Trustworthy Long-Tailed Classification

Bolian Li
Tianjin University
libolian@tju.edu.cn

Zongbo Han
Tianjin University
zongbo@tju.edu.cn

Haining Li
Xidian University
18200100006@stu.xidian.edu.cn

Huazhu Fu
IHPC, A*STAR
hzfu@ieee.org

Changqing Zhang*
Tianjin University
zhangchangqing@tju.edu.cn

Abstract

Classification on long-tailed distributed data is a challenging problem, which suffers from serious class-imbalance and accordingly unpromising performance especially on tail classes. Recently, the ensembling based methods achieve the state-of-the-art performance and show great potential. However, there are two limitations for current methods. First, their predictions are not trustworthy for failure-sensitive applications. This is especially harmful for the tail classes where the wrong predictions is basically frequent. Second, they assign unified numbers of experts to all samples, which is redundant for easy samples with excessive computational cost. To address these issues, we propose a Trustworthy Long-tailed Classification (TLC) method to jointly conduct classification and uncertainty estimation to identify hard samples in a multi-expert framework. Our TLC obtains the evidence-based uncertainty (EvU) and evidence for each expert, and then combines these uncertainties and evidences under the Dempster-Shafer Evidence Theory (DST). Moreover, we propose a dynamic expert engagement to reduce the number of engaged experts for easy samples and achieve efficiency while maintaining promising performances. Finally, we conduct comprehensive experiments on the tasks of classification, tail detection, OOD detection and failure prediction. The experimental results show that the proposed TLC outperforms existing methods and is trustworthy with reliable uncertainty.

1. Introduction

Data in real-world applications are usually long-tailed distributed over a series of categories [28, 34, 37, 44, 50, 51]. The frequencies of different categories vary a lot, with the head classes abundant in training samples, and the tail classes having only few training samples. Besides, there may also be new categories which models have not seen before [37], exceeding the tail of long-tailed distribution and being termed as out-of-distribution (OOD) data [32]. The long-tailed classification is very challenging since models need to handle the few-shot learning problem (and even with OOD data sometimes) for the tail classes, and the overall class-imbalance (models are trained on much more head samples than tail samples) would also deviate the models to focus extremely on the head classes [7]. These problems cause the models to perform unpromisingly especially on the tail classes [5, 19].

Existing algorithms address long-tailed classification mainly by rebalancing the training of different classes to assign larger importance to tail samples [7, 10, 33, 52], transferring knowledge between the head and tail classes [37, 57], ensembling statically sampled data groups [53, 55] (complementary ensembling), or ensembling individual classifiers in a multi-expert framework [51] (redundant ensembling). The redundant ensembling achieves the state-of-the-art performance mainly by reducing the model variance to obtain robust predictions [51]. However, there are two major limitations for redundant ensembling methods. First, they are usually vulnerable to yielding unreliable prediction (i.e., over-confident prediction). This also prevents the ensembling methods from perceiving the wrong predictions and OOD samples, and is especially harmful for the tail classes where the predictions have averagely more errors than the head classes [5, 19]. Consequently, their deployment in some failure-sensitive applications (e.g., disease diagnosis [2], automatic driving [54] and robotics [12]) is limited. Second, redundant ensembling usually assumes that all classifiers should be trained on all samples [51], which is static and often induces excessive computational cost by uniformly assigning experts to all classes. The expert redundancy is severe especially on head classes, where competitive classification performance can be achieved with

---

*Corresponding author.

much fewer experts.

For these issues, we propose a novel Trustworthy Long-tailed Classification (TLC) method to jointly conduct classification and uncertainty estimation in a unified framework. First, we introduce the evidence and its associated uncertainty under the Dempster-Shafer Evidence Theory (DST) [13]. With the help of evidence-based uncertainty (EvU), our model can perceive hard samples in long-tailed classification, promoting the trustworthiness by detecting the tail and OOD samples, and identifying potentially wrong predictions. Second, we propose to combine the evidences from different experts with a uncertainty-based multi-expert fusion strategy under the Dempster's rule. We leverage the advantages of multiple experts to obtain accurate uncertainty and robust prediction. Moreover, we propose to reduce the number of engaged experts dynamically for the easy samples to jointly promote the efficiency while maintaining promising performances. For example, the actually needed number of experts for the head classes is less than that for tail classes (the head classes contain more easy samples). Therefore, we need to dynamically assign fewer experts in the training of head classes for efficiency. We achieve the dynamic expert engagement by incrementally adding experts when the previously added experts are all uncertain about their predictions. The main contributions are summarized as follows:

- We introduce the evidence-based uncertainty (EvU) to promote the trustworthiness of long-tailed classification. To the best of our knowledge, the proposed TLC is the first work asserting trustworthiness in long-tailed classification.

- We propose a multi-expert fusion strategy based on the uncertainty of each expert under the Dempster-Shafer Evidence Theory (DST), which promotes the classification performance and trustworthiness by reliably perceiving hard samples.

- We achieve efficiency in training multiple experts by dynamically reducing the engaged experts with uncertainty, and obtain promising performances meanwhile.

- We conduct experiments on classification, tail & OOD sample detection and failure prediction, and evaluate the results with diverse metrics, which validates that the proposed TLC outperforms existing methods in the above tasks and is trustworthy with reliable uncertainty. The code[1] is publicly available.

2. Related Work

Long-tailed classification. Traditional long-tailed classification methods include under-sampling [36], over-

---

sampling [17] and data augmentation [8, 26, 35]. Rebalancing methods [7, 10, 33, 39, 52] focus more on the tail classes. OLTR [37] and inflated memory [57] transfer the knowledge between different class regions. BBN [55] learns the head and tail patterns separately. LFME [53] distills multiple teacher models respectively for class regions. RIDE [51] and ACE [6] ensembles multiple experts to obtain robust predictions. TDE [48] adopts casual inference to eliminate the biases of tail classes. These methods do not fully explore the uncertainty for perceiving hard samples in the predictions.

