chiang, and Vyas Sekar. Pareto GAN: Extending the representational power of gans to heavy-tailed distributions. In *International Conference on Machine Learning*, pages 4523–4532. PMLR, 2021.

[40] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020.

[41] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019.

[42] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33:1513–1524, 2020.

[43] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498, 2021.

[44] Dong Cao, Xiangyu Zhu, Xingyu Huang, Jianzhu Guo, and Zhen Lei. Domain balancing: Face recognition on long-tailed domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5671–5679, 2020.

[45] Yaoyao Zhong, Weihong Deng, Mei Wang, Jiani Hu, Jianteng Peng, Xunqiang Tao, and Yaohai Huang. Unequal-training for deep face recognition with long-tailed noisy data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7812–7821, 2019.

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

[47] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11662–11671, 2020.

[48] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Jun Hao Liew, Sheng Tang, Steven C. H. Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *European Conference on Computer Vision*, volume 12359 of *Lecture Notes in Computer Science*, pages 728–744. Springer, 2020.

[49] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S Dhillon. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining*, pages 3163–3171, 2020.

[50] Jiong Zhang, Wei-Cheng Chang, Hsiang-Fu Yu, and Inderjit Dhillon. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 34:7267–7280, 2021.

[51] Hsiang-Fu Yu, Jiong Zhang, Wei-Cheng Chang, Jyun-Yu Jiang, Wei Li, and Cho-Jui Hsieh. Pecos: Prediction for enormous and correlated output spaces. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4848–4849, 2022.

[52] Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. Uncovering main causalities for long-tailed information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 9683–9695. Association for Computational Linguistics, 2021.

[53] Aleksandar Bojchevski and Stephan Günnemann. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *International Conference on Learning Representations*, 2018.

[54] C Lee Giles, Kurt D Bollacker, and Steve Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98, 1998.

[55] Julian McAuley, Rahul Pandey, and Jure Leskovec. Inferring networks of substitutable and complementary products. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2015.

[56] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, pages 697–706. ACM, 2007.

[57] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a web of open data. In *Proceedings of the International The Semantic Web and Asian Conference on Asian Semantic Web Conference*, page 722–735, Berlin, Heidelberg, 2007. Springer-Verlag.

[58] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[59] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the international conference on World Wide Web*, pages 22–32, 2005.

[60] Simon Dooms, Toon De Pessemier, and Luc Martens. Movietweetings: a movie rating dataset collected from twitter. In *Workshop on Crowdsourcing and human computation for recommender systems, CrowdRec at RecSys*, volume 2013, page 43, 2013.

[61] Xiao Yu, Xiang Ren, Yizhou Sun, Quanquan Gu, Bradley Sturt, Urvashi Khandelwal, Brandon Norick, and Jiawei Han. Personalized entity recommendation: A heterogeneous information network approach. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, page 283–292, New York, NY, USA, 2014. Association for Computing Machinery.

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

[63] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.

[64] Mengmeng Li, Tian Gan, Meng Liu, Zhiyong Cheng, Jianhua Yin, and Liqiang Nie. Long-tail hashtag recommendation for micro-videos with graph convolutional network. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, page 509–518. Association for Computing Machinery, 2019.

[65] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, 2012.

[66] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[67] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[68] Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. AttentionXML: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 32, 2019.

[69] Lin Xiao, Xiangliang Zhang, Liping Jing, Chi Huang, and Mingyang Song. Does head label help for long-tailed multi-label text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14103–14111, 2021.

[70] Anshul Mittal, Kunal Dahiya, Sheshansh Agrawal, Deepak Saini, Sumeet Agarwal, Purushottam Kar, and Manik Varma. DECAF: Deep extreme classification with label features. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 49–57, 2021.

[71] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.

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

[73] Deepak Saini, Arnav Kumar Jain, Kushal Dave, Jian Jiao, Amit Singh, Ruofei Zhang, and Manik Varma. GalaXC: Graph neural networks with labelwise attention for extreme classification. In *Proceedings of the Web Conference*, pages 3733–3744, 2021.

[74] Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. Sparse local embeddings for extreme multi-label classification. *Advances in neural information processing systems*, 28, 2015.

[75] Yang Lu, Yiu-Ming Cheung, and Yuan Yan Tang. Bayes imbalance impact index: A measure of class imbalanced data set for classification problem. *IEEE transactions on neural networks and learning systems*, 31(9):3525–3539, 2019.