Uncertainty estimation. Traditional uncertainty estimation algorithms are discussed in [1, 3]. BNN [4] models uncertainty by replacing the deterministic parameters with distributions. MC Dropout [16] approximates BNN with dropout. MCP [20] obtains uncertainty from the softmax distribution. TCP [9] learns an extra module to yield uncertainty. EDL [45] models uncertainty under the subjective logic. Ensembling methods like [30] obtains uncertainty from the diverse predictions. DUQ [49] estimates the RBF distances as uncertainty. GP [11] models uncertainty as the similarity between samples using non-parametric kernel function.

Evidence theory. The Dempster-Shafer Evidence Theory (DST) was first proposed by [13]. It was later generalized as a framework to model the epistemic uncertainty [47]. The DST formulates the Bayesian inference with the subjective logic [14]. The DST allows the beliefs from different sources to be combined into a joint belief [22, 46].

3. Preliminaries

A long-tailed dataset consists of an imbalanced training set and a balanced test set. Formally, we define an input $\boldsymbol{x}_i \in \mathbb{R}^d$, its corresponding label $y_i \in \{1, 2, \cdots, K\}$, and the class-conditional distribution $p(\boldsymbol{x}|y)$. For the training set, the following relationships holds:

$$\begin{cases} \int p(\boldsymbol{x} \mid y = k_1)d\boldsymbol{x} \geq \int p(\boldsymbol{x} \mid y = k_2)d\boldsymbol{x}, \ \forall \ k_1 \leq k_2 \\ \lim_{k \to \infty} \int p(\boldsymbol{x} \mid y = k)d\boldsymbol{x} = 0 \end{cases},$$

(1)

indicating that the class volumes decay successively with the ascending class indexes and finally approach zero in the last few classes. The classes can be separated into the head, medium and tail regions based on different numbers of samples. For the test set, following the setting in most exising works on long-tailed problem [7, 10, 37, 51], the class frequencies are equal for the fairness among categories:

$$\int p(\boldsymbol{x} \mid y = k_1)d\boldsymbol{x} = \int p(\boldsymbol{x} \mid y = k_2)d\boldsymbol{x}, \ \forall \ k_1, k_2. \quad (2)$$

Additionally, for datasets that are not naturally long-tailed, a widely used transformation is to sample a subset with a spe-
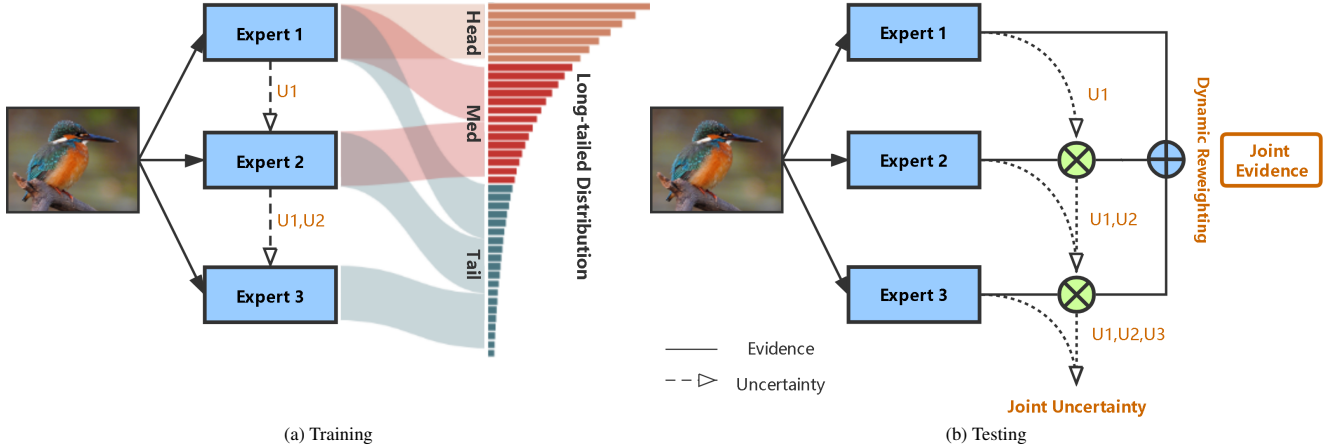
(a) Training   (b) Testing

Figure 1. Overview of the proposed Trustworthy Long-Tailed Classification (TLC). U1, U2 and U3 are the uncertainties of expert 1, 2 and 3 respectively. In training (a), we provide an example of collaborating in different class groups for multiple experts. TLC dynamically assigns averagely more experts to the samples in tail classes than those in head classes. This assignment is achieved automatically by identifying hard samples with uncertainty. In testing (b), the joint uncertainty is formed with the Dempster's rule, and the joint evidence is obtained by uncertainty-based dynamic reweighting.

cific decay distribution (e.g., exponential distribution [10] and Pareto distribution [37]).

4. Method

In this section, we introduce how to estimate uncertainty with Dempster-Shafer Evidence Theory in Sec. 4.1, propose to form joint uncertainty and joint evidence with the Dempster's rule in Sec. 4.2, and show the training process with dynamic expert engagement in Sec. 4.3.

4.1. Estimating Evidence-based Uncertainty

In long-tailed classification, perceiving hard samples with uncertainty can reduce the cost of trusting wrong predictions, which is especially important in tail classes with few training samples. However, existing methods suffer from over-confidence [40, 49] or excessive computational cost [4, 9, 16]. Therefore, for trustworthy long-tailed classification, we introduce the evidence-based uncertainty (EvU) under the Dempster-Shafer Evidence Theory (DST) to promote trustworthiness and efficiency simultaneously.