[76] B. German. Glass Identification. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5WW2P.

[77] Haojun Tang, Wenda Zhao, Guang Hu, Yi Xiao, Yunlong Li, and Haipeng Wang. Text-guided diverse image synthesis for long-tailed remote sensing object classification. *IEEE Trans. Geosci. Remote. Sens.*, 62:1–13, 2024.

[78] Mengke Li, Yiu-Ming Cheung, and Yang Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6919–6928. IEEE, 2022.

[79] Warwick Nash, Tracy Sellers, Simon Talbot, Andrew Cawthorn, and Wes Ford. Abalone. UCI Machine Learning Repository, 1995. DOI: https://doi.org/10.24432/C55C7W.

[80] Dean De Cock. Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3), 2011.

[81] Péter Mernyei and Cătălina Cangea. Wiki-CS: A wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901*, 2020.

[82] Hao Yin, Austin R. Benson, Jure Leskovec, and David F. Gleich. Local higher-order graph clustering. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 555–564. Association for Computing Machinery, 2017.

[83] Sheng Fu, Piao Chen, and Zhisheng Ye. Simplex-based proximal multicategory support vector machine. *IEEE Trans. Inf. Theory*, 69(4):2427–2451, 2023.

[84] Prateek Jain and Ashish Kapoor. Active learning for large multi-class problems. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 762–769. IEEE Computer Society, 2009.

[85] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. Deep learning for extreme multi-label text classification. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval*, pages 115–124, 2017.

[86] Miao Fan, Yeqi Bai, Mingming Sun, and Ping Li. Large margin prototypical network for few-shot relation classification with fine-grained features. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, page 2353–2356. Association for Computing Machinery, 2019.

[87] Haotao Wang, Aston Zhang, Yi Zhu, Shuai Zheng, Mu Li, Alex J Smola, and Zhangyang Wang. Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In *International Conference on Machine Learning*, pages 23446–23458. PMLR, 2022.

[88] Jianhong Bai, Zuozhu Liu, Hualiang Wang, Jin Hao, YANG FENG, Huanpeng Chu, and Haoji Hu. On the effectiveness of out-of-distribution data in self-supervised long-tail learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[89] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6959–6969, 2022.

[90] Xuan Kou, Chenghao Xu, Xu Yang, and Cheng Deng. Attention-guided contrastive hashing for long-tailed image retrieval. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 1017–1023. International Joint Conferences on Artificial Intelligence Organization, 2022.

[91] Tao He, Lianli Gao, Jingkuan Song, Jianfei Cai, and Yuan-Fang Li. Learning from the scene and borrowing from the rich: Tackling the long tail in scene graph generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 587–593. International Joint Conferences on Artificial Intelligence Organization, 2020.

[92] Harsh Rangwani, Naman Jaswani, Tejan Karmali, Varun Jampani, and R Venkatesh Babu. Improving gans for long-tailed data through group spectral regularization. In *European Conference on Computer Vision*, pages 426–442. Springer, 2022.

[93] Weitao Wang, Meng Wang, Sen Wang, Guodong Long, Lina Yao, Guilin Qi, and Yang Chen. One-shot learning for long-tail visual relation detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12225–12232, 2020.

[94] Sherif Abdelkarim, Aniket Agarwal, Panos Achlioptas, Jun Chen, Jiaji Huang, Boyang Li, Kenneth Church, and Mohamed Elhoseiny. Exploring long tail visual relationship recognition with large vocabulary. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15921–15930, 2021.

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

[96] Mengke Li, HU Zhikai, Yang Lu, Weichao Lan, Yiu-ming Cheung, and Hui Huang. Feature fusion from head to tail for long-tailed visual recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13581–13589, 2024.

[97] Xing Zhang, Zuxuan Wu, Zejia Weng, Huazhu Fu, Jingjing Chen, Yu-Gang Jiang, and Larry S Davis. Videolt: large-scale long-tailed video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7960–7969, 2021.

[98] Tai-Yu Pan, Cheng Zhang, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, and Wei-Lun Chao. On model calibration for long-tailed object detection and instance segmentation. *Advances in Neural Information Processing Systems*, 34:2529–2542, 2021.

[99] Tong Wang, Yousong Zhu, Chaoyang Zhao, Wei Zeng, Jinqiao Wang, and Ming Tang. Adaptive class suppression loss for long-tail object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3103–3112, 2021.

[100] Haohui Wang, Baoyu Jing, Kaize Ding, Yada Zhu, Wei Cheng, Si Zhang, Yonghui Fan, Liqing Zhang, and Dawei Zhou. Mastering long-tail complexity on graphs: Characterization, learning, and generalization. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3045–3056. ACM, 2024.

[101] Yuzhe Yang, Kaiwen Zha, Ying-Cong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11842–11851. PMLR, 2021.

[102] Luís Torgo, Rita P. Ribeiro, Bernhard Pfahringer, and Paula Branco. SMOTE for regression. In *Progress in Artificial Intelligence - Portuguese Conference on Artificial Intelligence*, volume 8154 of *Lecture Notes in Computer Science*, pages 378–389. Springer, 2013.

[103] Paula Branco, Luís Torgo, and Rita P. Ribeiro. SMOGN: a pre-processing approach for imbalanced regression. In *International Workshop on Learning with Imbalanced Domains: Theory and Applications, LIDTA@PKDD/ECML*, volume 74 of *Proceedings of Machine Learning Research*, pages 36–50. PMLR, 2017.

[104] Inderjeet Mani and I Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets, ICML*, volume 126, pages 1–7, 2003.

[105] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 715–724, 2021.

[106] Mohammad S Sorower. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18(1):25, 2010.

[107] Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *ArXiv*, abs/2008.05756, 2020.

[108] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[109] Dawei Zhou, Jingrui He, Hongxia Yang, and Wei Fan. SPARC: self-paced network representation for few-shot rare category characterization. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2807–2816. ACM, 2018.

[110] Haohui Wang, Yuzhen Mao, Yujun Yan, Yaoqing Yang, Jianhui Sun, Kevin Choi, Balaji Veeramani, Alison Hu, Edward Bowen, Tyler Cody, and Dawei Zhou. EvoluNet: Advancing dynamic non-iid transfer learning on graphs. In *International Conference on Machine Learning*, 2024.

[111] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019.

[112] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[113] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.

[114] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[115] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020.

[116] Fan Liu, Zhiyong Cheng, Lei Zhu, Chenghao Liu, and Liqiang Nie. An attribute-aware attentive gcn model for attribute missing in recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 34(9):4077–4088, 2022.

[117] Arjan Reurink. Financial fraud: A literature review. *Contemporary Topics in Finance: A Collection of Literature Surveys*, pages 79–115, 2019.

[118] Jonathan M Karpoff. The future of financial fraud. *Journal of Corporate Finance*, 66:101694, 2021.

[119] David M Weinstock, Larisa V Gubareva, and Gianna Zuccotti. Prolonged shedding of multidrug-resistant influenza a virus in an immunocompromised patient. *New England Journal of Medicine*, 348(9):867–868, 2003.

[120] Jiequan Cui, Zhisheng Zhong, Zhuotao Tian, Shu Liu, Bei Yu, and Jiaya Jia. Generalized parametric contrastive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Checklist

1. For all authors...

   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

   (b) Did you describe the limitations of your work? [Yes] Please see the supplementary materials.

   (c) Did you discuss any potential negative societal impacts of your work? [N/A]

   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [N/A]

   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please see the supplementary materials for the implementation details. The released toolbox is available at `https://github.com/SSSKJ/HeroLT`

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please see the supplementary materials.

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report the average value and standard deviations over 10 runs.

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please see the supplementary materials.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes]

   (b) Did you mention the license of the assets? [Yes] All the datasets we use are publicly available.

   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include our toolbox.

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We use publicly available datasets.

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The dataset has no personally identifiable information or offensive content.

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A   Long-Tailed Data Distributions

In this section, we will give some real-world examples with long-tailed distributions. Cora-Full dataset consists of 19,793 scientific publications classified into one of seventy categories [53]. As shown in Figure 3(a), this dataset exhibits a prominent long-tailed distribution, wherein the number of instances belonging to the head categories far surpasses that of the tail categories. Similarly, Amazon_Eletronics dataset [55] also exhibits long-tailed distribution (see Figure 3(b)), where each product is considered as a node belonging to a product category in "Electronics." Despite the emergence of machine learning methods aimed at facilitating accurate classification, further solutions are called due to the challenges of long-tailed distribution.