The DST is a generalization of the Bayesian theory of subjective probability [14]. While based on DST, subjective logic (SL) explicitly takes epistemic uncertainty and source trust into account [23]. The DST assigns **belief masses** to the possible sets of class labels for a prediction, measuring the chances to find the true class labels in these sets [45]. When a belief mass is assigned to all class labels, these classes are equally likely. Therefore, such belief mass can represent the **uncertainty** of the entire prediction [23]. Formally, the subjective logic defines the belief assignment

over a Dirichlet distribution [27]:

$$D(\boldsymbol{p} \mid \boldsymbol{\alpha}) = \begin{cases} \dfrac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k - 1} & \text{for } \boldsymbol{p} \in \mathcal{S}_K \\ 0 & \text{otherwise} \end{cases}, \quad (3)$$

where $\boldsymbol{\alpha}$ are parameters of the distribution, $B(\cdot)$ is the beta function, and $\mathcal{S}_K = \{\boldsymbol{p}| \sum_{k=1}^{K} p_k = 1 \text{ and } 0 \leq p_k \leq 1, \forall k\}$ is the K-dimensional unit simplex. The uncertainty and belief masses are determined by the parameters as:

$$u = \frac{K}{S} \quad \text{and} \quad b_k = \frac{\alpha_k - 1}{S}, \quad (4)$$

where α_k is the Dirichlet parameter for the k-th class and $S = \sum_{k=1}^{K} \alpha_k$ is the Dirichlet strength. In this way, the uncertainty is less likely to suffer from over-confidence since it avoids merely regarding the probability of labeled class as uncertainty, and takes into account the comparative values against other classes [45].

In DST, the **evidence** is a measure of the support derived from data in favor of a sample to fall into a certain class [45]. The evidence of each class $\boldsymbol{e} = [e_1, e_2, \cdots, e_K]$ can be obtained directly from the output of neural networks by replacing the softmax layer with a non-negative activation function. Then, the parameters of Dirichlet distribution in Eq. 3 can be computed by using $\alpha_k = e_k + 1$, and therefore, the uncertainty and belief masses can be quantified with Eq. 4. Moreover, it is obvious that the total amount of uncertainty and belief masses is a constant, i.e., $u + \sum_{k=1}^{K} b_k = 1$ (implying a $K + 1$-dimensional unit simplex). When the evidences on all classes are insufficient for a prediction, the belief masses assigned to all classes will also be low, and

meanwhile, the uncertainty for this output will be high to indicate a high probability of erroneous prediction.[2]

The strength of EvU (evidence-based uncertainty) lies in its modeling based on the Dirichlet distribution, which parameterizes the density of belief assignments directly from the outputs of neural networks. EvU models the uncertainty and high-order probabilities for a prediction [18]. Further, EvU also theoretically avoids the over-confident problem (common in traditional uncertainty estimation algorithms [20]) by obtaining the uncertainty from the overall belief masses. It is noteworthy that the EvU can be obtained directly with Eq. 4, which is efficient and reasonable.

4.2. Combining Experts with Dempster's Rule

We employ a multi-expert framework with each expert guided by the DST introduced in Sec. 4.1. [51] shows that integrating multiple classifiers will reduce the model variance, which is beneficial to the robustness of long-tailed classification. On top of ensembling, we combine the uncertainties and evidences of multiple experts under the Dempster's rule (shown in Fig. 1b).

Combining uncertainties. We combine the uncertainties of multiple experts in an incremental fashion (e.g., first combine expert 1 and 2, and then add on expert 3). First, we formalize the pair-wise Dempster's combination rule as:

$$u^1 \oplus u^2 = \frac{1}{1-C} u^1 u^2, \tag{5}$$

where $C = \sum_{i \neq j} b_i^1 b_j^2$ is the conflict factor. When two experts agree on most of belief masses (i.e., C is small), the combined uncertainty will be relatively low.[2] Second, we apply the combination rule to sequentially combine multiple experts, and the final combination is induced as:

$$u = u^1 \oplus u^2 \oplus \dots \oplus u^M = \frac{\prod_{m=1}^{M} u^m}{\prod_{m=1}^{M}(1-C^m)}, \tag{6}$$

where $C^m = \sum_{i \neq j} b_i^m b_j^{m-1}$ is the conflict factor between two consecutive experts and $C^1 = 0$ (the first expert do not have former result to compare with). The uncertainty combination considers both independent uncertainty from each expert and the agreement of beliefs between different experts.

Combining evidences. We dynamically reduce the engaged experts on easy samples in training stage (detailed discussion in Sec. 4.3). Therefore, at test stage, the engaged experts for easy samples should also be limited. For example, for the head classes, we primarily consider the evidence of the first few experts, while for the tail classes, the evidences of all experts should be considered. First, we define

the **prefix weights** of each expert with the following rules:

$$\begin{cases} w^1 = 1; \\ w^2 = u^1; \\ w^{m+1} = w^m \oplus u^m = \frac{1}{1-C^m} w^m u^m, \\ \qquad \text{for } m = 2, 3, \cdots, M-1. \end{cases} \tag{7}$$

The prefix weight w^m is a measure of the overall uncertainty from experts previous to expert m. It accords with the combining process of the joint uncertainty in Eq. 6, and regards the intermediate combination results as weights. When the experts previous to expert m are already certain about their evidences (indicating that it is an easy sample), the prefix weight w^m will be low (indicating that e^m is not obliged to consider). Second, we apply the prefix weights to combine the evidences at inference time:

$$e = \frac{\sum_{m=1}^{M} \exp\{w^m/\eta\} \cdot e^m}{\sum_{m=1}^{M} \exp\{w^m/\eta\}}, \tag{8}$$

where exponentiation is adopted for non-maximum suppression [43], and η is a temperature factor which adjusts the sensitivity of the prefix weights. We also apply the prefix weights for training efficiency in Sec. 4.3.

4.3. Learning Experts with Dynamic Engagement

In our multi-expert framework, each expert can capture the evidence from input to induce a classification opinion [25]. We propose to jointly learn experts under the subjective logic, while dynamically reducing the number of engaged experts for easy samples.

Learning single expert. For a single expert, we formulate the objective with the Type II Maximum Likelihood (Empirical Bayes) [21]. First, we obtain the evidence e_i and convert class label y_i into a one-hot vector y_i. Second, we treat an adjusted Dirichlet distribution $\widetilde{D}(p_i|e_i)$ as the prior of multinomial likelihood $P(y_i|p_i)$ (the classification opinion) and then compute the negative logarithm of the marginal likelihood:

$$\begin{aligned} \mathcal{L} &= -\log\left[\int \prod_{k=1}^{K} p_{ik}^{y_{ik}} \frac{1}{B(e_i)} \prod_{k=1}^{K} p_{ik}^{e_{ik}-1} dp_i\right] \\ &= \sum_{k=1}^{K} y_{ik}\left[\log\left(\sum_{k=1}^{K} e_{ik}\right) - \log(e_{ik})\right]. \end{aligned} \tag{9}$$

However, the objective in Eq. 9 only ensures that the correct class will be assigned with more evidence than other classes, while there is no support for the low overall evidences on the incorrect classes. In another word, the uncertainty may be unreasonably low due to the high overall

---

[2]It is discussed in the supplementary material.

evidences. We address this problem by introducing the following Kullback-Leibler divergence [45]:

$$\mathcal{L}_{kl} = KL(D(\boldsymbol{p}_i \mid \widetilde{\boldsymbol{\alpha}}_i) \parallel D(\boldsymbol{p}_i \mid \mathbf{1}))$$

$$= \log \frac{\Gamma(\widetilde{S}_i)}{\Gamma(K) \prod_{k=1}^{K} \Gamma(\widetilde{\alpha}_{ik})} \qquad (10)$$

$$+ \sum_{k=1}^{K} (\widetilde{\alpha}_{ik} - 1) \left[\psi(\widetilde{\alpha}_{ik}) - \psi\left(\widetilde{S}_i\right) \right],$$

where $\widetilde{\boldsymbol{\alpha}}_i = \mathbf{1} + (\mathbf{1} - \boldsymbol{y}_i) \odot \boldsymbol{e}_i$ is the adjusted Dirichlet parameters, $\widetilde{S}_i = \sum_{k=1}^{K} \widetilde{\alpha}_k$ is the adjusted Dirichlet strength, $\Gamma(\cdot)$ is the gamma function, and $\psi(\cdot)$ is the digamma function. This KL divergence regulates the evidences of incorrect classes to 0 by minimizing the distance between the adjusted distribution and the target distribution, and thus can avoid the uncertainty to be unreasonably low. Finally, the objective for a single expert is

$$\mathcal{L}_{single} = \mathcal{L} + \lambda_{kl}(t)\mathcal{L}_{kl}, \qquad (11)$$

where $\lambda_{kl}(t) = min\{1, t/T\}$ is the annealing factor (t is the current epoch). We gradually increase the the KL divergence to prevent the expert to learn a flat uniform evidence in the early stage of training.

Learning multiple experts dynamically. We argue that there is no necessity in learning easy samples (usually in head classes) with all experts, since using fewer experts can also achieve competitive performances for these samples. We show the experimental support for this in Sec. 5.3. To this end, we propose to apply the prefix weights of all experts (a measure of joint uncertainty from a group of experts defined in Eq. 7) to dynamically remove the losses on easy samples. For example, if experts $1, 2, \cdots, m-1$ are all certain about a sample (i.e., $w^m \leq \tau$), the loss of expert m on this sample will be removed. Therefore, the overall expert engagement should be in a descending pattern. For example, the first expert is responsible for all classes, the second expert for the classes except the head classes, and the last expert only focuses on the tail classes (as shown in Fig. 1a).

Additionally, to enhance the diversity of experts, we form the output distribution $P(\boldsymbol{p}_i|\boldsymbol{\alpha}_i^m)$ with the normalized Dirichlet parameters: $P(p_{ik}|\boldsymbol{\alpha}_i^m) = \alpha_{ik}^m / S_i^m$, and push different experts apart by the following KL divergence:

$$\mathcal{L}_{div} = -\frac{1}{M} \sum_{m=1}^{M} KL(P(\boldsymbol{p}_i \mid \boldsymbol{\alpha}_i^m) \parallel P(\boldsymbol{p}_i \mid \overline{\boldsymbol{\alpha}}_i)), \qquad (12)$$

where $\overline{\boldsymbol{\alpha}}_i = \sum_{m=1}^{M} \boldsymbol{\alpha}_i^m / M$ are the averaged Dirichlet parameters.

Finally, the joint objective to learn evidences in the multi-expert framework is given by adding up the objective of each expert:

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{m=1}^{M} \mathbb{1}\{w_i^m > \tau\} \mathcal{L}_{single} + \lambda_{div} \mathcal{L}_{div}. \qquad (13)$$

The overall training process is summarized in the supplementary material.

5. Experiments

In this section, we conduct experiments to answer the following questions:

- **Q1 (Effectiveness):** Does the proposed TLC outperform the state-of-the-art methods in long-tailed classification? (Sec. 5.2)

- **Q2 (Trustworthiness I):** How to validate the trustworthiness of TLC, and is the estimated uncertainty reliable? (Sec. 5.2 and Sec. 5.4)

- **Q3 (Efficiency I):** Why is it reasonable to reduce the engaged experts for easy samples? (Sec. 5.3)

- **Q4 (Trustworthiness II):** Is the estimated uncertainty good at discerning the head, medium and tail classes? (Sec. 5.3)

- **Q5 (Efficiency II):** Does the actual expert engagement accord with our expectation in Sec. 4.3? (Sec. 5.3)

Specifically, we show the configurations in Sec. 5.1, quantitative and qualitative results in Sec. 5.2 and Sec. 5.3 respectively, and ablation studies in Sec. 5.4.