(a) Cora-Full

(b) Amazon_Eletronics
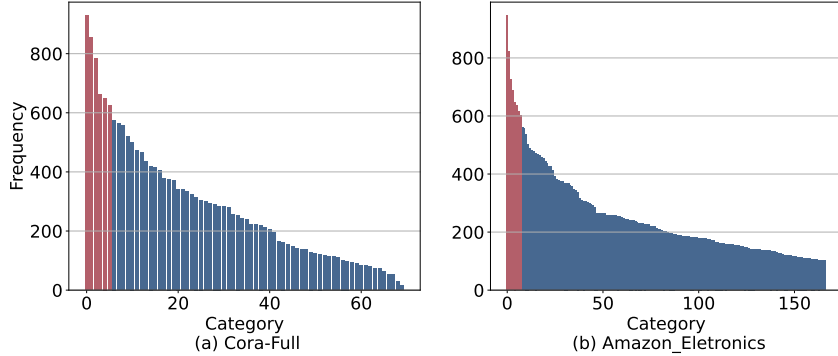
Figure 3: The data distributions on two commonly used datasets exhibit prominent long-tailed distributions.

In addition, we present a motivative application of the recommendation system [116] as shown in Figure 4, which naturally exhibits long-tailed data distributions coupled with data complexity [2] (e.g., tabular data and relational data) and task heterogeneity (e.g., user profiling [1] and recommendation [2]). Additionally, heterogeneous long-tailed learning has various real-world applications, such as financial fraud detection [117, 118] and ARG prediction [119].
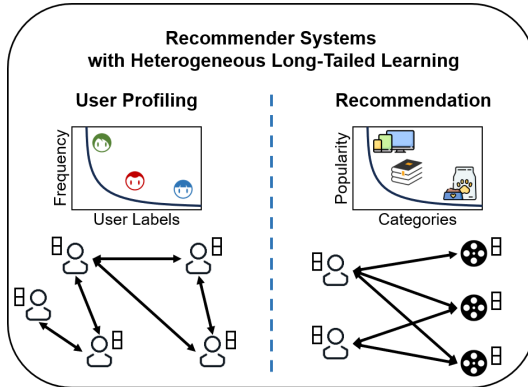

Figure 4: The data distributions on two commonly used datasets exhibit prominent long-tailed distributions.

B   More Details on HEROLT

B.1   Long-tailedness Metrics

To measure the long-tailedness of a dataset, a commonly used metric is the imbalance factor, and Gini coefficient is used as a measurement in [18]. In this section, we analyze the strengths and weaknesses of each metric and introduce Pareto-LT Ratio to evaluate the long-tailedness of a dataset.

Imbalance Factor. To measure the skewness of long-tailed distribution, [72] first introduces the imbalance factor as the size of the largest majority class to the size of the smallest minority class:

$$IF = n_1/n_C \tag{1}$$

where $n_c, c = 1, 2, \ldots, C$, represents the size of each category following descending order. The range of imbalance factor is $[1, \infty)$. Intuitively, the larger the value of IF indicates a more imbalanced dataset.

Gini Coefficient. Gini coefficient is a measure of income inequality used to quantify the extent to which the distribution of income among a population deviates from perfect equality. [18] propose to use Gini coefficient as a long-tailedness metric since long-tailedness is similar to inequality between each category.

$$Gini = \frac{\sum_{i=1}^{C} \sum_{j=1}^{C} |n_i - n_j|}{2nC} \tag{2}$$

Gini coefficient ranges from 0 to 1, where a larger value indicates that the dataset is more imbalanced.

Pareto-LT Ratio. We propose a new metric named Pareto-LT Ratio to measure the long-tailedness. The design of this metric is inspired by the Pareto distribution, which is defined as:

$$Pareto - LT = \frac{C - Q(0.8)}{Q(0.8)} \tag{3}$$

where $Q(p) = min\{y : Pr(\mathcal{Y} \leq y) = p, 1 \leq y \leq T\}$ is the quantile function of order $p \in (0, 1)$ for $\mathcal{Y}$. The numerator represents the number of categories to which the last 20% instances belongs, and the denominator represents the number of categories to which the other 80% instances belongs in the dataset. Intuitively, the higher the skewness of the data distribution, the larger the ratio will be; the more classes, the larger the long-tailedness ratio.
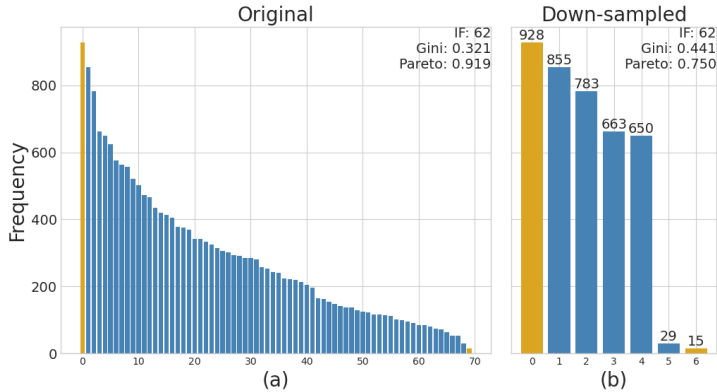

Figure 5: An example to compare the three long-tailedness metrics.