5.1. Experimental Setup

Tasks. Along with classification, to show the effect of uncertainty in long-tailed problem, we conduct the following tasks: tail detection, out-of-distribution (OOD) detection and failure prediction [20]. These tasks all use the estimated uncertainty for binary classification. Specifically, in tail and OOD detection, uncertainties are used to distinguish the tail/OOD samples from others, and in failure prediction, uncertainties are used to distinguish between incorrect and erroneous predictions. The metrics for evaluation are similar to those used in binary classification and confidence calibration (e.g., AUC [38], FPR-95 [31] and ECE [41]).

Datasets. We use three long-tailed datasets (CIFAR-10-LT, CIFAR-100-LT and ImageNet-LT) and three balanced OOD datasets (SVHN [42], ImageNet-open and Places-open). CIFAR-10-LT and CIFAR-100-LT [10] are sampled from the original CIFAR [29] dataset over exponential distributions [10]. ImageNet-LT [37] is sampled from the ImageNet-2012 [15] dataset over Pareto distributions with the power value $\alpha = 6$. It contains 115.8K images in 1,000 classes. ImageNet-open is the additional classes of images in the ImageNet-2010 dataset [37]. Places-open [37] is the test images from the Places-Extra69 dataset [56].

Compared methods. We compare the proposed TLC with re-balancing methods including Focal Loss [33], LDAM-DRW [7], τ-norm and cRT [24], knowledge transfer method

OLTR [37], and ensemble learning method RIDE [51]. We also compare the evidence-based uncertainty with other widely used uncertainty estimation algorithms including the Maximal Class Probability (MCP) [20], Gaussian Process (GP) [11], and Monte Carlo Dropout (MCD) [16].

5.2. Quantitative Evaluation

Classification (Q1). We evaluate the performances on classification with diverse metrics. Along with the Top-1 accuracy, we also report the **regional accuracy** which computes the frequency of predictions falling into the correct class region (e.g., whether tail samples are classified into tail classes[3]). Higher regional accuracy implies better trustworthiness for long-tailed classification. The evaluation results are listed in Table. 1. We run each experiment five times to report the average ACC and standard deviation. The proposed TLC outperforms the compared methods on all datasets, and improves the regional and tail ACC significantly.

Tail & OOD detection (Q2). We evaluate the performances on tail detection and OOD detection in terms of AUC scores [38]. For tail detection, we label the tail classes as positive and the others as negative. For OOD detection, we jointly use the in-distribution and OOD samples by labeling the in-distribution as negative and the OOD as positive. We use the MCP [20] (maximal value in the softmax distribution) to quantify the uncertainty for the compared methods. The evaluation results are listed in Table. 2. Our proposed TLC outperforms the compared methods, and is especially better at identifying OOD samples in large image datasets (i.e., ImageNet-open and Places-open).

Failure prediction (Q2). We evaluate the performances on failure prediction in terms of AUC [38], the FPR at 95% TPR (FPR-95) [31] and the Expected Calibration Error (ECE) [41] respectively for the head, medium and tail classes. We also use the MCP [20] to quantify uncertainty for the compared methods. The evaluation results are listed in Table. 3. Our TLC outperforms the compared methods, and performs much better especially in terms of ECE.

5.3. Qualitative Evaluation

Number of experts (Q3). We visualize the accuracy on three class regions (head, medium and tail) on CIFAR-100-LT with ascending numbers of experts in Fig. 2. We find that using more experts is beneficial to the tail classes, but do not have significant effect for the head classes. This observation justifies our motivation that assigning the same number of experts is redundant for easy samples.

---

[3]When they fail, they are still likely to be trusted due to the lower averaged uncertainty of head classes, but even if they fall into other tail classes erroneously, the model is still uncertain about them, and thus reduce the potential threat.
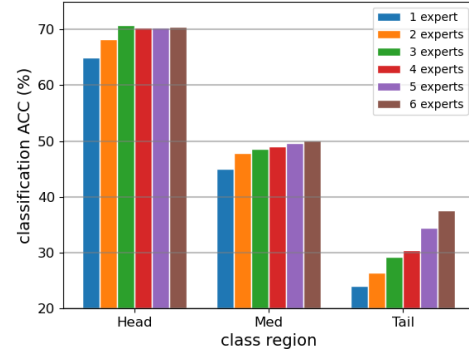

Figure 2. Classification accuracy on CIFAR-100-LT for the head, medium and tail classes with different number of experts.
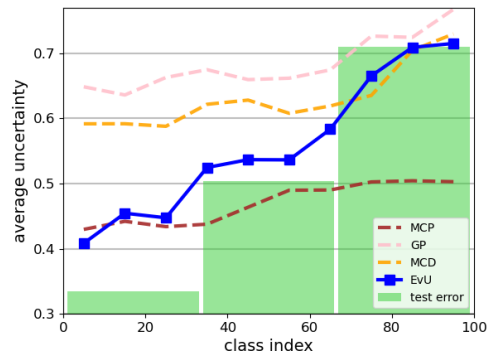

Figure 3. Average uncertainty on CIFAR-100-LT from different uncertainty estimation algorithms (also with the test errors of 3 class regions as benchmarks).

Uncertainty for each class (Q4). We visualize the uncertainty of each class of the CIFAR-100-LT test data with various uncertainty estimation algorithms in Fig. 3 (using 3 experts for the evidence-based uncertainty, EvU for short). We compute the averaged uncertainties of every 10 classes. Among the compared uncertainties, EvU is the most consistent with the real test errors. Therefore, it is easy to distinguish the head, medium and tail classes with EvU (practically, low uncertainty indicates the head classes and high uncertainty indicates the tail classes).

Expert engagement (Q5). We visualize the percentage of samples using diverse number of experts respectively on the head, medium, tail and all classes on CIFAR-100-LT in Fig. 4. We set the maximal number of experts as 4 and the threshold $\tau = 0.54$ according to the hyperparameter settings of quantitative evaluations. The samples using 4 experts dominate the tail classes and the samples using 1 expert dominate the head classes. Overall, the hard samples are assigned with more experts to learn the patterns, which is consistent with our motivation in Sec. 4.3.