Imbalance factor is intuitive and easy to calculate, but it only considers the imbalance between the largest and smallest classes. Gini coefficient indicates the overall degree of category imbalance, unaffected by extreme samples or absolute data size. However, the two metrics pay more attention to data imbalance and may not reflect the number of categories. Pareto-LT Ratio is proposed to characterize two properties of long-tailed datasets: (1) data imbalance and (2) extreme # of categories. For better understanding, we give a specific example of the three long-tailedness metrics (as shown in Figure 5). As the number of categories increases, the difficulty of classifying a long-tailed dataset therefore increases. For example, we down-sampled 7 categories from the original Cora-Full dataset. Although the two datasets are clearly different, the imbalance factor remains the same (62 for both the original and down-sampled datasets) as the number of samples in the most majority and most minority categories does not change. Gini coefficient indicates the overall degree of category imbalance and thus shows a large increase of value in the down-sampled dataset (0.321 for original and 0.441 for down-sampled). However, the data complexity also includes the number of categories, i.e., 70 classes in Fig(a) v.s 7 classes in Fig(b), which is not reflected in imbalance factor and Gini coefficient. As the number of categories of the down-sampled dataset decreases dramatically, Pareto-LT Ratio better

characterizes the differences between the original Cora dataset (0.919) and its down-sampled dataset (0.750) by a small decrease of the metric value.

B.2 HEROLT Algorithm List

In this section, we introduce 18 popular and recent methods for solving long-tailed problems in our benchmark according to whether they can solve the problems of data imbalance and an extreme number of categories. The baseline algorithms are selected due to the following reasons: (1) Data Complexity: The selected methods provide comprehensive coverage of heterogeneous data (i.e, tabular, sequential, grid, and relational data) in long-tailed learning problems. (2) Task Heterogeneity: we explore a range of techniques for long-tailed learning approaches (i.e., data augmentation, meta-learning, decoupled training, mixup), designing for various tasks (i.e, object recognition, multi-label text classification, image classification, instance segmentation, node classification, and regression). (3) SOTA Performance: Most of the selected methods are the SOTAs, which are recently published and highly cited. (4) Open Source: All of the selected methods are open-sourced in GitHub. In addition, our toolbox is still being updated, and we will include more algorithms in the future.

SMOTE [19] generates synthetic samples by feature interpolation minority samples with their nearest.

NearMiss [104] is a downsampling method that selects samples based on the distance of majority class samples to minority class samples.

X-Transformer [49] consists of three components: semantic label indexing, deep neural matching, and ensemble ranking. Semantic label indexing decomposes extreme multi-label classification (XMC) problems into a feasible set of subproblems with much smaller output space by label clustering; deep neural matching fine-tunes a Transformer model for each subproblem; ensemble ranking is conditionally trained based on the instance-cluster assignment and embeddings from the Transformer and is used to aggregate scores from subproblems. X-Transformer solves data imbalance and an extreme number of categories.

XR-Transformer [50] is a transformer-based XMC framework that fine-tunes pre-trained transformers recursively leveraging multi-resolution objectives and cost-sensitive learning. The embedding generated at the previous task is used to bootstrap the non pre-trained part for the current task. XR-Transformer solves data imbalance and an extreme number of categories.

XR-Linear [51] is designed for XMC problem. It consists of recursive linear models that traverse an input from the root of a hierarchical label tree to a few leaf node clusters and return top-k relevant labels within the clusters as predictions. XR-Linear solves data imbalance and an extreme number of categories.

OLTR [41] learns from natural long-tailed open-end distributed data. It consists of two main modules: dynamic meta embedding and modulated attention. The former combines a direct image feature and an associated memory feature to transfer knowledge between head and tail classes, and the latter maintains discrimination between them. Therefore, OLTR solves the challenges of data imbalance and an extreme number of categories.

BALMS [46] presents Balanced Softmax, an unbiased extension of Softmax, to accommodate the label distribution shift between training and testing. In addition, it applies a Meta Sampler to learn the optimal class re-sample rate by meta learning. Therefore, BALMS can address data imbalance.

TDE [42] is a framework for considering long-tailed problems using causal inference. It constructs a casual graph with four variables: momentum, object feature, projection of head direction, and model prediction, using casual intervention in training and counterfactual reasoning in inference to preserve "good" feature while cut off "bad" confounding effect. TDE solves data imbalance.

Decoupling [26] decouples the learning procedure into representation learning and classification. The authors find that instance-balance sampling gives more generalizable representations and can achieve state-of-art performance after properly adjusting the classifiers. Decoupling can address the challenge of data imbalance.

BBN [40] consists of two branches: the conventional learning branch with a uniform sampler for learning universal patterns for recognition and the re-balancing branch with a reversed sampler for modeling the tail data. Then the predicted outputs of these bilateral branches are aggregated in

the cumulative learning part using an adaptive trade-off parameter, which first learns the universal features from the original distribution and then gradually pays attention to the tail data. By different data samplers (such as mixup which is popular for solving long-tailed problems) and cumulative learning strategy, BBN can address data imbalance.

MiSLAS [43] decouples representation learning and classifier learning. It uses mixup and designs label-aware smoothing to handle different degrees of over-confidence for classes and improve classifier learning. It also uses shift learning on the batch normalization layer to reduce dataset bias in the decoupling framework. MiSLAS can solve data imbalance.

PaCo [105] presents parametric contrastive learning. To mitigate the problem of contrastive loss bias on head categories from an optimization perspective, the authors propose a set of parametric class-wise learnable centers, which adaptively change the intensity of pushing samples belonging to the same category close to each other. The follow-up work GPaCo [120] removes the momentum encoder, which achieves better model performance and robustness. PaCo can solve data imbalance and the extreme number of categories.

GraphSMOTE [13] is the first work considering node class imbalance on graphs proposed in 2021. It first uses a GNN-based feature extractor to learn node representations and then applies SMOTE to generate synthetic nodes for the minority classes. Then an edge generator pre-trained on the original graph is introduced to model the existence of edges among nodes. The augmented graph is used for classification with a GNN classifier. By generating nodes for minority classes, GraphSMOTE increases the number of labeled samples for these classes, thus addressing the data imbalance.

ImGAGN [30] is an adversarial learning-based approach. It uses a generator to generate a set of synthetic minority nodes and topological structures, then uses a discriminator to discriminate between real and fake (i.e., generated) nodes and between minority and majority nodes, which can solve the challenges of data imbalance. However, ImGAGN sets the smallest class as the minority class and the residual classes as the majority class, which fails when the dataset contains the large number of categories.

Tail-GNN [31] Due to the specificity of rational data, the long-tailed problem on graphs includes the long-tailed of category and the long-tailed of node degree. Tail-GNN [31] focuses on solving the long-tailed problem of degree by introducing transferable neighborhood translation to capture the relational tie between a node and its neighboring nodes. Then it complements missing information of the tail nodes for neighborhood aggregation. Tail-GNN learns robust node embeddings by narrowing the gap between head and tail nodes in terms of degree and addresses the challenges of data imbalance and an extreme number of categories in degree level.

LTE4G [32] splits the nodes into four balanced subsets considering class and degree long-tailed distributions. Then, it trains an expert for each balanced subset and employs knowledge distillation to obtain the head student and tail student for further classification. Lastly, LTE4G devises a class prototype-based inference. Because LTE4G uses knowledge distillation across the head and tail and considers the tail classes together, it can solve data imbalance and an extreme number of categories.

SmoteR [102] modifies the well-known SMOTE algorithm to handle regression tasks, where the target variable is continuous. It generates synthetic samples and applies over-sample and under-sample, helping to balance the distribution of the training data.

SMOGN [103] generates synthetic samples by combining an under-sampling strategy with two over-sampling strategies to address the challenges of imbalanced regression. It adjusts the training data distribution to handle rare and extreme values of a continuous target variable.

B.3 HEROLT Dataset List

Long-tailed challenges exist for various real-world data, such as tabular data, sequential data, grid data, and rational data. In this section, we briefly describe the collection of datasets selected for the initial version of our benchmark. In addition, we will show more datasets in past long-tailed studies on our toolbox page, including the statistics information (*e.g.*, size, #categories) and the long-tailedness (*e.g.*, imbalance factor, Gini coefficient, and Pareto-LT Ratio).