Table 1. Performance comparison on long-tailed classification in terms of ACC (in percentage).

Dataset	Method	All	Region	Head	Med	Tail
CIFAR-10-LT	Focal Loss	68.6±0.2	73.2±0.4	84.8±0.2	67.9±0.9	49.1±0.8
	OLTR	78.7±0.6	80.5±0.3	86.1±0.1	77.5±0.6	69.8±1.7
	LDAM-DRW	78.4±1.0	82.5±0.4	**89.6±0.1**	74.0±1.5	72.4±2.0
	τ-norm	79.6±1.0	83.5±0.4	87.7±0.2	76.2±1.6	73.6±1.4
	cRT	79.2±0.3	83.0±0.4	87.1±0.1	77.3±0.8	71.5±0.9
	RIDE	80.2±0.3	83.4±0.2	87.4±0.1	77.2±0.7	75.0±0.5
	TLC(2 experts)	80.3±0.4	84.2±0.3	86.0±0.1	77.8±0.5	75.4±0.8
	TLC(3 experts)	80.3±0.4	84.2±0.3	85.9±0.1	77.2±0.8	**75.9±0.6**
	TLC(4 experts)	**80.4±0.2**	**84.4±0.2**	85.7±0.1	**78.1±0.5**	75.6±0.5
CIFAR-100-LT	Focal Loss	42.3±1.3	55.4±0.4	70.3±1.7	40.7±1.6	15.9±1.9
	OLTR	43.4±0.8	59.9±0.2	64.6±2.0	44.8±1.5	20.9±2.4
	LDAM-DRW	44.4±1.2	61.4±0.2	64.8±1.5	43.8±1.3	24.6±1.8
	τ-norm	45.4±1.2	62.3±0.6	68.0±1.6	47.2±1.4	21.0±2.0
	cRT	45.6±0.3	62.3±0.5	67.8±2.4	47.1±2.1	21.8±1.6
	RIDE	48.3±0.5	62.8±0.1	68.8±1.2	**49.0±0.7**	27.1±1.4
	TLC(2 experts)	47.2±0.7	62.8±0.3	69.4±1.2	46.6±1.0	25.7±1.5
	TLC(3 experts)	49.0±0.4	64.0±0.2	70.9±0.8	47.9±0.9	28.1±1.3
	TLC(4 experts)	**49.8±0.8**	**64.5±0.2**	**71.1±1.0**	48.4±1.1	**29.7±1.6**
ImageNet-LT	Focal Loss	45.6±2.1	67.0±0.6	69.2±3.2	41.5±2.7	26.1±3.1
	OLTR	50.7±1.2	68.0±0.5	67.8±1.9	53.3±1.8	31.0±2.4
	LDAM-DRW	49.8±0.7	66.9±0.5	63.3±2.1	50.2±2.2	36.0±1.3
	τ-norm	47.9±1.2	67.8±0.3	60.3±1.8	50.6±1.3	33.0±1.8
	cRT	48.4±1.3	67.5±0.5	64.4±2.4	50.5±1.4	30.3±1.8
	RIDE	54.6±0.9	68.4±0.3	**70.6±1.3**	54.8±0.9	38.3±1.4
	TLC(2 experts)	54.1±0.6	68.4±0.3	68.7±1.2	55.4±1.2	38.3±1.4
	TLC(3 experts)	54.6±0.5	69.1±0.3	69.3±1.2	**56.7±0.8**	37.9±1.8
	TLC(4 experts)	**55.1±0.7**	**69.9±0.2**	68.9±1.2	55.7±1.5	**40.8±0.8**

Table 2. Performances comparison on tail detection and OOD detection in terms of AUC (in percentage).

Training	CIFAR-10-LT				CIFAR-100-LT				ImageNet-LT	
Testing	Tail	SVHN	ImageNet-open	Places-open	Tail	SVHN	ImageNet-open	Places-open	Tail	ImageNet-open
Focal Loss	36.3	64.0	70.0	70.6	35.4	54.0	53.5	53.2	26.8	43.1
OLTR	55.9	55.9	78.2	77.1	37.2	53.7	54.1	52.8	27.5	42.1
LDAM-DRW	54.9	55.8	78.2	76.5	36.9	**54.1**	53.1	54.7	26.4	42.6
τ-norm	56.2	56.0	79.7	77.9	36.5	52.0	54.3	52.3	28.1	43.3
cRT	56.1	55.5	81.2	75.1	36.8	53.8	50.1	52.3	**28.6**	43.5
RIDE	56.2	77.1	80.5	79.8	35.4	45.9	54.5	55.9	**28.6**	44.6
TLC(2 experts)	55.6	**83.8**	85.8	84.2	36.8	**54.1**	52.9	55.9	27.9	44.6
TLC(3 experts)	**56.9**	74.9	**87.0**	**86.2**	36.3	53.4	52.9	53.7	28.1	43.9
TLC(4 experts)	56.5	80.5	82.8	84.7	**37.3**	54.1	**54.6**	**56.5**	**28.6**	**44.7**

5.4. Ablation Study

Effectiveness of components. We compare different combinations of the components ($\mathcal{L}$, $\mathcal{L}_{kl}$ and the prefix weight w) on CIFAR-100-LT in terms of classification, OOD detection and failure prediction (all using 3 experts). The results are listed in Table. 4, where we also include the results of full objective for reference. It is easy to conclude: i) adding $\mathcal{L}_{kl}$ is beneficial to obtaining more reliable uncertainty (comparing line 2 against line 1 and line 3 against line 4), and ii) dynamically reducing engaged experts (with w) does not significantly affect the performance on the three tasks (comparing line 2 with line 4).