Glass [76] is a dataset from USA Forensic Science Service. Motivated by criminological investigation, the glass left at the scene of the crime can be classified into 6 types based on its oxide content.

Abalone [79] is a dataset for predicting the age of abalone from physical measurements, such as length, diameter, height, and weight.

EURLEX-4K [68] consists of legal documents from the European Union, and the number of instances in the training and test sets are 15,499 and 3,865.

AMAZONCat-13K [68] contains product descriptions from Amazon, and the number of instances in the training and test sets are 1,186,239 and 306,782.

Wiki10-31K [68] is a collection of Wikipedia articles, and the number of instances in the training and test sets are 14,146 and 6,616.

ImageNet-LT [41] is a long-tailed version sampled from the original ImageNet-2012, which is a large-scale image dataset constructed based on the WorldNet structure. The train set has 115,846 images from 1000 categories with maximally 1280 images per class and minimally 5 images per class. The test and valid sets have 50,000 and 20,000 samples and are balanced.

Places-LT [41] is a long-tailed version of scene classification dataset Places-2. The train set contains 62,500 images from 365 categories with maximally 4980 images per class and minimally 5 images per class. The test and valid sets are balanced with 100 and 20 images per class accordingly.

iNatural 2018 [62] is a species classification dataset with a train set with 437,513 images for 8142 classes. The class frequencies follow a natural power law distribution with a maximum number of 4,980 images per class and a minimum number of 5 images per class. The test and valid sets contain 149,394 and 24,426 images, respectively.

CIFAR 10-LT and **CIFAR 100-LT** [72]. The original CIFAR dataset has two versions: CIFAR 10 and CIFAR 100, where the former has 10 classes, 6,000 images in every class, while the latter has 100 classes with 600 samples in each class. CIFAR 10-LT and CIFAR 100-LT are two long-tailed versions of CIFAR dataset (named semi-synthetic long-tailed datasets in this paper), where the number of samples in each class is determined by a controllable imbalance factor. The commonly used imbalance factors are 10, 50, 100. The test set remains unchanged with even distribution.

LVIS v0.5 [63] is a large vocabulary instance segmentation dataset with 1231 classes. It contains a 693,958 train set, and a relatively balanced test/ valid set.

Housing [80] is a dataset designed to predict the selling price. It contains 79 explanatory variables that describe various aspects of residential homes in Ames.

Cora-Full [53] is a citation network dataset. Each node represents a paper with a sparse bag-of-words vector as the node attribute. The edge represents the citation relationships between two corresponding papers, and the node category represents the research topic.

Email [82] is a network constructed from email exchanges in a research institution, where each node represents a member, and each edge represents the email communication between institution members.

Wiki [81] is a network dataset of Wikipedia pages, with each node representing a page and each edge denoting the hyperlink between pages.

Amazon-Clothing [55] is a product network that contains products in "Clothing, Shoes and Jewelry" on Amazon, where each node represents a product and is labeled with low-level product categories. The node attributes are constructed based on the product's description, and the edges are established based on their substitutable relationship ("also viewed").

Amazon-Electronics [55] is another product network constructed from products in "Electronics". The edges are created with the complementary relationship ("bought together") between products.

C   Details on Experiment Setting

C.1   Hyperparameter Settings

Here we provide the details of parameter settings. We implement X-Transformer, XR-Transformer, and XR-Linear by applying the best hyperparameter settings from their original papers. In particular, for XR-Linear, we set the number of clusters (*i.e.*, beam size) predicted by the matcher to 10 and chose teacher-forced negation as the hard negation sampling scheme. We implement all experiments in

PyTorch and use ResNet-50 for ImageNet-LT dataset, ResNet-152 for Places-LT dataset, ResNet-50 for iNatural 2018 dataset, and ResNet-32 for CIFAR 10-LT and CIFAR 100-LT as the backbones for all methods. For method-related hyperparameters, we use the default settings for all methods on all datasets following the original papers. For GraphSMOTE, we set the weight of edge reconstruction loss to $1e-6$ as in the original paper. For LTE4G, we adopt the best hyperparameter settings reported in the paper. For Tail-GNN, we set the degree threshold to 5 (*i.e.*, nodes having a degree of no more than 5 are regarded as tail nodes), which is the default value in the original paper.