Comparison of uncertainties (Q2). We compare various uncertainty estimation algorithms on failure prediction. We use 1 expert as the backbone model and compute MCP [20], GP [11], MCD [16] and the EvU on CIFAR-100-

LT dataset. According to the results in Table. 5, the EvU outperforms the other algorithms on all tasks especially on the tail classes, which validates that the EvU is more reliable than other compared uncertainty estimation algorithms.

6. Conclusion

In this paper, we propose the Trustworthy Long-tailed Classification (TLC), which estimates evidence and uncertainty in a multi-expert framework. The estimated evidence and uncertainty of each expert are combined under the Dempster-Shafer Evidence Theory (DST). The TLC can dynamically reduce the number of engaged experts for easy samples, which ensures efficiency while preserving promising performances. We evaluate the TLC on multiple tasks with diverse metrics, where it outperforms existing methods and is trustworthy with reliable uncertainty.

Table 3. Performance comparison on failure prediction (in percentage).

| Dataset	Method	AUC ↑				FPR-95 ↓				ECE ↓			
		All	Head	Med	Tail	All	Head	Med	Tail	All	Head	Med	Tail
CIFAR-10-LT	Focal Loss	75.7	80.5	75.6	85.1	79.8	72.5	80.9	80.3	20.1	11.7	19.7	33.3
	OLTR	**83.9**	79.6	83.9	**85.9**	79.6	72.9	82.4	80.7	18.8	11.2	19.6	33.2
	LDAM-DRW	83.1	79.6	86.3	85.2	69.3	71.7	72.9	62.1	18.9	12.4	22.0	24.5
	τ-norm	83.8	79.5	85.2	83.5	67.9	71.2	71.9	**59.7**	17.8	12.0	20.7	22.3
	cRT	83.7	79.8	84.1	85.3	67.4	71.2	70.3	62.1	18.4	11.5	19.8	21.3
	RIDE	82.9	81.7	85.4	84.2	68.7	71.9	70.5	62.3	15.9	**9.8**	17.9	22.4
	TLC(2 experts)	83.5	**84.3**	85.6	83.4	68.9	66.7	68.0	71.2	12.8	10.6	13.1	**15.8**
	TLC(3 experts)	83.7	83.0	87.1	83.6	68.0	68.7	60.8	72.7	13.1	11.3	12.4	16.8
	TLC(4 experts)	**83.9**	84.0	**87.8**	83.8	**65.7**	66.3	58.2	68.2	**12.5**	11.4	**11.3**	15.9
CIFAR-100-LT	Focal Loss	73.3	83.7	72.9	53.3	78.9	66.4	81.2	89.5	24.2	16.0	22.3	35.0
	OLTR	73.5	85.6	79.2	56.3	79.5	69.5	79.6	90.4	23.6	16.2	22.2	34.5
	LDAM-DRW	72.7	**85.7**	75.5	55.6	81.8	68.9	76.7	92.1	30.9	18.7	30.8	43.6
	τ-norm	73.9	85.5	75.1	54.8	78.9	66.0	83.0	89.3	29.8	17.2	29.7	42.5
	cRT	74.1	83.4	**79.7**	53.5	78.6	64.5	78.9	**89.2**	30.2	19.8	28.9	43.8
	RIDE	76.3	85.5	79.5	**60.0**	79.5	66.2	80.1	89.7	24.1	**14.5**	23.8	34.3
	TLC(2 experts)	**77.9**	**85.7**	78.7	**60.0**	**78.3**	**64.3**	77.7	91.2	23.2	27.8	23.0	**22.4**
	TLC(3 experts)	76.9	84.5	78.8	57.5	79.8	67.1	**76.5**	90.6	22.8	24.8	21.9	24.6
	TLC(4 experts)	76.7	85.3	77.7	58.6	80.5	66.8	80.6	89.8	**21.2**	21.7	**20.6**	25.6
ImageNet-LT	Focal Loss	65.4	73.7	62.8	43.4	83.9	68.4	86.2	94.5	35.3	28.3	33.2	45.1
	OLTR	66.0	72.8	67.0	42.3	82.9	67.1	85.0	96.3	34.8	25.6	32.6	42.3
	LDAM-DRW	66.8	**76.8**	67.5	47.6	82.8	70.9	81.7	96.1	35.1	28.6	40.7	50.4
	τ-norm	66.1	71.6	59.7	48.5	82.9	68.4	83.1	94.7	33.0	27.8	38.1	50.4
	cRT	66.4	70.2	63.3	49.2	**82.6**	**65.1**	83.2	95.3	32.5	27.2	37.4	48.8
	RIDE	66.2	75.8	70.3	47.1	84.5	68.2	85.1	94.7	31.9	24.5	35.7	42.4
	TLC(2 experts)	66.7	75.2	**71.2**	48.2	84.5	65.9	**80.5**	94.5	32.8	26.3	**30.7**	40.3
	TLC(3 experts)	66.7	75.7	68.3	48.8	84.6	67.0	80.8	97.2	**31.3**	24.7	31.6	**39.6**
	TLC(4 experts)	**67.2**	76.2	68.7	**49.4**	**82.6**	65.2	81.5	**94.3**	31.9	**23.8**	30.8	40.1
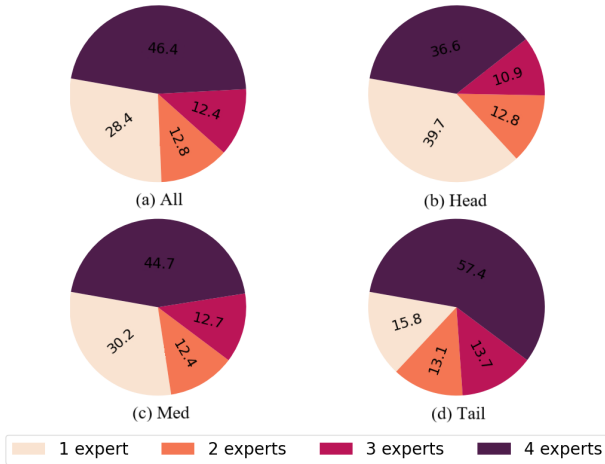
Figure 4. Visualization of expert engagement on CIFAR-100-LT (the percentage of samples using specific number of experts are marked on the pie charts).

Table 4. Ablation study on different combination of objective components of the proposed TLC with CIFAR-100-LT dataset.

$\mathcal{L}$	$\mathcal{L}_{kl}$	w	ACC classification	AUC OOD detection	AUC failure prediction
✓			48.9	45.8	63.7
✓	✓		**49.1**	**53.2**	**76.2**
✓		✓	48.7	42.4	61.6
✓	✓	✓	49.0	53.4	76.9

Table 5. Ablation study on uncertainty, comparing different uncertainties on CIFAR-100-LT dataset (in percentage).

Method		MCP	Entropy	MCS	EvU
AUC ↑	All	74.3	75.1	76.4	**77.9**
	Head	83.9	84.6	84.4	**85.7**
	Med	79.2	78.9	**80.9**	78.7
	Tail	57.0	57.6	58.2	**60.0**
FPR-95 ↓	All	79.5	79.4	80.9	**78.3**
	Head	66.0	65.6	65.4	**64.3**
	Med	76.1	76.8	**75.7**	77.7
	Tail	89.7	**89.1**	90.1	91.2
ECE ↓	All	24.1	25.0	23.9	**23.2**
	Head	**14.5**	18.5	19.7	27.8
	Med	23.8	24.3	23.1	**23.0**
	Tail	34.3	34.7	32.6	**22.4**

Acknowledgements

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 2021. 2

[2] Chunyan Ao, Shunshan Jin, Hui Ding, Quan Zou, and Liang Yu. Application and development of artificial intelligence and intelligent disease diagnosis. *Current pharmaceutical design*, 26(26):3069–3075, 2020. 1

[3] John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. Confidence estimation for machine translation. In *Coling 2004: Proceedings of the 20th international conference on computational linguistics*, pages 315–321, 2004. 2

[4] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015. 2, 3

[5] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. 1

[6] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 112–121, 2021. 2

[7] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 1567–1578, 2019. 1, 2, 5

[8] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 694–710. Springer, 2020. 2

[9] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 2902–2913, 2019. 2, 3

[10] 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. 1, 2, 3, 5

[11] Andreas Damianou and Neil D Lawrence. Deep gaussian processes. In *Artificial intelligence and statistics*, pages 207–215. PMLR, 2013. 2, 6, 7

[12] Brian Davies. A review of robotics in surgery. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, 214(1):129–140, 2000. 1

[13] AP Dempster et al. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38(2):325–339, 1967. 2

[14] Arthur P Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968. 2, 3

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2, 3, 6, 7

[17] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005. 2

[18] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2020. 4

[19] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 1

[20] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017. 2, 4, 5, 6, 7

[21] Tahira Jamil and Cajo JF ter Braak. Selection properties of type ii maximum likelihood (empirical bayes) in linear models with individual variance components for predictors. *Pattern Recognition Letters*, 33(9):1205–1212, 2012. 4

[22] Audun Jøsang and Robin Hankin. Interpretation and fusion of hyper opinions in subjective logic. In *2012 15th International Conference on Information Fusion*, pages 1225–1232. IEEE, 2012. 2

[23] AUDUN. JSANG. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer, 2018. 3

[24] 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*, 2019. 5

[25] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 4

[26] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13896–13905, 2020. 2

[27] Samuel Kotz, Narayanaswamy Balakrishnan, and Norman L Johnson. *Continuous multivariate distributions, Volume 1: Models and applications*. John Wiley & Sons, 2004. 3

[28] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome:

Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 1

[29] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. 5

[30] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[31] Shiyu Liang, Yixuan Li, and R Srikant. Principled detection of out-of-distribution examples in neural networks. *arXiv preprint arXiv:1706.02690*, pages 655–662, 2017. 5, 6

[32] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 1

[33] 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. 1, 2, 5

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[35] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2970–2979, 2020. 2

[36] 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, 2008. 2

[37] 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. 1, 2, 3, 5, 6

[38] Donna Katzman McClish. Analyzing a portion of the roc curve. *Medical decision making*, 9(3):190–195, 1989. 5, 6

[39] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2020. 2

[40] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *international conference on machine learning*, pages 7034–7044. PMLR, 2020. 3

[41] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. 5, 6

[42] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *NIPS Work-shop on Deep Learning and Unsupervised Feature Learning*, 2011. 5

[43] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 850–855. IEEE, 2006. 4

[44] William J Reed. The pareto, zipf and other power laws. *Economics letters*, 74(1):15–19, 2001. 1

[45] Murat Şensoy, L Kaplan, and M Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 2018. 2, 3, 5

[46] Kari Sentz, Scott Ferson, et al. *Combination of evidence in Dempster-Shafer theory*, volume 4015. Sandia National Laboratories Albuquerque, 2002. 2

[47] Glenn Shafer. *A mathematical theory of evidence*. Princeton university press, 1976. 2

[48] 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, 2020. 2

[49] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pages 9690–9700. PMLR, 2020. 2, 3

[50] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017. 1

[51] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2020. 1, 2, 4, 6

[52] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, pages 162–178. Springer, 2020. 1, 2

[53] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, pages 247–263. Springer, 2020. 1, 2

[54] Seiji Yasunobu and Ryota Sasaki. An auto-driving system by interactive driving knowledge acquisition. In *SICE 2003 Annual Conference (IEEE Cat. No. 03TH8734)*, volume 3, pages 2935–2938. IEEE, 2003. 1

[55] 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. 1, 2

[56] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 5

[57] Linchao Zhu and Yi Yang. Inflated episodic memory with region self-attention for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4344–4353, 2020. 1, 2