C.2   Evaluation Metrics

Considering the long-tailed distribution, accuracy, precision, recall, balanced accuracy, mean average precision, mean-average-error, mean-squared-error, Pearson correlation, error geometric mean, and time are used as the evaluation metrics.

Table 10: Ten metrics for evaluating long-tailed algorithms, where TP, TN, FP, FN stand for true positive, true negative, false positive, false negative, AP_i is average precision for class i, T is the total number of classes, y_i and $\hat{y}_i$ are the actual and predicted values of the i-th data point, $\bar{y}_i$ and $\bar{\hat{y}}_i$ are the means of the actual and predicted values, n is the total number of data points. For classification tasks (*e.g.*, object recognition, multi-label text classification, image classification, instance segmentation, and node classification), we give computations of two-class classification, which are slightly different for different tasks in our benchmark.

Metric name	Task	Computation	Description
Acc	Classification	$\frac{TP+FN}{TP+TN+FP+FN}$	Correct predictions of the algorithm on the dataset.
Precision	Classification	$\frac{TP}{TP+FP}$	Fraction of correctly predicted positive instances against total positively predicted instances.
Recall	Classification	$\frac{TP}{TP+FN}$	Fraction of correctly predicted positive instances against total positively classified instances.
bAcc	Classification	$\frac{TP/(TP+FN)+TN/(TN+FP)}{2}$	Arithmetic mean of the recalls for all classes.
MAP	Classification	$\frac{1}{T}\sum_{i=1}^{T} AP_i$	Average over APs for all classes.
MAE	Regression	$\frac{1}{n}\sum_{i=1}^{n} |e_i|$	Average of absolute prediction errors.
MSE	Regression	$\frac{1}{n}\sum_{i=1}^{n} e_i^2$	Average of squared prediction errors.
Pearson	Regression	$\frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}}$	Linear relationship between actual and predicted values.
GM	Regression	$\left(\prod_{i=1}^{n} |e_i|\right)^{\frac{1}{n}}$	Geometric mean of absolute prediction errors.
Time	All	-	Time (training time/inference time) of the algorithm.

Accuracy (Acc) [106] provides an overall measure of the model's correctness in predicting the entire test set. Acc favours the majority class as each instance has the same weight and contributes equally to the value of accuracy. In image classification and instance segmentation tasks, besides the overall accuracy across all classes, we follow previous work to comprehensively assess the performance of each method. We calculate the accuracy of three distinct subsets: many-shot classes (classes with over 100 training samples), medium-shot classes (classes with 20 to 100 training samples), and few-shot classes (classes with under 20 training samples). In the task of multi-label learning, each instance can be assigned to multiple classes, making predictions fully correct, partially correct, or fully incorrect. To capture the quality of these predictions, we utilize ACC@k($k = 1, 3, 5$) as evaluation metrics, which measure the top-k accuracy.

Precision [107] measures the number of correctly predicted positive instances against all the instances that were predicted as positive. In this paper, we use macro-precision, which is calculated by averaging the precision scores for each predicted class. The Macro approach considers all the classes equally, regardless of their size, ensuring the effect of majority classes is considered as important as that of minority classes.

Recall [107] is a metric that measures the proportion of true positive instances out of all the positive instances. In our evaluation, we utilize macro-recall, which is calculated by averaging recall for each actual class. Notably, recall equals accuracy when the test set is balanced.

Balanced accuracy (bAcc) [107] is the arithmetic mean of recall values calculated for each individual class. It is insensitive to imbalanced class distribution, as it treats every class with equal weight and importance and ensures minority classes have a more than proportional influence.

Mean average precision (MAP) [108] is a ranking-based metric that is the average of Average Precision (AP) over different classes, where AP is the area under the precision-recall curve.

Mean absolute error (MAE) [101] measures the average magnitude of errors between predicted and actual values without considering their direction.

Mean squared error (MSE) [101] calculates the average of the squared differences between predicted and actual values. It gives a higher weight to larger errors due to the squaring operation, making it particularly useful for identifying models that make fewer large mistakes.

Pearson correlation coefficient [101] quantifies the linear relationship between the actual and predicted values. It ranges from -1 to 1, where values closer to 1 or -1 indicate a strong positive or negative correlation, respectively, while a value near 0 indicates no linear relationship.

Error geometric mean (GM) [101] represents the geometric mean of the errors for better prediction fairness. It provides an indication of the central tendency of multiplicative differences.